The Science and Applications of Microbial Genomics: Workshop Summary

ISBN
978-0-309-26819-6

428 pages
6 x 9
PAPERBACK (2013)

Eileen R. Choffnes, LeighAnne Olsen, and Theresa Wizemann, Rapporteurs; Forum on Microbial Threats; Board on Global Health; Institute of Medicine

THE SCIENCE AND APPLICATIONS OF MICROBIAL GENOMICS

WORKSHOP SUMMARY

Eileen R. Choffnes, LeighAnne Olsen, and Theresa Wizemann,
Rapporteurs

Forum on Microbial Threats

Board on Global Health

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

Financial support for this project was provided by the U.S. Department of Health and Human Services: National Institutes of Health, National Institute of Allergy and Infectious Diseases, Centers for Disease Control and Prevention, and the Food and Drug Administration; U.S. Department of Defense, Department of the Army: Global Emerging Infections Surveillance and Response System, Medical Research and Materiel Command, and the Defense Threat Reduction Agency; U.S. Department of Veterans Affairs; U.S. Department of Homeland Security; U.S. Agency for International Development; Uniformed Services University of the Health Sciences; the Alfred P. Sloan Foundation, American Society for Microbiology; sanofi pasteur; Burroughs Wellcome Fund; GlaxoSmithKline; Infectious Diseases Society of America; and the Merck Company Foundation. The views presented in this publication do not necessarily reflect the views of the organizations or agencies that provided support for this project.

International Standard Book Number-13: 978-0-309-26819-6
International Standard Book Number-10: 0-309-26819-2

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

For more information about the Institute of Medicine, visit the IOM home page at: **www.iom.edu.**

Printed in the United States of America

The serpent has been a symbol of long life, healing, and knowledge among almost all cultures and religions since the beginning of recorded history. The serpent adopted as a logotype by the Institute of Medicine is a relief carving from ancient Greece, now held by the Staatliche Museen in Berlin.

Cover image: Derivation of genome trees from the comparative analyses of complete genomes. (2005) PLoS Computational Biology Issue Image | Vol. 1(7) December 2005. *PLoS Comput Biol* 1(7): ev01.i07. doi:10.1371/image.pcbi.v01.i07. Genome graphic representation completed with GenomeViz software, provided by Rohit Ghai. Fractal tree obtained with the FractalTrees X software, provided by Simon Woodside. Compiled by Edouard Yeramian.

Suggested citation: IOM (Institute of Medicine). 2013. *The Science and Applications of Microbial Genomics: Workshop Summary.* Washington, DC: The National Academies Press.

"Knowing is not enough; we must apply.
Willing is not enough; we must do."
—Goethe

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

Advising the Nation. Improving Health.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

PLANNING COMMITTEE FOR A WORKSHOP ON THE SCIENCE AND APPLICATIONS OF MICROBIAL GENOMICS[1]

BRUCE BUDOWLE, University of North Texas Health Science Center, Fort Worth, Texas
ARTURO CASADEVALL, Albert Einstein College of Medicine, Bronx, New York
JONATHAN EISEN, University of California, Davis, California
CLAIRE FRASER, University of Maryland School of Medicine, Baltimore, Maryland
PAUL KEIM, Northern Arizona University, Flagstaff, Arizona
DAVID RELMAN, Stanford University, Stanford, California

[1] Institute of Medicine planning committees are solely responsible for organizing the workshop, identifying topics, and choosing speakers. The responsibility for the published workshop summary rests solely with the workshop rapporteurs and the institution.

FORUM ON MICROBIAL THREATS[1]

[1] Institute of Medicine Forums and Roundtables do not issue, review, or approve individual documents. The responsibility for the published workshop summary rests with the workshop rapporteurs and the institution.

[2] Forum member until February 8, 2013.

[3] Forum member since January 1, 2013.

CAROLE A. HEILMAN, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland
DAVID L. HEYMANN, Health Protection Agency, London, United Kingdom
ZHI HONG, GlaxoSmithKline, Research Triangle Park, North Carolina
PHILIP HOSBACH, sanofi pasteur, Swiftwater, Pennsylvania
STEPHEN ALBERT JOHNSTON, Arizona BioDesign Institute, Arizona State University, Tempe, Arizona
KENT KESTER, Uniformed Services University of the Health Sciences, Bethesda, Maryland
GERALD T. KEUSCH, Boston University School of Medicine and Boston University School of Public Health, Boston, Massachusetts
RIMA F. KHABBAZ, Centers for Disease Control and Prevention, Atlanta, Georgia
STANLEY M. LEMON, School of Medicine, University of North Carolina, Chapel Hill, North Carolina
MARGARET McFALL-NGAI, University of Wisconsin, Madison, Wisconsin
EDWARD McSWEEGAN, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland
MARK A. MILLER,[4] Fogarty International Center, Bethesda, Maryland
PAULA J. OLSIEWSKI,[5] the Alfred P. Sloan Foundation, New York, New York
JULIE PAVLIN, Armed Forces Health Surveillance Center, Silver Spring, Maryland
GEORGE POSTE, Complex Adaptive Systems Initiative, Arizona State University, Tempe, Arizona
DAVID RIZZO, Department of Plant Pathology, University of California, Davis, California
GARY A. ROSELLE, Veterans Health Administration, Department of Veterans Affairs, Cincinnati, Ohio
ALAN S. RUDOLPH,[6] Defense Threat Reduction Agency, Fort Belvoir, Virginia
KEVIN RUSSELL, Armed Forces Health Surveillance Center, Silver Spring, Maryland
JANET SHOEMAKER, American Society for Microbiology, Washington, DC
P. FREDERICK SPARLING, University of North Carolina, Chapel Hill, North Carolina
MURRAY TROSTLE,[7] U.S. Agency for International Development, Washington, DC
MARY E. WILSON, Harvard School of Public Health, Harvard University, Boston, Massachusetts

[4] Forum member until August 31, 2012.
[5] Forum member since January 30, 2013.
[6] Forum member until February 8, 2013.
[7] Forum member until December 31, 2012.

IOM Staff

EILEEN CHOFFNES, Scholar and Director
LEIGHANNE OLSEN, Program Officer
KATHERINE McCLURE, Senior Program Associate
REBEKAH HUTTON, Research Associate
PAMELA BERTELSON,[8] Senior Program Assistant

[8] Staff member until February 15, 2013.

BOARD ON GLOBAL HEALTH[1]

[1] Institute of Medicine boards do not review or approve individual workshop summaries. The responsibility for the content of the workshop summary rests with the workshop rapporteurs and the institution.

GUY PALMER, Regents Professor of Pathology and Infectious Diseases, Director of the School for Global Animal Health, Washington State University, Pullman, Washington

THOMAS C. QUINN, Associate Director for International Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Professor of Medicine, International Health, Epidemiology, and Molecular Biology and Immunology, Johns Hopkins University School of Medicine, Baltimore, Maryland

JENNIFER PRAH RUGER, Associate Professor, Division of Health Policy and Administration, Yale University School of Public Health, New Haven, Connecticut

IOM Staff

PATRICK KELLEY, Director

ANGELA CHRISTIAN, Program Associate

Reviewers

This workshop summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published workshop summary as sound as possible and to ensure that the workshop summary meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the process. We wish to thank the following individuals for their review of this workshop summary:

Roger G. Breeze, Lawrence Livermore National Laboratory, Livermore, California

James M. Hughes, Rollins School of Public Health, Emory University, Atlanta, Georgia

Tim Stearns, Department of Biology, Stanford University, Department of Genetics, Stanford School of Medicine, Stanford, California

Mary E. Wilson, Harvard School of Public Health, Harvard University, Boston, Massachusetts

Although the reviewers listed above have provided many constructive comments and suggestions, they did not see the final draft of the workshop summary before its release. The review of this workshop summary was overseen by **Dr.**

Melvin Worth. Appointed by the Institute of Medicine, he was responsible for making certain that an independent examination of this workshop summary was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this workshop summary rests entirely with the rapporteurs and the institution.

Acknowledgments

The Forum on Emerging Infections was created by the Institute of Medicine (IOM) in 1996 in response to a request from the Centers for Disease Control and Prevention (CDC) and the National Institutes of Health (NIH). The purpose of the Forum is to provide structured opportunities for leaders from government, academia, and industry to regularly meet and examine issues of shared concern regarding research, prevention, detection, and management of emerging, reemerging, and novel infectious diseases in humans, plants, and animals. In pursuing this task, the Forum provides a venue to foster the exchange of information and ideas, identify areas in need of greater attention, clarify policy issues by enhancing knowledge and identifying points of agreement, and inform decision makers about science and policy issues. The Forum seeks to illuminate issues rather than resolve them. For this reason, it does not provide advice or recommendations on any specific policy initiative pending before any agency or organization. Its value derives instead from the diversity of its membership and from the contributions that individual members make throughout the activities of the Forum. In September 2003, the Forum changed its name to the Forum on Microbial Threats.

The Forum on Microbial Threats and the IOM wish to express their warmest appreciation to the individuals and organizations who gave their valuable time to provide information and advice to the Forum through their participation in the planning and execution of this workshop. A full list of presenters, and their biographical information, may be found in Appendixes B and E, respectively.

The Forum gratefully acknowledges the contributions of the members of the planning committee[1]: Bruce Budowle (University of North Texas Health Science Center), Arturo Casadevall (Albert Einstein College of Medicine), Jonathan Eisen (University of California, Davis), Claire Fraser (University of Maryland School of Medicine, Baltimore), Paul Keim (Northern Arizona University), and David Relman (Stanford University).

The Forum is also indebted to the IOM staff who tirelessly contributed throughout the planning and execution of the workshop and the production of this workshop summary report. On behalf of the Forum, we gratefully acknowledge these efforts led by Dr. Eileen Choffnes, Scholar and Director of the Forum; Dr. LeighAnne Olsen, Program Officer; Katherine McClure, Senior Program Associate; Rebekah Hutton, Research Associate; and Pamela Bertelson,[2] Senior Program Assistant, for dedicating much effort and time to developing this workshop's agenda and for their thoughtful and insightful approach and skill in planning for the workshop and in translating the workshop's proceedings and discussion into this workshop summary report. We would also like to thank the following IOM staff and consultants for their valuable contributions to this activity: Daniel Bethea, Laura Harbold DeStefano, Julie Wiltshire, Theresa Wizemann, and Sarah Ziegenhorn.

Finally, the Forum wishes to recognize the sponsors that supported this activity. Financial support for this project was provided by the U.S. Department of Health and Human Services: National Institutes of Health, National Institute of Allergy and Infectious Diseases, Centers for Disease Control and Prevention, and the Food and Drug Administration; U.S. Department of Defense, Department of the Army: Global Emerging Infections Surveillance and Response System, Medical Research and Materiel Command, and the Defense Threat Reduction Agency; U.S. Department of Veterans Affairs; U.S. Department of Homeland Security; U.S. Agency for International Development; Uniformed Services University of the Health Sciences; the Alfred P. Sloan Foundation; American Society for Microbiology; sanofi pasteur; Burroughs Wellcome Fund; GlaxoSmithKline; Infectious Diseases Society of America; and the Merck Company Foundation. The views presented in this workshop summary are those of the workshop participants and have been summarized by the rapporteurs. They do not necessarily reflect the views of the Forum on Microbial Threats, its sponsors, or the IOM.

[1] Institute of Medicine planning committees are solely responsible for organizing the workshop, identifying topics, and choosing speakers. The responsibility for the published workshop summary rests solely with the workshop rapporteurs and the institution.

[2] Staff member until February 15, 2013.

Contents

Tables, Figures, and Boxes

TABLES

FIGURES

BOXES

Workshop Overview[1]

THE SCIENCE AND APPLICATIONS OF MICROBIAL GENOMICS: PREDICTING, DETECTING, AND TRACKING NOVELTY IN THE MICROBIAL WORLD

Over the past several decades, new scientific tools and approaches for detecting microbial species have dramatically enhanced our appreciation of the diversity and abundance of the microbiota and its dynamic interactions with the environments within which these microorganisms reside. The first bacterial genome[2] was sequenced in 1995 and took more than 13 months of work to complete. Today (2012), a microorganism's entire genome can be sequenced in a few days. Much as our view of the cosmos was forever altered in the 17th century with the invention of the telescope (Nee, 2004), these genomic technologies, and the observations derived from them, have fundamentally transformed our appreciation of the microbial world around us.

Nucleic acid sequencing technologies now provide access to the previously "unculturable"—and thus, undetected—microorganisms that comprise the majority of microbial life. Rapid and inexpensive sequencing platforms make it

[1] The planning committee's role was limited to planning the workshop, and the workshop summary has been prepared by the workshop rapporteurs (with the assistance of Pamela Bertelson, Rebekah Hutton, and Katherine McClure) as a factual summary of what occurred at the workshop. Statements, recommendations, and opinions expressed are those of individual presenters and participants, and are not necessarily endorsed or verified by the Institute of Medicine, and they should not be construed as reflecting any group consensus.

[2] For the purposes of this summary, the genome is defined as the complete set of genetic information in an organism. In bacteria, this includes the chromosome(s) and plasmids (extrachromosomal DNA molecules that can replicate autonomously within a bacterial cell) (Pallen and Wren, 2007).

commonplace to sort through the genomes of dozens of strains of a single microbial species or to conduct "metagenomic" analyses of vast communities of the microbiota from a wide variety of environments. These technical advancements and concurrent investments in the fields of microbial ecology, evolution, forensics, and epidemiology have transformed our ability to use genomic sequence information to explore the origins, evolution, and catalysts associated with historical, emergent, and reemergent disease outbreaks. The ability to "read" the nucleic acid sequence of microbial genomes has provided important insights into this previously hidden, unculturable world by revealing the vast diversity and complexity of microbial life around us, and their myriad interactions with their abiotic and biotic environmental niches.

Recent examples of the use of "whole genome" sequencing to investigate outbreaks of emerging, reemerging, and novel infectious diseases illustrate the potential of these methods for enhancing disease surveillance, detection, and response efforts. Using slight sequence differences between isolates to discriminate between closely related strains, investigators have tracked the evolution of isolates in a disease outbreak, traced person-to-person transmission of a communicable disease, and identified point sources of disease outbreaks. When genomic information about related strains or past disease outbreaks is available, the genome sequence of outbreak strains has proved useful in identifying factors that may contribute to the emergence, virulence, or spread of pathogens, as well as in speeding diagnostic tool development. In a recent development, fast genome sequencing was used to halt the spread of a methicillin-resistant *Staphylococcus aureus* (MRSA) infection in a neonatal ward in a hospital in Cambridge, United Kingdom (Harris et al., 2012)

Statement of Task

On June 12 and 13, 2012, the Institute of Medicine's (IOM's) Forum on Microbial Threats convened a public workshop in Washington, DC, to discuss the scientific tools and approaches being used for detecting and characterizing microbial species, and the roles of microbial genomics and metagenomics to better understand the culturable and unculturable microbial world around us.[3] Through invited presentations and discussions, participants examined the use of microbial genomics to explore the diversity, evolution, and adaptation of microorganisms in

[3] A public workshop will be held to explore new scientific tools and methods for detecting and characterizing microbial species and for understanding the origins, nature, and spread of emerging, reemerging, and novel infectious diseases of humans, plants, domestic animals, and wildlife. Topics to be discussed may include microbial diversity, evolution, and adaptation; microbial genomic, epidemiology, and forensic tools and technologies; infectious disease detection and diagnostic platforms in clinical medicine, veterinary medicine, plant pathology, and wildlife epidemiology; development of microbial genomic and proteomic databases; and strategies for predicting, mitigating, and responding to emerging infectious diseases.

a wide variety of environments; the molecular mechanisms of disease emergence and epidemiology; and the ways that genomic technologies are being applied to disease outbreak trace back and microbial surveillance. Points that were emphasized by many participants included the need to develop robust standardized sampling protocols, the importance of having the appropriate metadata (e.g., the sequencing platform used, sampling information, culture conditions), data analysis and data management challenges, and information sharing in real time.

Organization of the Workshop Summary

This workshop summary was prepared by the rapporteurs for the Forum's members and includes a collection of individually authored papers and commentary. Sections of the workshop summary not specifically attributed to an individual reflect the views of the rapporteurs and not those of the members of the Forum on Microbial Threats, its sponsors, or the IOM. The contents of the unattributed sections of this summary report provide a context for the reader to appreciate the presentations and discussions that occurred over the 2 days of this workshop.

The summary is organized into sections as a topic-by-topic description of the presentations and discussions that took place at the workshop. Its purpose is to present information from relevant experience, to delineate a range of pivotal issues and their respective challenges, and to offer differing perspectives on the topic as discussed and described by the workshop participants. Manuscripts and reprinted articles submitted by some but not all of the workshop's participants may be found, in alphabetical order, in Appendix A.

Although this workshop summary provides a distillation of the individual presentations, it also reflects an important aspect of the Forum's philosophy. The workshop functions as a dialogue among representatives from different sectors and disciplines and allows them to present *their* views about which areas, in their opinion, merit further study. This report only summarizes the statements of participants over the course of the workshop. This summary is not intended to be an exhaustive exploration of the subject matter, nor does it represent the findings, conclusions, or recommendations of a consensus committee process.

GLIMPSES OF THE MICROBIAL WORLD

Microbiologists investigate a largely hidden world, laboring to understand the structure and function of organisms that are essentially invisible to the naked eye. Critical methodological advances—from microscopy through metagenomics—have made the staggering diversity of the microbial worlds on this planet easier to study and have brought them into focus (Table WO-1). Over the past several centuries, these approaches have provided ever-expanding views of the extraordinary organismal, metabolic, and environmental diversity of microorganisms.

TABLE WO-1 Some Major Methods for Studying Individual Microbes Found in the Environment

Method	Summary	Comments
Microscopy	Microbial phenotypes can be studied by making them more visible. In conjunction with other methods, such as staining, microscopy can also be used to count taxa and make inferences about biological processes.	The appearance of microbes is not a reliable indicator of what type of microbe one is looking at.
Culturing	Single cells of a particular microbial type are grown in isolation from other organisms. This can be done in liquid or solid growth media.	This is the best way to learn about the biology of a particular organism. However, many microbes are uncultured (i.e., have never been grown in the lab in isolation from other organisms) and may be unculturable (i.e., may not be able to grow without other organisms).
rRNA-PCR	The key aspects of this method are the following: (a) all cell-based organisms possess the same rRNA genes (albeit with different underlying sequences); (b) PCR is used to make billions of copies of basically each and every rRNA gene present in a sample; this amplifies the rRNA signal relative to the noise of thousands of other genes present in each organism's DNA; (c) sequencing and phylogenetic analysis places rRNA genes on the rRNA tree of life; the position on the tree is used to infer what type of organism (a.k.a. phylotype) the gene came from; and (d) the numbers of each microbe type are estimated from the number of times the same rRNA gene is seen.	This method revolutionized microbiology in the 1980s by allowing the types and numbers of microbes present in a sample to be rapidly characterized. However, there are some biases in the process that make it not perfect for all aspects of typing and counting.
Shotgun genome sequencing of cultured species	The DNA from an organism is isolated and broken into small fragments, and then portions of these fragments are sequenced, usually with the aid of sequencing machines. The fragments are then assembled into larger pieces by looking for overlaps in the sequence each possesses. The complete genome can be determined by filling in gaps between the larger pieces.	This has now been applied to over 1,000 microbes, as well as some multicellular species, and has provided a much deeper understanding of the biology and evolution of life. One limitation is that each genome sequence is usually a snapshot of one or a few individuals.

TABLE WO-1 Continued

Method	Summary	Comments
Metagenomics	DNA is directly isolated from an environmental sample and then sequenced. One approach to doing this is to select particular pieces of interest (e.g., those containing interesting rRNA genes) and sequence them. An alternative is ESS, which is shotgun genome sequencing as described above, but applied to an environmental sample with multiple organisms, rather than to a single cultured organism.	This method allows one to sample the genomes of microbes without culturing them. It can be used both for typing and counting taxa and for making predictions of their biological functions.

SOURCE: Eisen (2007).

There are three recognized domains of life: the Archaea, the bacteria, and the eukarya. Microorganisms are now recognized as the primary source of diversity for life on Earth and its inhabitants (Figure WO-1). Even more astonishing, perhaps, is what still remains to be discovered about the microbiota on this planet. As Fraser et al. (2000) have observed, "The genetic, metabolic and physiological diversity of microbial species is far greater than that found in plants and animals. *The diversity of the microbial world is largely unknown, with less than one-half of 1% of the estimated 2–3 billion microbial species identified* [emphasis added]." Moreover, while there are well over 10 million species of "known" bacteria only a few thousand have been formally described (Eisen, 2007). With the advent of genomic technologies, we are entering a new era of scientific discovery that holds great promise for revealing the breadth of diversity and depth of complexity inherent to the microbial world.

From Animalcules to Germs

Until just over 300 years ago, the microscopic world that we share the planet with was largely unseen and unknown. In the 17th century, Antonie van Leeuwenhoek provided the first detailed glimpses of the "animalcules" in the microbial world when he developed viewing techniques and magnifying lenses with sufficient power to see microorganisms. Van Leeuwenhoek obtained these organisms, as illustrated in Figure WO-2, from a variety of environmental sources, ranging from rain and pond water to plaque biofilms scraped from teeth. Their simple morphologies prevented the precise identification and classification of these organisms, but through detailed descriptions and illustrations in his letters to the Royal Society of England, van Leeuwenhoek brought the invisible world of microscopic life forms to the attention of scientists (Handlesman, 2004).

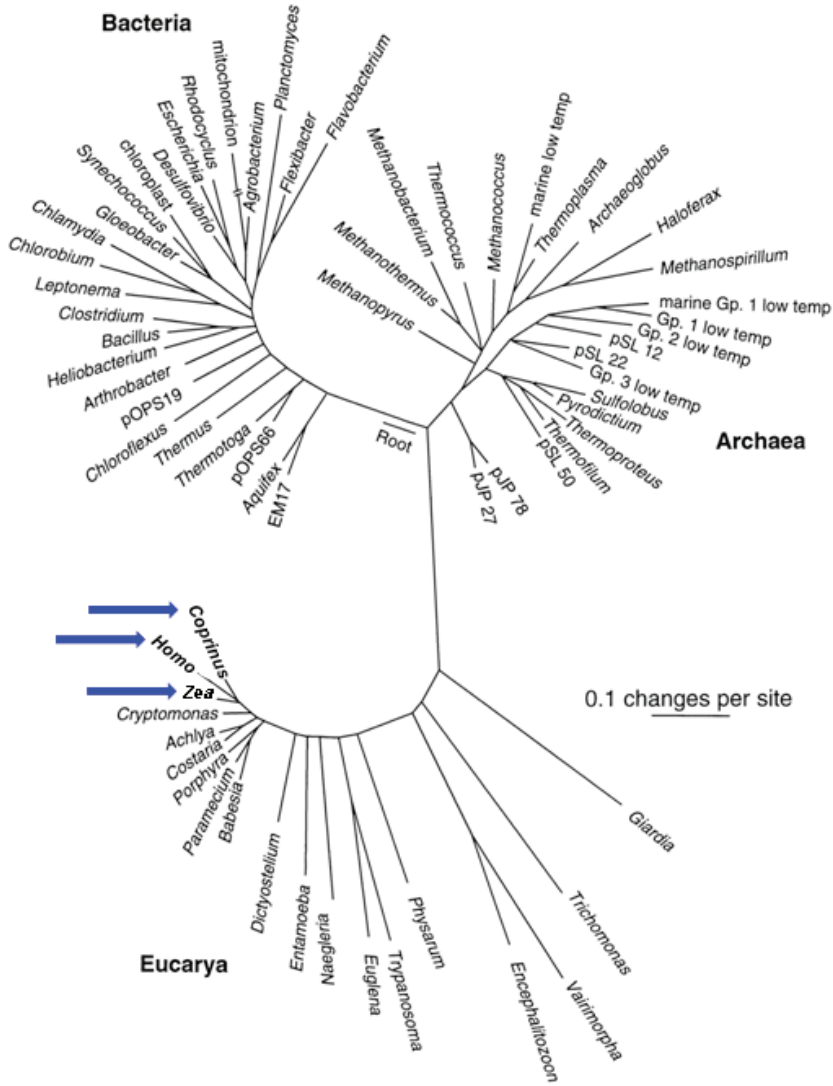


FIGURE WO-1 Universal tree of life based on a comparison of nucleic acid (RNA) sequences found in all cellular life (small subunit ribosomal RNA). "A sobering aspect of large-scale phylogenetic trees such as that shown in Figure WO-1 is the graphical realization that most of our legacy in biological science, historically based on large organisms, has focused on a narrow slice of biological diversity. Thus, we see that animals (represented by *Homo*), plants (*Zea*), and fungi (*Coprinus*) (see blue arrows) constitute small and peripheral branches of even eukaryotic cellular diversity" (Cracraft and Donoghue, 2004). NOTE: The scale bar corresponds to 0.1 changes per nucleotide position.
SOURCE: From Pace, N. R. 1997. A Molecular View of Microbial Diversity and the Biosphere. *Science* 276:734-740. Reprinted with permission from AAAS.

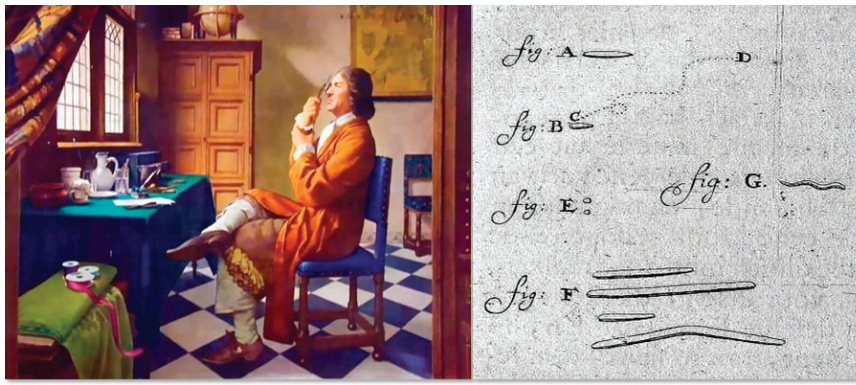


FIGURE WO-2 First glimpses of the microbial world. Panel A, Antonie van Leeuwenhoek was probably the first person to observe live microorganisms. Panel B, van Leeuwenhoek's drawings of "animalcules" from the human mouth.
SOURCE: Dobell (1932).

Careful observation of microorganisms by scientists such as Louis Pasteur revealed the connections between microorganisms and practical phenomena. The production of beer and vinegar, for example, depended upon the presence of yeast for the conversion of sugar to alcohol and the fermentation of alcohol into acetic acid, respectively. Until the development of standardized culturing techniques in the late 19th century researchers could do little more than observe these creatures as a mixture of organisms in complex matrices. Pasteur also examined the connections between microorganisms and diseases of plants, animals, and humans, becoming an early proponent of the "germ theory" of disease (de Kruif, 1926).

In 1884, Robert Koch and Friedrich Loeffler formalized the germ theory of disease by outlining a series of tests designed to determine whether a specific microorganism was the causative agent of a specific disease. These tests, known as Koch's postulates (Box WO-1), required the isolation and propagation of "pure cultures" of microorganisms. Koch initially applied these tests to establish the infectious etiology of anthrax and tuberculosis (de Kruif, 1926). Using these techniques, researchers could conduct experimental investigations of specific microorganisms under controlled conditions.

Our current understanding of microbe–host interactions have been influenced by more than a century of research, sparked by the germ theory of disease and rooted in historic notions of contagion that long preceded the research and intellectual syntheses of Pasteur and Koch in the 19th century (Lederberg, 2000). The success of this approach to the identification of the microbial basis of disease launched generations of "microbe hunters" who began a systematic search for disease-causing microbes that could be isolated and cultured under controlled laboratory conditions. Their work set a new course for the study and treatment of infectious disease-causing organisms. The "power and precision" of their studies

BOX WO-1
Koch's Postulates

1. The parasite occurs in every case of the disease in question and under circumstances that can account for the pathological changes and clinical course of the disease.
2. The parasite occurs in no other disease as a fortuitous and nonpathogenic parasite.
3. After being fully isolated from the body and repeatedly grown in pure culture, the parasite can induce the disease anew.

SOURCES: Fredericks and Relman (1996), Koch (1891), and Rivers (1937).

using pure culture established these methods as the standard laboratory microbiology technique (Lederberg, 2000). At the same time, this disease-centric approach to microbe discovery has, for the past century and a half, not only influenced our collective perceptions of what microbes do "to" rather than "for" their hosts but also biased the database of the tree of life to one that, until relatively recently, has been focused almost entirely on disease-causing, culturable microorganisms.

This pathogen-centric bias attributed disease entirely to the actions of invading microorganisms, thereby drawing battle lines between "them" and "us," the injured hosts (Casadevall and Pirofski, 1999). Although it was recognized in Koch's time that some microbes did not cause disease in previously exposed hosts (e.g., milk maids who had been exposed to cowpox did not become infected with smallpox), the fact that his postulates could not account for microbes that did not cause disease in all hosts was not generally appreciated until the arrival of vaccines and the subsequent introduction of immunosuppressive therapies in the 20th century (Casadevall and Pirofski, 1999; Isenberg, 1988). By then, the paradigm of the systematized search for the microbial basis of disease, followed by the development of antimicrobial and other therapies to eradicate these pathogenic agents, had been firmly established in clinical practice.

THE CULTIVATION BOTTLENECK, GENOMICS, AND THE UNIVERSAL TREE OF LIFE

In the 1950s and 1960s this focus on a few easily cultured organisms produced an explosion of information about microbial physiology and genetics that overshadowed efforts to understand the ecology and diversity of the microbial world (Pace, 1997). As the workhorses of the emerging field of molecular biology bacteria, such as *Escherichia coli* and *Bacillus subtilis* and their viruses (bacteriophages) became perhaps some of the best characterized microorganisms

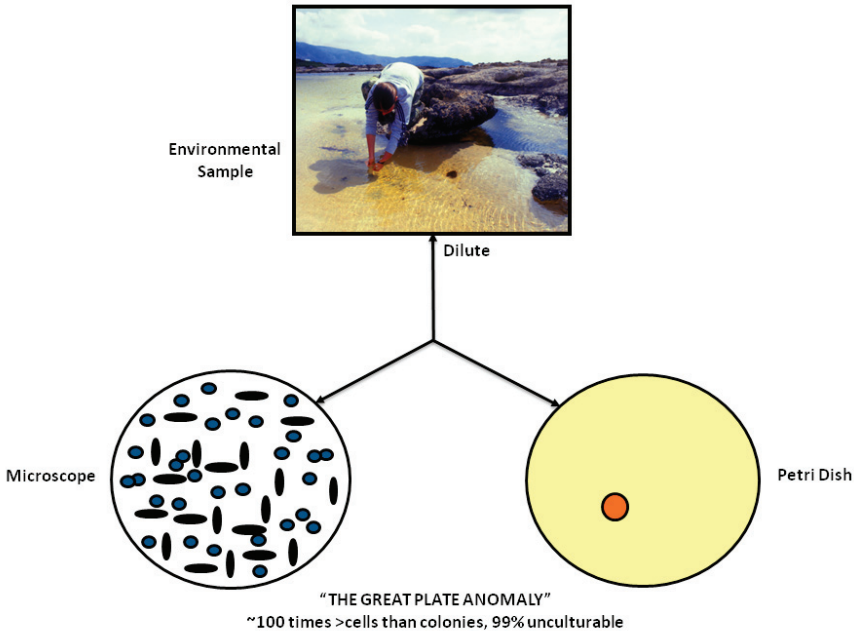


FIGURE WO-3 The great plate count anomaly.
SOURCE: Lewis (2011). Figure by Kim Lewis, Courtesy of Moselio Schaechter, Small things Considered, The Microbe Blog.

in biological research. While a rich source of discovery and knowledge, this focus on readily cultured organisms limited most researchers' appreciation of the diversity and ubiquity of microbial life.

The predisposition toward discovery, isolation, and characterization of microorganisms that could be readily cultured[4] in the laboratory is known as the "cultivation bottleneck" and is evident in the substantial difference in population counts of microorganisms present in a sample depending on whether they are conducted using microscopy or culturing techniques—a phenomenon known as the "great plate anomaly" (see Figure WO-3). This difference is attributed to the fact that the vast majority of microorganisms, 99 percent by some estimates, cannot be isolated and cultured[5] using standard laboratory techniques (Handelsman, 2004).

[4] The ease of isolation and culturing of certain organisms reflects an organism's ability to grow rapidly into colonies on high-nutrient artificial growth media, typically under aerobic conditions. This had led some to characterize these species as the "weeds" of the microbial world (Hugenholtz, 2002).

[5] Microorganisms may be unculturable because of the inability to replicate important nutritional or environmental requirements for growth, including the services provided by other microorganisms that may be present in natural settings.

SEQUENCE-BASED DETECTION AND DISCOVERY

Pace and colleagues (1985) used sequence-based methods to investigate the composition of all constituents of the microbial biosphere. These culture-independent surveys led to the discovery of previously unknown and diverse lineages of organisms from habitats across the Earth, including bacterial and parasitic pathogens in the human body (Handelsman, 2004; Pace, 1997; Relman et al., 1990; Santamaria-Fries et al., 1996). The polymerase chain reaction[6] (PCR) technique, developed in 1983 by Kary Mullis, aided these studies by allowing researchers to easily amplify single copies of a particular DNA sequence into thousands or millions of copies. This advance enabled investigators to rapidly and comprehensively catalog the diversity of life forms in the microbial world. Initial molecular phylogeny studies demonstrated that this "unseen world" of microorganisms could be studied and confirmed that the number of organisms represented in the unculturable world far exceeded the size of the culturable world.

While culture-based techniques remain the gold standard for disease detection, outbreak investigations, and infectious disease epidemiology, over the past several decades a range of sequence-based methods—including broad-range PCR, high-throughput sequencing technologies, microarrays, and shotgun metagenomics—have been applied to improve the detection and discovery of pathogens and other microorganisms. rRNA gene sequences may also be used to phylogenetically identify microbes that are otherwise uncharacterizable by other methods and approaches.

Broad-Range PCR

Some conserved genes and their encoded molecules have properties that render them useful as "molecular clocks." These conserved genes, such as the 16S rRNA gene in bacteria, can be amplified from any member of a phylogenetic group using consensus primers.[7] The sequences of the amplified, intervening gene regions with variable composition are then determined, in order to identify known or previously uncharacterized members of the group, and their evolutionary relationships to all other organisms revealed. This approach has been used to discover previously uncharacterized bacterial, viral, and parasitic pathogens (Nichol et al., 1993; Relman, 1993, 1999, 2011; Relman et al., 1990, 1992).

[6] The polymerase chain reaction (PCR) is a biochemical technology in molecular biology that amplifies a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

[7] Primers whose sequences are found in all known, and presumably unknown, members of the group.

High-Throughput Sequencing Technologies[8]

Nucleic acid sequencing technologies have dramatically enhanced our understanding of the diversity of the microbiota and their dynamic interactions with the environments they reside in. The genomes of thousands of organisms from all three domains of life, as well as those of quasi-life forms such as viruses, have been sequenced. Metagenomics has taken this approach a step further by cataloging the genomic components of microbes living in complex environmental matrices, from soil samples, to the ocean, coral reefs, and the human body (Mardis, 2008). The conventional or first-generation technology of automated Sanger sequencing produced all of the early microbial sequence data. Next-generation[9] sequencing technologies, which were introduced in 2005, have decreased the cost and time necessary for sequence production.

Sequence data have been used for a number of applications, including:

- *De novo* assembly of entire genomes to produce primary genetic sequences and to support the detailed genetic analysis of an organism.
- Whole genome "resequencing" for the discovery of variants that differ in sequence to known genome sequences of a closely related strain.
- Species classification and the identification of predicted coding sequences and novel gene discovery in genomic surveys of microbial communities (metagenomics).
- "Seq-based" assays that determine the sequence content and abundance of mRNAs, non-coding RNAs, and small RNAs (collectively called RNA-seq); or measure genomewide profiles of DNA-protein complexes (ChIP-seq), methylation sites (methyl-seq), and DNase I hypersensitivity sites (DNase-seq) (Metzker, 2010).

Microarrays

Microarray technology runs the gamut from assays that contain hundreds to those containing millions of probes. Probes can be designed to distinguish differences in sequence variation that allow for pathogen speciation, or to detect thousands of agents across the tree of life. Arrays comprising longer probes (e.g., > 60 nucleotides) are more tolerant of sequence mismatches and may detect agents that have only modest similarity to those already known. Two longer probe array platforms are in common use for viral detection and discovery: the

[8] These are large-scale methods to purify, identify, and characterize DNA, RNA, proteins, and other molecules. These methods are usually automated, allowing rapid analysis of very large numbers of samples. http://www.learner.org/courses/biology/glossary/through_put.html (accessed November 13, 2012).

[9] As more advanced technologies are introduced, these technologies are sometimes referred to as "second generation" technologies. Nearly all current sequencing is "next generation" (i.e., not Sanger methodology).

GreeneChip and the Virochip. Although they differ in design, both employ random amplification strategies to allow a relatively unbiased detection of microbial targets.

Shotgun Metagenomics

In 1995, The Institute for Genomic Research (TIGR) used a "shotgun sequencing" strategy coupled with Sanger sequencing and advanced bioinformatics methods to produce the first whole genome sequence of a free-living organism, *Haemophilus influenzae*[10] (Fleischmann et al., 1995). Shotgun sequencing refers to the fragmentation of an organism's genome into small pieces that can then be sequenced in parallel using automated sequencing platforms. It is now used routinely for producing whole genome sequences. Individual sequence fragments are then additively assembled into larger units (known as "contigs") of the genome. The resulting "draft" typically represents more than 99 percent of the genome (Pallen and Wren, 2007). Draft sequence data may be sufficient for surveying species and metabolic diversity in communities of microorganisms that cannot be grown in culture, or for comparative studies if a complete sequence is available for a closely related strain or species and can be used to order and orient contigs (Fraser et al., 2002).

Finishing a genome-sequencing project is a costly and time-consuming process in which gaps in the assembly are closed and sequence errors are resolved. For this reason, many sequences are left in draft form (MacLean et al., 2009). Finished sequences provide complete genomic information, including the overall organization of a genome and the presence of particular genes on plasmids versus chromosomes (Fraser et al., 2002).

Improvements in sequencing methods and the development of automated systems have contributed to significant decreases in the cost and time it takes to produce a completed genome. The genome of *Haemophilus influenzae* Rd required 13 months of work. Today, draft bacterial genomes can be sequenced in days. In addition, the cost of sequencing the human genome has dropped by three orders of magnitude, from about $1 million per genome to about $1,000 (JASON, 2010; Figure WO-4). Over the past several decades these advances have led to a proliferation of genome sequencing projects of bacteria, eukaryotes, and of entire microbial communities (metagenomes) that have resulted in a number of completed genomes for a variety of microorganisms (Figure WO-5).

[10] The *Haemophilus influenzae* genome was selected for its genome size (1.8 million base pairs), which was typical for bacteria, its G + C base composition (38 percent) was close to that of the human genome, and the fact that a physical clone map did not exist.

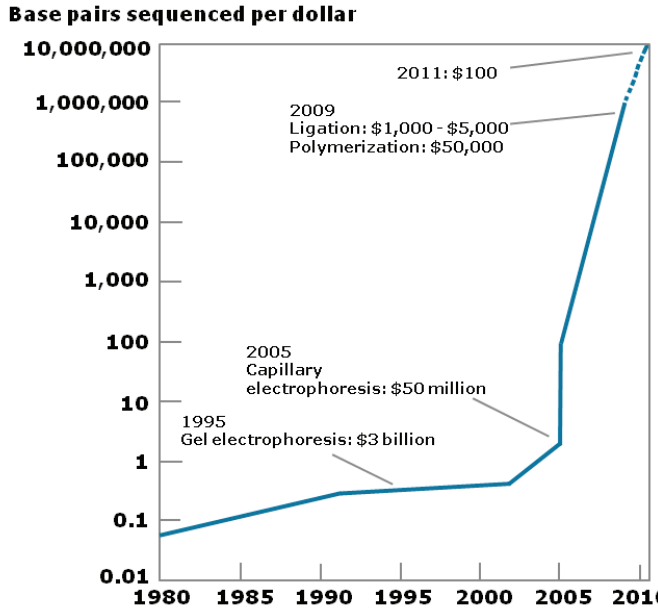


FIGURE WO-4 The improvements in DNA sequencing efficiency over time. Costs excludes equipment and personnel.
SOURCE: JASON (2010).

Viral Diversity Discovery

Studies of viral diversity and genomics have only recently come into their own. Because there is no single gene that is common to all viral genomes, "total uncultured viral diversity cannot be monitored using approaches analogous to ribosomal DNA profiling" (Edwards and Rohwer, 2005). The introduction of high-throughput sequencing and metagenomic analyses are now providing insights into the composition and diversity of cultured viral species and environmental viral communities. These analyses are still limited by current capacity to match sample sequences to sequences stored in databases, but the initial efforts have demonstrated that we have only begun to scratch the surface of virus discovery (Lipkin, 2010; Figure WO-6).

MICROBIOLOGY IN THE POST-GENOMIC ERA

As of mid-2011, complete genome sequences had been published for 1,554 bacterial species (the majority of which are pathogens), 112 archaeal species, and 2,675 virus species. Within these species, sequences exist for tens of thousands of strains; there are approximately 40,000 strains of flu viruses and more than

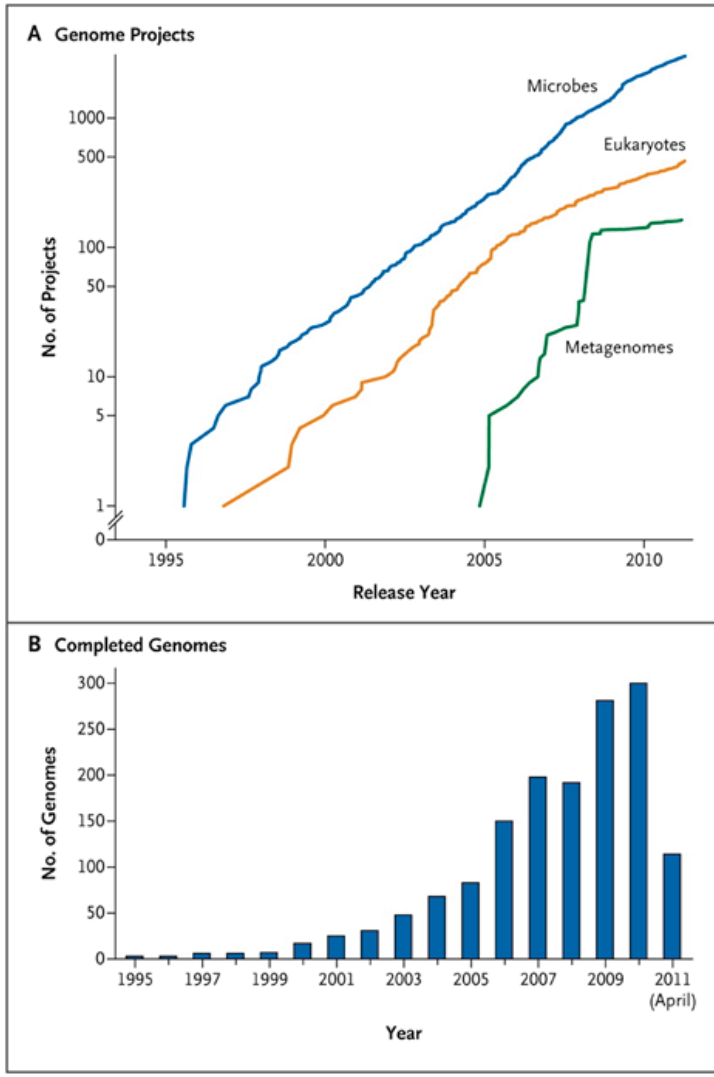


FIGURE WO-5 Genome projects and complete genomes since 1995. Panel A shows a cumulative plot of the number of genome projects (involving microbial [bacterial and archaeal], eukaryotic, and viral genomes) and metagenome projects, according to the release year at the National Center for Biotechnology Information since 1995. Panel B shows the number of completed microbial genome sequences according to year (the most recent data were collected on April 21, 2011).
SOURCE: Relman (2011). From *The New England Journal of Medicine*, David A. Relman, Microbial Genomics and Infectious Diseases, 365, 347-357. Copyright © 2011 Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

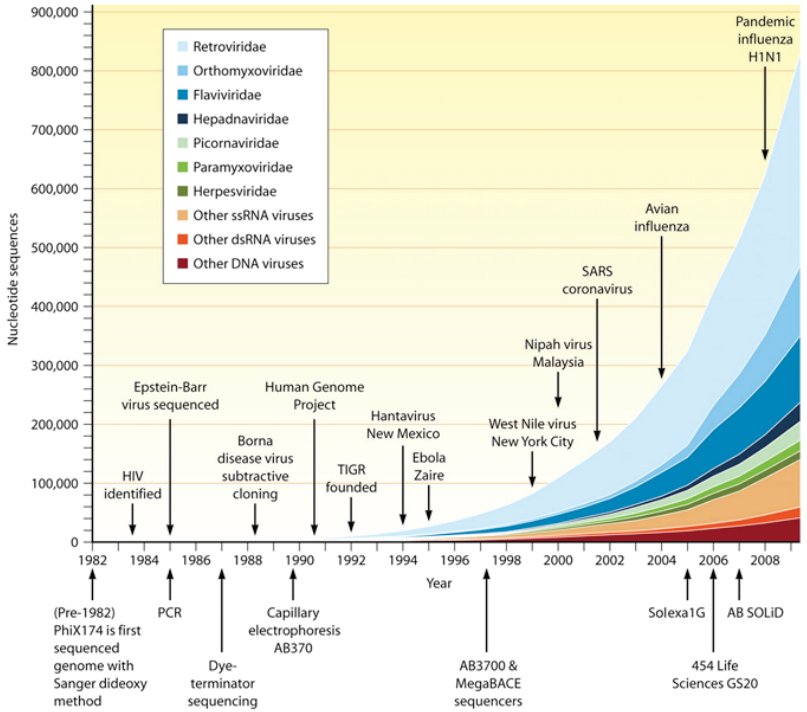


FIGURE WO-6 Growth of the viral sequence database mapped to seminal discoveries and improvements in sequencing technology.
SOURCE: (2010) Copyright © American Society for Microbiology, Lipkin, W. I.(2010). Microbe Hunting. *Microbiology and Molecular Biology Reviews* 74 (3):363-377:doi:10.1128/MMBR.00007-10. Reproduced with permission from American Society for Microbiology.

300,000 strains of HIV, for example (Relman, 2011). As the fidelity and resolution of nucleic acid sequencing technologies have improved, so has the ability of investigators to explore the diversity and predicted function of microorganisms and the composition and dynamics of the communities they form. These advances offer the hope that we can one day channel some of the activities of microorganisms for improvements to the health and well-being of plants, animals, humans, and ecosystems.

USE OF WHOLE GENOME SEQUENCING IN OUTBREAK INVESTIGATIONS

Recent examples, discussed below, of the use of whole genome sequencing to investigate outbreaks of emerging, reemerging, and novel infectious diseases

illustrate the potential of these methods for enhancing disease surveillance, detection, and response efforts. Using slight sequence differences between isolates to discriminate between closely related strains, investigators have tracked the evolution of isolates in a disease outbreak; traced person-to-person transmission; and identified point sources of disease outbreaks. When genomic information about related strains or past disease outbreaks is available, the genome sequence of outbreak strains has proved useful in identifying factors that may contribute to the emergence, virulence, or spread of pathogens, as well as in speeding diagnostic tool development. For example:

- Investigators used genomic sequencing to investigate, and find the source for, the cholera outbreak in Haiti in 2010, a disease that had been absent from the island of Haiti for almost a century. Twenty-four *Vibrio cholerae* isolates from Nepal were found to belong to a single monophyletic group that also contained isolates from Bangladesh and Haiti. These findings (Hendriksen et al., 2011) supported the epidemiological conclusion that cholera was introduced into Haiti by soldiers from Nepal, who served as United Nations' peacekeepers in the aftermath of the 2010 earthquake (Chin et al., 2011; Frerichs et al., 2012; Piarroux et al., 2011).
- The Black Death, which swept through Europe in the 14th century, was one of the most devastating pandemics in human history. In order to investigate the origins of this pandemic, investigators compared the genomes of today's bubonic plague bacteria (*Y. pestis*), obtained from plague-endemic countries, to "plague" obtained from victims who were buried in mass graves in the 14th century. These investigations were able to confirm that *Y. pestis* was the cause of the Black Death and that it originated from China, more than 1,000 years ago (Bos et al., 2011; Haensch et al., 2010; Morelli et al., 2010).
- Some strains of MRSA are resistant to almost all commonly available antibiotics. Through sequencing and comparing the genomes of MRSA, researchers have been able to trace the origins of this "superbug" to Europe in the 1960s, tracked its global spread, and established a previously unknown link among five patients from a single hospital in Thailand (Harris et al., 2010).
- The 2011 European outbreak of *E. coli* O104:H4 (discussed by Pallen on pages 86-87) was the deadliest outbreak of food poisoning on record. Thousands were sickened and more than 50 died, many due to a deadly complication of this food-borne infection that can lead to hemolytic uremic syndrome. Comparison of the genomic sequences of the outbreak strain and 11 related strains of *E. coli* revealed the presence of an unusual combination of virulence factors, which may help to account for the high frequency of hemolytic uremic syndrome associated with this outbreak (Scheutz et al., 2011).

- A 2011 outbreak of a highly drug-resistant strain of *Klebsiella pneumoniae* proved extremely difficult to treat. By comparing the genome of the outbreak strain to the genomes of 300 previously isolated strains of *K. pneumoniae*, researchers were able to identify a stretch of DNA that was unique to the outbreak strain. These sequences were then used to develop a rapid diagnostic test for screening patients for this dangerous pathogen (Kumarasamy et al., 2010).

Microbes and Human History

The workshop opened with keynote remarks by Paul Keim of Northern Arizona University, who observed that we are moving toward studying microbial diversity on unprecedented scales, using novel methods that we have never had before (Dr. Keim's contribution to the workshop summary report may be found in Appendix A, pages 207-229). According to Keim, our understanding of microbial diversity has been severely biased because of our inability to culture the vast majority of microorganisms. This means that what we know about microorganisms, and microbiology generally, comes from a very, very, small subset of the microbial universe. Moreover, we have a very anthropocentric view of the microbial world and tend to focus on those microorganisms that cause illness or death in people. Non-human disease reservoirs are very important in disease ecology, but they are often difficult to identify and study because of their sometimes cryptic and transient nature within their "host" environments—making sampling extremely difficult. It is hoped that the use of whole genome sequencing will expand our understanding of the evolution and population structure of all microbes, including pathogens.

Keim discussed *Yersinia pestis*, the causative agent of plague,[11] as an example of how the emergence of a highly fit microbial clone can alter human history, and how population genetics can help us understand this disease. The dogma has been that there were three major pandemics of plague (reviewed in Perry and Fetherston, 1997, and illustrated in the plague map in Figure WO-7, Morelli et al., 2010).

The first—the "Plague of Justinian"—spread across the eastern Mediterranean and parts of the Middle East and Central Asia from AD 547 to 767, decimating the Byzantine Empire with population losses estimated to be 50 to 60 percent. The second pandemic, referred to as "the Black Death," began in the Middle Ages and persisted into the 19th century, spanning North Africa, Europe, and parts of Asia. Keim noted that an estimated 17 to 28 million people, or 30 to 40 percent of the European population, died as a result of successive waves of this pandemic. The third pandemic began in the late 1850s and continues to this day. Starting in China and initially spread by steamships, this pandemic has been

[11] The Black Death.

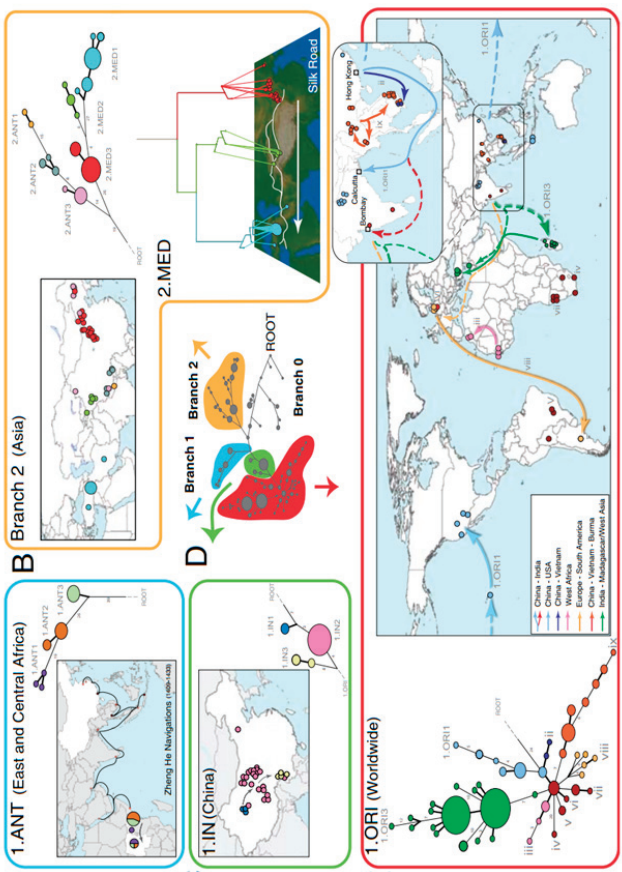


FIGURE WO-7 Postulated routes of spread. Panel A: Sources of 1.ANT isolates in Central Africa and routes of navigation (black lines) by Zheng He from China to Africa in 1409–1433. Panel B: Upper left—sources of isolates from branch 2 superimposed on a map of Asia. Lower right—sources of 2.MED isolates superimposed on a topographical map of the Silk Road (white lines; 200 BC–AD 1400). Panel C: Direction of spread of 1.IN nodes within China from northwest to south. Panel D: A condensed MSTree. Panel E: Postulated routes of migration of 1.ORI since 1894. ii–ix—Radiations with few isolates.
SOURCE: Morelli et al. (2010). Reprinted by permission from Macmillan Publishers Ltd: NATURE GENETICS. Morelli, G., Y. et al. 2010. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics* 42(12):1140-1143, copyright 2010.

responsible for millions of deaths worldwide. Modern hygiene (e.g., rat control) and antibiotics have largely controlled—but have not eradicated—this pandemic.

Plague Ecology

Basic plague ecology involves a bacterial pathogen, *Y. pestis*, that moves back and forth between a warm-blooded host (almost always rodents) via an arthropod flea vector. On a larger scale, Keim explained that plague ecology involves different hosts and different vectors at different times (Figure WO-8).

Y. pestis continues to evolve out of sight, for decades or even centuries, in a "reservoir" or "cryptic" phase called the enzootic cycle. Sampling during an epizootic cycle or during human pandemics provides evidence for the changes occurring in the reservoir phase. Outbreaks of plague in other "indicator" species (generally rodents) occur during epizootic cycles. Other species, including humans, are also part of the complex ecology of plague. Phenotypic manifestation in humans can be bubonic, septicemic, or pneumonic. Pneumonic plague is highly contagious via respiratory aerosols. Study of the enzootic cycle is extremely difficult; however, sampling during an epizootic cycle or during human pandemics provides evidence for the changes occurring in the reservoir phase.

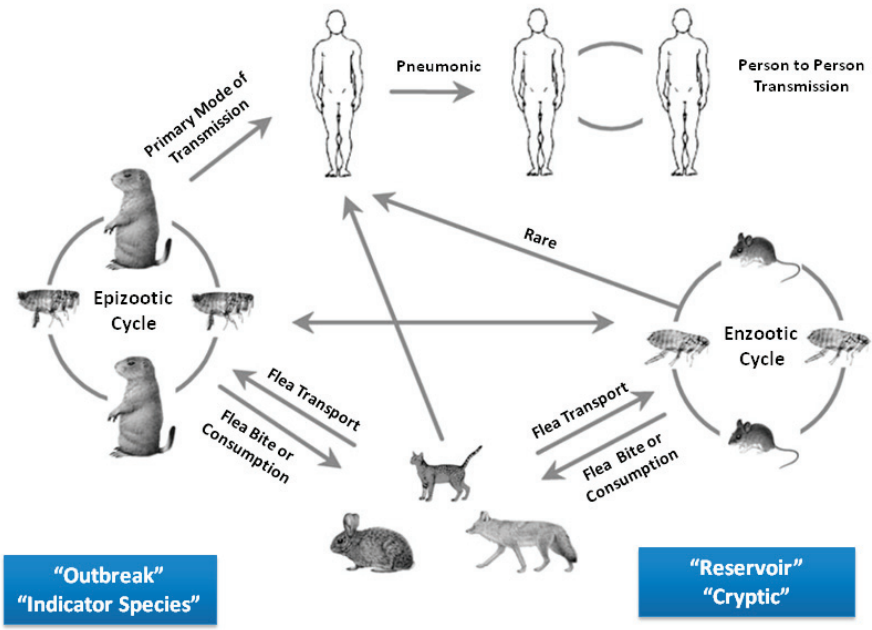


FIGURE WO-8 Plague ecology.
SOURCE: Gage and Kosoy (2005).

Y. pestis *Pathogenomics*

A 1951 publication by Devignat first linked the three historical pandemics with three different metabolic phenotypes, or biovars of *Y. pestis* (antigua, mediaevalis, and orientalis) defined by their ability to ferment glycerol and reduce nitrate (Devignat, 1951). These phenotypes are the result of successive losses of function and, as discussed below, there is no real concordance with phylogenetic[12] information that is now available.

Y. pestis is a relatively young, recently emerged, organism. Single nucleotide polymorphisms (SNPs)[13] in the core genome are better than 99.9 percent conserved, consistent with clonal propagation. Keim explained that *Y. pestis* generates diversity by accumulating mutations in a sequential fashion over time. One can select for these mutations in order to assemble a phylogenetic reconstruction of the organism's history.

Keim cited the collaborative work of Achtman et al. (2004) and Morelli et al. (2010), who used whole genome sequencing and SNP typing to develop a phylogeny for *Y. pestis*. Their analysis demonstrated that *Y. pestis* emerged from *Y. pseudotuberculosis* and acquired new genes in order to become a highly fit clone (Achtman et al., 2004; Morelli et al., 2010). Instead of the antigua, mediaevalis, and orientalis biovar structure, they offer a new type of structure that provides a detailed, high-resolution population genetics map of *Y. pestis* based on an analysis of 933 SNPs from 282 carefully selected isolates representing the diversity of *Y. pestis* across the globe (Figure WO-9). Their conclusion is that *Y. pestis* originated in China and has reemerged from the region in a series of pandemics—more than just three.

The cause of the second plague pandemic in the Middle Ages remains controversial, with some speculating that the cause was not *Y. pestis* but some other organism(s). Keim cited the work of Bramanti and collegues (Haensch et al., 2010) who studied ancient DNA samples taken from victims of the Black Death buried in mass graves in sites across Europe. They concluded that distinct clones of *Y. pestis* were in fact associated with the Black Death, and that there were multiple, distinct, waves of *Y. pestis* coming out of China during the Middle Ages.

A more recent study by Bos and colleagues (2011) reconstructed the ancient genome of *Y. pestis* from DNA samples obtained from plague victims buried in mass graves that were known to be used from 1348 through 1350 in London. Their findings were similar to the earlier work of Bramanti (Haensch et al., 2010) and consistent with the idea that the Black Death during the Middle Ages was a series of epidemics. Keim noted that only a very small number of SNPs differ

[12] The study of evolutionary relationships among groups of organisms (e.g., species, populations), which is discovered through molecular sequencing data and morphological data matrices.

[13] SNPs are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered.

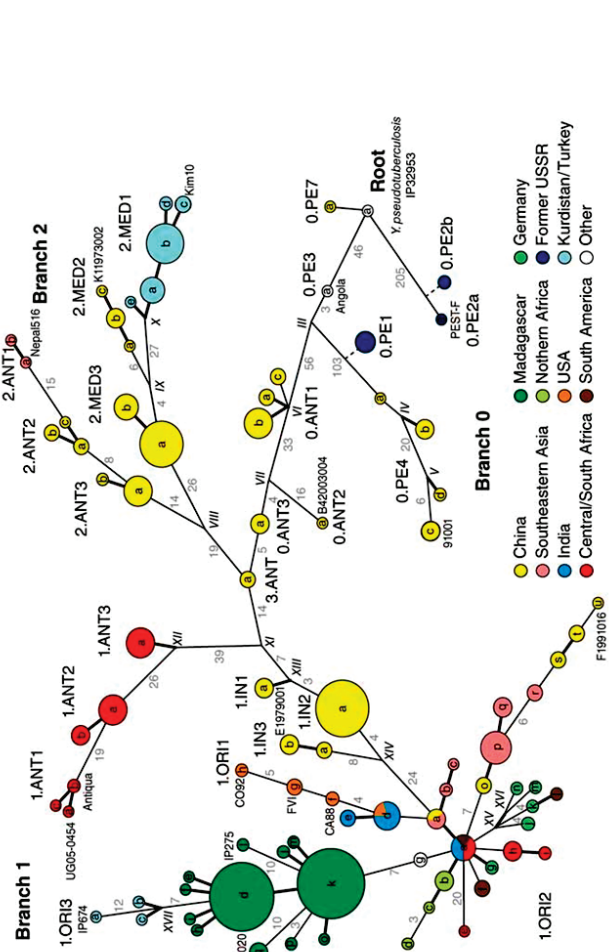


FIGURE WO-9 Fully parsimonious minimal spanning tree of 933 SNPs for 282 isolates of *Y. pestis* colored by location.
NOTE: Large, bold text: Branches 1, 2 and 0. Smaller letters: populations (e.g., 1.ORI3). Lowercase letters: nodes (e.g., 1.ORI3.a). Strain designations near terminal nodes: genomic sequences. Roman numbers: hypothetical nodes. Grey text on lines between nodes: numbers of SNPs, except that one or two SNPs are indicated by thick and thin black lines, respectively. Six additional isolates in 0.PE1 and 0.PE2b (blue dashes) were tested only for selected, informative SNPs.
SOURCE: Morelli et al. (2010). Reprinted by permission from Macmillan Publishers Ltd: NATURE GENETICS. Morelli, G., Y. et al. 2010. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics* 42(12):1140-1143, copyright 2010.

between the whole genome sequence from the 14th-century plague and what was observed by Morelli.

The third pandemic, which is ongoing, arrived in North America by first appearing in Hawaii in 1899, and later through mainland port cities. From localized outbreaks of rat-borne plague beginning in the port cities of the West Coast in the early 1900s, *Y. pestis* then spread to native ground squirrels and became ecologically established across the American West and migrated east through the mid-1940s (Link, 1955; Pollizter, 1951). Capitalizing on whole genome sequences and SNPs from U.S. isolates of *Y. pestis*, Keim concluded that plague in the United States is likely the result of a single introduction from nonnative rodents (i.e., rats on ships) to native rodents.

The population structure of *Y. pestis* in the United States suggests introduction through a genetic bottleneck, followed by radiation of different lineages across the landscape in a strictly clonal fashion. There do not appear to be adaptive benefits for any given lineage, and there was no sequential wave across the landscape. Rather, the transmission pattern across North America is complex and suggests that some places in North America were colonized more than once, that diverse populations coexisted in the same geographical location, and that dispersal was initially west to east but with some east to west reintroductions.

Studying Y. pestis *Evolution in Real Time*

Keim and colleagues investigated a plague outbreak in colonial ground squirrels (prairie dogs) that occurred over the course of several months around Flagstaff, Arizona, in 2001, and developed a mutation-rate-based model for assessing plague transmission patterns in real time in order to better understand how plague spread so quickly in the United States (Girard et al., 2004). By collecting fleas from the prairie dog holes in plague-infested areas, Keim's group was able to directly genotype *Y. pestis* from DNA extracted from flea vectors without the need for culturing the organism. Studying variable number tandem repeats (VNTRs),[14] Keim's research team developed a phylogenetic tree for *Y. pestis* in Arizona that suggests that plague entered the state in the late 1930s or early 1940s, swept across the landscape from west to east, and became established in rodent reservoirs where it continues to coevolve with its vector and mammalian host(s) to the present day. In certain years, plague emerges, causing epidemics in prairie dog colonies, resulting in rapid geographic dispersal of *Y. pestis*. Interestingly, phylogenetically distinct types of *Y. pestis* were observed in different geographically clustered reservoirs in the Flagstaff area, resulting in a star phylogeny (many short branches off of a single node, rather than a dichotomous tree).

[14] VNTRs are short nucleotide sequences that are present in multiple copies at a particular locus in the genome. The number of repeats can vary from individual to individual, making analysis of VNTRs useful for subtyping of microorganisms.

Plague is also endemic in the highlands of Madagascar, resulting in at least 100 human cases of plague each year. Plague first arrived in Madagascar in the late 1800s, and outbreaks in rats and humans occurred in the port city of Mahajanga in the early 1900s. Plague moved to the highlands and became ecologically established around 1926, and it did not recur in Mahajanga for more than 60 years, reappearing in 1991. Genetic analysis suggests that plague was reseeded in Mahajanga from one of the endemic foci in the highlands. Analysis of hypervariable sequences, whole genome sequences, and SNPs from about 40 samples from Mahajanga revealed an unusual, linear phylogeny (in contrast to the star radiation observed in Arizona), suggesting multiple introductions (Vogler et al., 2011).

Although the SNPs in the ancient *Y. pestis* DNA from the Black Death and in DNA from the current plague are different, *Y. pestis* remains very effective at killing. As such, Keim suggested that pandemicity has less to do with the organism's pathogenicity, and more to do with the ecological situation it found itself in. Clonal propagation can reseed a reservoir, and it can also lead to a massive increase in the number of organisms, dispersal, and an increase in fitness. As a clonal pathogen, *Y. pestis* is not taking in genetic material, but Keim suggested that perhaps clonal organisms are contributing to the diversity of the ecosystem. Keim also observed that the saprophytic soil organism, *Burkholderia pseudomallei*[15] has a set of genes that encode for fimbriae that appear to have been horizontally transferred from a *Yersinia*-like organism.

Microbial Forensics

Forensic evidence is a continuum, and the quality of information, potential errors, and uncertainty influence the power of and confidence in the analysis, interpretation, and inferences made. Microbial forensics is not a new discipline; epidemiologists have been practitioners of the science and art of forensic microbiology since at least the 19th century, identifying the agent, exposed population, the source of exposure, and the extent of contamination, with the goal of disease identification, containment, and treatment of ill populations. The development of genomic sequencing technologies and platforms was stimulated in part by the law enforcement communities to apply these tools and approaches for use in forensic analysis. Speaker Bruce Budowle of the University of North Texas Health Science Center defined microbial forensics as the analysis of evidence from an act of bioterrorism, biocrime, or inadvertent microorganism/toxin release for attribution purposes (Dr. Budowle's contribution to the workshop summary report may be found in Appendix A, pages 117-133). A microbial forensic investigation, according to Budowle, is more about attribution, determining the agent, source,

[15] An organism on the U.S. Department of Health and Human Services select agent list. For a complete list of select agents as of 2012, see www.selectagents.gov (accessed November 1, 2012).

and perpetrator, and interpreting and presenting evidence to investigators, the courts and policy makers. In addition, evidence can be used to eliminate certain sources. Traditional trace evidence including DNA, hairs, fibers, and fingerprints may also be involved.

Any infectious agent may be deployed offensively as a biological weapon against a suitable living or nonliving target. According to Budowle, there are more than 1,000 agents—bacteria, viruses, fungi, and protozoa—that are known to infect humans, plants, and animals, along with emerging pathogens and potentially bioengineered organisms. A forensic investigation seeks to gather as much information as possible about the threat agent and compare it to known samples to characterize the organism and/or its processing (e.g., engineering, production, geolocation, date) and delivery. Microbial genomics, including phylogenetics, can help to narrow the focus of the investigation.

Challenges

Source attribution for a "biocrime" requires more circumspection than predicting a source for research purposes. Budowle raised some concerns about interpretation, which hinges on the sensitivity and reliability of the analysis. Missing data are also a concern, and inferences must be made in situations where there is vast uncertainty. Budowle, and others throughout the workshop, emphasized the need for appropriate databases. A particular challenge for forensics is that when a case is under scrutiny in court, it may only be possible to say that isolates are closely related, and experts for the opposition may challenge that the references or databases are insufficient or inappropriate.

With regard to technology limitations, Budowle noted that not all microbial forensic evidence is suitable for genetic analysis using next-generation sequencing. Some samples will be limited in quantity, highly degraded, and/or contaminated. A challenge is to extract as much genetic information as possible from limited materials and nonviable organisms. In forensics, sequencing errors will inflate differences between samples creating a degree of uncertainly. As such, defining and quantifying the error rates associated with the sequencing platform and chemistry are critically important. Again, the quality of sequence data and the results of bioinformatics analyses must be as high as possible.

Budowle also emphasized the need for standard reference and test materials. Today's databases will be the test panels and reference samples of tomorrow, but some data are woefully inadequate. As sequencing capability moves into application-oriented laboratories, we need to consider quality control, and perhaps something along the lines of proficiency testing, to ensure high-quality data in databases. Best practices need to be established regarding what qualifies as a reference sample, as well as standards for preparation, validation, and characterization (including metadata). It is not clear where the responsibility lies for generating such standard reference materials, because there are numerous stakeholders (Table WO-2).

TABLE WO-2 Interests in Reference Collections and Management

Agency or Organization	Role in National Biodefense System	Interests Related to Pathogen Collections
DHS	R&D in detection and microbial forensics; central laboratory role in microbial forensics investigations	Research access to isolates for assay development, validation, and other R&D; Archives for comparisons to forensic samples
FBI (DOJ)	Investigation and prosecution of crimes using biological agents	Quickly identifying possible sources for a pathogen used in an attack; development of standards and controls
CDC (HHS)	Epidemiological investigation of disease outbreaks; administrator of the Select Agent Program	Tracking endemic strains and identifying sources of outbreak strains
NIAID (HHS-NIH)	Understanding disease mechanisms and host-pathogen interaction; development of treatments and diagnostic assays	Research access to isolates for studies of pathogenicity mechanisms, host-pathogen response, and other R&D related to medical countermeasures
FDA (HHS)	(With CDC) investigation of food-borne disease outbreaks	Tracking endemic strains and identifying outbreak strains and their sources
USDA	Plant and animal pathogens	Disease tracking, identifying outbreak strains and their sources, developing treatments and vaccines
AFMIC, DTRA, Services (DOD)	Military force protection	Tracking foreign disease that may impact military operations; medical countermeasures Tracking potential biothreat agents
ATCC BEI Resources SDO	Biomaterials resource	Development of biothreat and EID standard biomaterials that meet ISO and ANSI guidelines

NOTE: DHS, Department of Homeland Security; FBI, Federal Bureau of Investigation; DOJ, Department of Justice; CDC, Centers for Disease Control and Prevention; HHS, Department of Health and Human Services; FDA, Food and Drug Administration; USDA, U.S. Department of Agriculture; AFMIC, Armed Forces Medical Intelligence Center; DTRA, Defense Threat Reduction Agency; DOD, Department of Defense; ATCC, American Type Culture Collection; SDO, Standards Development Organization; EID, emerging infectious disease; ISO, International Organization for Standardization; ANSI, American National Standards Institute.
SOURCE: Budowle (2012).

Forensics Case Example: The Amerithrax Investigation

Claire Fraser of the Institute for Genome Sciences provided an overview of the genomic approaches used in the Amerithrax investigation.[16] Fraser reminded

[16] Amerithrax is the case name assigned by the Federal Bureau of Investigation (FBI) to the 2001 anthrax letter attacks.

participants that in 2001 there was only one sequencing platform available (the Sanger 3730), the cost for sequencing was $200,000 to $300,000 per genome, and it took nearly a year to completely sequence a genome. While the technology has changed dramatically over the past decade, our understanding of the dynamics of microbial genomes is still quite limited.

In collaboration with the overall FBI Amerithrax investigation, the goal of the scientific investigation, according to Fraser, was to explore whether a genomics-based approach could be used to attribute the spore preparation of a genetically homogeneous species (used in the letters) to a potential source (Keim et al., 1997; Rasko et al., 2011; Read et al., 2002). Could genetically unique features be identified using traditional DNA sequencing-based analysis that would be useful for the purposes of attribution? Fraser reiterated the point made by Budowle that attribution falls on a continuum; it may not be an exact match, and exclusion may also be important.

The starting material was the *Bacillus anthracis* spore preparations obtained from the letters that were mailed to the office of Senator Leahy and to the *New York Post*, supplemented by related material obtained from other sources including the Hart Senate Office Building, postal workers from the Brentwood (Washington, DC) post office facility, and people exposed to anthrax spores in the Hart Senate Office Building. Gross examination of the physical characteristics of the samples suggested there were at least two different preparations. It became quickly apparent that traditional genotyping methods were not achieving sufficient resolving power for the purposes of this investigation. Based on VNTR analysis by the Keim laboratory, it appeared that all of the isolates initially collected as part of this investigation were the Ames strain of *B. anthracis*.[17]

At the same time, TIGR was in the final stages of assembling the first genome sequence of *B. anthracis* and was asked to partner with the FBI and other laboratories to determine whether the complete genome sequence would be useful for purposes of attribution. With funding from the National Science Foundation, TIGR began sequencing a colony of *B. anthracis* recovered from the spinal fluid of Mr. Robert Stevens of American Media in Florida, the first victim to die as a result of exposure to anthrax spores mailed to him (referred to as *B. anthracis* Florida). Because no appropriate reference strain was available, TIGR also sequenced the genome of the *B. anthracis* Ames ancestral strain. (Fraser explained that the strain TIGR had initially sequenced was obtained from a facility in the United Kingdom that had been cured of its two virulence plasmids—pXO1 and pXO2—making it an inappropriate reference strain for comparative purposes.)

According to Fraser, SNP analysis of the reference *B. anthracis* Ames ancestor and the Florida isolate found no differences in more than 5 million base pairs assessed. Similarly, no polymorphisms were found when the wild-type isolates

[17] The Ames strain was originally isolated from a dead Beefmaster heifer in Texas in 1981. It quickly became a standard laboratory research strain used worldwide for vaccine challenge studies.

from the letters to the *New York Post* and Senator Leahy's office were compared to the reference Ames ancestor sequence (Rasko et al., 2011). While *B. anthracis* is highly monomorphic, Fraser noted that it was somewhat surprising to find absolutely no sequence differences, and it raised questions about whether genomics would be useful in the investigation after all.

At the same time, researchers at the U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) began to notice some *B. anthracis* colonies with distinct, and apparently heritable, morphology as the spore preparations from four anthrax-laced letters were passaged in culture. Examination of the colonies formed on sheep blood agar (SBA) resulted in the identification of four distinct morphological variants (morphotypes)—designated A, B, C/D, and E—from each of the material analyzed (Rasko et al., 2011). These morphotypes are illustrated in Figure WO-10.

According to Fraser, these phenotypic variants were all found to be altered with regard to their ability to sporulate under different conditions, potentially linking the events in New York and Washington. This new information inspired a population genomics approach to this investigation. Could the population composition (rather than the wild-type) be used to make a match?

Morphologic variants from the Leahy and *New York Post* letters were sequenced and compared to both wild-type and the Ames ancestor strains. Morphological variant A was the most different from the wild-type in terms of sequence, although the sequence variability represented a very small portion of the genome.

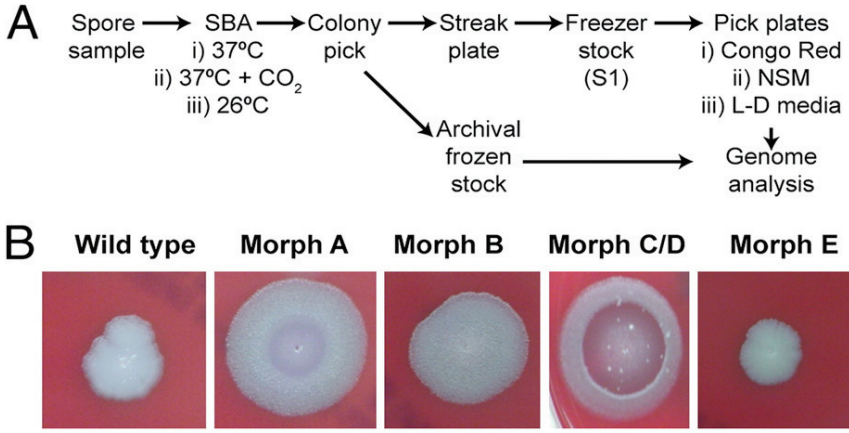


FIGURE WO-10 Microbiological identification of morphological variants of *B. anthracis* Ames. Panel A, Flowchart used to process the evidentiary material for the identification of the morphological variants. In all cases, the morphotypes are altered in the sporulation phenotype and colony morphology. Panel B, Image of a representative colony of each of the morphotypes grown on SBA.
SOURCE: Rasko et al. (2011).

Fraser noted that it was initially believed the sequence was identical to the wild-type. However, using paired-end sequencing, it was possible to look at mate pairs that were present in the assembly at distances that would not be expected based on the insert sizes that went into the cloned library. This led to the discovery of a number of chimeric reads in the assembly that ended up being a tandem duplication, and exhaustive PCR analysis was done to establish confidence in the finding. Analysis of three morphotype A mutants present in all the letters showed that while they were identical phenotypically, they were genetically different.

Isolates of morphological variant B have SNPs in a sporulation gene, variants C/D have two sequence variants in a histidine kinase sensor gene (C has a SNP, and D has an indel[18]), and opaque variants have an indel of either 9 or 21 base pairs in a response regulator gene (i.e., all are mutations along the sporulation pathway). Fraser went on to explain that each of these genetic variants was converted into a quantitative PCR assay and used to screen a repository of nearly 1,100 samples collected by the FBI. All four of these mutations were found in a sample from a single source, a flask at USAMRIID labeled RMR-1029. Other samples that also contained the four mutations could trace their provenance back to RMR-1029. An assortment of other samples were also found to have subsets of these four mutations, but not all of them. RMR-1029 was a heterogeneous mixture cultivated for vaccine trials in the late 1990s and flasks were stored at USAMRIID in Maryland and the Dugway Proving Ground in Utah.

In summary, it was population genomics that provided the unique signature that facilitated attribution in the Amerithrax case. The minor subpopulation was unique to the spore preparations recovered from the letters. Fraser noted that these polymorphisms were used to screen batch cultures, not single *B. anthracis* colonies on a plate, which was very different from all of the clonal genome projects that had been completed to that point in time.

Jumping forward a decade, what might be different today in the era of metagenomics? Clearly, the process could occur more quickly and at much lower cost. With current technologies investigators would be working with shorter sequence reads and, Fraser added, it is not possible to know if the gene duplication would have been as easy to identify from short reads as it was from the 800-plus-base-pair Sanger reads that were used at the time. What would community-level analysis with very deep coverage provide versus what was done by looking at single colonies? Morphotype A was present in all RMR-1029 samples, but not necessarily at the same low level of abundance. What does it mean in terms of being able to say that something is the same or not? Fraser and Budowle emphasized that a significant gap, that still persists today, is the lack of appropriate, standardized, criteria (thresholds, confidence limits, etc.) that would lead one to conclude that a given microbial sample was or was not derived from the same source, to answer the question of what makes a match with confidence.

[18] Indel refers to an insertion or deletion mutation.

Fraser noted that the experience with *B. anthracis* is in no way generalizable to other pathogens or potential agents of bioterrorism. Had it been an organism with horizontal gene transfer and genome rearrangements over periods of time, it would likely have been in a very different situation. Budowle added that the Amerithrax forensic investigation was somewhat unique because as many samples as could be collected from the letter attacks were collected, and more inferences could be made in this case than might be possible in many other cases.

Microbial Evolution: Studying Genomes, Pangenomes, and Metagenomes

The ability to sequence and compare whole genomes of many related microorganisms has prompted a deeper understanding of the biology and evolution of microorganisms. The completeness of finished microbial genomes is particularly powerful in a comparative context. Differences in genomic content such as the presence or absence of genes or changes in gene order or sequence, from SNPs to large indels,[19] may have important phenotypic consequences. The comparison of multiple, related genome sequences offers insights into an organism's evolutionary history—including the relative importance that natural selection attaches to specific gene functions (Eisen et al., 1997; Fraser-Ligget, 2005).

Comparative analyses have revealed that the microbial genome is a dynamic entity shaped by multiple forces including gene loss/genome reduction, genome rearrangement, expansion of functional capabilities through gene duplication, and acquisition of functional capabilities through lateral or horizontal gene transfer, as shown in Figure WO-11 (Fraser-Liggett, 2005). Three main forces shape bacterial genomes: gene gain, gene loss, and gene change. All three can take place in a single bacterium. Some of the changes that result from the interplay of these forces are shown in the following illustration. Several natural processes carry genetic information from one species to another. DNA can be transported by viruses (transduction), via bacterial mating (conjugation), and through the direct uptake of DNA from the environment (transformation). Genes that must function together are transferred together as genomic islands (e.g., pathogenicity islands) (Hacker and Kaper, 2000).

The frequent gain and loss of genomic information exhibited by many bacterial species makes it difficult to trace bacterial phylogenies and has strained the species concept (Bentley, 2009). Among genetically variable bacterial species, it is clear that a single strain rarely typifies an entire species. Instead, researchers sequence multiple strains of a species to compile the "pan-genome" or global gene repertoire of a bacterial species (Medini et al., 2005) (Figure WO-12). The pan-genome can be divided into three elements: the core genome (housekeeping genes shared by all strains); a set of strain-specific genes that are unique to various isolates; and a set of dispensable genes that are shared by some but not

[19] Insertions and deletions.

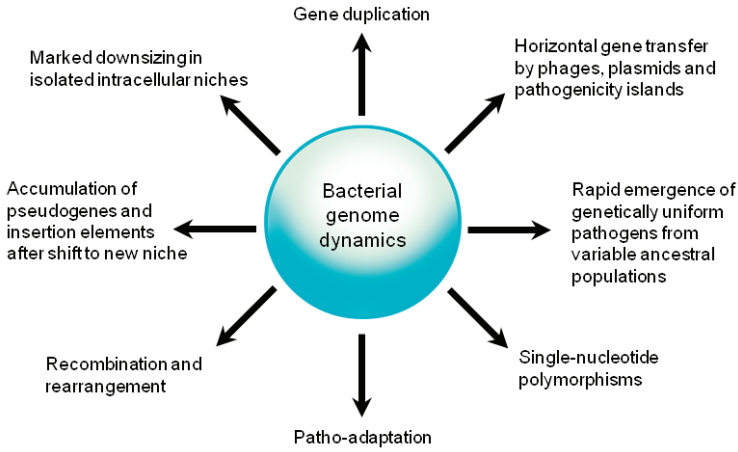


FIGURE WO-11 Bacterial genome dynamics.
SOURCE: Pallen and Wren (2007). Reprinted by permission from Macmillan Publishers Ltd: NATURE. Pallen, M. J., and B. W. Wren. 2007. Bacterial pathogenomics. *Nature* 449:835-842, copyright 2007.

all isolates. These latter, dispensable sequences are associated with high rates of nucleotide sequence variability and contribute to phenotypic diversity within bacterial populations (Medini et al., 2008).

The comparison of eight genomes from strains of group B *Streptococcus* (GBS) found an average of 1,806 genes in the core genome and 439 dispensable or strain-specific genes. Based on these data, models predict that the GBS pan-genome is "open," with unique genes continuing to emerge even after hundreds or thousands of genomes are sequenced. Indeed, many bacterial species have extensive genetic diversity, with an average of 20 to 35 percent of genes being unique for a single strain. In contrast, as discussed by Fraser, other organisms appear to be monomorphic—with a "closed" pan-genome. In the case of *B. anthracis* four genome sequences completely characterize the species (Medini et al., 2008).

A species' pan-genome likely reflects selective pressure to adapt to specific environmental conditions. Species with an open pan-genome typically "colonize multiple environments and have many ways of exchanging genetic material." By contrast, monomorphic species with a closed pan-genome "live in isolated niches and have a low capacity to acquire foreign genes" (Medini et al., 2008). In natural settings, bacteria and other microorganisms interact with each other and with their surroundings to form complex communities that occupy diverse environmental niches. The shuttling of genes between species via horizontal gene transfer (HGT) plays an important role in a species' ability to adapt to environmental change. As discussed in the section that follows, ecological factors may strongly influence the acquisition and loss of genes via HGT (Smillie et al., 2011). The pan-genome concept suggests the presence of a large microbial gene pool in the environment

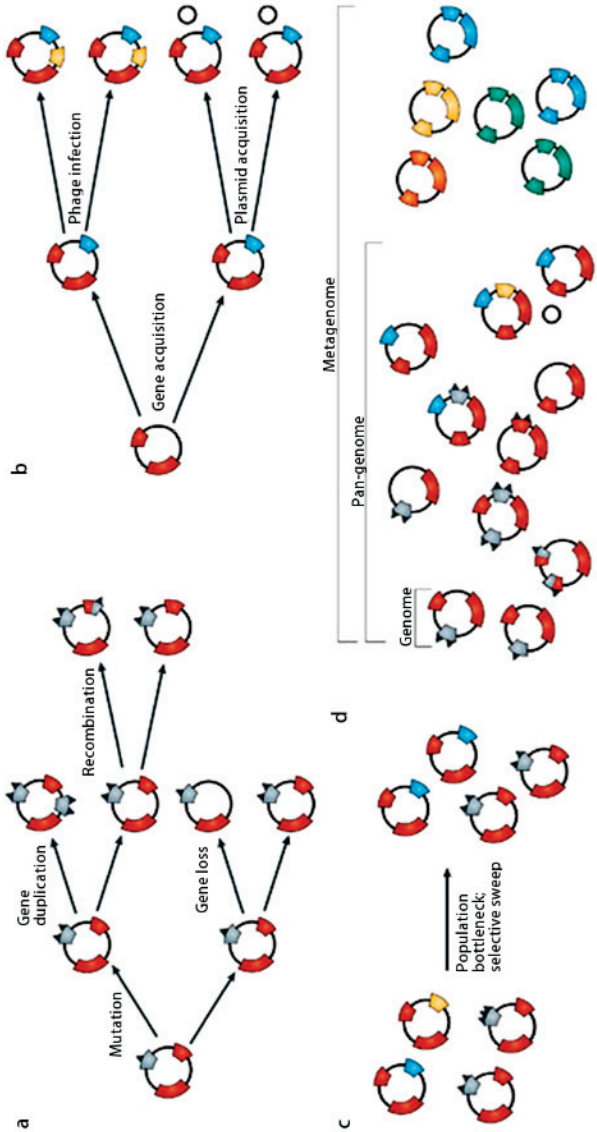


FIGURE WO-12 Molecular evolutionary mechanisms that shape bacterial species diversity: one genome, pan-genome, and metagenome. Intraspecies (a), inter-species (b), and population dynamic (c) mechanisms manipulate the genomic diversity of bacterial species. For this reason, one genome sequence is inadequate for describing the complexity of species genera and their inter-relationships. Multiple genome sequences are needed to describe the pan-genome, which represents, with the best approximation, the genetic information of a bacterial species. Metagenomics embraces the community as the unit of study and, in a specific environmental niche, defines the metagenome of the whole microbial population (d).
SOURCE: Medini et al. (2008). Reprinted by permission from Macmillan Publishers Ltd: NATURE REVIEWS MICROBIOLOGY. Medini, D. et al. 2008. Microbiology in the post-genomic era. *Nature Reviews Microbiology* 6: 419-430, copyright 2008.

that drives microbial evolution, with HGT providing microorganisms with rapid access to genetic innovation. HGT can enable beneficial traits (such as resistance to antimicrobial drugs or other environmental perturbations) to spread through entire populations (Medini et al., 2005, 2008).

Pathogenomics

The vast majority of microbes do not produce overt illness in their hosts, but may, instead, establish themselves as persistent colonists that can be described as either low-impact parasites (e.g., causes of asymptomatic infection), commensals (i.e., organisms that "eat from the same table," deriving benefit without harming their hosts), or symbionts (establishing a mutually beneficial relationship with the host) (Blaser, 1997; Merrell and Falkow, 2004). These states, while separate, represent part of a continuum extending to pathogenesis and disease that may be occupied at any point by a specific microbial species through the influence of environmental and genetic factors (Casadevall and Pirofski, 2000, 2002, 2003). Persistent colonization of a host by a microbe is rarely a random event; such coexistence depends upon a relationship between host and microbe that can be characterized as a stable equilibrium (Blaser, 1997).

Over the course of the past century, the identification of increasing numbers of microbial pathogens and the characterization of the diseases they cause have begun to reveal the extraordinary complexity and individuality of host–microbe relationships. In the case of microbes that cause persistent, asymptomatic, infection, physiological, or genetic changes in either host or microbe may disrupt this equilibrium and shift the relationship toward pathogenesis, resulting in illness and possibly death for the host (Merrell and Falkow, 2004). As a result, it has become exceedingly difficult to identify what makes a microbe a pathogen.

Virulence as an Emergent Property

The question of "why some microbes cause disease and others do not" has puzzled microbiologists for centuries. Genomics is a new and useful tool for exploring this question, but it has its limitations, according to Arturo Casadevall of the Albert Einstein College of Medicine (Dr. Casadevall's contribution to the workshop summary report may be found in Appendix A, pages 134-140). As illustrated by several examples discussed by Casadevall in his prepared remarks, the complexity of host–microbe interactions complicates researchers' ability to link genotypic information with phenotypic expression of this genomic information. This complexity has important implications for the development of predictive tools to identify microbial threats.

Research associating certain microorganisms with activities that could be beneficial or harmful to human health has lead scientists to a central question: are pathogenic microbes inherently different from nonpathogens? Most

microbiologists in the early 20th century reasoned that pathogenic—disease-causing—microbes must differ from nonpathogenic microorganisms in the expression of traits associated with virulence. Others, including the Belgian immunologist Jules Bordet,[20] argued that there could be no difference. He based his argument on two observations of the context dependence of virulence:

1. The same organism can exist in both virulent and nonvirulent states. For example, isolates of *Neisseria meningitidis* associated with a meningococcal epidemic lose virulence when maintained in laboratory culture and regain virulence after passage[21] through a mouse that resulted in the selection for characteristics that allowed survival in the mammalian hosts and thus reenabled virulence.
2. In an infected but immunized host, a pathogenic organism exists in a nonpathogenic state.

Indeed, during the early 20th century, many common infectious diseases disappeared as a result of immunization, and microorganisms that were not previously considered pathogenic were increasingly associated with disease later in the century. The microbes did not change, noted Casadevall, "what happened was that we changed the host."

Casadevall went on to emphasize that the concept of a "pathogen" is flawed, because it assumes that pathogenicity is an intrinsic, immutable characteristic of a microorganism. Neither pathogenicity nor virulence is an independent microbial property; according to Casadevall both are characteristics that are expressed only in a susceptible host. Labeling a microbe a "pathogen" endows it with properties that are not its own. In Casadevall's view, there are only "microbes" and "hosts"—what is truly important is the outcome of interactions between the microbe and its host environment(s).

Casadevall and his colleague Liise-anne Pirofski developed the "damage-response framework" to provide an integrated theory that accounts for the contribution of both the host and the pathogen to pathogenesis (Casadevall and Pirofski, 2003). Within this framework, a pathogen is defined as a microorganism that is capable of causing disease, and pathogenicity is the capacity of a microbe to cause damage in a host. The damage-response framework defines a virulence factor as a microbial component that damages the host and virulence as the *relative capacity* of a microbe to cause damage in a host (Casadevall and Pirofski, 1999). Damage is thus an expression of microbe–host interactions, which for most host–pathogen interactions, can be graphed as a parabola, as illustrated in

[20] Jules Bordet was awarded the Nobel Prize in physiology or medicine in 1919 for his discoveries relating to immunity.

[21] In microbiology, "passage" refers to the successive transfer of cultures of microorganisms across various nutrient mediums or the reinoculation of one animal with pathogenic microbes from another infected animal.

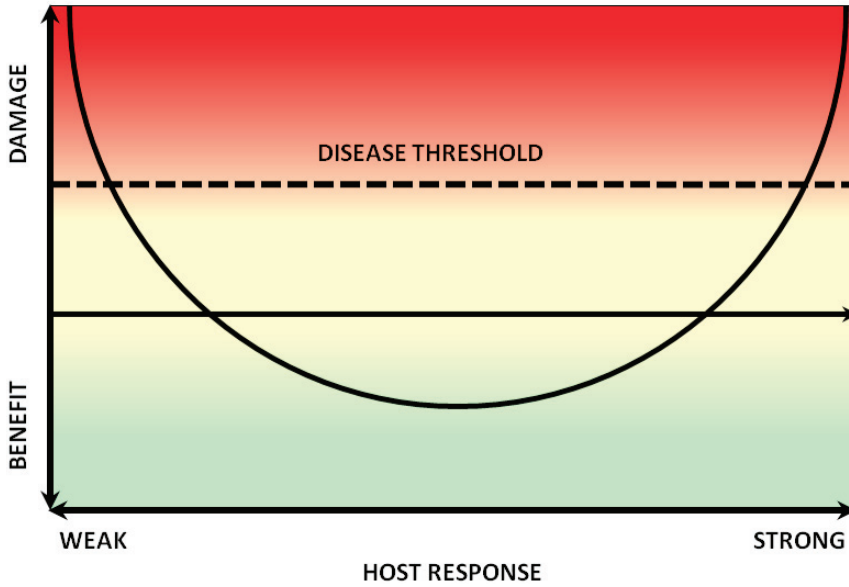


FIGURE WO-13 Damage-response framework and the case of *S. cerevisiae*. SOURCE: Casadevall (2012).

Figure WO-13 (Casadevall and Pirofski, 2001). As an individual becomes immunosuppressed, damage can occur, and once a certain threshold is reached, disease may occur. The same organism might also elicit an untoward immune response resulting in disease (Casadevall and Pirofski, 2003).

For example, *Saccharomyces cerevisiae*, commonly used in baking and brewing, can cause disease in immunocompromised patients with HIV; vaginitis in normal women that is indistinguishable from candidiasis; and lung nodules in bakers as a result of hypersensitivity responses in the lung. *S. cerevisiae* cannot be defined as food, commensal organism, opportunistic pathogen, or primary pathogen without taking into account the host. As noted by Casadevall, a reductionistic approach—whether the microbe-centric view of many microbial geneticists, who focus on virulence factors, or the host-centric view of many immunologists, who focus on factors affecting host susceptibility—provides an incomplete picture of this continuum of outcomes.

Host- vs. environment-acquired microbes The diversity of possible outcomes associated with many host–microbe interactions is also evident when one considers virulence factors associated with microbes acquired from another host or directly from the environment. Organisms acquired from another host include all viruses, many parasites, most bacteria, and a few fungi. These are generally communicable diseases with a limited host range. These organisms are not free

living, and there is likely to be selective pressure on the microbe to coexist with the host. Disease often results from the disruption of the host–microbe relationship (Casadevall and Pirofski, 2007).

Environmentally acquired microbes include bacteria, fungi, and some parasites. They are not communicable, have a very broad host range, and are free living. The selective pressures in the environmentally acquired microorganisms for causing disease are unknown. Because they have no host requirement for survival, these are the only organisms that are known to cause extinction.[22] Disease often manifests in hosts with impaired immunity, or when there are large microbial inocula (Casadevall and Pirofski, 2007). The fungus *Cryptococcus neoformans* is an example of a soil-dwelling, environmentally acquired" microorganism that infects a wide range of hosts—including plants, animals, and humans—but only rarely causes disease. Casadevall noted that "everyone in this room is infected [with *C. neoformans*], but you have a one-in-a-million chance of getting disease unless you become immunosuppressed." The virulence of *C. neoformans* is complex, explained Casadevall, because the organism did not evolve to cause disease in these hosts. Instead, this organism was selected for properties that allowed it to survive in the soil, and "by the luck of the draw" it happens to have the traits necessary to cause disease in some hosts. Rather than being a special property of only certain microorganisms, virulence is an emergent property.

The challenges of studying an emergent property Casadevall defined an emergent property as a novel property that unpredictably comes from a combination of two simpler constituents—in essence the "whole is greater than the sum of its parts" (Casadevall et al., 2011). In this case, a host and a microbe are the components, and the novelty may be expressed as either virulence and pathogenicity, mutualism, commensalism, or even the death of either party.

Emergent properties abound in the natural world, and while they can be understood after the fact, emergence is not reducible or predictable. We understand the structure of water, for example, and we can explain surface tension when we see it, but we cannot predict surface tension from individual water molecules. We understand the physics of small particles, but we cannot predict sand dunes. Emergent properties cannot be reduced to either component. In this regard, Casadevall suggested that research focused on either the host or the microbe may produce interesting results, but it may not be relevant to understanding outcomes of host–microbe interactions.

To illustrate the limits of a reductionistic approach to understanding virulence, Casadevall described recent research on *C. neoformans*. Applying a mathematical model to compute the relative contribution of microbial virulence factors, Casadevall estimated that the majority of cryptococcal virulence in mice

[22] The environmentally acquired fungi *Geomyces destructans* and *Batrachochytrium dendrobatidis,* for example, currently pose a significant threat to populations of New World bats and amphibians worldwide (IOM, 2011).

can be attributed to the *C. neoformans* capsule and cell wall melanin (McClelland et al., 2006). Casadevall then evaluated the relative virulence of *C. neoformans* and two closely related cryptococcal species in *Galleria mellonella* moth larvae. As expected, *C. neoformans* was pathogenic in this system. However, the cryptococcus species that has a capsule and makes melanin was not pathogenic, and the cryptococcus species that has no capsule and no melanin was pathogenic in the larvae.

Casadevall challenged workshop participants to consider whether virulence is a chaotic system—like weather. If it is, then the outcome of host–microbe interactions, including virulence, may be inherently unpredictable. There are limits to what we can know and predict through reductionism (i.e., through a focus only on one component or the other, the host or the microbe). Still, despite emergence and potentially chaos, progress is still possible. Casadevall noted that the use of computer analysis has improved the accuracy of weather forecasts. Looking ahead, Casadevall suggested that the focus of future research should be on developing probabilistic models for host–pathogen interactions.

Microbial Genomics: Epidemiology and the Mechanisms of Disease Emergence

Comparative genomics has helped to inform our understanding of host–microbe interactions from that of a "war metaphor" ("the only good bug is a dead bug") to one that places these interactions into a broader ecological and evolutionary context. These insights have profound implications for detecting, diagnosing, and anticipating infectious disease emergence, including:

- Bacterial genome sequence data have challenged the simplistic views that pathogens can be understood solely by identifying their virulence factors, and that pathogens often evolve from "nonpathogenic" organisms through the acquisition of virulence genes from plasmids, bacteriophages, or pathogenicity islands. Metagenomic surveys conducted in diverse environments have improved our understanding of the biodiversity and biogeography of microbes and have underscored the important role of environmental factors in disease emergence and spread.
- Simply studying a pathogen without understanding its biotic and abiotic environmental contexts may lead to false confidence in our ability to detect it. Microbial detection will be most effective if there is sufficient basic scientific information concerning microbial genetics, evolution, physiology, and ecology.
- Likewise, strain sub-typing will be difficult to interpret if we do not understand some of the basic evolutionary mechanisms and population diversity of pathogens and nonpathogens alike (JASON, 2009).

The following case examples explore some of the different kinds of systems in which genomics has been used to study microorganisms or microbial communities associated with disease. These studies have revealed important insights into the mechanisms of variation and genome change as well as the role of host–microbe–environment interactions on the evolution and adaptation of pathogens and nonpathogens alike.

Comparative Genomics: E. coli *Including* Shigella

Speaker David Rasko of the University of Maryland School of Medicine discussed how comparative genomics can be used to improve diagnostic methods. According to Rasko whole genome sequence analysis has rapidly advanced researchers' understanding of pathogenic variants of *E. coli*—an organism that has been intensely studied for almost 40 years.

Diarrheagenic *E. coli* (DEC) are food-borne and water-borne pathogens associated with approximately 300,000 deaths annually, primarily in the developing world. Pathogenic variants of DEC (referred to as pathovars or pathotypes) exhibit diverse characteristics and pathogenic mechanisms, only some of which have been well characterized (Figure WO-14).

With well over 100 serotypes, the phylogeny and evolution of DEC is diverse and complex. Even though *E. coli* has been studied for decades, Rasko noted that current diagnostic and typing methods are inadequate and there are no approved vaccines. Genomic analyses provide new ways to characterize organism diversity and to identify novel virulence factors—information that will enhance methods for outbreak and strain identification and current understanding of pathogen emergence.

The genome structure of *E. coli* is highly conserved, yet between 20 to 25 percent of the DNA in any strain can be novel (i.e., 1.5 megabases of the total 5 megabase). The widely used technique for categorizing *E. coli* strains according to sequence similarity—multi-locus sequence typing (MLST)—compares a subset of sequence from each strain's genome (~3,400 bases). As noted by Rasko, the resulting phylotypes[23] are not well resolved, and pathotypes are not restricted to any one phylotype.

Whole genome sequencing allows researchers to compare a much larger proportion of the genome (2.3–2.7 million bases) and provides greater discriminating power than MLST. The greater resolution produced by whole genome sequencing is dependent upon the quality and number of genomes and isolates to which strains can be compared. Whole genome phylogenies are more robust and clearly distinguish each phylotype. Rasko added that there is practically no cost difference between MLST and whole genome phylogeny. Rasko underscored the importance of this "context" by noting that when there was an outbreak of

[23] Classification of an organism by its evolutionary relationship to other organisms.

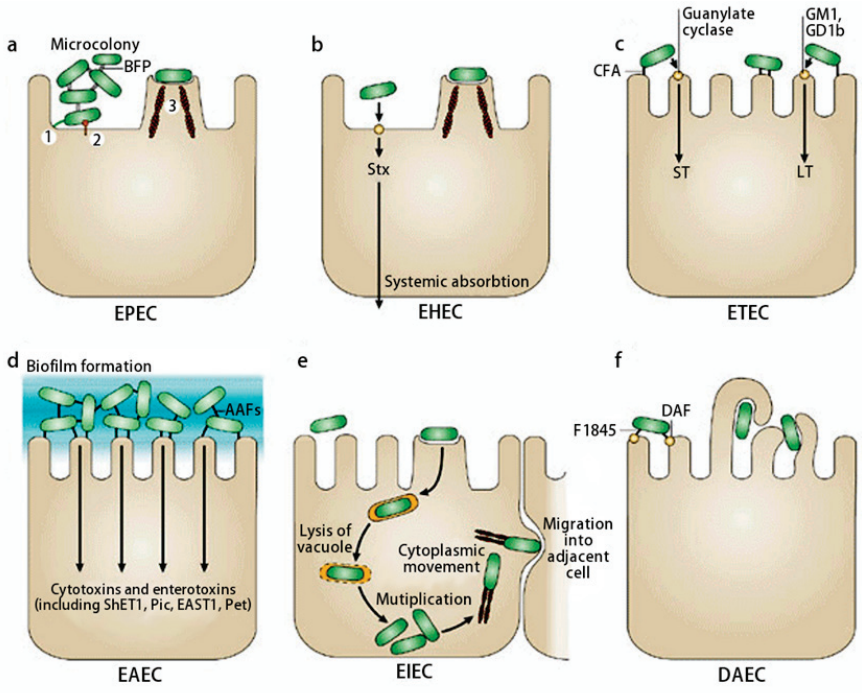


FIGURE WO-14 Six distinct pathogenic variants of diarrheagenic *E. coli*. Panel a, Enteropathogenic *E. coli* (EPEC) form microcolonies via the association of bundle-forming pili (BFP) (1), bind to the small intestinal epithelium (2), and cause the reorganization of cytoskeletal actin, pedestal formation, and attaching-and-effacing lesions (3). Panel b, Enterohemorrhagic *E. coli* (EHEC) infect the large intestine, form attaching-and-effacing lesions, and release Shiga toxin (Stx). Panel c, Enterotoxigenic *E. coli* (ETEC) release heat-stable toxin (ST) and heat-labile toxin (LT) into host cells. Panel d, Enteroaggregative *E. coli* (EAEC) form a biofilm on HEp-2 cells in vitro; however, they do not have a common virulence factor that can be used to identify the entire group. Panel e, Enteroinvasive *E. coli* (EIEC), including *Shigella*, are the only *E. coli* that grow intracellularly. Panel f, Diffusely adhering *E. coli* (DAEC) have one adhesion factor that has been characterized thus far, but other mechanisms of pathogenesis are not yet known.
SOURCE: Reprinted from Kaper et al. (2004). Reprinted by permission from Macmillan Publishers Ltd: NATURE REVIEWS MICROBIOLOGY. Kaper, J. B., J. P. Nataro, and H. L. Mobley. 2004. Pathogenic Escherichia coli. *Nature Reviews Microbiology* 2(2):123-140, copyright 2004.

a rare strain of *E. coli* O104:H4—first identified in northern Germany in May 2011—"because we had a good sequence database and collection of isolates we could very quickly and accurately place the [*E. coli* O104:H4] outbreak as being an enteroaggregative *E. coli* that had acquired Shiga toxin, [rather than] an enterohemorrhagic *E.coli*."

Case example Attaching-and-effacing *E. coli* include the pathotypes enteropathogenic *E. coli* (EPEC) and enterohemorrhagic *E. coli* (EHEC) (see Figure WO-14 on page 38). According to Rasko, "These groups are lumped together because the community has had difficulty over the past 20 to 30 years actually defining each of these pathotypes distinctively." This diverse group of *E. coli* has a large number of phage and is variable in the presence of Shiga toxin. They are defined by their common locus of enterocyte effacement (LEE), which codes for the type III secretion system[24] and other elements involved in pedestal formation and the development of attaching-and-effacing lesions.

In order to develop more accurate diagnostics for this group of *E. coli*, Rasko's lab sequenced and aligned 136 genomes (113 attaching-and-effacing strains and 23 reference isolates spanning all of the pathotypes) to create a whole genome phylogeny. Rasko noted that while virulence factors have been the basis for isolate typing for the past 30 to 40 years, it is now clear that these characteristics do not always match the genome phylogenies (Figure WO-15). Rasko went on to observe that there is a mismatch between phylogeny based on the core genome, and phylotype based on virulence factors when characterizing which strains produce Shiga toxin (indicative of EHEC) and which have a bundle-forming pilus (indicative of EPEC).

In evaluating phylogenetic clusters in comparison to virulence factors, Rasko found that the majority of attaching-and-effacing strains contain the LEE pathogenicity island. Rasko observed that the LEE regions encoding the type III secretion system were highly conserved; the areas coding for the secreted effectors were more variable. Isolates can be loosely grouped based on secreted effectors. Rasko is now investigating possible associations between those differences and virulence or disease severity. Interestingly, there is genome conservation and virulence factor conservation even though the majority of these secreted effectors are not encoded by the LEE region, but rather on other phage in the genome.

Using whole genome alignments to identify novel genome features and possible biomarkers While comparative methods are adequate for the pairwise comparison of a limited number of genomes, identifying novel regions of interest across hundreds of genomes presents a challenge. To help address this analytical

[24] A protein appendage found in several Gram-negative bacteria. In pathogenic bacteria, the needle-like structure is used as a sensory probe to detect the presence of eukaryotic organisms and secrete proteins that help the bacteria infect them.

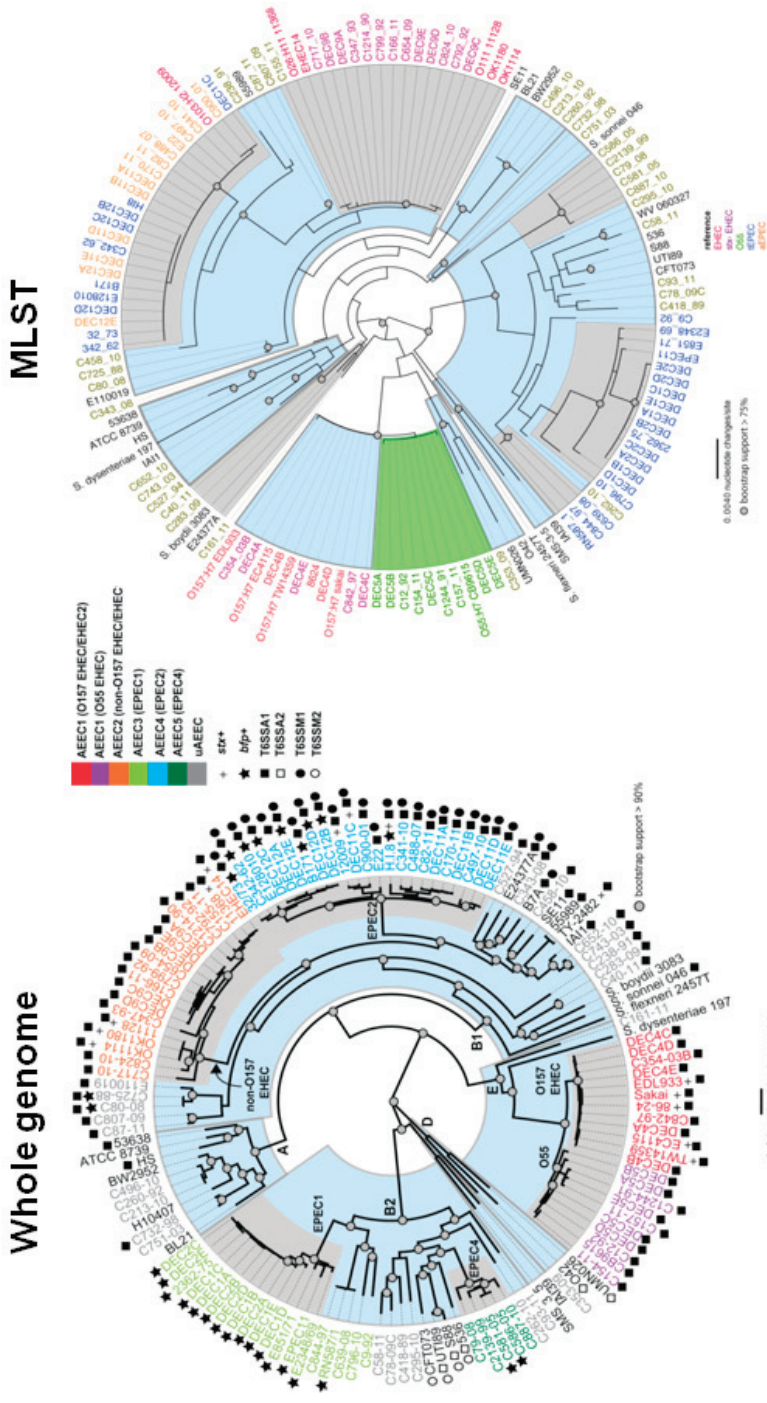


FIGURE WO-15 Inconsistency of typing with whole genomes or MLST.
SOURCE: Rasko (2012).

weakness, Rasko is using what he terms "genomic epidemiology" to look for unique genome signatures that are present in EHEC but not EPEC, or vice versa. This is based on alignment of the genomes, he explained, and is gene independent. Out of the genome features that were identified as being different between the groups, about 10 percent were known virulence factors unique to one of the groups. Functional analysis of the previously unknown and unique features is under way.

Genomic information is also informing the development of group-specific biomarkers. Detection of virulence factors using PCR is commonly used to identify an isolate. Rasko explained that virulence factors are often not the best identification markers. All attaching-and-effacing *E. coli* isolates will have genes for components of the type III secretion system, while the presence of genes for other virulence factors (Shiga toxin 1 and 2, bundle-forming pilus) is extremely variable. In addition, virulence factors tend to be mobile. Novel genome regions that are unique to a specific group provide a more effective way to rapidly identify the phylotype of an isolate using PCR. In his presentation, Rasko shared an algorithm for the identification of different attaching-and-effacing pathogenic *E. coli* (Figure WO-16).

The primary obstacle for developing diagnostics, vaccines, and therapeutics for EPEC, according to Rasko, is characterizing the pathogen beyond serotype and virulence factors. While the sequencing of isolates is relatively easy, there is a need for case-control studies with well-defined parameters, and patient and isolate metadata to inform the analysis and interpretation of genomic data. In addition, observed Rasko, our understanding of the population structure of these organisms is woefully inadequate to the task at hand. For example, how many distinct EPEC isolates are within one individual? What is the rate of variation within that host and within the environment? The development of rapid diagnostics requires appropriate comparison to close relatives, Rasko concluded, including both pathogens and commensals (the majority of *E. coli* in the gut are not pathogens). Rasko concurred with Casadevall that virulence is only expressed in a susceptible host, and he reiterated the importance of looking at population structure. We can develop probabilistic models in terms of whether an organism is likely to cause disease, he said; whether or not it does is entirely contextual.

Signatures of Selection and Transmission: Staphylococcus aureus, Streptococcus pneumoniae, Vibrio cholerae

The limitations of current typing methods—such as MLST—to discriminate between closely related organisms poses a challenge to tracking the transmission and adaptation of very recently evolved strains of pathogenic bacteria. Julian Parkhill of the Sanger Institute in Cambridge (UK) reviewed several examples of how high-throughput genomics can provide the higher-resolution view required in order to better understand the epidemiology and microevolution of strains that

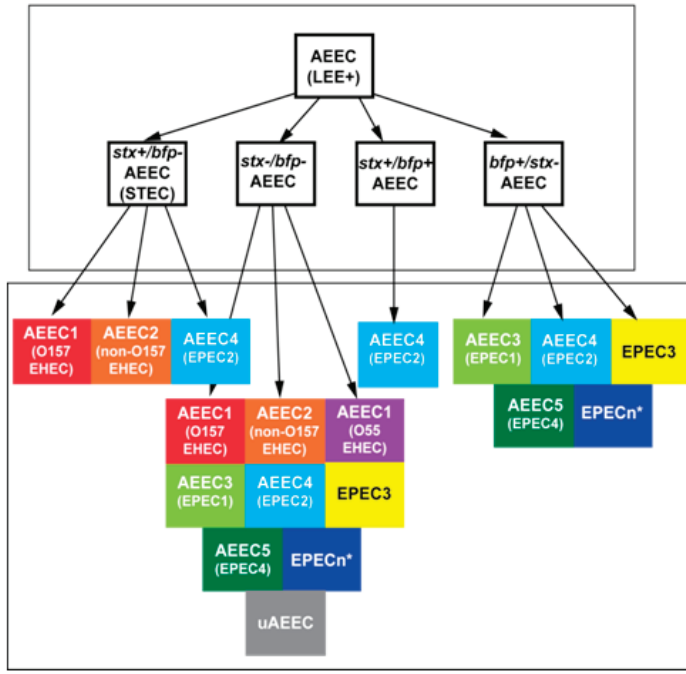


FIGURE WO-16 Algorithm for the identification of attaching-and-effacing *E. coli* (AEEC) pathogens. Classically, LEE-positive isolates are assayed for the presence of other virulence factors and subgrouped accordingly (top box). The introduction of phylogenetic markers allows for more rapid and accurate subgrouping (bottom box).
NOTE: *stx*, Shiga toxin gene; *bfp*, bundle-forming pilus structural gene; STEC, Shiga toxin-producing *E. coli*.
SOURCE: Rasko (2012).

have evolved and spread within the past 30 to 40 years (Dr. Parkhill's contribution to the workshop summary report may be found in Appendix A, pages 257-269).

Parkhill noted the multiple evolutionary processes that have an effect on bacterial genomes—including random mutation, homologous DNA exchange, acquisition and loss of genes, genetic drift, and Darwinian selection. Parkhill emphasized that these processes are acting simultaneously on different time scales with different strengths. Whole genome data help researchers to distinguish between the effects of these different processes and to identify underlying signatures of selection and transmission.

Identifying signatures of selection Parkhill discussed one strain—sequence type 239 (ST239)—of the globally important human pathogen methicillin-resistant *Staphylococcus aureus* (MRSA). This multi-antibiotic-resistant strain

arose in Europe and, later, spread globally (Harris et al., 2010).[25] According to Parkhill, although 63 different isolates of this strain are indistinguishable by current typing techniques (including MLST and pulsed-field gel electrophoresis), whole genome data has revealed a great amount of variation—with 4,310 SNPs in the core sequence alone.

Parkhill observed that the isolates' SNPs were almost entirely random, which would suggest the absence of immediate selection. SNPs are randomly distributed across the genome and exhibit a random rate for acquisition over time. Parkhill's group constructed a maximum likelihood phylogeny based on the 4,310 sites in the core genome of ST239 that contained one or more SNPs. Charting root-to-tip distance against the year of strain isolation suggested that this strain arose in the late 1960s (corresponding with the first administration of methicillin in 1959 and the subsequent emergence of MRSA in Europe). Parkhill noted that the rate of SNPs acquisition of approximately 3×10^{-6} per site per year is about 1,000 times faster than the accepted mutation rate of bacterial SNPs; it is, however, consistent with what is typically found in studies of very recently evolved organisms (Croucher et al., 2011; Mutreja et al., 2011).

Researchers use the ratio of nonsynonymous changes to synonymous changes (dN/dS) as a measure of selection. For ST239 the ratio is close to 1 (0.68).[26] Parkhill noted that in this case the rate is very close to 1, not because there is neutral selection, but because selection has not had time to act. Subsequent studies provided insights into how the rate of change, and therefore the rate of selection, varied with time (as measured by similarity between genomes) and for different regions of the genome (Castillo-Ramirez et al., 2011). As described by Parkhill and depicted in Figure WO-17, "The more closely related the genomes are, the closer the dN/dS ratio is to 1. That means that as the SNPs occur, there is no selection going on. dN/dS is 1 simply because nonsynonymous changes and synonymous changes are occurring randomly at the same rate." Over time, Parkhill continued, selection starts to act. "You can see selection acting as the dN/dS falls away." Parkhill emphasized that the rate of change is different between the core and the noncore genes. The noncore genes may contain mobile elements or DNA that has been exchanged with other strains. This has two effects; the changes are effectively "older," which increases the likelihood that mutations have undergone selection, and the changes are drawn from a larger effective population size, which increases the effectiveness of selection.

[25] "By analyzing whole genome data of a collection of MRSA ST239, we have gained new insights into fundamental processes of evolution in an important human pathogen. By creating a precise and robust phylogeny for the collection, we now have a highly informative perspective on the evolution of the clone" (Harris et al., 2010).

[26] Synonymous changes are considered "silent" because the base change within an exon of a gene coding for a protein does not result in changes to the protein's amino acid sequence. Nonsynonymous changes result in altered amino acid sequences. A dN/dS ratio greater than 1 suggests diversifying (or disruptive) selection, while a ratio below 1 is associated with purifying (or stabilizing) selection, and a ratio close to 1 indicates neutral selection.

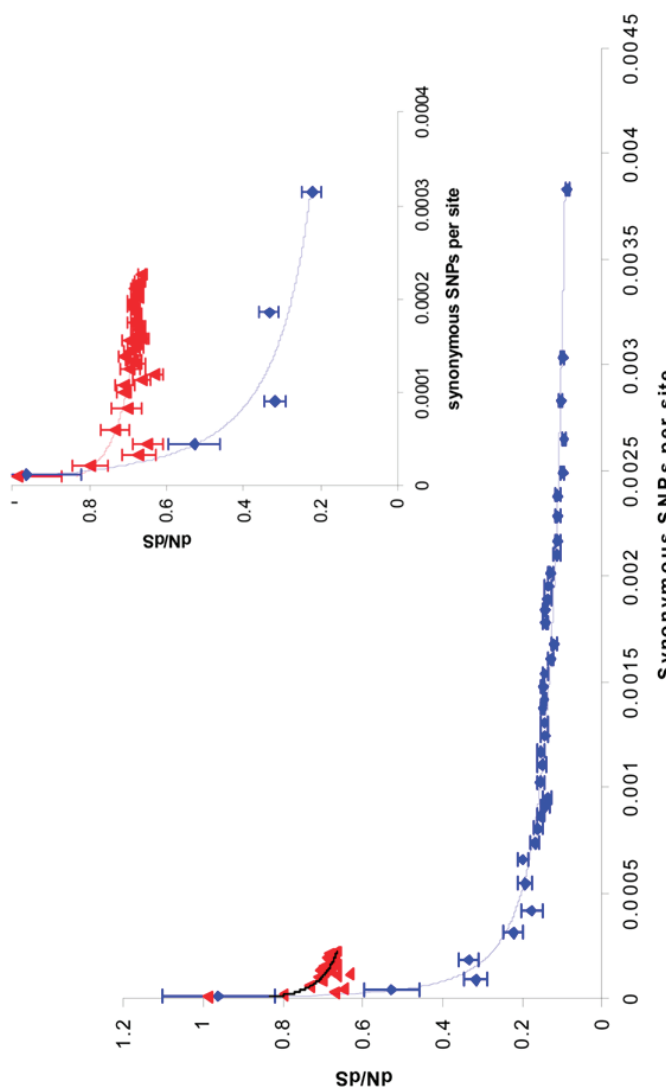


FIGURE WO-17 Time dependence of dN/dS in the core and non-core genome. dN/dS ratio for 1,953 pairwise ST239 strain comparisons for the two sets of genes. The number of synonymous SNPs per synonymous site is used as a measure of divergence. The data for the core are shown in red, the non-core in blue. To clarify the plot for the core data, which is far less diverged than the non-core, the figure was rescaled (inset). SOURCE: Castillo-Ramierez et al. (2011).

The random nature of mutation in the core genome allowed Parkhill to use these sequences to build a robust and congruent phylogenetic tree. In so doing, he was also able to identify a subset of sites that were "acting unusually," suggesting that they were under selection. Of the 4,310 SNPs in the tree, 38 loci were homoplasic, that is, they appeared independently on different branches as a result of convergent evolution (Harris et al., 2010). Parkhill explained that the lack of recombination in the tree overall suggests that these homoplasies are under recent and very strong selection. Eleven of these loci correspond to known drug resistance mutations suggesting an association with antibiotic use in clinical practice. When most SNPs are random and overall recombination is moderate, homoplasy works very well as an alternative to dN/dS and other classical measures to identify selection.

Parkhill noted the use of homoplasy to identify selective pressures acting on isolates of other pathogenic bacteria, including *Mycobacterium tuberculosis* and *Clostridium difficile*. "By looking, not for dN/dS or all those classical measures of selection, but for things that don't fit the tree," Parkhill said, researchers can see selective pressures such as drug resistance and compensatory mutations, as well as evidence of selection on surface proteins, two-component sensor/regulators, and other genes that are likely to be under diversifying selection from host immune pressure.

Parkhill also described an approach that makes it possible to use homoplasy to identify selection in organisms where recombination is very common such as *Streptococcus pneumoniae*.[27] Because they are derived from other strains of *S. pneumoniae*, recombined regions are representative of the wider species, noted Parkhill, and therefore have an older date of origin. Constructing a coherent, consistent phylogenetic tree of the multi-drug resistant clone of *S. pneumoniae* (sequence type 81 and serotype F) that emerged in the 1970s required the identification and elimination of recombined regions so that only the vertically transmitted, random point mutations remained. Parkhill and colleagues were then able to put the recombination events back onto the tree to see where, when, and why they occurred (for more information see Croucher et al., 2011).

Identifying transmission Parkhill explained that the development of a congruent phylogenetic tree—such as that developed using the *S. aureus* ST329 data set—can also be used to define transmission events—from intercontinental transmissions over the course of decades to person-to-person transmission within a

[27] "The ability to distinguish vertically acquired substitutions from horizontally acquired sequences is crucial to successfully reconstructing phylogenies for recombinogenic organisms such as *S. pneumoniae*. Phylogenies are, in turn, essential for detailed studies of events such as intercontinental transmission, capsule type switching, and antibiotic-resistance acquisition. Although current epidemiological typing methods have indicated that recombination is frequent among the pneumococcal population, they cannot sufficiently account for its impact on relations between strains at such high resolution" (Croucher et al., 2011).

hospital; a robust tree even allows researchers to identify ongoing transmission events (Harris et al., 2010).

Parkhill concluded his presentation by demonstrating how whole genome approaches have clarified the origin and international spread of a globally important pathogen, *Vibrio cholerae*. The El Tor strain of this pathogen is associated with the seventh cholera pandemic that originated in the early 1960s and persists today—including the 2010 Haitian cholera outbreak. Before whole genome sequencing, cholera typing was done on the basis of the presence or absence, or variable sequences, of mobile elements. This approach identified a superficial diversity among cholera strains when, in fact, the core genome—excluding possible recombination events—forms a robust transmission tree (Mutreja et al., 2011).

For example, SXT is a mobile element that encodes multi-drug resistance. Looking only at the tree generated with the core genome, its presence in the El Tor chromosome appears to have resulted from a single acquisition. However, generating a tree using 3,000 SNPs in the 60 kb SXT element itself, reveals that the phylogeny of the element is different from that of the core genome. This phylogeny also reveals an apparent rate of SNP accumulation in the SXT element that is 100 times greater than that of the core El Tor genome. Parkhill noted that this elevated rate of mutation acquisition is "very, very unlikely to be true." Rather, this result suggests that the SXT element is much older and is evolving outside of El Tor. Indeed, comparing the SXT tree with a tree based on the core El Tor genome, novel versions of the SXT element appear to have entered the El Tor chromosome at least five times (Mutreja et al., 2011). "Trying to understand the transmission of cholera based on [the analysis of] mobile elements like SXT is doomed to failure," Parkhill said, "because this approach is reporting the movements of older elements in and out of the organism's genome, rather than the phylogeny of the core genome, which should represent the ancestry of the organism itself. To understand transmission, we need to strip out all of these mobile elements and recombinations and look only at the core genome."

The consistent rate of SNP accumulation of the core genome among isolates also illuminated the origin and spread of El Tor as part of the seventh cholera pandemic. The El Tor core genome tree shows three groups of isolates, which correlate with three different time periods. When coupled to the physical origin of each isolate, Parkhill said, you can see that over the past 40 to 50 years, "three independent waves [of El Tor strains] have spread around the world from a single location, almost certainly around the Bay of Bengal" (Mutreja et al., 2011). The El Tor strain of *V. cholerae* has been introduced multiple times into different parts of the world. New outbreaks are derived from the trunk of the tree, then expand and die out (Figure WO-18) (Mutreja et al., 2011).

In summary, Parkhill said that different processes are acting on different time scales with different strengths. Whole genome sequencing allows investigators to identify and separate out the varying effects of mobile elements, recombination, and point mutation. The resulting high-resolution phylogenies can provide

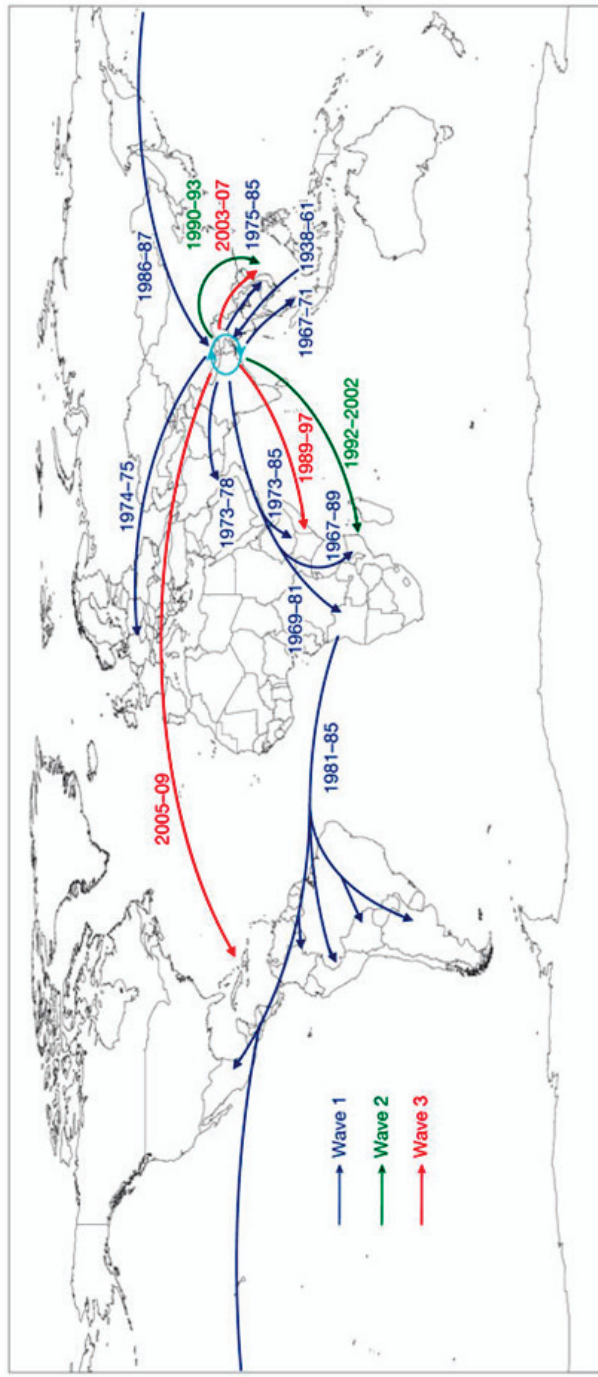


FIGURE WO-18 *V. cholerae*—repeated global transmission. Transmission events inferred from the phylogenetic tree of the seventh pandemic lineage of *V. cholerae* based on SNP differences across the whole core genome, excluding probable recombination events.
SOURCE: Mutreja et al. (2011). Reprinted by permission from Macmillan Publishers Ltd: NATURE. Mutreja, A. et al. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477(7365):462-465, copyright 2011.

the basis for important epidemiological and phenotypic analyses and insights to illuminate patterns of gene selection and transmission.

Intra-Host Diversity, Selection, and Evolution: Influenza Virus

Investigators are also exploiting next-generation sequencing approaches to understand how viral genetic diversity changes within a host and during transmission between hosts. Most of what is known about the intra-host diversity of viruses is derived from research on viruses that cause persistent infections (e.g., HIV, hepatitis C virus), where accumulation of replication errors and recombination over time leads to high levels of diversity. Speaker Elodie Ghedin of the University of Pittsburgh School of Medicine elaborated on her interest in intra-host diversity, selection, and evolution of influenza viruses, which cause acute infections (Dr. Ghedin's contribution to the workshop summary report may be found in Appendix A, pages 151-165). Her work focuses on the following questions:

- Does natural selection occur within individual hosts?
- How big is the population bottleneck at transmission?
- What is the extent of mixed infection?
- What is the mutational spectrum within an individual host?
- What is the fitness distribution of these mutations?

Influenza is a negative strand RNA virus with an eight-segmented genome. Coinfection of a host cell with different strains can result in reassortment of the segments and a mixed population. Ghedin noted that during the H1N1 influenza pandemic, researchers were able to track some of the segments to two different types of swine flu. Whole genome sequencing of influenza generally produces a consensus sequence with one genome that is representative of the strain an individual host carries. In the case of influenza, even though the strain could have high diversity, according to Ghedin you are only looking at what is dominant, so you just have a consensus of each of the segments. The use of next-generation sequencing and methodologies will be important to capture this diversity, Ghedin said.

Influenza can adapt to its host after transmission, or changes may occur before transmission, facilitating movement into a different host species. Immune status of the host (e.g., whether the host is immunocompromised or vaccinated) can also influence the evolution and co-adaptation of a virus. In an immunocompromised host, influenza infection persists well beyond the typical 5 to 7 days, in some cases as long as 5 to 6 months. In the case of a persistent infection, what happens to viral diversity?

Ghedin described the case of an immunocompromised boy infected with pandemic H1N1 influenza virus for 35 days. He was treated with Tamiflu (an antiviral neuraminidase inhibitor) on day 2, and at some time between when viral

RNA was sampled and sequenced on days 6 and 13, a Tamiflu-resistant variant had emerged. This variant may have been present in the population before treatment was started, or it could have been the result of a *de novo* mutation. Current thinking is that because the drug-resistance mutation also reduces viral fitness, the emergence of resistance is more likely to result from a *de novo* mutation that is favored under drug pressure. Still, as Ghedin pointed out, the use of consensus sequences does not allow researchers to distinguish between these two possibilities.

In contrast to consensus sequencing, deep sequencing identifies minor variants (above the background of error) and reports the percentage of reads that have the dominant or minor codons at a given position. This approach is illustrated in Figure WO-19, which depicts the alignment of short sequence reads from a sample to the gene sequence of interest—in this case the neuraminidase gene sequence of influenza. Whereas consensus assembly would read the codon at position 275 as CAC (coding for histidine), deep sequencing reveals that some sequences read TAC (coding for tyrosine), which is a codon associated with drug resistance. In the case of this boy, the variant codon (TAC) was present in a very

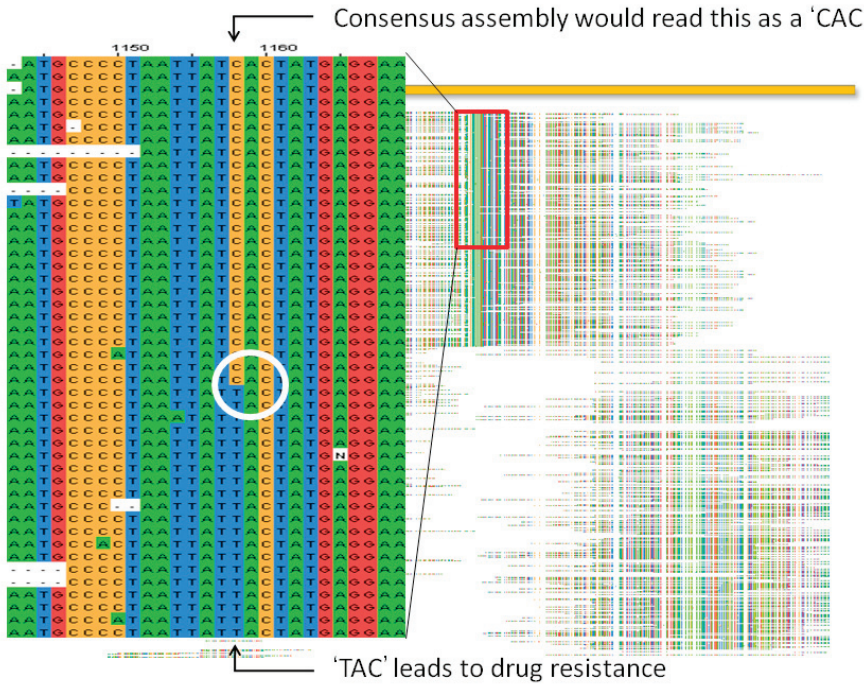


FIGURE WO-19 Next generation sequencing and new methodologies will help to capture intra-host diversity.
SOURCE: Ghedin (2012).

low percentage of reads in samples taken at day 1 (pre-treatment) and day 6; by day 13, this variant was the dominant codon.

When the 2009 H1N1 influenza A virus emerged in the United States, there were multiple clades of viral strains that, while antigenically similar, were clearly clustered by geographic region. As Ghedin observed, "We know a lot about the circulation of different strains, but we know very little about what is actually carried by individuals." Ghedin's research has also explored whether the variants observed in a patient resulted from a *de novo* mutation in the wild-type or from a mixed infection in which the individual is infected with multiple variants either simultaneously or sequentially.

By the second wave of the pandemic, there was a complete mixing geographically, and a single viral clade became dominant across the country, while others appeared to fade away. This is a typical pattern, Ghedin noted. She noted that given this observed tendency toward one dominant viral clade, mixed infections[28] (which can occur and are an important driver of genotypic diversity) are generally considered to be rare (Ghedin et al., 2011).

Consensus sequencing of samples taken from an immunosuppressed patient during the second wave of the pandemic showed that the patient's virus was drug sensitive on both day 1 (pre-treatment) and on day 14 (after antiviral treatment early in infection). Deep sequencing, however, identified three distinct variants. Partial reconstruction of sequence reads into different viral genomes placed two variants into two phylogenetically distinct clades of the pandemic H1N1/2009 virus, a result that strongly suggests a mixed infection (Ghedin et al., 2011). Ghedin concluded that "when we see strains disappear, it doesn't mean they are completely gone. They can still be present at a low level and may lead to emergence [of new phenotypes]."

Having identified the presence of multiple viral variants within an individual, Ghedin sought to examine how transmission affected diversity. She cited a clear case of transmission of influenza from a son to his father in which the father was prophylactically treated with Tamiflu at the same time the son started therapeutic treatment (Baz et al., 2009). Six days after the start of his Tamiflu treatment, the father had a completely resistant virus. Deep sequencing revealed that more than 2 percent of the son's viral population was drug-resistant; further analysis demonstrated that multiple variants were being transmitted (i.e., the transmission bottleneck was not especially narrow) (Ghedin et al., 2012). While some variants are lost at transmission, others attain a higher frequency in the recipient.

Ghedin suggested that this improved understanding of the role of transmission in shaping the genetic diversity of influenza may help to refine influenza transmission models. In contrast to consensus sequences, deep sequencing can provide the fine-grained genomic mapping needed to identify significant changes

[28] Infection in which the individual is infected with multiple variants either simultaneously or sequentially.

in the distribution in the intra-host population. These approaches may also help to reconstruct chains of transmission for an epidemic—insights which may inform future response strategies.

Evolution of Novelty and Pathogenicity: Chytrid Fungal Pathogen of Amphibians

Novel microbial pathogens do not just appear, they evolve, according to speaker Erica Bree Rosenblum of the University of California, Berkeley (Dr. Rosenblum's contribution to the workshop summary report may be found in Appendix A, pages 291-311). Genomics, she continued, can provide a useful tool for studying a pathogen's evolutionary history. Rosenblum's presentation described an emerging fungal pathogen of amphibians—*Batrachochytrium dendrobatidis*—and underscored the importance of studying novelty at multiple levels, because a pathogen may be novel in multiple ways including changes in geographic range, host range, virulence, function, etc.. She noted that because of recent technological advances, genomic studies can now be conducted for time-critical studies in non-model species—even in ecological systems where there has not been much previous research.

***Bd* and amphibian declines** *Batrachochytrium dendrobatidis*, referred to as *Bd*, is a chytrid fungus that has been implicated in worldwide declines and possible extinctions of amphibian populations. With the exception of Antarctica, *Bd* occurs on every continent, infecting more than 500 species of amphibians. Genetic and spatial-temporal data demonstrate that *Bd* is a novel, emerging, pathogen that has quickly spread around the world. Data from many parts of the world trace *Bd*'s arrival and subsequent spread in a wave-like fashion throughout amphibian populations (Lips et al., 2006). Rosenblum noted the rejection of an early hypothesis that *Bd* had long existed as a commensal organism and that this relationship had changed due to shifts in the environment.

Chytrids are basal fungi, and almost all known chytrids are saprobes—organisms that live on decaying organic matter in leaf litter or aquatic environments. *Bd* is unique among chytrids because it is the only species known to infect vertebrates. Rosenblum noted that there appears to have been a very recent change in *Bd* that allows it to exploit this new host niche. *Bd* kills frogs by disrupting the structure and function of their skin such that important skin functions, including osmoregulation and electrolyte balance, are compromised. Recent studies suggest that *Bd* may also have an immune evasion or suppression strategy, and there is speculation that *Bd* may also release toxins (Rosenblum et al., 2009, 2012).

Although *Bd* was discovered and described more than 10 years ago, several persistent unanswered questions remain about the origin and spread of *Bd* and the interaction between *Bd* and its amphibian hosts. Simply stated: where did it come

from, and, what makes it so deadly? Rosenblum noted that ecological analyses alone cannot answer many of the questions about *Bd* and, as yet, *Bd* cannot be manipulated for cellular and molecular analyses. Rosenblum is using comparative and functional genomics to try to understand *Bd* evolution and pathogenicity.

Using genomics to understand evolutionary novelty of *Bd* Rosenblum first explored novelty at the phylogenetic level looking for important genetic or functional variation within isolates from around the world, which are collectively described under the single species name, *Bd*. Sequencing of the genomes of these 28 *Bd* isolates, as well as the genome of the closest known nonpathogenic chytrid, produced more than 100,000 SNPs that could be used for understanding the evolutionary history of *Bd*.

Given the rapid spread of *Bd* around the world and its presumed recent origin, Rosenblum said she expected samples from all over the world to appear all mixed up along very short branches, or a tree reflecting a linear progression from a basal group to subsequent radiations into new geographic or host-specific clusters. Instead, the evolutionary history of *Bd* was much more complex. As illustrated in Figure WO-20, there is no clear point-source for the origin of *Bd* or any linear history of how it has spread around the world. The absence of geographic or host-specific population structure confirms the rapid spread and broad host range of *Bd*. Yet the tree has more structure than expected, exhibiting two highly divergent lineages: a basal lineage with isolates from Latin America and a large clade with significant global diversity. Rosenblum suspects that with more geographic sampling, *Bd* will continue to look more like a "they" than an "it."

Using genomics to understand functional novelty of *Bd* As previously noted, *Bd* exhibits functions that no other chytrid has acquired. Comparative chytrid genomics has revealed that while most fungal genomes had one or several copies of genes from different protease gene families, the *Bd* genome had massive expansions of protease gene families—dozens of fungalysin metallopeptidases and serine protease genes as just two examples (Rosenblum et al., 2008). This is interesting, Rosenblum explained, because these two protease gene families have been implicated as potential pathogenicity factors in other fungal pathogens that infect vertebrates including for example, tinea and ringworm, as well as other dermatophytic fungi that infect vertebrate skin. To follow up on these observations Rosenblum again used genomic approaches to look for functional and evolutionary evidence that these proteases may have been important to *Bd*'s transition from saprobe to amphibian pathogen. Functional genomics studies of the presence or absence of gene expression at different life stages or in response to different nutrient conditions (nutrient broth versus frog skin) have identified gene copies of particular interest—those that show higher levels of expression in the host tissue and during life stages in which *Bd* is known to infect amphibians (Rosenbaum et al., 2012).

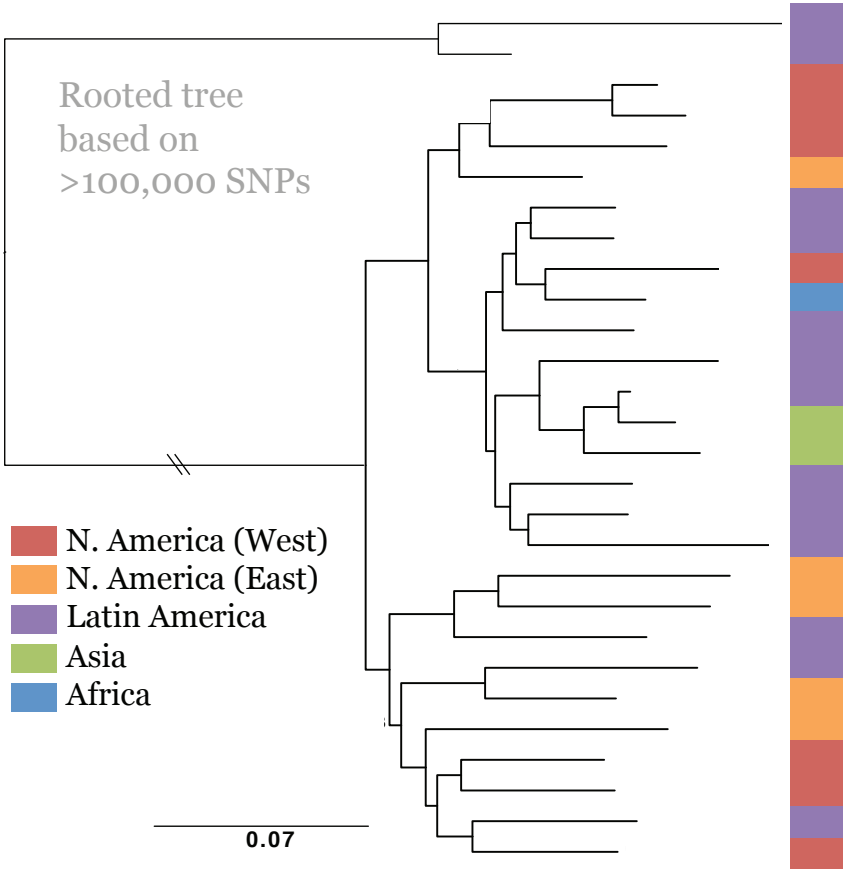


FIGURE WO-20 Novelty of *Bd* at the phylogenetic level. Phylogenetic analysis of 28 *Bd* isolates from around the world revealed two very divergent lineages: (1) a basal lineage containing isolates from Latin America that are very divergent from everything else and (2) a large clade containing most of the global diversity.
SOURCE: Rosenblum (2012).

Rosenblum and colleagues compared the genomes of *Bd* and its closest relative, the nonpathogenic saprobe *Homolaphlyctis polyrhiza*, in order to investigate whether the protease gene family expansions accompanied the evolution of pathogenicity. Results suggested these gene family expansions are recent, and generally *Bd* specific—for example, *Bd* was found to have 38 fungalysin peptidase gene family copies where *H. polyrhiza* has 5 (Joneson et al., 2011). This study also revealed that *Bd* has 62 Crinkler-like proteins, which are proteins that have never before been described in any fungal species. What is interesting

about this observation, according to Rosenblum, is that the closest *Bd* relative that has Crinklers are the oomycetes, in which they [Crinklers] are believed to be virulence effectors.

Rosenblum concluded that genomic approaches have been invaluable in understanding the complex history of *Bd* and identifying its evolutionary transition points. Genomic approaches have also assisted in developing hypotheses about functional aspects of *Bd*'s lifestyle. Using these genomic tools and approaches to *Bd* has revealed that this fungal vertebrate pathogen has a much more complex and deeper evolutionary history than had been previously appreciated. These same approaches, according to Rosenblum, may also help investigators identify key transition points in the evolutionary history of this pathogen.

Microbial Communities in Coral Health and Disease

Corals are very simple animals that harbor diverse microorganisms in and on their tissues—including archaea, bacteria, viruses, and zooxanthellae (Figure WO-21). Endosymbiotic dinoflagellate algae (*Symbiodinium*, or more commonly, Zooxanthellea) live within the endothelial tissue of the coral. The symbiotic relationship between the coral animal and this single-celled algae has been well studied, but far less is known about the bacterial communities that are associated with the algae and the coral, explained speaker Kim Ritchie of the Mote Marine Laboratory (Dr. Ritchie's contribution to the workshop summary report may be found in Appendix A, pages 269-290).

Molecular methodologies and metagenomic sequencing confirm the broad diversity of bacterial species associated with corals. Some bacteria—*Vibrio* species, *Serratia* species, *Aspergillus sydowii*, *Aurantimonas coralicida*—are opportunistic pathogens that have been implicated in coral diseases when the coral communities are stressed. Other bacteria may provide beneficial services to the corals. Ritchie reviewed current research using culture-based approaches that explore the nature and specificity of coral-bacterial associations, and the ways that these associations are maintained over time.

Bacteria and early life stages of coral Ritchie also reviewed recent research exploring bacterial colonization of developing coral tissue and the role of bacterial biofilms in this symbiotic association. There are two general types of corals; broadcast spawners and brooding corals. Broadcast spawners produce eggs and sperm that are externally fertilized, and the eggs therefore acquire *Symbiodinium* algae, horizontally from the water column. Brooding corals have internal fertilization and release planula larvae that already have the parent *Symbiodinium*. Bacterial colonization of most corals (both brooders and broadcast spawners) occurs during planula larvae or postlarval settlement stages[29] much like the Hawaiian

[29] According to Ritchie, an exception is the Caribbean coral *Porites astreoides*, in which bacteria are passed directly from parent to offspring (vertical transmission).

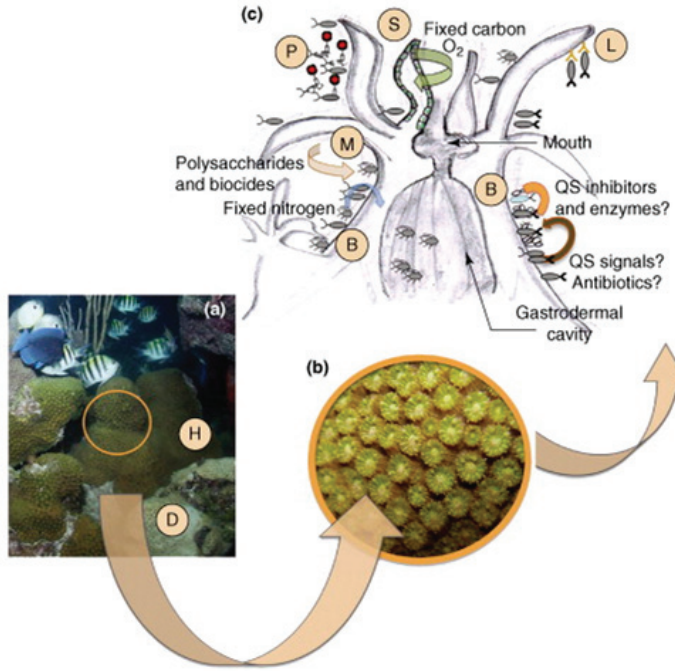


FIGURE WO-21 Interactions within coral reef communities. Panel a, Healthy coral reefs support a diverse ecosystem, as shown here by a healthy *Montastraea faveolata* colony (H) and a dead colony (D). Panel b, A coral colony comprises genetically identical polyps that are connected by a living tissue. Panel c, Functionally and taxonomically diverse microbial communities are found on the coral surface, within the gastrodermal cavity and in the intracellular spaces of the polyp. Coral-associated bacteria (B) are thought to supply the host with fixed nitrogen. Bacteria produce antimicrobials with a potential function in host defense. Symbiotic dinoflagellates (*Symbiodinium* spp., [S]) are housed within cells of the ectoderm. *Symbiodinium* provides photosynthate to its animal host, which then supplements its diet by feeding on plankton. In addition to satisfying the energy needs of the coral animal host, the photosynthate from *Symbiodinium* is converted into glycoprotein-rich mucus (M) within specialized cells of the polyp, called mucocytes. Mucus is then excreted onto the coral surface. Corals actively manipulate their associated microbial communities, and the mucus contains nutrients that support growth of bacteria and eukaryotes. Antimicrobials in coral mucus may also contribute to the coral microbiome. Coral lectins (L) bind and coagulate bacteria. Populations of coral-associated bacteria are also controlled by phages (P)—viruses of bacteria that may function as biocontrol agents for opportunistic pathogens of corals.
NOTE: Reproduced with permission from Jonathan Onufryk (a) and Erich Bartels (b). Schematic drawing (c) not to scale.
SOURCE: Teplitski and Ritchie (2009). Reprinted from *Trends in Ecology & Ecolution*, 24, Max Teplitski and Kim Ritchie, How feasible is the biological control of coral diseases? 24(7):378-385, Copyright 2009, with permission from Elsevier.

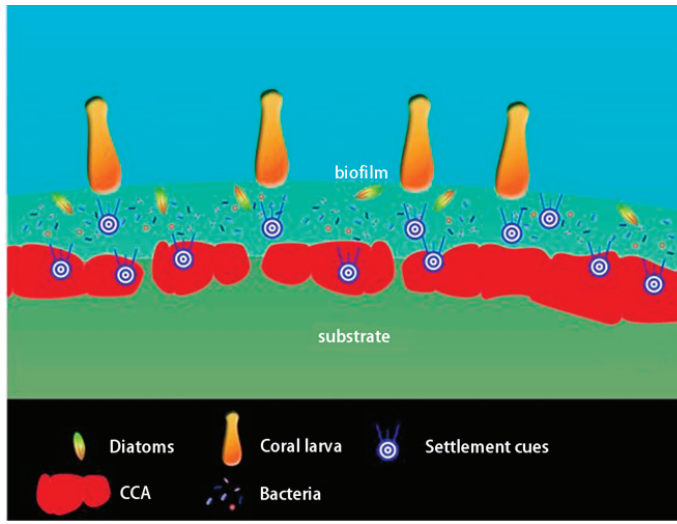


FIGURE WO-22 Microbial biofilms are necessary for larval settlement. Crustose coralline algae (CCA) and biofilm microbial communities facilitate attachment and settlement of coral larvae via inductive compounds (settlement cues) produced by the CCA or by recruiting specific bacteria that release these cues.
SOURCE: Sharp and Ritchie (2012). Sharp, K. H., and K. B. Ritchie. 2012. *Biological Bulletin* 174:319-329. Used with permission from the Marine Biological Laboratory, Woods Hole, MA.

bobtail squid *Euprymna scolopes* forms a persistent association early in its life cycle with the Gram-negative luminous bacterium *Vibrio fischeri* (McFall-Ngai et al., 2012; Sharp and Ritchie, 2012; Sharp et al., 2010).

According to Ritchie, *Roseobacter* clades are present in the early life stages of many corals. Their consistent detection in seawater during coral spawning suggests their potential importance for mediating larval settlement and survival. Ritchie cited a recent study that suggests that coral larvae need a microbial biofilm on some type of substrate, preferably calcium, for settlement—as illustrated in Figure WO-22 (Sharp and Ritchie, 2012). *Roseivivax* (in the *Roseobacter* clade) and *Marinobacter* were among the bacteria found to encourage coral larval settlement.

Ritchie also reviewed the recent finding that all of the *Symbiodinium*-associated roseobacters tested produced gene transfer agents (GTAs) (McDaniel et al., 2010, 2012; Paul et al., *in review*). Ritchie explained that GTAs resemble bacteriophage, packaging random pieces of host DNA and transferring them to other bacteria (McDaniel et al., 2010). Interestingly, gene transfer via GTAs was found to be 100 million times higher in the reef environment than in open oceans and in Tampa Bay. There are also more roseobacters in the reef environment,

Ritchie added, particularly during coral spawning, and experiments suggest that *Roseobacter* and gene transfer agents may increase larval settlement.

Microbial community regulation of coral disease development When stressed by environmental conditions—such as elevated sea surface temperatures—corals often "bleach." As corals begin to bleach, the population of *Vibrio* species that are normally associated with healthy corals and their symbiotic algae increase while the population of other types of bacteria decline. Studies by Ritchie to explore the coral bleaching phenomenon have revealed that the surface mucus of healthy elkhorn coral, *Acropora palmata*, may be a potent defense against disease. *In vitro*, it inhibits the growth of a range of potentially invasive microorganisms. Ritchie identified a number of antibiotic-producing bacterial species that were isolated from surface mucus that could contribute to this defense. Mucus collected during a bleaching event (i.e., when the coral were unhealthy or stressed) were found to lack this antimicrobial activity (Ritchie, 2006).

By sampling corals monthly, Ritchie discovered that when sea surface temperatures increase, the population of antibacterial-producing bacteria decreases, while the population of potentially pathogenic bacteria (including *Vibrio* species) increases (Ritchie, 2006). A mathematical model based on these findings suggests that there is a lag time in coral recovery following a warming event (Mao-Jones et al., 2010). Consistent with this prediction, Ritchie has detected overgrowth and persistence of different types of *Vibrio* species in the coral mucus long after environmental conditions were once again favorable for healthy coral populations and their associated microbiota. This lag in response and recovery, or hysteresis, may explain why corals are more susceptible to disease and bleaching both during and after ocean warming events, Ritchie said.

Ritchie used both culture-based and molecular methods to identify bacteria associated with a diverse collection of *Symbiodinium* samples. Bacteria isolated were primarily members of three bacterial groups: the *Roseobacter* clade, marinobacters (oceanospirillates group), and the *Cytophaga-Flavobactrium-Bacteroides* group. Emerging evidence suggests that *Roseobacters* are present in the early life stages of corals and are abundant on coral reefs when coral spawn. Simple laboratory experiments have revealed that roseobacters produce antimicrobial compounds, enhance the growth of *Symbiodinium*, and reverse symptoms of disease caused by the putative coral pathogen, *Serratia marcescens* (Krediet et al., *in press;* Ritchie, 2011). Together, these data suggest that roseobacters are likely to be critical in the maintenance of a healthy coral ecosystem.

Many of the beneficial bacteria isolated by Ritchie produce compounds associated with a cell-to-cell communication system used by a variety of bacterial species to mediate host–microbe interactions. These signals have been shown to be produced *in situ* on the surface of corals and to inhibit swarming and biofilm formation by coral pathogens (Alagely et al., 2011). Ritchie is using the polyp anemone, *Aiptasia pallida*, as a model system (there are no adequate coral

models) to explore interactions between bacterial community members, their coral hosts, and potential pathogens. As reported by Sharp and Ritchie (2012), *Aiptasia pallida* presents an opportunity to integrate a model systems approach with novel technologies from the "omics age" to learn more about multipartner interactions in corals in a moment of great environmental change.

METAGENOMICS

Exploring Microbial Diversity

As emphasized throughout the workshop, microbial genomics has supported the exploration of the vast diversity of the unseen microbial world. During the past decade, the scope and scale of these studies have increased from studies of low-complexity microbial communities (such as those found in acid mine drainage biofilms) to broad surveys of how complex microbial communities vary across space and time (Hugenholtz and Tyson, 2008) (Figure WO-23). The Human Microbiome Project, for example, seeks to map microbial communities associated with the different environments on and in the human body (e.g., gut, mouth, skin, vagina), track how these communities differ by individual and his or her health status, and identify how the microbiota (individual/community) contributes to states of health and disease (Turnbaugh et al., 2007). The Earth Microbiome Project is a metagenomics survey that will collect natural samples and analyze microbial communities from around the world; it is anticipated that 200,000 environmental samples will be collected from across the various biomes of Earth and sequenced for taxonomic and functional analysis (Gilbert and DuPont, 2011).

Metagenomic studies have provided insights into the rich and untapped genetic potential of microorganisms, the functional (metabolic) potential of microbial communities, and the structure (species richness and distribution) of communities in a wide range of environments. Examples include:

- The analysis by Venter et al. (2004) of seawater from the Sargasso Sea resulted in the identification of more than 1.2 million new genes, including more than 700 new rhodopsin-like photoreceptors. These proteins are now thought to be a major source of energy flux in the world's oceans (Hugenholtz and Tyson, 2008).
- Tyson et al. (2004) were able to reconstruct two almost complete genome sequences of *Leptospirillum* group II and *Ferroplasma* type II and the partial sequence of three other species from a low-complexity acid mine drainage biofilm. Genome analysis for each organism revealed specified pathways for carbon and nitrogen fixation and energy generation. More recently, Denef and Banfield (2012) analyzed samples taken at this site over a 9-year period to measure the evolutionary rate of free-living

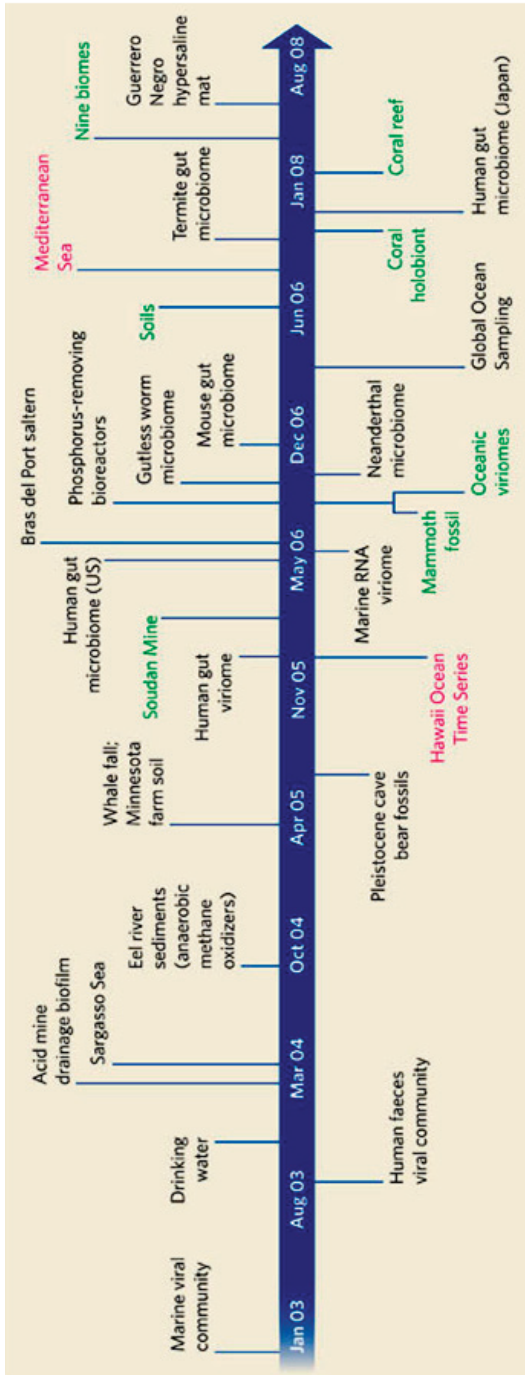


FIGURE WO-23 Timeline of sequence-based metagenomic projects showing the variety of environments sampled since 2003. The oceanic viromes (all viruses in a habitat) (August 2006) were from the Sargasso Sea, Gulf of Mexico, coastal British Columbia, and the Arctic Ocean. The nine biomes (March 2008) were stromatolites, fish gut, fish pond, mosquito virome, human-lung virome, chicken gut, bovine gut, and marine virome. The different technologies used are dye-terminator shotgun sequencing (black), fomid library sequencing (pink), and pyrosequencing (green).
NOTE: Graphic based on data sets represented at www.genomesonline.org.
SOURCE: Hugenholtz and Tyson (2008). Reprinted by permission from Macmillan Publishers Ltd: NATURE. Hugenholtz, P. and G. W. Tyson. 2008. Metagenomics. *Nature* 455:481-483, copyright 2008.

microorganisms in the wild. Such studies will illuminate how ecologic and evolutionary forces interact to shape microbial population dynamics.

- A recent effort to catalog the genes of the adult human gut microbiome suggested that the genes contained within the gut flora outnumber those contained within our own genome by 150-fold. An individual likely harbors at least 160 bacterial species in his or her lower digestive tract (Qin et al., 2010). These communities appear to be carefully calibrated enterprises, and the disruption of this delicate balance may contribute to a variety of diseases including obesity, autoimmune diseases, and asthma (Ley et al., 2008).
- Tringe et al. (2008) used a gene-centric analysis to assess DNA isolated from highly complex and nutrient-rich environments: soil and three isolated deep-sea "whale fall" carcasses. Comparisons of gene abundance may provide habitat-specific fingerprints that reflect known characteristics of the sampled environment and "hint at certain nutrition conditions, novel genes, and systems contributing to a particular life style or environmental interactions."

Workshop speakers elaborated on several current metagenomics projects that seek to provide insights into the interactions among microorganisms within a community.

Human-Associated Microbial Communities: Links to Health and Disease

Just as microbes colonize the bobtail squid's light organ shortly after hatching, microbes colonize the human body internally and externally during its first weeks to years of life and become established in relatively stable communities in a variety of microhabitats (Dethlefsen et al., 2007). Research to date suggests that the site-specific microbial communities—known as microbiota or microbiomes[30]—that inhabit the skin, intestinal lumen, mouth, teeth, and so on of most individuals contain characteristic microbe families and genera. The species and strains of microbes present on or in any given individual may be as unique as a fingerprint (Dethlefsen et al., 2007). The microbiota of other terrestrial—and possibly aquatic—vertebrates are dominated by microbes that are related to, but distinct from, those found in humans. This suggests that host species and their microbial colonists have uniquely coevolved with and adapted to one another.

As discussed previously, the complexity of the human microbiome is astounding. Each community has its own unique structure and ecosystem that is

[30] The term *microbiome* is attributed to the late Joshua Lederberg, who suggested that a comprehensive genetic view of the human as an organism should include the genes of the human microbiome (Hooper and Gordon, 2001). Because most of the organisms that make up the microbiome are known only by their genomic sequences, the microbiota and the microbiome are from a practical standpoint largely one and the same (IOM, 2009).

shaped by and actively influences the habitat within which it resides. As speaker George Weinstock of the Genome Institute at Washington University observed, "in all microbiomes, the organisms are talking to each other. They are talking to their hosts; they are doing things to the environment" (Dr. Weinstock's contribution to the workshop summary report may be found in Appendix A, pages 357-378). He noted that their dynamic nature makes them challenging to study, but that understanding these dynamics may shed light on "the emergent and amazing properties that microbiomes bestow on their hosts and environments." According to Weinstock, "the fundamental goal of human microbiome research is to measure the structure and dynamics of microbial communities, the relationships between their members, what substances are produced and consumed, the interaction with the host, and differences between healthy hosts and those with disease" (Weinstock, 2012a).

He offered four vignettes from research associated with the Human Microbiome Project to illustrate the challenges and insights emerging from studies of variation in the human microbiome:

- discovery of new bacterial taxa (using 16S rRNA sequence variation)
- discovery of new strains of known taxa (looking for within species variation)
- health versus disease (variation of community structure)
- tracking organisms (genetic variation in populations)

Discovery of new bacterial taxa The Human Microbiome Project has generated a massive amount of 16S rRNA gene sequence data.[31] 16S rRNA differs for each bacterial species (Weinstock, 2012a). A bacterial species is hard to define, but it is often thought of as organisms with 16S rRNA gene sequences having at least 97 percent identity—an operational taxonomic unit (OTU) (Weinstock, 2012a).

In order to identify organisms in the human body that had not been described before, Weinstock chose 11 subjects from the Human Microbiome Project and analyzed their stool samples for sequences that had < 97 percent identity to any known 16S rRNA sequence.[32] Stringent criteria were applied to subject selection[33] and analysis of sequence data to distinguish truly novel sequences from artifacts (e.g., sequencing errors, chimeric sequences). After validation of the can-

[31] The 16S rRNA gene is used for phylogenetic studies because it is highly conserved between different species of bacteria and archaea. In addition to these, mitochondrial and chloroplastic rRNA are also amplified.

[32] The Human Microbiome Project has sampled 18 body sites in 300 subjects.

[33] In response to a question from a Forum member, Weinstock elaborated on the criteria for what a healthy individual was and inclusion/exclusion criteria for this study. Exclusion criteria included everything from not having had an antibiotic treatment for a period of time prior to sample collection to not having rashes and having most of your teeth. Three days before sample collection, the subjects were also given packs of soap and toothpaste so that the microbiome being sampled was the microbiome of humans who use these products and who [are otherwise healthy].

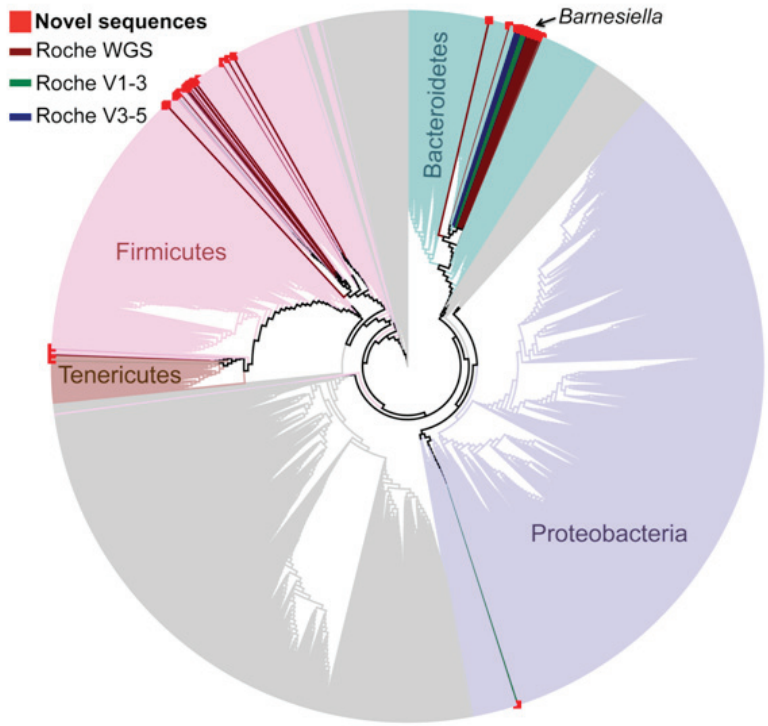


FIGURE WO-24 Novel taxa are found in three of the major phyla associated with human stool samples.
SOURCE: Wylie et al. (2012a).

didate taxa by sequencing with multiple platforms and approaches, 26 sequences remained as candidate novel OTUs (i.e., genetically distinct groups of microbes). As illustrated in Figure WO-24, taxonomic analysis suggested that they are most likely new genera or species that are most similar to uncultured bacteria in the phyla Firmicutes, Bacteroidetes, and Proteobacteria. Further analysis of stool sample data from100 additional Human Microbiome Project participants demonstrated that most of the novel taxa were found in multiple individuals sampled at two different geographic locations (Wylie et al., 2012a).

Discovery of new strains of known taxa Novel taxa remain to be discovered, Weinstock observed. In this survey, novel taxa were low in abundance within individuals, but were found in multiple people. The novel taxa identified are related, but not identical to, previously identified organisms, and are most similar to uncultured organisms. Looking for variation within species has led to the discovery of new strains of known species of bacteria and the association of these new strains with disease phenotypes.

Weinstock described a study, in collaboration with Huiying Li at the University of California, Los Angeles, that looked for a correlation between the composition of the human microbiome and acne (Fitz-Gibbon et al., 2013). Samples taken from the epithelial skin pores of the noses of 101 participants—half with and half without acne—were sequenced and cultured. 16S rRNA gene sequencing revealed similar community structure on a species level, with samples from both groups dominated by *Propionibacterium acnes*. This observation is illustrated in Figure WO-25.

Full length sequencing of more than 31,000 16S rRNA gene clones allowed investigators to take a more detailed examination of community structure, which revealed variation at the subspecies level. Weinstock presented data for the 10 most abundant ribotypes. Compared against the most abundant strain, the sequence of the 16S rRNA gene of the other 9 strains differed by one or two nucleotides. Six of those nine were clearly more abundant in individuals with acne than those without (84 to 100 percent of clones with these ribotypes came from participants with acne). These 10 ribotypes clustered into five microbiome types, two that were primarily found in acne patients and three that were found in both groups. Whole genome sequencing of 71 cultured strains showed that two ribotypes highly associated with acne are distinct because they have two chromosomal regions that lack the other ribotypes (Figure WO-26). In addition, these two ribotypes have plasmid DNA.

Genomewide SNP clustering of the 71 *P. acnes* strains demonstrated that strains that tend to be associated with acne cluster together, as do the strains that tend to be associated with healthy people. Weinstock remarked that these results were striking; the identification of a one- or two-nucleotide difference in 16S rRNA has led to the discovery of what appear to be potentially pathogenic strains of *P. acnes* and regions of the *P. acnes* genome that might be relevant to acne pathogenesis. Experiments are now being designed to test this intriguing hypothesis. Weinstock further observed that subspecies-level variation may be a widespread driver for other diseases, such as bacterial vaginosis, in which clinical presentation of disease does not always show a relationship with community variation at the species level.

Health vs. disease: Variation between communities A primary interest of the Human Microbiome Project is defining changes in a microbiome that are associated with disease. However, the intrinsic variation raises several important questions related to study design:

- How does one approach hypothesis testing, and power and sample size determination?
- What kind of a distribution best fits metagenomic data?
- How do you derive a number or a metric that conveys how similar or different two people are?

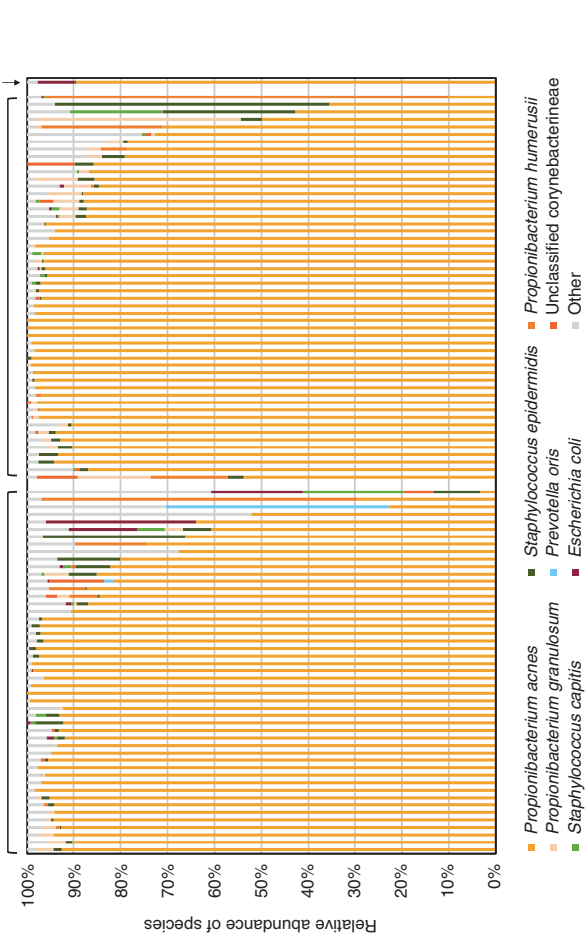


FIGURE WO-25 *P. acnes* is dominant in pilosebaceous units in both acne patients and individuals with normal skin. By 16S rDNA sequencing, *P. acnes* sequences accounted for 87 percent of all the clones. Species with a relative abundance > 0.35 percent are listed in order of relative abundance. As shown on the far-right column, species distribution from a metagenomic shotgun sequencing of pooled samples from normal individuals confirmed the high abundance of *P. acnes* in pilosebaceous units.
SOURCE: Fitz-Gibbon et al. (2013). Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL OF INVESTIGATIVE DERMATOLOGY] (Fitz-Gibbon et al., 2013), copyright (2013).

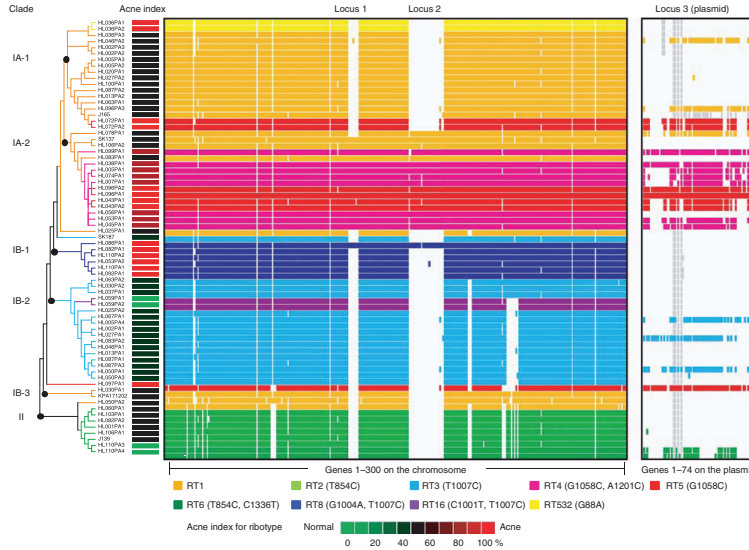


FIGURE WO-26 Genomic differences at the subspecies level (*P. acnes*). Genome comparison of 71 *P. acnes* strains shows that the genomes of RT4 and RT5 are distinct from others. Two chromosomal regions, loci 1 and 2, are unique to clade IA-2 and one other genome HL086PA1. Clade IA-2 consists of mainly RT4 and RT5 that are highly enriched in acne. The presence of a plasmid (locus 3) is also characteristic of RT4 and RT5. Each row represents a *P. acnes* genome colored according to the ribotypes. Rows are ordered by the phylogeny calculated based on the SNPs in *P. acnes* core genome. Only the topology is shown. The clades were named based on their *recA* types (IA, IB, and II). Columns represent 16 predicted open reading frames (ORFs) in the genomes and are ordered by ORF positions along the finished genome HL096PA1, which encodes a 55 Kb plasmid. Only the first 300 ORFs on the chromosome (on the left) and all the ORFs on the plasmid (on the right) are shown. The colored plasmid regions represent genes on contigs that match exclusively to the HL096PA1 plasmid region. The genes that fall on contigs that clearly extend beyond the plasmid region are likely to be chromosomally located and are colored in grey. Acne index for the ribotypes was calculated based on the percentage of clones of each ribotype found in acne.
SOURCE: Fitz-Gibbon et al. (2013). Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL OF INVESTIGATIVE DERMATOLOGY] (Fitz-Gibbon et al., 2013), copyright (2013).

Weinstock reviewed the Dirichlet-multinomial distribution for modeling and comparing species abundance distributions in human microbiota samples. This approach allows for quantification of community similarity by overdispersion (Figure WO-27), a metric that can be used to factor variation in community structure between healthy subjects (i.e., normal variation) and variation between states of health and disease into study design (La Rosa et al., 2012).

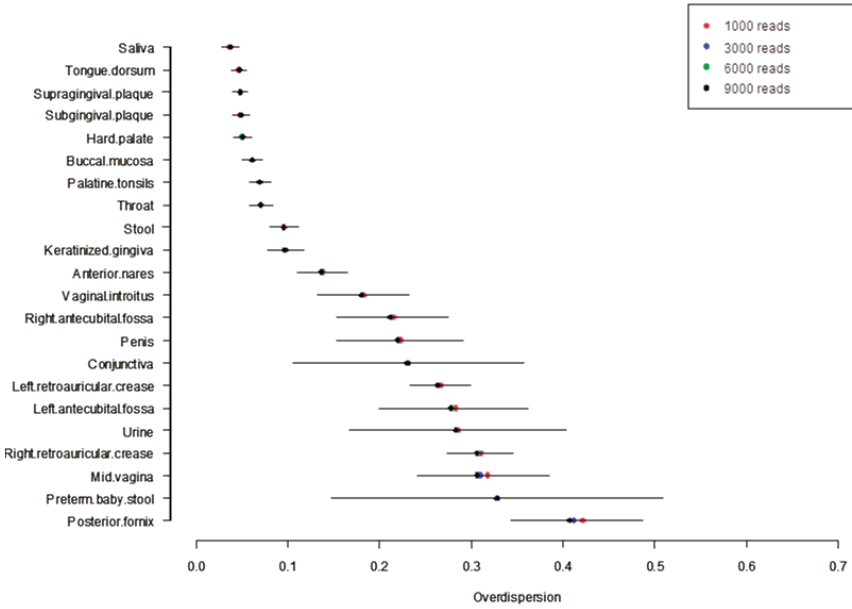


FIGURE WO-27 Overdispersion of 22 sample types from Dirichlet-multinomial distribution showing similarity between subjects (mean +/– SD). The closer the mean is to zero, the less variation (or the greater similarity) between the microbiomes of two subjects. SOURCE: Zhou et al. (2013).

Understanding the intrinsic variation in any two samples is essential for designing meaningful comparative studies. Weinstock noted that investigators are just beginning to define different classes of communities within a body site; such as high and low bacterial communities of lactobacillus in the vagina.[34] A somewhat controversial area in human microbiome research, Weinstock said, is whether individuals or groups of people can be uniquely identified by their gut microbiomes. Three distinct gut community "enterotypes" have been proposed to group people and communities based on the predominance of one of three classes of organisms, Ruminococcus, Bacteroides, or Prevotella. Weinstock cautioned that for such community classifications, "time will tell whether these really reflect some kind of difference in people." He noted however, that "we certainly know how to do an analysis that teases [apart] differences and allows you to test certain [hypotheses]."

Weinstock offered two examples of studies of microbial community variation associated with states of "disease" and "health." The first, a collaboration

[34] According to Weinstock, Jacques Ravel, Larry Forney, and others who have investigated this for a long time think there may be as many as five different community classes of the vaginal [microbiome].

with Homer Twigg of the Indiana University Medical Center in Indianapolis, was a longitudinal study of the effect of HIV on the lung microbiome, which suggests that variation between these communities increases with HIV infection and is reduced when patients are treated with highly active antiretroviral therapy (HAART). These observations are illustrated in Figure WO-28. After HIV is removed by antiretroviral therapy, it takes a long time for the disturbed microbiome to return to a healthy equilibrium state and for the host to re-establish immunological homeostasis. Interestingly, *Tropheryma whippelei* (the causative agent of Whipple's disease, a rare gastrointestinal malabsorptive condition) was the dominant organism in lung samples during HIV infection. The significance of this finding is yet to be determined, but according to Weinstock, it underscores how little we know about the microbiome, and the potential for metagenomic analyses to generate new insights into human health and disease.

The second study discussed by Weinstock surveyed babies 36 months of age or younger that had been admitted to the hospital emergency room with a high fever (Wylie et al., 2012b). Standard diagnostic tests could not identify a causative agent in more than 30 percent of these patients. An analysis of nasal and plasma samples obtained from both febrile and afebrile children using shotgun

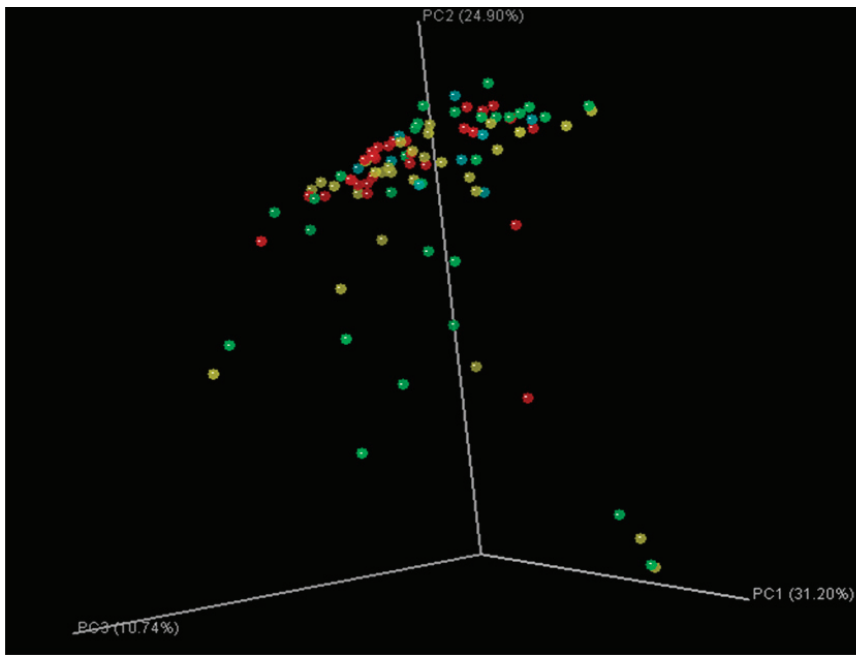


FIGURE WO-28 Change in HIV communities with HAART.
SOURCE: in collaboration with Homer Twigg of the Indiana University Medical School in Indianapolis (unpublished).

sequencing[35] demonstrated that, while numerous DNA and RNA viruses were detected in all patients, virus sequences were significantly more abundant in febrile children. Weinstock highlighted the intriguing result that healthy (or at least asymptomatic) children harbor many viruses, that these viruses have also been detected in nasal, vaginal, gastrointestinal, skin, and oral samples from healthy adults, and these infections are stable (i.e., not transient).

Tracking organisms: Genetic variation in populations Weinstock discussed insights emerging from research on genetic variation in metagenomic populations. Pooling of stool sample sequence data from the Human Microbiome Project and MetaHIT[36] (together totaling 252 samples obtained from 207 people) created a massive amount of sequence information on the gut microbiome. When aligned with a reference set of approximately 1,500 finished bacterial genomes, this data set produced high-quality sequence information for 101 bacterial species, which could be further studied for genetic variation (Schloissnig et al., 2013).

This project has now created a catalog of how these 101 different species may fluctuate in different individuals, Weinstock explained, identifying 10.3 million SNPs, 1.1 million short insertions and deletions, and more than 4,000 structural variants in these species. Phylogenetic distribution and abundance of the 101 samples in the combined Human Microbiome Project-MetaHIT cohort were consistent with the usual community structure observed for the gut. By assessing the presence or absence of specific SNPs in samples collected from the same patient at different times, researchers have been able to show that strains present in an individual were fairly stable over time—leading to the conclusion that a person has his or her own, unique, microbial communities that are distinct from the strains from another individual.

In another study, SNP analysis of samples from three different oral sites—buccal mucosa, supragingival plaque, and the tongue dorsum, revealed site-specific strains of *Streptococcus oralis*. An analysis of the community structure of the mouth may simply identify *S. oralis* as a member of the mouth microbiota. It appears, however, that different strains of *S. oralis* reside in different sites—distinct ecological niches—in the mouth. Weinstock reiterated the importance of subspecies variation with respect to the pathogenicity of *P. acnes*, and wondered what else might be different between these three strains of *S. oralis* and what role those differences might play in oral health and disease.

[35] "Shotgun sequencing" randomly shears genomic DNA into small pieces that are cloned into plasmids and sequenced on both strands, thus eliminating the BAC (bacterial artificial chromosome) step from the HGP's (Human Genome Project's) approach. Once the sequences are obtained, they are aligned and assembled into finished sequence.

[36] MetaHIT, or Metagenomics of the Human Intestinal Tract, is a academic-industry consortium funded by the European Commission to study the association of bacterial genes in the human intestinal microbiota with human health and disease. See www.metahit.eu.

Phylogenetic and Functional Diversity in Deep-Sea Microbial Communities

Occupying 80 percent of our biosphere, the deep ocean[37] remains a largely unexplored *terra incognita*. Located at an average depth of 4,000 meters (~2.5 miles) below the sea surface, the deep-sea floor includes the largest mountain range on earth—the 50,000 mile long mid-ocean ridge system—and is the site of ~90 percent of the volcanic activity on the planet. According to speaker Peter Girguis of Harvard University, the ocean is also dominated by microorganisms. About 50 percent of the oxygen in Earth's atmosphere—as well as about 50 percent of Earth's primary production—is produced by marine photosynthetic microbes. Marine microbes also play a major role in other biogeochemical cycles critical to our biosphere. The grand challenge for investigators interested in environmental microbes, Girguis said, is linking microbial identity to functional potential and activity, including elucidating the role a given microbe plays in global biogeochemical cycles.

The deep-sea floor influences Earth's heat and energy budgets, primarily through hydrothermal circulation. The ocean circulates through the crust and underlying sediments on the sea floor, known collectively as the ocean crustal aquifer.[38] With an estimated 10^{29} microbes in ocean sediments alone, the crustal aquifer harbors an extensive, and very poorly understood, community of microorganisms (Whitman et al., 1998). These microbes are thought to play a major role in marine biogeochemical cycles that influence the entire ocean, such as metal cycling.

Much of the fluid circulating through the crustal aquifer is discharged at deep-sea hydrothermal vents. These vents are prominent features along mid-ocean ridges, and emit hot, chemically reduced fluids. The buoyant fluid emerges through the crust, and metals precipitate to form chimneys or "smokers." Temperatures range from 4 degrees Celsius in ambient water to more than 350 degrees Celsius inside smokers. Pressure at vents is around 250 atmospheres, or 4,000 psi. Vent fluid contains millimolar concentrations of hydrogen sulfide and high concentrations of heavy metals such as arsenic, and it has an acidic pH (similar to vinegar) (Figure WO-29). According to Girguis, hydrothermal vent water is "pretty hostile stuff" that can melt plastics, dissolve tin and pewter, char wood, and melt glass. "But, what is amazing" he continued "is that the communities around these hydrothermal vents are immense in terms of their biomass. These are highly productive communities." Indeed, chemoautotrophic[39] microorganisms flourish around these vents and support abundant communities of flora and fauna, reaching densities comparable to those of rainforests.

[37] Deeper than 1,000 meters.

[38] The entire volume of this aquifer is expelled every 2,000 to 5,000 years, and the entire ocean circulates through this aquifer every 70,000 to 200,000 years (Fisher and Von Herzen, 2005).

[39] Organisms that use energy derived from the oxidation of inorganic compounds to fix carbon.

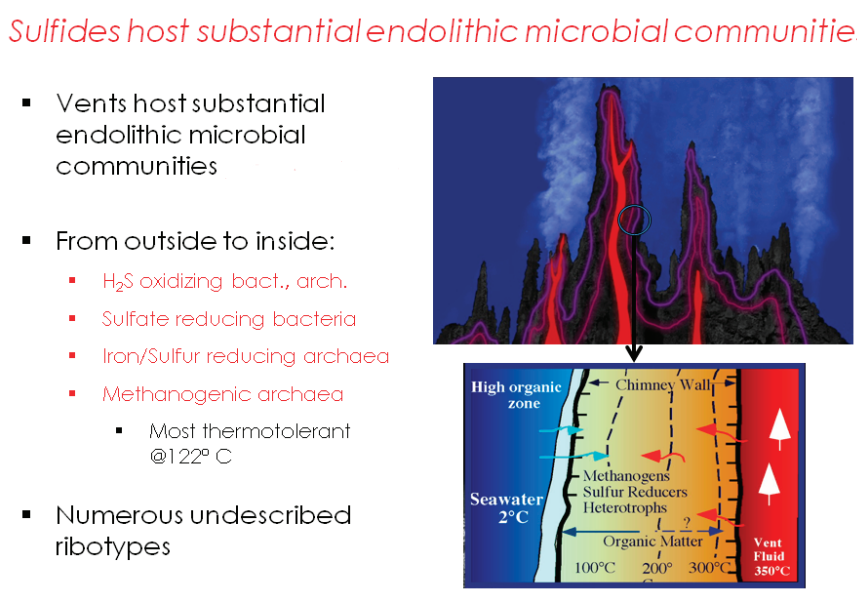


FIGURE WO-29 Hydrothermal vents host substantial endolithic microbial communities. SOURCE: Girguis (2012); Schrenk et al. (2003).

The basis of the food chain at hydrothermal vents is chemoautotropic microbes, but according to Girguis, "we don't really know who is doing what, and how they are doing it." Studying microbes in the laboratory is hampered by the fact that more that 99 percent of known microorganisms have eluded cultivation. For the study of deep-sea organisms specifically, it is difficult to replicate the conditions of the deep-ocean trenches on Earth's surface in order to conduct *in situ* chemical measurement and experiments. Girguis has designed novel instruments to collect contextual metadata in these remote sites (such as underwater mass spectrometers to look at dissolved gases). Measuring activity at the microbial scale (such as, microbial influence on pH, temperature, CO_2) is a slow and tedious process, and the nature, quality, and abundance of contextual data available to link specific microbes to biogeochemical processes varies widely. As such, there is an "impedance mismatch" between the volume and nature of data available from high-throughput sequencing technologies and the volume and nature of the contextual data obtained from geochemical sensors. Omics-derived data that inform us about who is there and what they are potentially capable of doing, according to Girguis, will be helpful in guiding the design and conduct of geochemical studies.

One paradox in vent microbial ecology is that the rate of sulfide oxidation in chimneys is higher than expected given the limited concentration of oxidants in vent effluents. As it turns out, metal sulfides in chimney walls (e.g., chalcopyrite, pyrite, wurtzite) are highly conductive, and vent-dwelling microbes are capable

of extracellular electron transfer, a process of shuttling electrons from anaerobic respiration inside the cell to solid phase insoluble remote oxidants outside the cell. In essence, the anaerobes are benefiting from electrical continuity to oxygen as an oxidant, without being in chemical continuity with that oxygen. Further studies suggest that, in fact, vent microbes are using extracellular electron transfer to both donate and accept electrons. Microbes are essentially sharing electron equivalents through the conductive vent matrix.

Girguis described studies of microbial extracellular electron transfer using "microbial fuel cells" that mimic mineral oxidants. In a lab-based artificial vent, he measured the amount of current across electrodes in vent fluid and in oxygenated seawater, which in the absence of bacteria was minimal. Inoculation of the vent fluid with microbes collected from hydrothermal vents resulted in a substantial, measurable increase in current, which Girguis noted, is continuous and sustained (over 6 months thus far). Pyrite in the system acts as a conductor and supports the establishment of a lush and diverse microbial community that is representative of what is seen in an active hydrothermal vent community freshly recovered from the sea floor (Schrenk et al., 2003; Takai et al., 2003) (Figure WO-30). In the absence of electrical continuity, the minimal community that forms on pyrite bears a greater resemblance to what is found on extinct sulfides (Sylvan et al., 2012). Girguis noted, "our literature [is] unfortunately populated with a lot of data that may not actually represent the conditions that the microbes

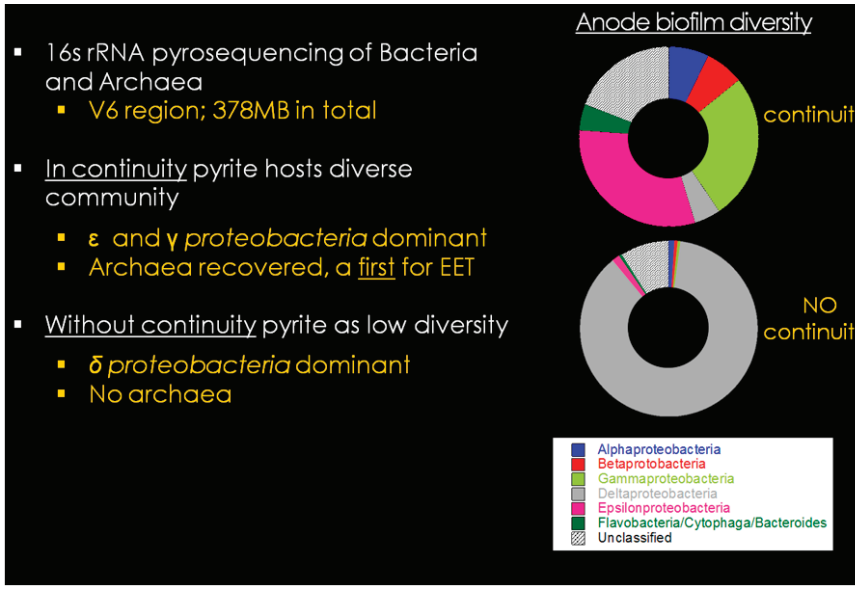


FIGURE WO-30 Electrical continuity yields diverse representative community. SOURCE: Girguis (2012).

are seeing *in situ*" because many researchers have studied sulfides from the sea floor without electrical continuity to oxygen.

Metagenomics suggests that microbial communities formed in the presence of electrical continuity have the potential for sulfide oxidation, hydrogen oxidation, and carbon fixation. Experiments conducted in the presence of electrical continuity result in substantially more carbon fixation—which rectifies the discrepancies seen in previous biogeochemical measurements. Using shotgun metagenomics to interrogate communities recovered from the interior and exterior of sulfide deposits has led to the identification and isolation of iron oxidizers that are accepting electrons and using it to fix carbons. Essentially, Girguis noted, these microbes are sharing electron equivalence through a conductive matrix—a finding that has reshaped the way we think about the communities growing at these hydrothermal vents.

From a biogeochemical perspective, extracellular electron transfer enables microbes to access remote oxidants, stimulates primary productivity, and influences local alkalinity. Extracellular electron transfer also reshapes the notion of anaerobic metabolism, Girguis said. Although these organisms are anaerobes (and do not grow in the presence of oxygen) it is clear they are coupled to the availability of remote oxidants via extracellular electron transfer. Girguis and collaborators are currently focused on characterizing the phylogenetic and functional diversity of microbes living in hydrothermal vents around the world.

Community Ecology and Adaptation to the Environment—Use of Genomic Approaches to Study Evolution in Real Time

Comparative analyses of microbial genomes have demonstrated that the microbial genome is a dynamic entity shaped by multiple forces including gene loss/genome reduction, genome rearrangement, expansion of functional capabilities through gene duplication, and acquisition of functional capabilities through lateral or horizontal gene transfer (HGT) (Fraser-Liggett, 2005). Several natural processes carry genetic information from one species to another. DNA can be transported by viruses (transduction), via bacterial mating (conjugation), and through the direct uptake of DNA from the environment (transformation). Genes that must function together are transferred together as genomic islands (e.g., pathogenicity islands) (Hacker and Kaper, 2000).

While microorganisms might superficially appear the same according to 16S gene sequence profiles, specific SNPs in the genome and particular genes that are acquired by bacteria can contribute to fine-scale diversity and adaptation to environments. Horizontal gene transfer studies have identified DNA that has been moved across different organisms in the environment quite recently as well as more than a million years ago. In his prepared remarks, Eric Alm of the Broad Institute described two microbial ecology projects that study how bacteria adapt to their environment in real time.

Global network of horizontal gene transfer The first project Alm described involved the horizontal transfer of nearly identical genes from one species of bacteria into an entirely different species. Alm was inspired by a study describing the transfer of a gene for an enzyme that breaks down algal polysaccharides. The gene appears to have transferred from the algae itself, to marine bacteria, to the gut microbiota of Japanese individuals (potentially through the consumption of an edible seaweed) (Hehemann et al., 2010). In each case, Alm noted, the acquisition of this gene provided a selective advantage to the recipient host.

The Alm laboratory searched GenBank and found nearly 11,000 different proteins for which the genes were close to 100 percent identical in two totally different bacterial species (suggesting recent transfer of the DNA). Even more striking, Alm said, was that despite screening equal amounts of environmental and human-associated bacteria, the vast majority of pairs of bacteria that shared a nearly identical gene were isolated from sites on the human body. The highest rates of transfer were between bacteria isolated from the same body part (albeit on two different individuals), that is, ecologically similar but physically separated sites (Figure WO-31). Horizontally transferred genes for antibiotic resistance are a particular concern, and there is evidence that transfer does occur between humans, the foods they consume, and the livestock they raise (Smillie et al., 2011).

Sympatric speciation Alm explained that there are two conflicting, empirically based observations for how sympatric speciation occurs.[40] The "ecotype" model hypothesizes that within a particular ecological niche, one bacteria will acquire a selective advantage and out compete everything else within, but not outside, that niche. Over time this type of clonal expansion leads to the tree structure that is commonly observed when looking at bacterial diversity (Gevers et al., 2005; Lipsitch et al., 2009). The "gene-centric" model states that there are environment-specific genes that are independent of species (Coleman and Chisholm, 2010).

It had been assumed that sympatric speciation was only theoretically possible when the number of adaptive loci was relatively small (Kondrashov and Mina, 1986). Alm pointed to ecological differentiation in animals to illustrate this point. In parts of Western Africa there are two variants of the malaria vector, *Anopheles gambiae*, that coexist in the same place and at the same time, differing by only a very few select loci in the genome (White et al., 2010).

To study ecological differentiation in microbial communities, Alm sequenced Vibrionaceae strains from coastal bacterioplankton together with Martin Polz at the Massachusetts Institute of Technology (Hunt et al., 2008) (Figure WO-32). They identified one clade of *Vibrio* that differentiated phylogenetically into 15 clusters that inhabited different parts of the water column. Upon sequencing chromosome 1[41] from 20 closely related strains in one particular cluster, it was

[40] Speciation in the absence of physical barriers to genetic exchange between incipient species.
[41] *Vibrio* have two circular chromosomes.

found that these strains partitioned further into two distinct, ecologically oriented groups. The sequence of chromosome 2, however, suggested significant recombination had occurred, which did not support the ecotype theory.

Alm concluded that most mutations across the genome supported an ecological tree, although there is a history of rampant recombination within and between

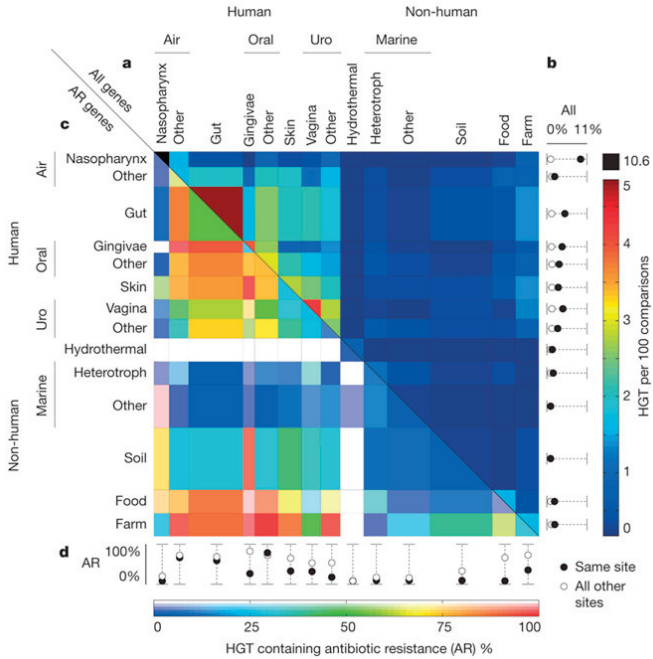


FIGURE WO-31 HGT is ecologically structured by functional class and at multiple spatial scales. The frequency of transfer between different environments is shown for all functional groups (Panels a, b) and for antibiotic resistance (AR) genes only (Panels c, d). Box widths indicate the number of genomes from each environment. Panel a, When all genes are considered (upper half), human isolates form a block of enrichment (upper left). Panel b, For every environment examined we observe more transfer within the same environment (black dots) than between environments (white dots). Panel c, The fraction of gene transfers that includes at least one AR gene for each environment. Statistical uncertainty in the proportion of AR transfer is indicated by decreased color saturation (see Methods). Panel d, AR genes comprise a significantly higher fraction of observed HGT between different environments (white dots) relative to that within the same environment (black dots) in contrast to Panel b**.**

NOTE: Uro, urogenital.

SOURCE: Smillie et al. (2011). Reprinted by permission from Macmillan Publishers Ltd: NATURE . Smillie, C. S. et al. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241-244, copyright 2011.

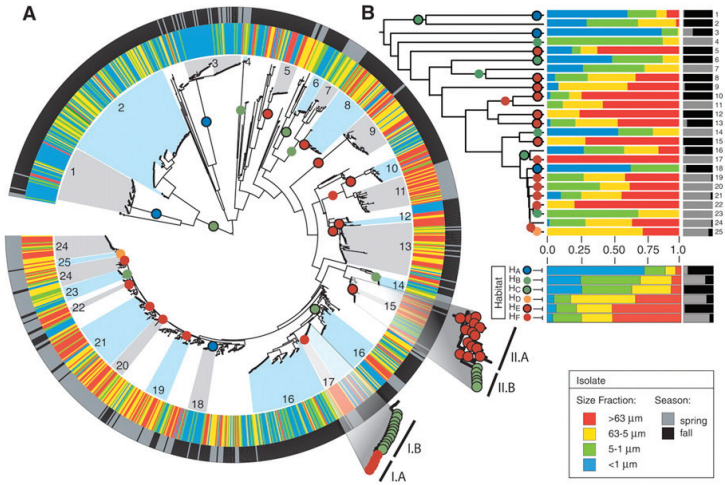


FIGURE WO-32 Season and size fraction distributions and habitat predictions mapped onto Vibrionaceae isolate phylogeny inferred by maximum likelihood analysis of partial *hsp60* gene sequences. Projected habitats are identified by colored circles at the parent nodes. Panel A, Phylogenetic tree of all strains, with outer and inner rings indicating seasons and size fractions of strain origin, respectively. Ecological populations predicted by the model are indicated by alternating blue and gray shading of clusters if they pass an empirical confidence threshold of 99.99 percent (see supporting online material for details). Panel B, Ultrametric tree summarizing habitat-associated populations identified by the model and the distribution of each population among seasons and size fractions. The habitat legend matches the colored circles in (A) and (B) with the habitat distribution over seasons and size fractions inferred by the model. Distributions are normalized by the total number of counts in each environmental category to reduce the effects of uneven sampling. The insets at the lower right of (A) show two nested clusters (I.A and I.B and II.A and II.B) for which recent ecological differentiation is inferred, including habitat predictions at each node. The closest named species to numbered groups are as follows: G1, *V. calviensis*; G2, *Enterovibrio norvegicus*; G3, *V. ordalii*; G4, *V. rumoiensis*; G5, *V. alginolyticus*; G6, *V. aestuarianus*; G7, *V. fischeri/logei*; G8, *V. fischeri*; G9, *V. superstes*; G10, *V. penaeicida*; G11 to G25, *V. splendidus*.
SOURCE: Hunt et al. (2008). From Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320(5879):1081-1085. Reprinted with permission from AAAS.

ecological groups, and just a few highly divergent, habitat-specific loci drive the genome-wide ecological signal (as originally predicted by Kondrashov in 1986). Alm also found evidence for the recent emergence of barriers to homologous recombination between habitats (i.e., bacteria tend to recombine with others from the same environment).

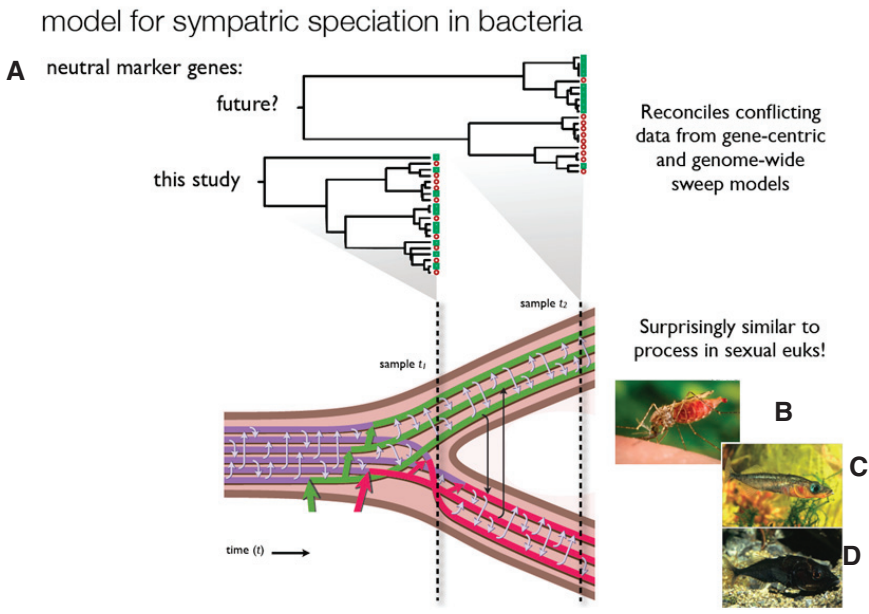


FIGURE WO-33 Ecological differentiation in recombining microbial populations. (A) Example genealogy of neutral marker genes sampled from the population(s) at different times. (B) Underlying model of ecological differentiation. Thin gray or black arrows represent recombination within or between ecologically associated populations. Thick colored arrows represent acquisition of adaptive alleles for red or green habitats.
Alm (2012). A: From Shapiro et al. (2012). Population genomics of early events in the ecological differentiationsof bacteria. *Science* 336(6077):48-51. Reprinted with permission from AAAS; B: CDC/James Gathany; C&D reproduced with permission from the International Institute for Applied Systems Analysis (IIASA).

Based on this work, Alm and colleagues proposed a new model for sympatric speciation in bacteria (see Figure WO-33) that reconciles the conflicting ecotype and gene-centric models described earlier (Shapiro et al., 2012). The new model begins with an ancestral population in which there was significant homologous recombination. At some point a new niche arose and a small number of alleles conferred fitness to one niche and not the other. Those alleles then selectively swept through each of those smaller populations.

If one were to sample a random housekeeping gene that is not one of these very few loci that contribute to fitness, there would be no clear ecological patterns in the phylogenetic tree (sample t_1 in the figure). Over long periods of time, if these gene pools remain separate, then there will be habitat-specific clusters that emerge (sample t_2 in the figure).

Cataloging and Characterizing the Earth's Microbiome

The Earth Microbiome Project seeks to provide a systematic characterization of microbial life on Earth, cataloging and comparing the microbial diversity across all Earth environments (e.g., air, soil, water, humans). Because the majority of Earth's microbes are at present not readily cultivable, the Earth Microbiome Project takes a metagenomic approach to tap into what speaker Jack Gilbert of the Argonne National Laboratory called the "great dark biosphere," extracting and sequencing total DNA from environmental samples (Figure WO-34) (Dr. Gilbert's contribution to the workshop summary report may be found in Appendix A, pages 166-188).

The Earth Microbiome Project involves more than 120 collaborators across more than 50 institutions in 25 countries. Gilbert noted that unlike the Human Microbiome Project, the Earth Microbiome Project is not very well funded at the moment. Beyond funding, one of the biggest technical challenges is collecting samples from around the globe. Obtaining samples from the Peruvian Amazon, the bottom of the Marianas Trench, the air above Colorado, and from humans across a vast swath of Africa requires a network of collaborators who have access to those samples. More than 50 researchers have pledged more than 60,000 samples to the project. Samples are selected on the basis of their position in environmental gradients—those chemical, physical, and biological gradients

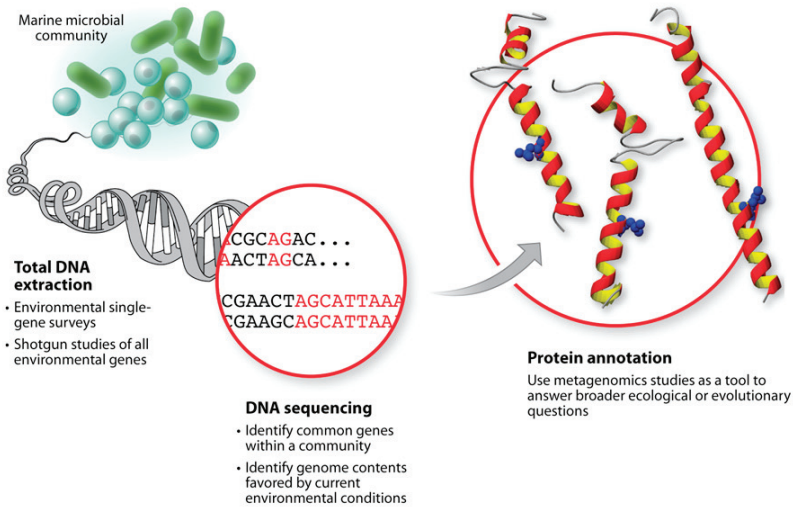


FIGURE WO-34 Metagenomic approach to studying the Earth's microbiome. Total DNA is extracted from an environmental sample, in this case, a marine microbial community. Taxonomy and protein function are inferred from DNA sequences.
SOURCE: Gilbert (2012) from Gilbert and Dupont (2011). Republished with permission of *Annual Reviews*, from Microbial metagenomics: Beyond the genome, Gilbert, J.A and Dupont C.L., 3, 2011; permission conveyed through Copyright Clearance Center, Inc.

that define the niche space within which bacteria, archaea, microbial eukaryotes, and viruses exist. Rich contextualization of that environment at the time of sample collection is important, including, for example, nutrient concentrations, pH, temperature, host, time of collection, latitude and longitude, and any other environmental, contextual data that can provide clues as to why that community exists where it exists, and why it exhibits a particular functional phenotype.

Gilbert provided an overview of some of the analyses that have been done to date on approximately 15,000 samples. By cataloging microbes across specific environmental gradients (in extremes of temperature, extremes of nutrient availability, extremes of pressure and light availability, etc.) it is possible to systematically catalog that community in a much more robust fashion. Sequences from the Earth Microbiome Project map to 82 percent of the Greengenes tree of life (compared to the Human Microbiome Project, which has thus far cataloged 17 percent of the tree). Gilbert emphasized that this is just what can be matched against what is already known from surveys of 16S rRNA genes. This known diversity accounts for only 5 or 6 percent of the total diversity cataloged in the Earth Microbiome Project. In the first 10,000 samples analyzed, ~600,000 bacterial OTUs, or putative bacterial species, were identified.

Gilbert observed that the environment with the most diverse microbial community composition sampled thus far is oil spill sediments in the Gulf of Mexico, followed by stream water, and soil. Interestingly, the insect microbiome (bacteria in ant and termite intestines) was the least diverse environment, but contained the greatest number of new species identified. If we want to catalog more new bacterial species, Gilbert said, we should be looking to those particular environments with the most novelty, as opposed to environments that are already the subject of systematic surveys such as ice cores, marine water, and human environments.

Genomic environmental monitoring The Genomic Observatories network was described by Gilbert as "a network of sites working to generate genomic observations that are well-contextualized and compliant with global data standards." Observatories will participate in long-term monitoring of environments across the globe, applying genomic technologies to monitor microbial, vertebrate, invertebrate, and eukaryotic diversity and interactions.

Gilbert reviewed one project exploring the seasonal microbial community structure in the English Channel. Researchers conducted microbial 16S rRNA gene surveys, followed by shotgun metagenomic and shotgun metatranscriptomic surveys to understand function, every month for 6 years—a total of 72 time points. Gilbert then developed a model to predict community structure from environmental parameter metadata (e.g., temperature, nitrate concentration, pH), predict functional metabolism from that community structure, and to then predict how that functional metabolism influences environmental parameters. The resulting positive feedback loop can be used for microbial forecasting (Larsen et al., 2012). The model takes into account biological interactions; if one particular

organism increases its relative abundance, then it will also have an impact upon the relative abundance of another organism. Gilbert described how he was able to extrapolate from the information gleaned from the 6-year survey of one location in the English Channel to predict the relative abundance of more than 100,000 different bacterial species at any given time and location in this environment. This extrapolated predictive capability may then be used to refine the sampling strategy.

The next step in these studies is what Gilbert referred to as "predictive relative metabolic turnover," in which functional dynamics of the predicted community are inferred from the metagenome and metatranscriptome information—for example, how that community responds to environmental change (Larsen et al., 2011). From the predicted functional ability of a community, we can predict its metabolic capacity. The ultimate goal is to study how changes in metabolism feed back and influence the environment within which the organisms are found.

Looking at CO_2 turnover within the English Channel environment over a 24-hour period, for example, Gilbert and colleagues saw an impressive 94 percent correlation between their ability to predict CO_2 generation or consumption as a semi-quantitative measure, and the quantitated, analyzed, and observed flux of CO_2 from the channel surface as reflected in the UNESCO Surface Ocean CO_2 atlas. These observations are depicted in Figure WO-35.

Other microbiome projects Gilbert sampled different environments on a couple (e.g., palm, heel, inside of nose) and their home including the kitchen counter, floor, and light switch; bedroom floor; front door knob; and bathroom door knob. These data are presented in Figure WO-36.

Gilbert went on to discuss several environment-specific microbiome projects under way, including the Home Microbiome Study that seeks to understand how humans interact with the microbes on their skin and the microbes in their home. We live in our spaces, but those spaces are also living, Gilbert said. When a person moves into a new house, does the person adopt the microbiome that already exists in the house, or does the house adopt the microbiome of its new inhabitant? Or perhaps there is a new state where the microbiomes of both house and inhabitant are modified?

He found that the microbiomes of the feet of both the man and the woman were very heavily dominated by Staphylococcaceae. When they moved in, the solid oak bedroom floor was dominated with Mycoplasmataceae; however, after 6 days the floor became repopulated with Staphylococcaceae. While bedroom floors reflect their inhabitants, kitchen countertops in general look very similar across houses because they are constantly in a state of dynamic flux due to frequent cleaning. Gilbert added that people living together tend to have similar microbiota.

The final project that Gilbert discussed was the Hospital Microbiome Project that is just getting under way. This study is taking advantage of a new hospital

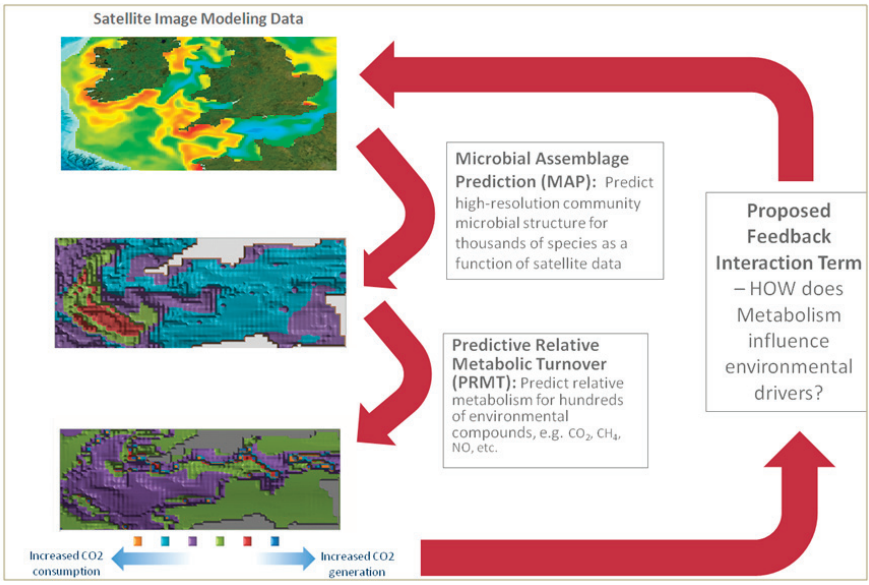


FIGURE WO-35 Extrapolating microbial community structure. Using satellite image modeling data (hyperspectral data from 5,000 49-km^2 grid cells in the English Channel), it is possible to predict the microbial community assemblage (relative abundance of each OTU), and then to predict the turnover of specific metabolites such as CO_2. Determining what impact the changes in microbial-mediated CO_2 have on the hyperspectral properties of the water will provide a feedback term.
SOURCE: Gilbert (2012).

pavilion being built at the University of Chicago to try to understand what happens to the microbiome of the hospital when the humans move in. Samples will be collected before the hospital opens in early 2013, and then every day for a year across patient rooms, nursing stations, hospital corridors, hallways, and people, including patients and staff. The goal is to better understand the dynamics of microbial populations over space and time and how these populations may lead to nosocomial infections.

PRACTICAL APPLICATIONS OF GENOMIC TECHNOLOGIES

Applications of microbial genomics, such as those discussed throughout the workshop, have expanded researchers' appreciation for the biology of microorganisms including their organismal, metabolic, and environmental diversity; the structure of microbial populations over space and time; the evolution of microbial species; and the acquisition of novel virulence factors and pathogenicity islands. These areas of inquiry are providing important insights into the "ground rules" of

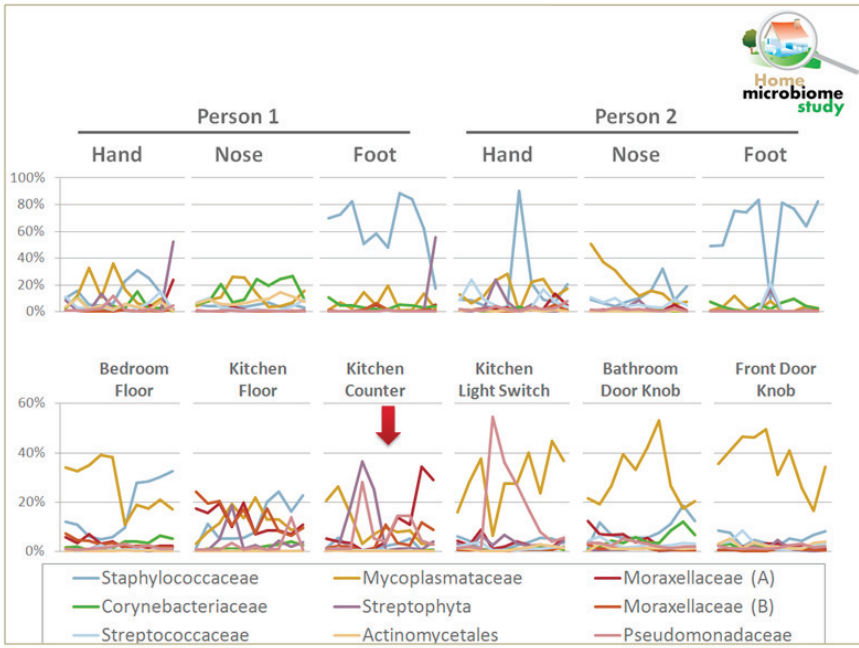


FIGURE WO-36 We change the microbial community of a house.
SOURCE: Gilbert (2012).

pathogen evolution and will inform the development of a "versatile platform for developing new responses to infectious disease" (Lederberg, 2000).

Tools for Microbial Detection, Surveillance, and Response

Although still very much a research enterprise at this stage, microbial genomics has great potential for practical application. As illustrated in Figure WO-37, microbial genomics is becoming an increasingly important tool for a wide range of applications. Relman has previously noted that, "[d]ifferences in the sequence and structure of genomes from members of a microbial population reflect the composite effects of mutation, recombination, and selection. With the increasing availability of genome sequences, these effects have become better characterized and more effectively exploited in order to understand the history and evolution of microbes and viruses and their occasionally intimate relationships with humans. The resulting insights have practical importance for epidemiologic investigations, forensics, diagnostics, and vaccine development" (Relman, 2011).

A genome sequence facilitates the development of a variety of tools and approaches for understanding, manipulating, and mitigating the overall effect of a microbe. The sequence provides insight into the population structure and

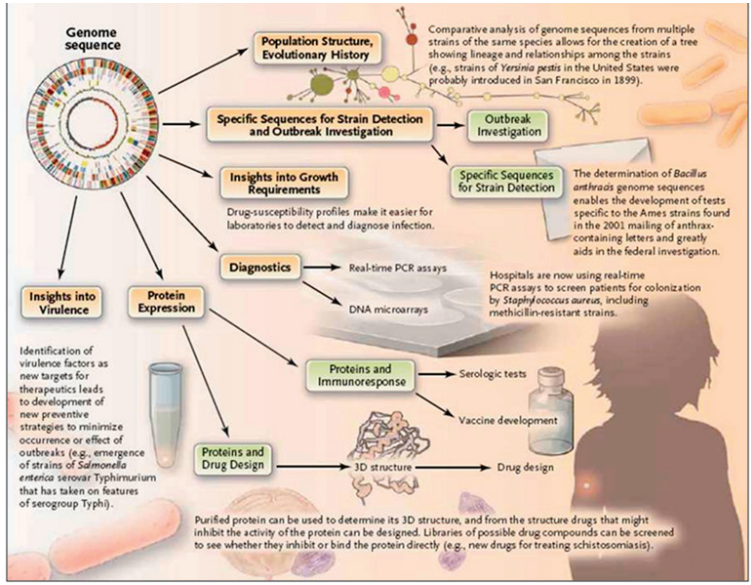


FIGURE WO-37 Microbial genomics and tool development.
SOURCE: Relman (2011). From *New England Journal of Medicine*, David A. Relman, Microbial Genomics and Infectious Diseases, 365(4):347-357. Copyright ©2011 Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

evolutionary history of a microbe for epidemiologic investigation, information that could be used to develop new diagnostic tests and cultivation methods, new targets of drug development, and antigens for vaccine development (Relman, 2011).

Microbial Genomics as a Frontline Public Health Tool

Understanding how a pathogen spreads from person to person or through populations is essential to developing an effective public health response. While we understand a great deal about the means of transmission for some pathogens—measles and influenza viruses are spread on droplets, HIV is a blood-borne disease, and West Nile virus is a vector-borne viral disease—public health epidemiologists do not yet fully understand how pathogens spread from person to person, from community to community, over space and time, and how underlying social network structures shape the movement of pathogens, said speaker Jennifer Gardy of the British Columbia Centre for Disease Control, Vancouver, Canada (Dr. Gardy's contribution to the workshop summary report may be found in Appendix A, pages 141-150).

Field epidemiology, or traditional "shoe leather" epidemiology, involves interviewing people—the "host"—and looking for commonalities in social contacts, attendance at particular locations, contact with foods, or other behaviors that might explain the appearance and spread of disease. Molecular epidemiology, by contrast, essentially interviews the pathogens, Gardy said, through serotyping and genetic fingerprinting, including the use of restriction fragment length polymorphism analysis and multilocus sequence typing.

Tuberculosis: Whole genome sequencing vs. genetic fingerprinting One of the problems with the current standard for outbreak investigations, which uses a combination of field epidemiology and genetic fingerprinting, is that different molecular epidemiology techniques have different resolutions, which influence the perception of an outbreak.

Recalling investigations of tuberculosis (TB) outbreaks in British Columbia, Gardy noted that when the provincial public health laboratory switched from using 12-loci MIRU-VNTR (a fingerprinting technique used for *Mycobacterium tuberculosis*) to 24-loci MIRU-VNTR (a higher resolution tool), some clusters of cases that appeared at first to be related were actually unrelated. Such differences influence the perceptions of how cases might be related to each other and how an outbreak is reconstructed, Gardy said. Another issue with genetic fingerprinting methods is that the order of transmission is not always easy to infer (i.e., who infected whom and when).

Even using the best genetic fingerprinting techniques, only a very small fraction of the genome is being sampled. For the average bacterium, noted Gardy, MLST examines short fragments of seven housekeeping genes, essentially 0.03 percent of the available DNA. She continued, "that means you are ignoring 99.97 percent of its genome that contains informative and interesting variation that might give you a high-resolution view of an outbreak and that might actually be able to tell you about the order of transmission." The whole genome sequence has often been referred to in molecular epidemiology as "the ultimate genotype." With next-generation genome sequencing technology, molecular epidemiologists may be able to sequence numerous isolates from an outbreak quickly and cheaply.

An emerging discipline of "genomic epidemiology" is employing whole genome sequences from outbreak isolates to track person-to-person spread of an infectious disease. Even in the space of a few days or a week, pathogens are measurably evolving and accruing mutations, and there can be enough informative variation (i.e., SNPs) to be able to distinguish isolates from each other. Whole genome sequencing, combined with information about the social or other relationships between cases, can facilitate "visualizing" the actual transmission events.

Gardy noted that this approach has been used twice so far in British Columbia for TB outbreaks (Gardy et al., 2011). In one example, 24-loci molecular fingerprinting suggested that all of the outbreak isolates were identical to each other. The social contact network for the extremely close-knit community hinted

at the source case of the outbreak; however, after the first two cases or so, there were too many possibilities for transmission pathways for any one person's infection. Even though all the isolates had the same mycobacterial interspersed repetitive unit-variable number tandem repeat (MIRU-VNTR) fingerprint, there was enough variation observed in whole genome sequencing to divide the isolates into two distinct phylogenetic clades, making that social network more amenable to interpretation (Figure WO-38).

The second TB outbreak discussed by Gardy was a location-based outbreak. The affected individuals did not necessarily have social connections to each other, but they all shared attendance at common locations—homeless shelters in the interior region of British Columbia. In addition, many of the strains isolated were resistant to low levels of isoniazid—the frontline antibiotic for the treatment of tuberculosis.

Whole genome sequencing suggested an early wave in the Vancouver area, including an individual in whom it is hypothesized that the resistance mutation arose, followed by the first wave of transmission associated with one of the shelters, then a second wave of infection at the second shelter that was likely seeded by one individual showing one very characteristic mutation. Gardy noted that the uses of microbial genomics described in her presentation are largely confined to the research environment, and that there is a long way to go before such an approach becomes a clinically validated technique used in reference laboratories. Gardy emphasized that in order to move forward with whole genome sequencing as a molecular epidemiology tool, it is important to remember that genomic data must be interpreted in the context of epidemiological and clinical data. It is not possible to reconstruct an outbreak from genome sequence alone. Gardy concluded that public health has the registries and the clinical expertise to complement the genome sequencing research, and collaboration and data sharing is key to maximizing the value of genomic data sets.

Microbial Diagnostics and Genomic Epidemiology

Speaker Mark Pallen of the University of Birmingham (UK) assessed the state of diagnostic microbiology by saying "We are now in the 21st century, but most of the time we are relying on 19th-century techniques" (Dr. Pallen's contribution to the workshop summary report may be found in Appendix A, pages 238-256). As illustrated by two case studies of genomic epidemiology of Gram-negative pathogens, *Acinetobacter* and *E. coli*, genomics provides a way to do things a bit differently and perhaps better.

***Acinetobacter*: Defining species, transmission, and resistance** Pallen first described how he used whole genome sequencing of *Acinetobacter baumannii* isolates to detect differences between isolates within an outbreak and to determine chains of transmission. *A. baumannii* is a Gram-negative bacillus associated with

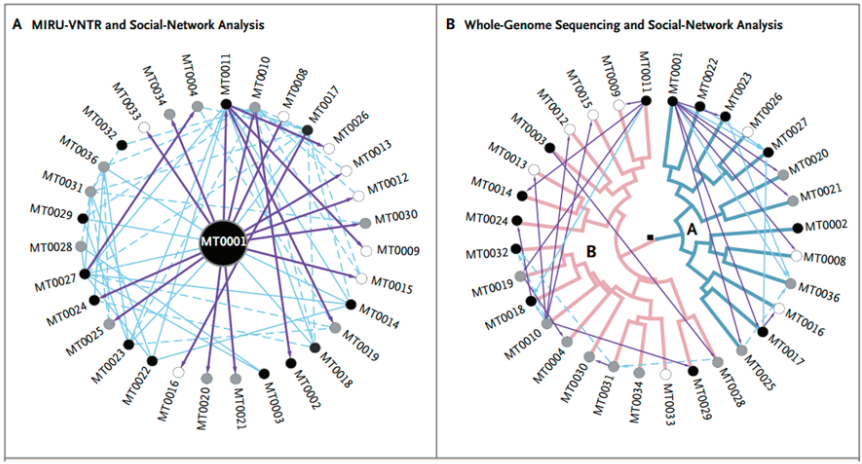


FIGURE WO-38 Putative transmission on networks constructed from genotyping data versus whole genome data for 32 patients. Genotyping data from analyses of mycobacterial interspersed repetitive unit-variable number tandem repeats (MIRU-VNTRs) were used in Panel A, and whole genome data were used in Panel B. Each panel shows patients (identified by case number) represented by circles colored according to smear status and clinical presentation as an index of infectivity: black circles indicate smear-positive pulmonary disease, gray circles smear-positive miliary disease or smear-negative pulmonary disease, and white circles smear-negative extrapulmonary disease. The cases are connected by arrows on the basis of reported social relationships representing plausible transmission attributable to a single case (purple arrows) or multiple potential sources of transmission (light blue lines), with dashed arrows indicating moderately infective patients and solid lines highly infective patients. In Panel A, the case with the earliest symptom onset was MT0001 (center), and the second-earliest case is shown at the 12 o'clock position, with the remaining cases listed in the clockwise direction in order of increasing time since symptom onset. When a clonal outbreak is assumed, the epidemiologic interpretation of the data suggests that most transmission events can be traced to the source case, MT0001. In Panel B, the cases are shown according to circular dendrograms based on whole genome data reflecting the tuberculosis lineage (A in blue and B in pink). This network provides a more accurate picture of transmission, with transmission restricted to each lineage, facilitating epidemiological interpretation of the underlying social-network data and revealing the role of the second and third source cases (MT0010 and MT0011).

SOURCE: Gardy et al. (2011). From *New England Journal of Medicine*, Jennifer L. Gardy, et al., Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak, 364, 730-739. Copyright ©2011 Massachusetts Medical Society. Reprinted with permission from Massachusetts Medical Society.

wound infections, ventilator-associated pneumonia, and bacteremia. Isolates are generally multi-drug resistant, with colistin and tigecycline the only reserve antibiotic agents in many cases. *A. baumannii* has been isolated from military personnel returning from Iraq and Afghanistan, and there have been cases of transmission from military to civilian patients in shared health care facilities.

Isolates from an outbreak in Birmingham Hospital, in 2008, were classified as indistinguishable by standard typing methods. Using whole genome sequencing of six isolates, Pallen's team was able to identify three loci that were SNP variations between isolates (Lewis et al., 2010). Pallen then developed and validated a genotyping scheme based on those SNPs and was able to map different aspects of the outbreak over space and time.

In collaboration with colleagues in London, Pallen found significant differences between *A. baumannii* isolates collected from a patient before and after tigecycline therapy. Prior to therapy the strain was drug-sensitive; following treatment it became drug-resistant. Whole genome sequencing of the post-treatment isolate detected 18 SNPs that were absent from the sequence of the pre-treatment isolate. Nine of the SNPs were non-synonymous (i.e., the change resulted in translation to a different amino acid), and one of them was actually in a gene called *adeS*, which is known to be part of a pair of genes that encode a two-component regulatory system that is involved in the resistance in this particular pathogen to tigecycline (Hornsey at al., 2011). In addition, there were three contigs in the pre-treatment isolate that were not found in the post-treatment isolate. It is often assumed that developing resistance involves acquisition of DNA. In this case, moving toward resistance was actually associated with a loss of DNA. One deletion resulted in a truncated *mutS* gene that is involved in DNA repair. Pallen hypothesized that loss of DNA repair led to an increase in mutation rate and primed the isolate to then acquire antibiotic resistance.

Based on phylogenetic studies, Pallen speculated that, for *Acinetobacter* species, it should be possible to define species by genome sequence alone, without the need for any phenotypic testing or any other screening methods like DNA-DNA hybridization to define a species. 16S sequences, which are commonly used as a taxonomic marker, were not capable of delineating the accepted species within the *Acinetobacter* genus. A core genome phylogenetic tree, however, was consistent with the currently accepted taxonomy and also identified three misclassifications within strains and strain collections. Pallen noted that phylogenetics can be very processor- and time-intensive. However, the use of the "average nucleotide identity" approach quickly delivered results that were consistent with the traditional and phylogenetic classifications (Hornsey et al., 2011).

***E. coli*: Crowdsourcing the genome** To illustrate the power of making data publicly available, Pallen described how the use of social media such as blogging and Twitter can augment the usual channels of scientific discovery and academic discourse. An outbreak of more than 4,000 cases of *E. coli* O104:H4 in Germany

in the summer of 2011 led to more than 50 deaths. This *E. coli* outbreak was characterized by a very high risk of hemolytic uremic syndrome (a complication of Shiga toxin–producing *E. coli*) and was ultimately linked to the consumption of sprouted fenugreek seeds.

Pallen's collaborators in Hamburg sent DNA from an outbreak isolate to BGI[42] for sequencing; BGI subsequently released the sequence data from five sequencing runs on the Ion Torrent platform into the public domain. Within 24 hours of release of the data, Nick Loman in Pallen's research group had assembled the genome and, through his blog post and Twitter account, called upon the bioinformatics community to analyze these data. Within 2 days, the genome had been assigned to an existing lineage of *E. coli*; within 5 days, additional analyses had given rise to a strain-specific diagnostic test. Within a week there were more than two dozen reports filed on the biology and evolution of this strain on an open-source wiki. Pallen underscored that contributors from around the world[43] to this example of open-source genomics were "not professional public health people, for the most part . . . they were people just doing this out of interest and good will . . . and because there was a clear and present need for it."

Tweets and blog posts, however, are no substitute for a peer-reviewed publication, noted Pallen. In collaboration with his German and Chinese collaborators, Pallen and his colleagues wrote up a case study of a family outbreak of O104:H4, which included a description of community research efforts that were coordinated via social media as well as confirmation of all reported analyses (Rohde et al., 2011). Pallen noted that this was just one example of a crowdsourced research project (e.g., www.crowdsourcing.org or www.ancientlives.org). Although everyday science is not likely to open its laboratory notebooks and share everything all the time, this approach could be useful and appropriate in times of a public health emergency.

This collaborative approach raises the question of what constitutes "published" for the purposes of avoiding duplicate publication of original material (the Ingelfinger rule). All of the sequence data and analysis had already been placed into the public domain (through Twitter and blogs) prior to submission of these research results to the *New England Journal of Medicine*, yet the manuscript was accepted for publication (Rohde et al., 2011).

[42] BGI (formerly known as the Beijing Genomics Institute) is a premier genomics and bioinformatics center with facilities around the world.

[43] "If you just look at the names . . . you can see the variety of input. . . . You have English names, Spanish names, Chinese names, Muslim names, Jewish names. . . . People from around the world, from multiple continents, were actually involved in this activity."

Metagenomic Screening for Novel Viruses

David Wang of Washington University in St. Louis, focused his remarks on how the application of viral metagenomics to outbreaks and other settings may provide new directions for investigations (Dr. Wang's contribution to the workshop summary report may be found in Appendix A, pages 311-338). Wang shared two case examples of the application of viral metagenomics: how the investigation of outbreaks can provide new directions for research on viruses in infectious disease, and how viruses may potentially be used as probes to elucidate fundamental mechanisms of innate immunity.

Application of genomics to diarrheal diseases Diarrheal diseases are a significant cause of morbidity and mortality worldwide. Nearly 2 million children die from diarrheal diseases annually, primarily children under the age of 5 in the developing world, who succumb to dehydration in the absence of medical interventions. In the developed world, fatalities from diarrheal diseases are almost nonexistent. Major viral causes of diarrhea include norovirus, rotavirus, adenovirus, and astrovirus. A number of other viruses have also been implicated in acute diarrhea; however, 40 percent of all cases of acute diarrhea have no known etiologic agents. According to Wang, viral metagenomics is a pivotal strategy for detecting potential pathogens in these unknown cases.

Wang described his efforts to identify novel viruses in stool samples from children with clearly diagnosed acute diarrhea that were negative in conventional assays for major known diarrheal viruses. Metagenomic screening provided evidence for many novel viruses, including a new astrovirus.[44] Wang noted that previously, only one species of human astrovirus with eight closely related serotypes (10 percent amino acid variation between serotypes) had been defined. Wang recovered seven astrovirus sequence reads from one sample that mapped to two different loci in the genome (Finkbeiner et al., 2008a). The complete genome of the virus was painstakingly sequenced. Phylogenetic analysis revealed that the newly detected astrovirus was highly divergent—exhibiting less than 60 percent amino acid identity. This was clearly a distinct astrovirus, Wang concluded, not a new serotype of the previously described human astrovirus (Finkbeiner et al., 2008b).

There are many unanswered questions about the role of novel astroviruses in human disease; first and foremost, are they in fact a causative agent of acute human diarrhea, or do they cause disease outside the gastrointestinal tract and are simply shed or transmitted by a fecal–oral route? Could this astrovirus be a

[44] Astroviruses are small, single-stranded, positive sense RNA viruses, 6 to 8 kb in length. They cause diarrhea in humans and other animals. When Wang's group did this study, there was essentially one species of human astrovirus that had been previously defined. The first member of the species was discovered in 1975. Subsequently, eight closely related viral serotypes have been described. At the amino acid sequence level, they vary by about 10 percent from each other.

commensal or a symbiotic part of the human virome? Or was the virus simply ingested by the child and passaged through the gut, having nothing to with human infection? Wang shared preliminary data that suggest that 100 percent of the normal healthy general population is seropositive for this astrovirus (i.e., has been infected at some point in time). Wang noted that viral metagenomic strategies can be used to identify candidate agents but that these leads must undergo a very long process of classic biological workup in order to understand "what the relevance is, what the role is, what this virus is doing—either in a particular set of candidate diseases or even in other diseases that we haven't thought about yet."

Wang described the application of next-generation sequencing to the investigation of an outbreak of acute gastroenteritis in a daycare center. Twenty-six children and teachers became ill over the course of several weeks, and the center was closed. Conventional testing was negative for all known enteric pathogens. Wang sequenced six fecal samples that were available and found another novel astrovirus (see Figure WO-39). He pointed out that while sequencing the entire genome of the novel astrovirus described above took several months, by using next-generation technologies this genome was sequenced in 10 hours. Real-time PCR demonstrated that this novel astrovirus—AstV-VA1—was present in high titer in three out of the six samples available from this outbreak. Wang reiterated that it is not known if this is, in fact, the causal agent of this outbreak, but it certainly generates new hypotheses.

An evaluation of the global diversity of viruses will provide researchers with important genomic libraries for future studies of novel viruses associated with disease outbreaks. Wang and colleagues are now evaluating more than 1,000 stool samples collected from patients with diarrhea or from healthy children, shared by collaborators from around the world. Numerous new viruses have been found, including six additional astroviruses. They are also taking a shotgun metagenomics approach to sequencing untreated raw sewage collected from different sites, integrating environmental metagenomics with clinical metagenomics. Data thus far show incredible diversity in terms of viruses. Wang reported that "the number of viruses that we knew about historically, which were mostly guided by culture-based methods, vastly underestimates the amount of virus that is in any given specimen or niche."

Using genomics to identify viruses that infect the nematode, *Caenorhabditis elegans* *C. elegans* is a genetically tractable model organism that has been used for more than 40 years in developmental biology and neuroscience research investigations. Many fundamental discoveries that were initially made in *C. elegans* have now been translated to mammals and humans. Current thinking is that *C. elegans*, as a primitive eukaryote with no known adaptive immunity and distinct innate immunity, may be a robust system to study host–virus interactions and identify novel antiviral immune pathways. The challenge is that no virus has ever been described that naturally infects *C. elegans*.

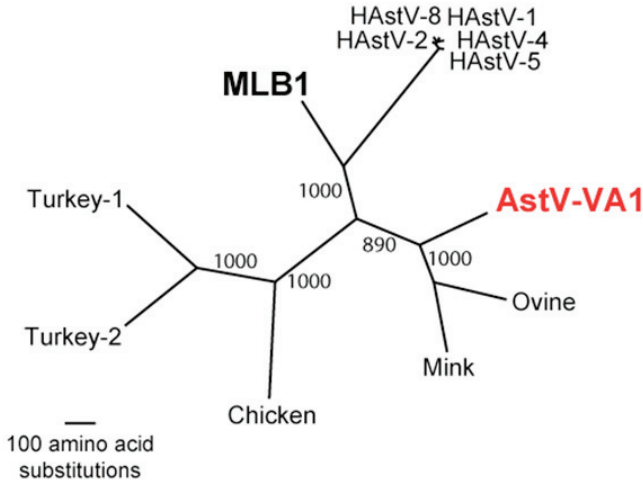


FIGURE WO-39 Phylogenetic analysis of novel astrovirus VA1[45] identified in an outbreak of gastroenteritis. stV-VA1 present in 3 out of 6 samples in outbreak.
SOURCE: Finkbeiner et al. (2008b).

Much of the work in *C. elegans* has been conducted using a reference laboratory strain that is not exposed to any natural pathogens.To identify natural viral pathogens of *C. elegans*, Wang sampled worms in collaboration with Marie-Anne Félix, a nematode ecologist, who collected wild isolates of *C. elegans* and *Caenorhabditis briggsae* from rotting fruits obtained in French orchards. Félix identified several worm isolates for further investigation that appeared sick (i.e., having an unusual intestinal cell morphology) and did not respond to treatment with antibiotics. Multiple isolates appeared to be infected by viruses, and sequencing identified three novel viruses, one from a wild *C. elegans* strain and two from a wild *C. briggsae* strain (Félix et al., 2011). The three viruses are distinct but related to each other, and they are distantly related to nodaviruses (i.e., positive sense RNA viruses that are significant pathogens of fish and also infect insects). The laboratory *C. elegans* strain could also be infected with virus isolated from the wild nematodes. The naïve worms developed the disease phenotype, fulfilling Koch's postulates.

Viral infection of *C. elegans* strains that had a defective RNAi pathway—a well-established antiviral mechanism in plants and insects—resulted in 50- to 100-fold greater accumulation of viral RNA in the mutant strains than in the wild-type *C. elegans* strain. Immunofluorescence assays using antibodies against

[45] Astrovirus VA1 is most closely related to ovine astrovirus.

a viral polymerase showed that the virus primarily infects cells of the nematode intestine.

These are the first data to clearly demonstrate in nematodes that RNAi is antiviral using a bona fide virus infection, Wang explained. More importantly, this demonstrates the feasibility of using the nematode as a model animal system for probing host–virus interactions and for capitalizing on the genetics that have already been worked out in the nematode system as a tool to identify novel modalities of antiviral immunity. Specifically, one can take mutants with defined deletions in genes and obtain phenotypes that are quantifiable as increased viral replication (Félix et al., 2011). Wang emphasized that viral metagenomics provide a useful starting point for biological investigations and that numerous biological questions can now be explored due to this revolution in genomic sequencing technology.

Assessing Abundance: The Rare Biosphere vs. Sequencing Errors

There are many similarities between bacterial communities across samples and environments. Speaker Susan Huse of the Marine Biological Laboratory in Woods Hole[46] said that profiles of microbial communities generated by sequencing of 16S rRNA hypervariable regions consistently have a small number of highly abundant organisms and a large number of low-abundance organisms (the "rare[47] biosphere") (Dr. Huse's contribution to the workshop summary report may be found in Appendix A, pages 188-207). A persistent question is how to determine what is truly rare in a complex mixture and what is sequencing error. For some projects, researchers address this question by simply eliminating all of the low-abundance organisms from the analysis. This approach is obviously not appropriate when studying diversity.

Huse described analysis of the relative abundance of the microbiota in samples from the Human Microbiome Project (Huse et al., 2012). Sequencing 16S rRNA hypervariable regions of 210 stool samples, for example, shows that distribution of abundance varies from patient to patient. Organisms that are present in all patients may be relatively rare in some and highly abundant in others. *Bacteroides*, for example, is the most abundant genus in the stool samples sequenced. Some patient samples contained almost no *Bacteroides*, while others contained nearly 100 percent *Bacteroides*. Huse noted that these results "speak to the importance of not throwing away those rare organisms in these patients, because it's part of a pattern with other patients, where it is more abundant."

A graph of the relative abundance of sequences for a variety of different stool samples shows how widely the observed rare biosphere can vary (Figure WO-40). This does not mean that everything that is rare is true, but rather that everything

[46] Dr. Huse has since relocated to Brown University.

[47] For the purposes of this talk, Huse defined rare as less than 0.1 percent.

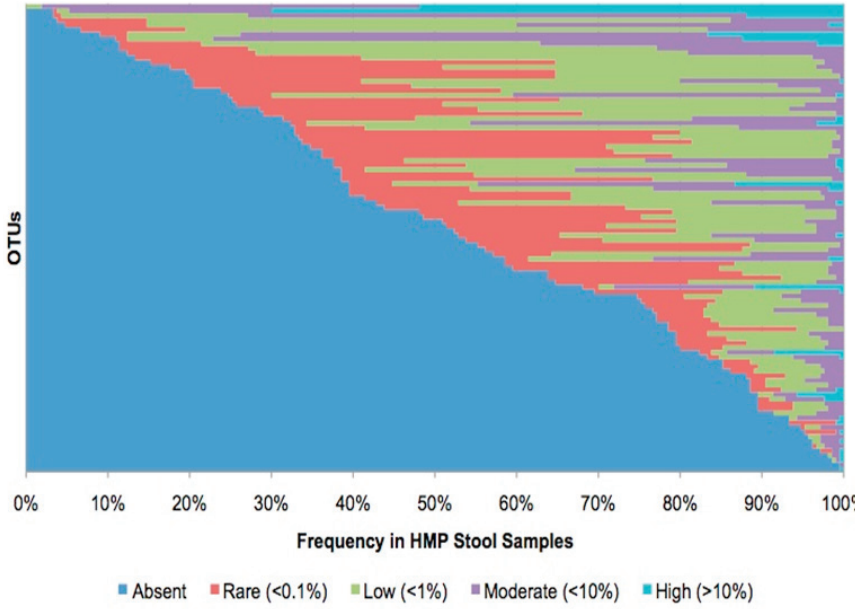


FIGURE WO-40 Distribution of OTU relative abundances across 210 Human Microbiome Project stool samples.
SOURCE: Huse (2012). Data from VandeWalle et al. (2012).

that is abundant can be rare somewhere else. Rather than throwing these data away, Huse noted that researchers need better ways of finding singletons or rare organisms and seeing them in a context of other samples where they are not so rare.

As Huse noted, there is no universal definition of "rare." For microorganisms, rare is not one or two, but tens of thousands, and even millions of cells. Huse observed that one estimate is that there are 1×10^8 bacterial cells in a milliliter of seawater. If "rare" is as low as 0.01 percent of a sample, then rare is about 10,000 cells in a milliliter of seawater. If the estimate is that there are 1×10^{10} bacterial cells in 1 gram of stool, then rare is 1 million cells in a 1 gram sample.

To be able to compare across samples and to identify truly rare organisms in a sample, researchers need to minimize errors and compensate for sequencing quality limitations. Huse suggested that before sequencing begins, a variety of primers should be designed with diverse affinity and minimal bias. If a primer has low affinity to some organisms, they may appear to be part of a rare biosphere, when in fact there was bias against them. The use of proofreading enzymes is important, as is catching PCR errors via sequencing replication, to reduce the impact of early rounding errors. Huse also cautioned that low microbial biomass

samples can be confounded by contaminants. The use of samples that have undergone whole genome amplification can also dramatically alter results.

Once obtained, data must be filtered based on the quality of the sequencing. Huse advised "when in doubt, leave it out," especially when working with 16S rRNA amplicons. The focus is on sequence quality, not simply removing things that are low abundance—because that will bias against truly low-abundance organisms. Quality filtering of next-generation sequencing reads could include, for example, omitting reads with *N*s (unspecified bases), with inexact matches to primers, with low-quality scores, or that do not meet a minimum length. Also useful are stringent chimeric filtering (to remove chimeras from PCR amplification and sequencing errors) and paired-end reads with overlap, in order to eliminate overlapping sequences that are not identical. The next step is to analyze the data at the taxonomic level, filter out anything that is not bacteria (e.g., archaeans, organelles), and aggregate similar sequences.

Because not all species have been named, another common approach is to aggregate the 16S sequences into groups of similar sequences, or OTUs. Huse noted, however, that results often vary by clustering method. Huse described a new clustering method involving a pre-clustering step that improves prediction of OTUs (Huse et al., 2010). Regardless of the platform used, and the quality filtering applied, there will be errant OTUs remaining, however rare the biosphere. Huse discussed several metrics to estimate diversity, including Shannon and Simpson diversity estimates for alpha diversity (within the community) and Jaccard, Bray-Curtis, and Morisita-Horn for beta diversity (across communities).

Huse explained that if a sequence is abundant elsewhere, or is present elsewhere, and follows an ecologically meaningful pattern, then it is highly likely that it is a true rare sequence and not an error. As an example, Huse described two *Acinetobacter* V6 ribosomal sequence tags that differed by a single nucleotide and demonstrated an ecological balance based on season (Figure WO-41). Researchers that only had samples from one time point might have thrown out the low-abundance taxa as an error and missed this potentially important pattern. Huse noted that her group always keeps rare sequences that cannot otherwise be identified as errors until all of the data can be analyzed for broad ecological patterns.

An Approach to Analyzing Metagenomic Data: Inferring Function with MG-RAST

As metagenomics enters into the era of "big data," the research community needs to find ways to drop the costs of data analysis; integrate the terabytes of data being generated; and make these data comparable, said speaker Folker Meyer of the Argonne National Laboratory (Dr. Meyer's contribution to the workshop summary report may be found in Appendix A, pages 230-238). Without doing so, these data will essentially be lost.

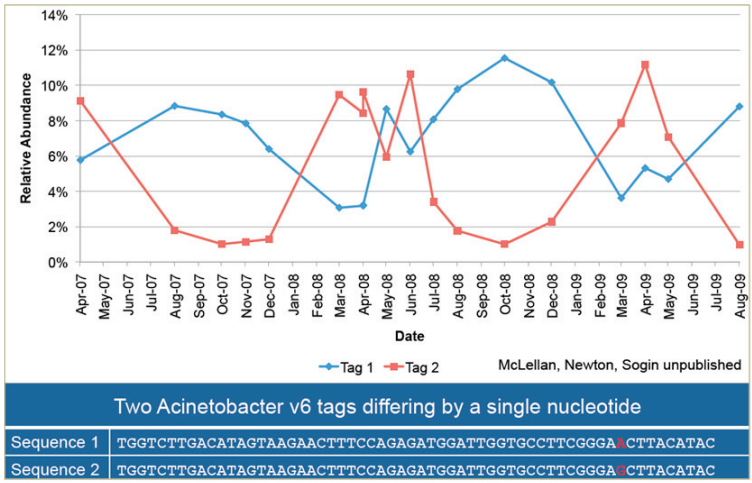


FIGURE WO-41 Two *Acinetobacter* V6 tags differing by a single nucleotide and demonstrating a seasonally balanced abundance.
SOURCE: Huse (2012). Data from VandeWalle et al. (2012).

Meyer illustrated the challenges of analyzing metagenomic data by reviewing his work with Metagenomics RAST (MG-RAST), an open-source, high-performance computing server for metagenomic sequence analysis (Meyer et al., 2008). Developed in 2007 as a simple tool that automated several procedures for data analysis, MG-RAST immediately attracted more than 100 user groups and 100 data sets. As of May 2012, MG-RAST included 49,600 data sets (metagenomes) submitted by users. Since its launch in 2007, more than 14 terabase pairs (or 10^{12} base pairs) of data and more than 120 billion sequences have been analyzed in the system. Meyer explained that user-submitted sequence data undergoes normalization and intensive quality analysis and are mapped against known annotations in numerous data sources (e.g., SEED [Overbeek et al., 2005], KEGG[48] [Kanehisa, 2002], COG[49] [Tatusov et al., 2003]) to predict features of interest (e.g., genes). The system further transforms these data into phylogenetic and functional profiles, which are useful for sample comparison.

MG-RAST requires the submission of metadata along with sequence data, and Meyer underscored the critical importance of metadata—where the data are from, what procedures were used in data generation, and how the data were recorded—to the interpretation of results. Meyer cited functional and taxonomic comparisons of 1,606 Human Microbiome Project shotgun metagenomes against

[48] Kyoto Encyclopedia of Genes and Genomes.
[49] Clusters of orthologous groups of proteins.

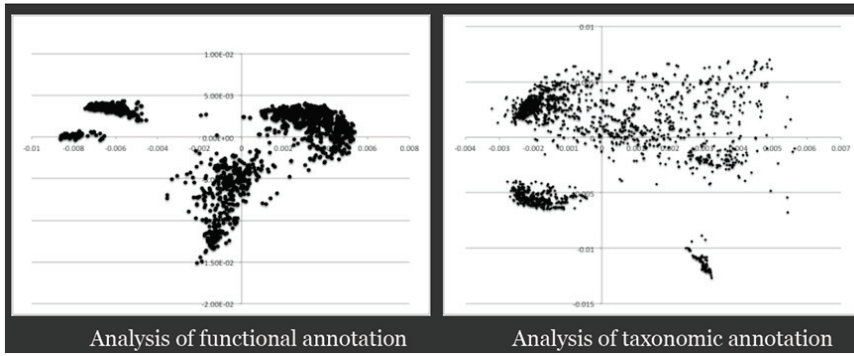


FIGURE WO-42 Principal coordinates analysis of 1,606 Human Microbiome Project shotgun metagenomes.
SOURCE: Meyer (2012).

known annotations that resulted in two different principal coordinates analysis (PCoA) outputs (Figure WO-42).

Outlier groups for both comparisons reflect the different sequencing technologies and platforms used for different samples (Figure WO-43). Meyer pointed out that this finding is only possible because there are metadata regarding the sequencing technology for the majority of the samples.

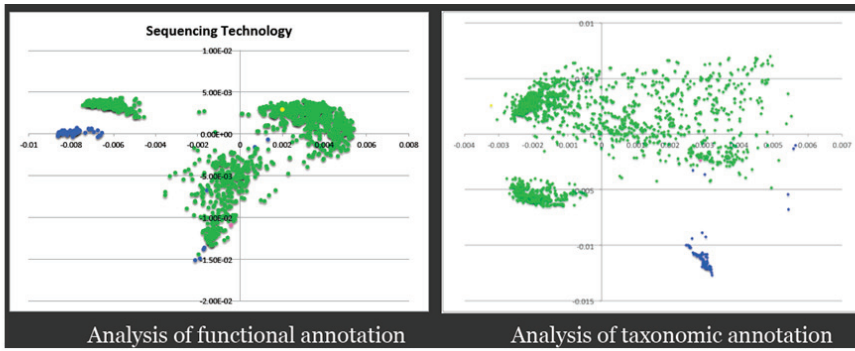


FIGURE WO-43 Principal coordinates analysis of 1,606 Human Microbiome Project shotgun metagenomes painted by sequencing technology. The purple dot is explained by a problem with the metadata and therefore should be green, but the metadata provided with the sample do not correctly identify it as Illumina (green). Blue: 454 GS FLX or GS FLX Titanium; Green: Illumina Genome Analyzer, Analyzer II, or HiSeq 2000; Yellow: unspecified.
SOURCE: Meyer (2012).

Quality control for *de novo* sequencing Data quality is tightly correlated to the ability to analyze and compare sequences, especially on automated platforms such as MG-RAST. Next-generation sequencing often produces "noisy" data. However, unlike the more mature field of phylogenetic surveys, which has developed "de-noising" approaches (discussed by Huse), the field of shotgun metagenomics does not have any "vendor-neutral" way to characterize data quality. Novel quality control approaches are needed to ensure the integrity of comparison between metagenomic datasets.

Meyer described a novel method of "duplicate read inferred sequencing error estimation" (DRISEE) that uses PCR artifacts in replicate reads to develop error profiles for shotgun metagenomic sequencing data sets (Keegan et al., 2012). He noted that when his group has used DRISEE, it has demonstrated that different experiments, and different samples in a single experiment, exhibit unique error profiles. He noted that this phenomenon was not related to the sequencing hardware but rather to operator error; some sequencing centers consistently produce high-quality data while others are more highly variable. Echoing Huse, Meyer noted that "errors are much, much higher in our estimate than what the vendors typically tell us they are." He concluded that "we as a community need to find a way to look at errors, at error profiles, [so we know when to] discard data sets or redo experiments."

Predicting features Meyer described two basic methods for predicting protein coding features from genome sequence data: (1) statistics-based approaches using codon usage and (2) similarity-based approaches using BLASTX searches. He noted that it is difficult to find novel proteins, genomic islands, and horizontally transferred genes using a statistics-based approach. Similarity-based approaches require substantial computational time, often weeks on multiple machines, and are correspondingly expensive (about 10 times the cost of generating the sequence data). In addition, novel proteins will not be identified through similarity searching. Again, novel computational approaches are needed.

Meyer and colleagues evaluated the reading frame prediction accuracy of five current gene computation algorithms—FragGeneScan, MetaGeneAnnotator, MetaGeneMark, Prodigal, and Orphelia. They found the tools were comparable in making predictions based on error-free sequence. When these tools were applied to simulated data sets containing errors, there were notable differences in performance (Trimble et al., 2012). Meyer concluded, "Even in the best of all cases, we only hit 60 percent of all possible gene features that we can see in our data."

Annotation and reproducibility There are multiple annotation databases that can be used to find similarities between sequencing data and known proteins. As such, MG-RAST supports a number of different databases to annotate function. Despite identical processing of data, choosing a different database as your basis for comparison significantly changes results, raising questions of data comparability using a different pipeline.

To address a similar question about result reproducibility from the same setting, Meyer presented data suggesting that results obtained from shotgun metagenomics experiments were reproducible. He also referred to the work of Zhou who found that analyses of 16S rRNA amplicons were not reproducible (Zhou et al., 2011). Meyer suggested that this is because abundance was not taken into account.

Assembly Metagenome assembly should be relatively straightforward, but it is fraught with challenges regardless of the assembly tool used. Meyer noted that algorithms designed to assemble clonal genomes cannot handle the subspecies or strain variation present in metagenomic samples. Moreover, he found that varying the size of the overlap that the assembler is looking for (the k-mer size) changed the results dramatically. Using short k-mers resulted in many short contigs, while long k-mers led to a few long contigs. According to Meyer, this is a problem because most metagenomic studies "picked an arbitrary point in that [k-mer] space, declared it to be true, and submitted that data." This suggests that "everything we have assembled and deposited in terms of metagenomics needs to be revisited, and we frequently do not have access to the raw reads."

If assembly does work, and the data set is of high quality, metagenomic data can support new types of research. Analyses can produce genomes that serve as a reference strain for mapping all of the reads in a metagenome, allowing investigators to conduct population genomic and genetic studies on single metagenomes—all for less than $1,000 (Meyer et al., *in preparation*).

The analysis bottleneck Computing costs of sequence analysis far outsize those associated with generating metagenomic sequencing data (Figure WO-44). Although new analytical tools are emerging, new approaches to data sharing and storage, as well as to building and maintaining community resources, will be needed (Thomas et al., 2012). Meyer emphasized that although the community relies upon curated databases of existing knowledge, annotation and reference databases are not well funded. Currently metagenomic data are not centralized; most are stored in private resources (e.g., the RAST server at Argonne National Laboratory) rather than in public resources such as GenBank. There need to be easier ways for the community to deposit and openly access data deposited and curated in central locations.

Meyer suggested that the challenge is so substantial that "[the field] needs to change the way we do business." In particular, Meyer called for standards that would support data integration and exchange. The community could also identify and make available "gold standard" data sets. Ideally, this would mean that data sets would only need to be analyzed once, and a common archive could be established to distribute the raw data and analytical results for further study (Desai et al., 2012).

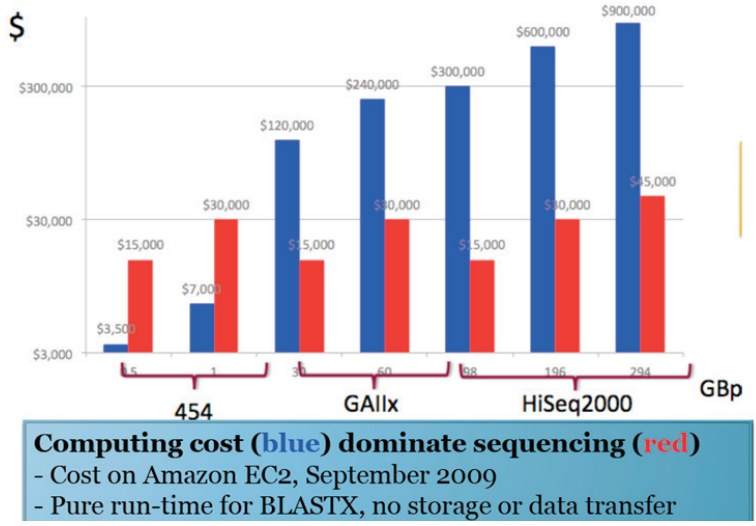


FIGURE WO-44 The costs associated with data analysis are becoming the bottleneck for metagenomic analyses.
SOURCE: Meyer (2012). Adapted from Wilkening et al. (2009).

Moving Forward: Challenges and Opportunities

While genomics has introduced a new era in microbiology with significant implications for improving our ability to detect, diagnose, treat, and even anticipate infectious disease emergence, there exist many challenges and caveats that must be addressed before these technologies will find widespread adoption and use beyond the research laboratory. Sequencing technologies that have supported this revolution continue to evolve, and while this evolution leads to ongoing advancement of the field, it also creates a moving target with new challenges. For Sanger sequencers, sequence production was the rate-limiting step. Next-generation technologies that became available beginning in 2005 are radically different from Sanger-based technology;[50] they are "massively parallel" systems that produce "sequence information from hundreds of thousands to hundreds of millions of DNA molecules simultaneously" (Mardis, 2011). Each new platform has struck a different balance between cost, read length, data volume, and rate

[50] Although the biochemistry of each platform is different, next-generation sequencing technologies feature simpler sample preparation steps, dispense with the need to create libraries of cloned sequences in bacteria, and rely on parallel, cyclic interrogation of sequences of clonal amplicons of DNA. An orchestrated series of repeating steps are performed and detected automatically (Mardis, 2011).

of data generation, but all produce far more sequence reads per instrument run than capillary sequencers and at significantly lower cost (Figure WO-45) (Mardis, 2011). The massive influx of data from next-generation sequencing technologies, with shorter read lengths and different error profiles, has also brought significant challenges to their analysis (Mardis, 2011).

Data Quality, Comparability, and Analysis

The explosion of sequencing technologies and applications is generating more results than ever, but it is unclear what these data may mean. Workshop speakers and participants discussed some of the key challenges in this rapidly advancing field, and potential methods to address them, focusing on the core themes of data quality and analysis.

Comparability Across Sequencing Platforms and Technologies

Perhaps the most formidable challenge for investigators using different sequencing technology platforms is data comparability across methods and

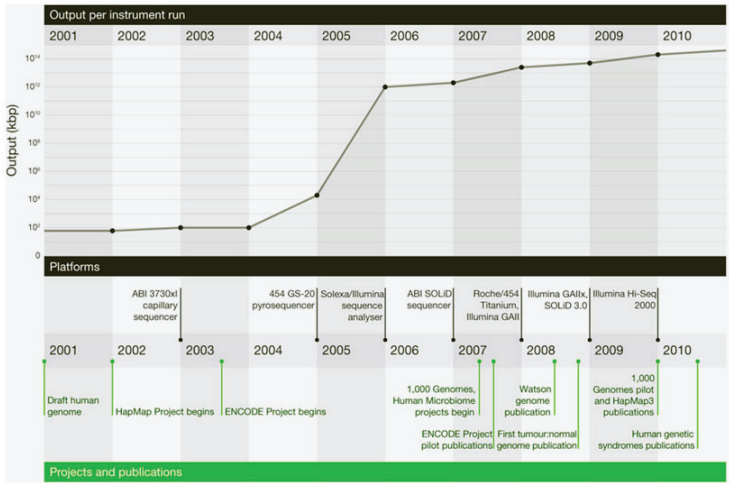


FIGURE WO-45 Changes in instrument capacity over the past decade, and the timing of major sequencing projects. Top panel, Increasing scale of data output per run plotted on a logarithmic scale. Middle panel, Timeline of major milestones in massively parallel sequencing platform introduction and instrument revisions. Bottom panel, timing of selected projects and milestones.
SOURCE: Mardis (2011). Reprinted by permission from Macmillan Publishers Ltd: NATURE. Mardis, E.R. 2011. A decade's perspective on DNA sequencing technology. *Nature* 470:198-203, copyright 2011.

approaches. Each sequencing platform uses a unique set of protocols for template preparation, amplification and sequencing, imaging, and data analysis (genome alignment and assembly methods) that determine the type and quality of data produced, and such protocols may influence the strength of the "signal" versus the "noise" and overall accuracy of these data.[51] Although each manufacturer provides quality scores and accuracy estimates, there is no consensus that a "quality base from one platform is equivalent to that from another platform" (Metzker, 2010). The cross-platform comparability of data also has implications for the reproducibility of research results.

As sequencers have become much smaller and less expensive, more and more laboratories are starting to perform their own sequencing. Benchtop sequencers are poised to revolutionize microbiology, but platforms do have strengths and weaknesses (see Loman et al., 2012a). In his presentation, Pallen described sequencing an isolate from the German *E. coli* O104:H4 outbreak on three different platforms and concluding that all three were "fit for the purpose," although each had particular issues; both the Ion Torrent PGM and the 454 GS Junior produced homopolymer-associated indel errors (Loman et al., 2012b). Pallen emphasized it is not only the sequencer used, but also the analytical approaches, particularly the assemblers employed, that may contribute to divergent results. None of these sequencing platforms or assemblers gives perfect results.

Training Gilbert observed that in addition to concordance between platforms, there are also issues of concordance within a platform. The same sample run multiple times on the same platform can give disparate results depending upon the technician/operator, and data quality is often more dependent on the people doing the sequencing than on the platform itself.

In addition, the adoption of these technologies to enable clinical applications will also require the development of a "gold standard" approach to sequencing, including training of new staff, or training of staff on new equipment. Good training is essential in order to ensure comparable results. Most training takes place on the job, and Gilbert noted that a laboratory can lose two or three sequence runs because the genomic libraries were constructed inappropriately. These types of errors can reduce customer confidence and set the laboratory back in terms of time and money. One approach is to send laboratory staff to a major sequencing center for training, and Weinstock observed that laboratories that purchase next-generation sequencing instruments often send their core facility person to the Genome Institute at Washington University in St. Louis, Missouri to learn the protocols.

[51] Errors may include insertion and deletion errors; mismatches/substitutions; and underrepresentation of certain regions—such as AT- or GC-rich repeat regions that reduce the accuracy of sequence reads.

Communication The sharing of experiences by sequencing community members is also important for fostering high-quality results across platforms and laboratories. Pallen noted that when his laboratory first acquired the 454 sequencer, an experienced postdoctoral fellow who began to use it obtained poor results. Pallen's first impulse was to blame the fellow. In discussing this problem with another laboratory, Pallen learned that the other laboratory was also obtaining poor results on a machine they had used for 18 months. Inasmuch as this machine was being operated by the same technicians, it soon became clear that, in fact, the source of the problem was "bad" reagents rather than operator error.

Many workshop participants agreed that there is a need for better communication among the platform user community. There is no central location to report and track problems with sequencing systems (along the lines of the Food and Drug Administration website for reporting adverse reactions to drugs). Users resort to informal communication media, such as Twitter, blogs, and community forums like SEQanswers to share and obtain information about sequencing issues. It was pointed out that manufacturers' websites generally have a page where problems can be posted and seen by all users. While posting an issue on a manufacturer's webpage garners a response from a representative to help solve the particular issue, it does not necessarily help to promote discussion within and across the community. Instead it is an approach to solving issues one laboratory at a time, for what may, in fact, be a community-wide problem.

Regulatory issues From a regulatory perspective, sequencing is primarily a research activity. A laboratory that performs sequencing that will be used in clinical medicine must first be approved under the Clinical Laboratory Improvement Amendments (CLIA). In addition, diagnostic kits must have approval from the FDA. In this regard, it is not clear whether sequencers used for diagnostics will be classified as Class III medical devices. Weinstock noted that the Genome Center at Washington University in St. Louis, Missouri, is currently seeking CLIA approval.

Data Quality: Careful Sample Preparation and Results Filtering

Much of data quality resides in careful planning before sequencing, including primer design and sample preparation, and quality filtering of raw sequence reads before assembly and analysis—approaches discussed earlier by Huse and Meyer. With regard to planning, participants noted that multiple samples are commonly assayed in a single sequencing run, and there exists the potential for data mix-ups due to bar code errors. Huse explained that if bar codes are within one base pair of one other, there can be sequencing errors that take data bioinformatically from one sample to another.

Weinstock suggested that for microbial genomics, the goal is to routinely be able to produce a perfect genome. The only sequences that are ambiguous in such a perfect genome are bases that are programmed to change; for example, in

the process of culturing the organism to harvest DNA for sequencing there will be a certain amount of variation (e.g., antigenic variation). Obtaining the perfect genome requires the use of multiple sequencing platforms for data comparability.

Weinstock emphasized data quality over quantity. One approach is to prepare sequence read pairs that overlap in the variable region of the 16S rRNA and to only use sequences where there is complete agreement over the entire length. This means that a lot of sequence data are discarded, but what remains is unlikely to have errors. Speaker Huse noted that if one is not careful with quality filtering, and only relies on a perfect overlap for example, then it is possible to end up with a range of perfect, but incorrect reads. Pallen emphasized the importance of human involvement and the expertise of the microbiologist looking at the sequence data to make observations, find errors, and draw conclusions. Participants also discussed the importance of training in this regard as well as the performance of the sequencing platforms.

Gilbert noted the situation is somewhat different for high-throughput metagenomic community analysis. The primary interest in doing a 16S rRNA survey is screening the samples to search for emergent properties in the ecological patterns that exist between samples. If there are patterns of interest, then the system can be characterized using more rigorous techniques. Weinstock concurred that 16S rRNA sequencing is mainly to compare organisms, and there is a tolerance for errors because it is possible to get a database hit without necessarily having a perfect sequence (as long as errors are consistent and do not introduce biases).

Sample collection and preparation Throughout the workshop presentations, the need for improvement in sample collection and preparation was repeatedly discussed and highlighted. Weinstock noted that as the capacity for sequencing increases, the challenge for laboratories is how to handle much larger numbers of samples in order to take advantage of the available sequencing capacity.

Another challenge is the volume of sample needed, and participants discussed how it would be helpful to have sequencing platforms that require much lower amounts of nucleic acid input. It was noted that there are new library preparation instruments in development that use nanochannel arrays. Reducing the amount of nucleic acid needed for sequencing would allow researchers to take advantage of sequencing technologies for samples with very small amounts of material (Woyke et al., 2010). One example of a nanopore sequencing technology under development is a portable, disposable, single molecule analysis device (Figure WO-46) (Pennisi, 2012) .

Improvements in sampling and the recovery of genomes from environmental samples will be particularly relevant to the analysis of samples from infectious disease outbreaks. The rapid sequencing of samples from the deadly epidemics of cholera in Haiti in 2010 and *E. coli* O104:H4 in Germany in 2011 (discussed earlier by Pallen) suggest that it may soon be feasible for clinicians to access the genomic content of an outbreak strain in close to real time (Otto, 2011). In this

FIGURE WO-46 Prototype of a nanopore sequencing technology currently under development.
SOURCE: Oxford Nanopore Technologies.

regard, Budowle suggested that a role for government is to develop sampling strategies, modeling outbreak and biocrime scenarios to determine the best approaches to gather information to enable rapid inferences (rather than focusing on technology or software, which industry is already working on).

Keim noted that the ability to collect and bring samples home from overseas—such as from China and Russia—has become much more difficult since the terrorist attacks of 9/11, because of changes in both U.S. and foreign regulations on the movement, export, and import of microbiological samples. The strategy now is to engage the international community and raise its technological capabilities and standards so that advanced genomics may be reliably performed locally. While we may not be able to transport organisms, SNP genotypes can be transferred instantaneously around the world via the Internet.

Standards for Data Sharing and Analysis

To date, most laboratories have deposited sequences in public sequence repositories. As the amount of sequencing data continues to grow, consideration is needed of how to usefully record sequence data as well as experimental metadata. The Genomics Standards Consortium[52] and the Sanger Institute have proposed standards for the description of genome sequences, including information about the origin, pathogenicity, host, growth conditions, estimated genome size, and

[52] The genomic standards consortium is a consortium of groups, including the DNA Data Bank of Japan, the European Bioinformatics Institute, the European Molecular Biology Laboratory, the Joint Genome Institute, and the National Center for Biotechnology Information.

characteristics related to growth, habitat, taxonomy, and genetics (MacLean et al., 2009). Standards development for metadata will be critical for comparative analysis of large metagenomic datasets.

As discussed earlier, it has become feasible for individual laboratories to undertake sequencing projects rather than sending samples to a few large genome centers. Current large-scale genome and metagenome sequencing projects are deploying multiple platforms and different sequencing chemistries in parallel. The additional challenge created by so many sequencers is the sheer volume of data. The production of billions of sequence reads places substantial demands on the existing information technology infrastructure in terms of data transfer, storage, quality control, computational analysis, and information management systems for sample tracking and process management (Metzker, 2010).

Investigators must increasingly struggle with the question of how relevant information will be most effectively extracted from the massive amounts of genomic data being generated, and what other methods, tools, data, and analytical approaches may be needed in order to effectively interpret and interrogate these data for a variety of applications (Mardis, 2011; Relman, 2011). There is a need for better bioinformatics algorithms, including assembly algorithms, that can integrate data from diverse sources; access to faster computing capabilities; larger and more efficient data storage devices; and careful capture of important metadata (e.g., date and location of isolation, culture passages in the laboratory, patient medical information).

Workshop participants discussed the challenges of maintaining curated data sets, particularly the lack of funding for this purpose. One option suggested was a "broker ecosystem" where one broker for metagenomics would have stable funding for a number of years and would provide services to the community (e.g., curate data, metadata, act as chaperone for the data of the specific communities). Another suggestion discussed was to develop applications that take a first pass at the data stream to classify it in some way (e.g., taxonomic, gene family, etc.) prior to comparisons with databases (rather than constant, expensive, all-versus-all BLAST searches)—more of an iterative, semi-distributed, and then semi-centralized analysis pipeline. The problems surrounding lack of data sharing were also noted. Developers are creating algorithms based on 5-year-old data from old sequencing technology. If data were available in a timely fashion to research groups, algorithms developed would be more relevant to the current data emerging from next-generation technologies.

Useful Metagenomic Data Sets: Experimental Design and Collection of Metadata

"For the data sets produced for metagenomic studies, the size of data sets, their heterogeneity, and lack of standardization for both metadata and gene descriptive data present significant challenges for comparative analyses" (DeLong,

2009). Gilbert et al. (2011) noted the need to balance observational studies with carefully designed experimental approaches. The authors cite the Global Ocean Sampling expedition as an example of an early and groundbreaking observational study that provided an unprecedented snapshot of the diversity and heterogeneity in naturally occurring microbial populations (Gilbert and DuPont, 2011; Rusch et al., 2007). The study collected 41 different samples from a wide variety of aquatic habitats from more than 8,000 km and resulted in 7.7 million sequencing reads.

The resulting data set is one of the largest metagenomic data sets ever collected, comprising more than 6.12 million predicted proteins including representatives from all previously known families of microbial proteins and 1,700 new ones (Hugenholtz and Tyson, 2008). This data set has, however, been criticized "for poor experimental design and the absence of appropriate metadata necessary for analysis of the influence of environment on microbial diversity" (Gilbert et al. 2011). As metagenomics moves from the description of apparent diversity to the genuine description of complexity and function, carefully designed experimental approaches are needed to deliver the true potential of metagenomics (Knight et al., 2012).

Improving the Genome Knowledge Base: Diversity and Meaningful Annotation

Annotation is the process of assigning meaningful information, such as the location or function of genes, to raw sequence data. Reliable and consistent annotations are thus essential for the analysis and interpretation of genome data (Berglund et al., 2009). Insights from genomics, such as functional prediction, depend upon the accessibility of existing, well-annotated gene sequences. One obstacle to the use of existing annotations is the bias in the genome and protein knowledge bases. Organisms selected for genomic sequencing are culturable, numerically dominant in habitats of interest (e.g., the human body), and predominantly associated with (human) disease (Relman, 2011).

To be useful, whole genome sequence data must be annotated with functional information and gathered from a diverse array of microorganisms. To aid the characterization of the human microbiota, for example, one of the goals of the human microbiome project is to sequence 3,000 bacterial genomes as a reference set, as well as other genomes from fungi and eukaryotic organisms (The Human Microbiome Jumpstart Reference Strains Consortium, 2010). As illustrated by Wu et al. (2009), a considerable amount of phylogenetic "dark matter" remains unsampled (Figure WO-47).

Whole genome sequence information is available for a small subset of the known phylogenetic diversity of bacteria and archaea (based on unique SSU rRNA gene sequences). Here the authors of the Genomic Encyclopedia of Bacteria and Archaea (GEBA) depict four subsets of phylogenetic diversity: organisms with sequenced genomes pre-GEBA (blue), the GEBA organisms (red), all cultured organisms (dark grey), and all available SSU rRNA genes (light grey).

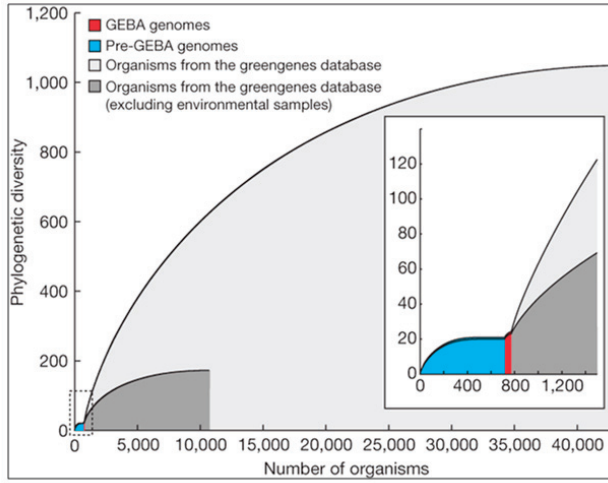


FIGURE WO-47 Phylogenetic "dark matter" left to be sampled.
SOURCE: Wu et al. (2009). Reprinted by permission from Macmillan Publishers Ltd: NATURE. Wu, D. et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056-1060, copyright 2009.

For each subtree, taxa were sorted by their contribution to the subtree phylogenetic diversity, and the cumulative phylogenetic diversity was plotted from maximal (left) to the least (right). The inset magnifies the first 1,500 organisms. Comparison of the plots shows the phylogenetic dark matter left to be sampled (Wu et al., 2009).

Looking Forward

"Once the diversity of the microbial world is catalogued, it will make astronomy look like a pitiful science."

—Julian Davies, Professor Emeritus,
Microbiology and Immunology,
University of British Columbia

In order to maximize the discovery and characterization of new gene families and their associated novel functions, Wu et al., (2009) suggest that phylogenetic considerations should guide the selection of genomes for sequencing. By focusing on the novelty of an organism (highly divergent lineages of bacteria or archaea that lack representatives with sequenced genomes) the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project seeks to "provide a phylogenetically balanced genomic representation of the microbial tree of life" (Wu et al., 2009). Sequencing the genomes of 1,520 phylogenetically selected isolates could include half of the phylogenetic diversity represented by known cultured bacteria and

archaea. The sequencing of an additional 9,218 genome sequences from currently uncultured species could capture 50 percent of this subset of recognized diversity. According to Wu et al. (2009), "such an undertaking will require the development of new approaches to culturing or processing of multi-species samples using methods such as . . . physical isolation of cells from mixed populations followed by whole genome amplification methods."

The field of microbiology has made tremendous strides over the past several decades in describing the microbial world glimpsed for the first time just 300 years ago. Comparative genomic studies of bacterial species and metagenomic analyses of microbial communities to date have revealed how vastly we have underestimated the diversity, variability, and complexity of the microbial world. Microbial genomics offers the potential to efficiently characterize the vast cosmos of microbial diversity and rewrite the microbial community's tree of life. Indeed, with the proliferation of culture-independent technologies and generation of enormous quantities of raw genomic sequences of microorganisms from diverse settings, the field of microbiology now suffers an "embarrassment of riches." As observed in a recent editorial in *Nature Reviews Microbiology* (Editorial, 2011), "[t]he scale of life in the microbial world is such that amazing numbers become commonplace. These numbers can be sources of inspiration for those in the field and used to inspire awe in the next generation of microbiologists."

WORKSHOP OVERVIEW REFERENCES

Achtman, M., G. Morelli, P. Zhu, T. Wirth, I. Diehl, B. Kusecek, A. J. Vogler, D. M. Wagner, C. J. Allender, W. R. Easterday, V. Chenal-Francisque, P. Worsham, N. R. Thomson, J. Parkhill, L. E. Lindler, E. Carniel, and P. Keim. 2004. Microevolution and history of the plague bacillus, *Yersinia pestis*. *Proceedings of the National Academy of Sciences USA* 101(51):17837-17842.

Alagely, A., C. J. Krediet, K. B. Ritchie, and M. Teplitski. 2011. Signaling-mediated cross-talk modulates swarming and biofilm formation in a coral pathogen *Serratia marcescens*. *International Society for Microbial Ecology Journal* 5(10):1609-1620.

Alm, E. 2012. *Session I: The Application of Computational/Theoretical and Experimental Approaches to Study the Evolution of Microorganisms*. Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 12.

Baz, M., Y. Abed, J. Papenburg, X. Bouhy, M. E. Hamelin, and G. Boivin. 2009. Emergence of oseltamivir-resistant pandemic H1N1 virus during prophylaxis. *New England Journal of Medicine* 361(23):2296-2297.

Bentley, S. 2009. Sequencing the species pan-genome. *Nature Reviews Microbiology* 7:258-259.

Berglund, E. C., B. Nystedt, S. G. E. Andersson. 2009. Computational resources in infectious disease: Limitations and challenges. *PLoS Computational Biology* 5(10):e1000481.

Blaser, M. 1997. Ecology of *Helicobacter pylori* in the human stomach. *Journal of Clinical Investigation* 100(4):759-762.

Budowle, B. 2012. *Session IV: Microbial Forensics.* Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine Forum on Microbial Threats, June 13.

Bos, K. I., V. J. Schuenemann, G. B. Golding, H. A. Burbano, N. Waglechner, B. K. Coombes, J. B. McPhee, S. N. DeWitte, M. Meyer, S. Schmedes, J. Wood, D. J. Earn, D. A. Herring, P. Bauer, H. N. Poinar, and J. Krause. 2011. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* 478(7370):506-510.

Casadevall, A. 2012. *Session III: Virulence as an Emergent Property*. Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 13.

Casadevall, A., and L. A. Pirofski. 1999. Host-pathogen interactions: Redefining the basic concepts of virulence and pathogenicity. *Infection and Immunity* 67(8):3703-3713.

———. 2000. Host-pathogen interactions: Basic concepts of microbial commensalism, colonization, infection, and disease. *Infection and Immunity* 68(12):6511-6518.

———. 2001. Host-pathogen interactions: The attributes of virulence. *Journal of Infectious Diseases* 184(3):337-344.

———. 2002. The meaning of microbial exposure, infection, colonization, and disease in clinical practice. *Lancet Infectious Diseases* 2(10):628-635.

———. 2003. The damage-response framework of microbial pathogenesis. *Nature Reviews Microbiology* 1(1):17-24.

———. 2007. Accidental virulence, cryptic pathogenesis, martians, lost hosts, and the pathogenicity of environmental microbes. *Eukaryotic Cell* 6:2169-2174.

Casadevall, A., F. C. Fang, and L. A. Pirofski. 2011. Microbial virulence as an emergent property: Consequences and opportunities. *PLoS Pathogens* 7(7):e1002136.

Castillo-Ramírez, S., S. R. Harris, M. T. Holden, M. He, J. Parkhill, S. D. Bentley, and E. J. Feil. 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathogens* 7(7):e1002129.

Chin, C. S., J. Sorenson, J. B. Harris, W. P. Robins, R. C. Charles, R. R. Jean-Charles, J. Bullard, D. R. Webster, A. Kasarkis, P. Peluso, E. E. Paxinos, Y. Yamaichi, S. B. Calderwood, J. J. Mekalanos, E. E. Schadt, and M. K. Waldor. 2011. The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* 364(1):33-42.

Coleman, M. L., and S. W. Chisholm. 2010. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proceedings of the National Academy of Sciences USA* 107(43):18634-18639.

Cracraft, J., and M. J. Donoghue. 2004. *Assembling the Tree of Life.* New York: Oxford University Press.

Croucher, N. J., S. R. Harris, C. Fraser, M. A. Quail, J. Burton, M. van der Linden, L. McGee, A. von Gottberg, J. H. Song, K. S. Ko, B. Pichon, S. Baker, C. M. Parry, L. M. Lambertsen, D. Shahinas, D. R. Pillai, T. J. Mitchell, G. Dougan, A. Tomasz, K. P. Klugman, J. Parkhill, W. P. Hanage, and S. D. Bentley. 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science* 331(6016):430-434.

de Kruif, P. 1926. *Microbe hunters.* Orlando, FL: Harcourt, Inc.

DeLong, E. F. 2009. The microbial ocean from genomes to biomes. *Nature* 459:200-206.

Denef, V. J., and J. F. Banfield. 2012. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* 336:462-466.

Desai, N., D. Antonopoulos, J. A. Gilbert, E. M. Glass, and F. Meyer. 2012. From genomics to metagenomics. *Current Opinion in Biotechnology* 23(1):72-76.

Dethlefsen, L., M. McFall-Ngai, and D. A. Relman. 2007. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* 449:811-818.

Devignat, R. 1951. Varieties of *Pasteurella pestis*; new hypothesis (in French). *Bulletin of the World Health Organization* 4(2):247-263.

Dobell, C. 1932 (reisssued 1960). *Antony van Leeuwenhoek and his "Little Animals."* New York: Harcourt, Brace and Company.

Editorial. 2011. Microbiology by numbers. *Nature Reviews Microbiology* 9:628.

Edwards, R. A., and R. Rohwer. 2005. Viral metagenomics. *Nature Reviews Microbiology* 3(6): 504-510.

Eisen, J. A. 2007. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biology* 5(3):0384-0388.

Eisen, J. A., D. Kaiser, and R. M. Myers. 1997. Gastrogenomic delights: A movable feast. *Nature Medicine* 3(10):1076-1078.

Félix, M. A., A. Ashe, J. Piffaretti, G. Wu, I. Nuez, T. Bélicard, Y. Jiang, G. Zhao, C. J. Franz, L. D. Goldstein, M. Sanroman, E. A. Miska, and D. Wang. 2011. Natural and experimental infection of *Caenorhabditis* nematodes by novel viruses related to nodaviruses. *PLoS Biology* 9(1):e1000586.

Finkbeiner, S. R., A. F. Allred, P. I. Tarr, E. J. Klein, C. D. Kirkwood, and D. Wang. 2008a. Metagenomic analysis of human diarrhea: Viral detection and discovery. *PLoS Pathogens* 4(2):e1000011.

Finkbeiner, S. R., C. D. Kirkwood, and D. Wang. 2008b. Complete genome sequence of a highly divergent astrovirus isolated from a child with acute diarrhea. *Virology Journal* 5:117.

Finkbeiner, S. R., Y. Li, S. Ruone, C. Conrardy, N. Gregoricus, D. Toney, H. W. Virgin, L. J. Anderson, J. Vinje, D. Wang, and S. Tong. 2009. Identification of a novel astrovirus (Astrovirus VA1) associated with an outbreak of acute gastroenteritis. *Journal of Virology* 83(20):10836-10839.

Fisher, A. T., and R. P. Von Herzen. 2005. Models of hydrothermal circulation within 106 Ma seafloor: Constraints on the vigor of fluid circulation and crustal properties, below the Madeira Abyssal Plain. Geochemistry Geophysics Geosystems. 6: Q11001. doi:10.1029/2005GC001013. doi:10.1029/2005GC001013.

Fitz-Gibbon, S., S. Tomida, B. H. Chiu, L. Nguyen, C. Du, M. Liu, D. Elashoff, M. C. Erfe, A. Loncaric, J. Kim, R. L. Modlin, J. F. Miller, E. Sodergren, N. Craft, G. M. Weinstock, and H. Li. 2013. *Propionibacterium Acnes* strain populations in the human skin microbiome associated with acne. *Journal of Investigative Dermatology* doi: 10.1038/jid.2013.21. [Epub ahead of print].

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L. I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Frichman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C.M. Fraser, H. O. Smith, J. C. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenza* rd. *Science* 269:496-512.

Fraser, C. M., J. A. Eisen, and S. L. Salzberg. 2000. Microbial genome sequencing. *Nature* 406: 799-803.

Fraser, C. M., J. A. Eisen, K. E. Nelson, I. T. Paulsen, and S. L. Salzberg. 2002. The value of complete microbial genome sequencing (you get what you pay for). *Journal of Bacteriology* 184(23):6403-6405.

Fraser-Ligett, C. M. 2005. Insights on biology and evolution from microbial genome sequencing. *Genome Research* 15:1603-1610.

Fredericks, D. N., and D. A. Relman. 1996. Sequence-based identification of microbial pathogens: A reconsideration of Koch's postulates. *Clinical Microbiology Reviews* 9:18-33.

Frerichs, R. R., P. S. Keim, R. Barrais, and R. Piarroux. 2012. Nepalese origin of cholera epidemic in Haiti. *Clinical Microbiology and Infection* 18(6):E158-E163.

Gage, K. L., and M. Y. Kosoy. 2005. Natural history of plague: Perspectives from more than a century of research. *Annual Review of Entomology* 50:505-528.

Gardy, J. L., J. C. Johnston, S. J. Ho Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. Jones, F. S. Brinkman, R. C. Brunham, and P. Tang. 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New England Journal of Medicine* 364(8):730-739.

Gevers, D., F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F. L. Thompson, and J. Swings. 2005. Opinion: Re-evaluating prokaryotic species. *Nature Reviews Microbiology* 3(9):733-739.

Ghedin, E. 2012. *Session II: Characterizing Intra-host Influenza Virus Populations to Predict Emergence*. Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 12.

Ghedin, E., J. Laplante, J. DePasse, D. E. Wentworth, R. P. Santos, M. L. Lepow, J. Porter, K. Stellrecht, X. Lin, D. Operario, S. Griesemer, A. Fitch, R. A. Halpin, T. B. Stockwell, D. J. Spiro, E. C. Holmes, K. St George. 2011. Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. *Journal of Infectious Diseases* 203(2):168-174.

Ghedin, E., E. C. Holmes, J. V. Depasse, L. T. Pinilla, A. Fitch, M. E. Hamelin, J. Papenburg, G. Boivin. 2012. Presence of oseltamivir-resistant pandemic A/H1N1 minor variants before drug therapy with subsequent selection and transmission. *Journal of Infectious Diseases* 206(10): 1504-1511.

Gilbert, J. 2012. *Session I: The Earth Microbiome Project: Modeling the Earth's Microbiome*. Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 12.

Gilbert, J. A., and C. L. DuPont. 2011. Microbial metagenomics: Beyond the genome. *Annual Review of Marine Science* 3:347-371.

Gilbert, J. A., R. O'Dor, N. King, and T. Vogel. 2011. The importance of metagenomic surveys to microbial ecology: Or why Darwin would have been a metagenomic scientist. *Microbial Informatics and Experimentation* 1:5.

Girard, J. M., D. M. Wagner, A. J. Vogler, C. Keys, C. J. Allender, L. C. Drickamer, and P. Keim. 2004. Differential plague-transmission dynamics determine *Yersinia pestis* population genetic structure on local, regional, and global scales. *Proceedings of the National Academy of Science USA* 101(22):8408-8413.

Girguis, P. 2012. *Session I: Population Diversity in Deep-Sea Microbial Communities*. Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 12.

Hacker, J., and J. D. Kaper. 2000. Pathogenicity islands and the evolution of microbes. *Annual Review of Microbiology* 54:641-679.

Haensch, S., R. Bianucci, M. Signoli, M. Rajerison, M. Schultz, S. Kacki, M. Vermunt, D. A. Weston, D. Hurst, M. Achtman, E. Carniel, and B. Bramanti. 2010. Distinct clones of *Yersinia pestis* caused the Black Death. *PLoS Pathogens* 6(10):e1001134.

Handelsman, J. 2004. Metagenomics. *Microbiology and Molecular Biology Reviews* 68(4):669-685.

Harris, S. R., E. J. Feil, M. T. G. Holden, M. A. Quail, E. K. Nickerson, N. Chantratita, S. Gardete, A. Tavares, N. Day, J. Lindsay, J. D. Edgeworth, H. de Lencastre, J. Parkhill, S. J. Peacock, and S. D. Bentley. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327:467-474.

Harris, S. R., E. J. P. Cartwright, M. E. Török, M. T. G. Holden, N. M. Brown, A. L. Ogilvy-Stuart, M. J. Ellington, M. A. Quail, S. D. Bentley, J. Parkhill, S. J. Peacock. 2012. Using whole genome sequencing to dissect the cause and effect of a meticillin-resistant *Staphylococcus aureus* outbreak: A descriptive study. *Lancet Infectious Diseases* (Epub ahead of print).

Hehemann, J. H., G. Correc, T. Barbeyron, W. Helbert, M. Czjzek, and G. Michel. 2010. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* 464(7290):908-912.

Hendriksen, R. S., L. B. Price, J. M. Shupp, J. D. Gillece, R. S. Kaas, D. M. Engelthaler, V. Bortolaia, T. Pearson, A. E. Waters, B. P. Upadhyay, S. D. Shrestha, S. Adhikari, G. Shakya, P. S. Keim, and F. M. Aarestrup. 2011. Population genetics of *Vibrio cholerae* from Nepal in 2010: Evidence on the origin of the Haitian outbreak. *mBio* 2(4):e00157-11.

Hornsey, M., N. Loman, D. W. Wareham, M. J. Ellington, M. J. Pallen, J. F. Turton, A. Underwood, T. Gaulton, T. P. Thomas, M. Doumith, D. M. Livermore, N. Woodford. 2011. Whole-genome comparison of two *Acinetobacter baumannii* isolates from a single patient, where resistance developed during tigecycline therapy. *Journal of Antimicrobial Chemotherapy* 66(7):1499-1503.

Hugenholtz, P. 2002. Exploring prokaryotic diversity in the genomic era. *Genome Biology* 3(2): 0003.1-0003.8.

Hugenholtz, P., and G. W. Tyson. 2008. Metagenomics. *Nature* 455:481-483.

The Human Microbiome Jumpstart Reference Strains Consortium. 2010. A catalog of reference genomes from the human microbiome. *Science* 328(5981):994-999.

Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320(5879):1081-1085.

Huse, S. 2012. *Session III: The Impact of Sequencing Errors on Estimates of Diversity in the Rare Biosphere (and Potential Solutions).* Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 13.

Huse, S. M., D. M. Welch, H. G. Morrison, and M. L. Sogin. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* 12(7):1889-1898.

Huse, S. M., Y. Ye, Y. Zhou, and A. A. Fodor. 2012. A core human microbiome as viewed through 16S rRNA sequence clusters. *PLoS One* 7(6):e34242.

Isenberg, H. D. 1988. Pathogenicity and virulence: Another view. *Clinical Microbiology Reviews* 1(1):40-53.

JASON. 2009. *Microbial forensics.* JSR-08-512.

———. 2010. *The $100 genome: Implications for the DoD.* JSR-11-320.

Joneson, S., J. E. Stajich, S. H. Shiu, E. B. Rosenblum. 2011. Genomic transition to pathogenicity in chytrid fungi. *PLoS Pathogens* 7(11):e1002338.

Kanehisa, M. 2002. The KEGG database. *Novartis Foundation symposium* 247:91-101; discussion 101-103, 119-128, 244-252.

Kaper, J. B., J. P. Nataro, and H. L. Mobley. 2004 Pathogenic *Escherichia coli*. *Nature Reviews Microbiology* 2(2):123-140.

Keegan, K. P., W. L. Trimble, J. Wilkening, A. Wilke, T. Harrison, M. D'Souza, and F. Meyer. 2012. A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. *PLoS Computational Biology* 8(6):e1002541.

Keim, P., A. Kalif, J. Schupp, K. Hill, S. E. Travis, K. Richmond, D. M. Adair, M. Hugh-Jones, C. R. Kuske, and P. Jackson. 1997. Molecular evolution and diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers. *Journal of Bacteriology* 179:818-824.

Knight, R., J. Jansson, D. Field, N. Fierer, N. Desai, J. Fuhrman, P. Hugenholtz, F. Meyer, R. Stevens, M. Bailey, J. I. Gordon, G. Kowalchuk, and J. A. Gilbert. 2012. Unlocking the potential of metagenomics through replicated experimental design. *Nature Biotechnology* 30(6):513-520.

Koch, R. 1891. Uber bakteriologische Forschung Verhandlung des X Internationalen Medichinischen Congresses, Berlin, 1890, 1, 35. August Hirschwald, Berlin. (In German.) Xth International Congress of Medicine, Berlin.

Kondrashov, A. S., and M. V. Mina. 1986. Sympatric speciation: When is it possible? *Biological Journal of the Linnean Society* 27:201-223.

Krediet, C., K. B. Ritchie, A. Alagely, M. Teplitski. (*In Press*). Coral commensal bacterial interference of metabolism and surface motility in a white pox pathogen during early colonization of coral surfaces. *ISME Journal*.

Kumarasamy, K. K., M. A. Toleman, T. R. Walsh, J. Bagaria, F. Butt, R. Balakrishnan, U. Chaudhary, M. Doumith, C. G. Giske, S. Irfan, P. Krishnan, A. V. Kumar, S. Maharjan, S. Mushtaq, T. Noorie, D. L. Paterson, A. Pearson, C. Perry, R. Pike, B. Rao, U. Ray, J. B. Sarma, M. Sharma, E. Sheridan, M. A. Thirunarayan, J. Turton, S. Upadhyay, M. Warner, W. Welfare, D. M. Livermore, N. Woodford. 2010. Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: A molecular, biological, and epidemiological study. *Lancet Infectious Diseases* 10(9):597-602.

La Rosa, P. S., J. P. Brooks, E. Deych, E. L. Boone, D. J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, and W. D. Shannon. 2012. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE* 7(12):e52078.

Larsen, P. E., F. R. Collart, D. Field, F. Meyer, K. P. Keegan, C. S. Henry, J. McGrath, J. Quinn, and J. A. Gilbert. 2011. Predicted relative metabolomic turnover (PRMT): Determining metabolic turnover from a coastal marine metagenomic dataset. *Microbial Informatics and Experimentation* 1(1):4.

Larsen, P. E., D. Field, and J. A. Gilbert. 2012. Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods* 9(6):621-625.

Lederberg. J. 2000. Infectious history. *Science* 288(5464):287-293.

Lewis, K. 2011. Presentation given at the March 14-15, 2011, public workshop, "Synthetic and Systems Biology," Forum on Microbial Threats, Institute of Medicine, Washington, DC.

Lewis, T., N. J. Loman, L. Bingle, P. Jumaa, G. M. Weinstock, D. Mortiboy, and M. J. Pallen. 2010. High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. *Journal of Hospital Infection* 75(1):37-41.

Ley, R. E., C. A. Lozupone, M. Hamady, R. Knight, and J. I. Gordon. 2008. Worlds within worlds: Evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology* 6:776-788.

Link, V. B. 1955. A history of plague in United States of America. *Public Health Monographs* 26:1-120.

Lipkin, W. I. 2010. Microbe hunting. *Microbiology and Molecular Biology Reviews* 74(3):363-377.

Lips, K. R., F. Brem, R. Brenes, J. D. Reeve, R. A. Alford, J. Voyles, C. Carey, L. Livo, A. P. Pessier, and J. P. Collins. 2006. Emerging infectious disease and the loss of biodiversity in a neotropical amphibian community. *Proceedings of the National Academy of Sciences USA* 103:3165-3170.

Lipsitch, M., C. Colijn, T. Cohen, W. P. Hanage, and C. Fraser. 2009. No coexistence for free: Neutral null models for multistrain pathogens. *Epidemics* 1(1):2-13.

Loman, N. J., C. Constantinidou, J. Z. Chan, M. Halachev, M. Sergeant, C. W. Penn, E. R. Robinson, and M. J. Pallen. 2012a. High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology* 10(9):599-606.

Loman, N. J., R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen. 2012b. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* 30(5):434-439.

MacLean, D. J., D. G. Jones, and D. J. Studholme. 2009. Application of "next-generation" sequencing technologies to microbial genetics. *Nature Reviews Microbiology* 7:287-296.

Mao-Jones, J., K. B. Ritchie, L. E. Jones, and S. P. Ellner. 2010. How microbial community composition regulates coral disease development. *PLoS Biology* 8(3):e1000345.

Mardis, E. R. 2008. Next-generation sequencing methods. *Annual Reviews Genomics and Human Genetics* 9:387-402.

———. 2011. A decade's perspective on DNA sequencing technology. *Nature* 470:198-203.

McClelland, E. E., P. Bernhardt, and A. Casadevall. 2006. Estimating the relative contributions of virulence factors for pathogenic microbes. *Infection and Immunity* 74(3):1500-1504.

McDaniel, L. D., E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, and J. H. Paul. 2010. High frequency of horizontal gene transfer in the oceans. *Science* 330(6000):50.

McDaniel, L. D., E. C. Young, K. B. Ritchie, and J. H. Paul. 2012. Environmental factors influencing gene transfer agent (GTA) mediated transduction in the subtropical ocean. *PLoS ONE* 7(8):e43506.

McFall-Ngai, M., E. A. Heath-Heckman, A. A. Gillette, S. M. Peyer, and E. A. Harvie. 2012. The secret languages of coevolved symbioses: Insights from the *Euprymna scolopes-Vibrio fischeri* symbiosis. *Seminars in Immunology* 24:3-8.

Medini, D., C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli. 2005. The microbial pan-genome. *Current Opinion in Genetics and Development* 15:589-594.

Medini, D., D. Serruto, J. Parkhill, D. A. Relman, C. Donati, R. Moxon, S. Falkow, and R. Rappuoli. 2008. Microbiology in the post-genomic era. *Nature Reviews Microbiology* 6:419-430.

Merrell, D. S., and S. Falkow. 2004. Frontal and stealth attack strategies in microbial pathogenesis. *Nature* 430(6996):250-256.

Metzker, M. L. 2010. Sequencing technologies—The next generation. *Nature Reviews Genetics* 11:31-45.

Meyer, F. 2012. *Session IV: Analyzing Metagenomic Data: Inferring Microbial Community Function with MG-RAST*. Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 13.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. 2008. The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386.

Morelli, G., Y. Song, C. J. Mazzoni, M. Eppinger, P. Roumagnac, D. M. Wagner, M. Feldkamp, B. Kusecek, A. J. Vogler, Y. Li, Y. Cui, N. R. Thomson, T. Jombart, R. Leblois, P. Lichtner, L. Rahalison, J. M. Petersen, F. Balloux, P. Keim, T. Wirth, J. Ravel, R. Yang, E. Carniel, and M. Achtman. 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nature Genetics* 42(12):1140-1143.

Mutreja, A., D. W. Kim, N. R. Thomson, T. R. Connor, J. H. Lee, S. Kariuki, N. J. Croucher, S. Y. Choi, S. R. Harris, M. Lebens, S. K. Niyogi, E. J. Kim, T. Ramamurthy, J. Chun, J. L. Wood, J. D. Clemens, C. Czerkinsky, G. B. Nair, J. Holmgren, J. Parkhill, and G. Dougan. 2011. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477(7365):462-465.

Nee, S. 2004. More than meets the eye. *Nature* 429:804-805.

Nichol, S. T., C. F. Spiropoulou, S. Morzunov, P. E. Rollin, T. G. Ksiazek, H. Feldmann, A. Sanchez, J. Childs, S. Zaki, and C. J. Peters. 1993. Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness. *Science* 262:914-917.

Otto, T. D. 2011. Real-time sequencing. *Nature Reviews Microbiology* 9:633.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, N. Diaz, H. Y. Chuang, M. Cohoon, V. de Crécy-Lagard, T. Disz, R. Edwards, et al. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* 33(17):5691-5702.

Pace, N. R. 1997. A molecular view of microbial diversity and the biosphere. *Science* 276:734-740.

Pace, N. R., D. A. Stahl, D. J. Lane, and G. J. Olsen. 1985. The analysis of natural microbial populations by ribosomal RNA sequences. *American Society for Microbiology News* 51:4-12.

Pallen, M. J., and B. W. Wren. 2007. Bacterial pathogenomics. *Nature* 449:835-842.

Paul, J. H., E. Young, L. McDaniel, K. B. Ritchie, C. Voolstra. (*In Review*). Gene transfer agent mediated horizontal gene transfer in the marine environment. *Proceedings of the National Academy of Sciences USA.*

Pennisi, E. 2012. Search for pore-fection. *Science* 336:534-537.

Perry, R. D., and J. D. Fetherston. 1997. *Yersinia pestis*—Etiologic agent of plague. *Clinical Microbiology Reviews* 10(1):35-66.

Piarroux, R., R. Barrais, B. Faucher, R. Haus, M. Piarroux, J. Gaudart, R. Magloire, and D. Raoult. 2011. Understanding the cholera epidemic, Haiti. *Emerging Infectious Diseases* 17(7):1161-1168.

Pollitzer, R. 1951. Plague studies. 1. A summary of the history and survey of the present distribution of the disease. *Bulletin of the World Health Organization* 4(4):475-533.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, MetaHIT Consortium, P. Bork, S. D. Ehrlich, and J. Wang J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59-65.

Rasko, D. 2012. *Session II: Comparative Genomics of E. coli and Shigella: Identification and Characterization of Pathogenic Variants Based on Whole Genome Sequence Analysis.* Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 12.

Rasko, D. A., P. L. Worshamb, T. G. Abshireb, S. T. Stanley, J. D. Bannand, M. R. Wilsond, R. J. Langhamc, R. S. Deckerc, L. Jianga,, T. D. Reade, A. M. Phillippy, S. L. Salzberg, M. Popf, M. N. Van Ertg, L. J. Kenefic, P. S. Keim, C. M. Fraser-Liggett, and J. Ravel. 2011. *Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation. *Proceedings of the National Academy of Sciences USA* 108(12):5027-5032.

Read, T. D., S. L. Salzberg, M. Pop, M. Shumway, L. Umayam, L. Jiang, E. Holtzapple, J. D. Busch, K. L. Smith, J. M. Schupp, D. Solomon, P. Keim, and C. M. Fraser. 2002. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* 296:2028-2033.

Relman, D. A. 1993. The identification of uncultured microbial pathogens. *Journal of Infectious Diseases* 168:1-8.

_____. 1999. The search for unrecognized pathogens. *Science* 284:1308-1310.

_____. 2011. Microbial genomics and infectious diseases. *New England Journal of Medicine* 365(4):347-357.

_____. 2012. Day 2 Welcoming Remarks and Summary of Day One. Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 13.

Relman, D. A., J. S. Loutit, T. M. Schmidt, S. Falkow, and L. S. Tompkins. 1990. The agent of bacillary angiomatosis: An approach to the identification of uncultured pathogens. *New England Journal of Medicine* 323:1573-1580.

Relman, D. A., T. M. Schmidt, R. P. MacDermott, and S. Falkow. 1992. Identification of the uncultured bacillus of Whipple's disease. *New England Journal of Medicine* 327:293-301.

Ritchie, K. B. 2006. Regulation of microbial populations by coral surface mucus and mucus-associated bacteria. *Marine Ecology Progress Series* 322:1-14.

_____. 2011. Bacterial symbionts of corals and symbiodinium. In: *Beneficial microorganisms in multicellular life forms,* E. Rosenberg and U. Gophna (Eds.). Berlin: Springer-Verlag Chapter 9, pp. 139-150.

Rivers, T. M. 1937. Viruses and Koch's postulates. *Journal of Bacteriology* 33:1-12.

Rohde, H., J. Qin, Y. Cui, D. Li, N. J. Loman, M. Hentschke, W. Chen, F. Pu, Y. Peng, J. Li, F. Xi, S. Li, Y. Li, Z. Zhang, X. Yang, M. Zhao, P. Wang, Y. Guan, Z. Cen, X. Zhao, M. Christner, R. Kobbe, S. Loos, J. Oh, L. Yang, A. Danchin, G. F. Gao, Y. Song, Y. Li, H. Yang, J. Wang, J. Xu, M. J. Pallen, J. Wang, M. Aepfelbacher, R. Yang. 2011. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *New England Journal of Medicine* 365(8):718-724.

Rosenblum, E. B. 2012. *Session II: Evolution and Pathogenicity in the Deadly Chytrid Pathogen of Amphibians*. Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 12.

Rosenblum, E. B., J. E. Stajich, N. Maddox, and M. B. Eisen. 2008. Global gene expression profiles for life stages of the deadly amphibian pathogen *Batrachochytrium dendrobatidis*. *Proceedings of the National Academy of Sciences USA* 105(44):17034-17039.

Rosenblum, E. B., T. J. Poorten, M. Settles, G. K. Murdoch, J. Robert, N. Maddox, and M. B. Eisen. 2009. Genome-wide transcriptional response of *Silurana (Xenopus) tropicalis* to infection with the deadly chytrid fungus. *PLoS ONE* 4(8):e6494.

Rosenblum, E. B., T. J. Poorten, S. Joneson, and M. Settles. 2012. Substrate-Specific Gene Expression in *Batrachochytrium dendrobatidis*, the Chytrid Pathogen of Amphibians. *PLoS One* 7(11): e49924.

Rosenblum, E. B., T. J. Poorten, M. Settles, and G. K. Murdoch. 2012. Only skin deep: Shared genetic response to the deadly chytrid fungus in susceptible frog species. *Molecular Ecology* 21(13):3110-3120.

Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter. 2007. The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* 5:e77.

Santamaria-Fries, M., L. F. Fajardo, M. L. Sogin, P. Olson, and D. A. Relman. 1996. Lethal infection by a previously unclassified metazoan parasite. *Lancet* 347:1797-1801.

Scheutz, F., E. M. Nielsen, J. Frimodt-Møller, N. Boisen, S. Morabito, R. Tozzoli, J. P. Nataro, and A. Caprioli. 2011. Characteristics of the enteroaggregtive Shiga toxin/verotoxin-producing *Escherichia coli* O104:H4 strain causing the outbreak of haemolytic uraemic syndrome in Germany, May to June 2011. *Eurosurveillance* 16(24):pii=19889.

Schloissnig, S., M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D. R. Mende, J. R. Kultima, J. Martin, K. Kota, S. R. Sunyaev, G. M. Weinstock, and P. Bork. 2013. Genomic variation landscape of the human gut microbiome. *Nature* 493(7430):45-50.

Schrenk, M. O., D. S. Kelley, J. R. Delaney, and J. A. Baross. 2003. Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. *Applied and Environmental Microbiology* 69(9):3580-3592.

Shapiro, B. J., J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabó, M. F. Polz, and E. J. Alm. 2012. Population genomics of early events in the ecological differentiation of bacteria. *Science* 336(6077):48-51.

Sharp, K. H., and K. B. Ritchie. 2012. Multi-partner interactions in corals in the face of climate change. *Biological Bulletin* 223:66-77.

Sharp, K. H., K. B. Ritchie, P. J. Schupp, R. Ritson-Williams, and V. J. Paul. 2010. Bacterial acquisition in juveniles of several broadcast spawning coral species. *PLoS One* 5(5):e10898.

Smillie, C. S., M. B. Smith, J. Friedman, O. X. Cordero, L. A. David, and E. J. Alm. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241-244.

Sylvan, J. B., B. M. Toner, and K. J. Edwards. 2012. Life and death of deep-sea vents: Bacterial diversity and ecosystem succession on inactive hydrothermal sulfides. *mBio* 3(1):e00279-11.

Takai, K., F. Inagaki, S. Nakagawa, H. Hirayama, T. Nunoura, Y. Sako, K. H. Nealson, and K. Horikoshi. 2003. Isolation and phylogenetic diversity of members of previously uncultivated ε-proteobacteria in deep-sea hydrothermal fields. *FEMS Microbiology Letters* 218(1):167-174.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, et al. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41, Epub.

Teplitski, M., and K. B. Ritchie. 2009. How feasible is the biological control of coral disease? *Trends in Ecology and Evolution* 24(7):378-385

Thomas, P. D., V. Wood, C. J. Mungall, S. E. Lewis, J. A. Blake; Gene Ontology Consortium. 2012. On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: A short report. *PLoS Computational Biology*8(2):e1002386

Trimble, W. L., K. P. Keegan, M. D'Souza, A. Wilke, J. Wilkening, J. Gilbert, and F. Meyer. 2012. Short-read reading-frame predictors are not created equal: Sequence error causes loss of signal. *BMC Bioinformatics* 13(1):183.

Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. 2008. Comparative metagenomics of microbial communities. *Science* 308:554-557.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. 2007. The human microbiome project. *Nature* 449:804-810.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. R. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37-43.

VandeWalle, J. L., G. W. Goetz, S. M. Huse, H. G. Morrison, M. L. Sogin, R. G. Hoffmann, K. Yan, and S. L., McLellan. 2012. *Acinetobacter*, *Aeromonas* and *Trichococcus* populations dominate the microbial community within urban sewer infrastructure. *Environmental Microbiology* 14(9): 2538-2552.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parson, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66-74.

Vogler, A. J., F. Chan, D. M. Wagner, P. Roumagnac, L. Lee, R. Nera, M. Eppinger, J. Ravel, L. Rahalison, B. W. Rasoamanana, S. M. Beckstrom-Sternberg, M. Achtman, S. Chanteau, and P. Keim. 2011. Phylogeography and molecular epidemiology of *Yersinia pestis* in Madagascar. *PLoS Neglected Tropical Diseases* 5(9):e1319.

Weinstock, G. 2012a. Genomic approaches to studying the human microbiota. *Nature* 489:13.

_____. 2012b. *Session I: Variation in Microbial Communities and Genomes*. Paper presented at the Forum on Microbial Threats Workshop, The Science and Applications of Microbial Genomics, Washington, DC, Institute of Medicine, Forum on Microbial Threats, June 12.

White, B. J., C. Cheng, F. Simard, C. Costantini, and N. J. Besansky. 2010. Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular Ecology* 19(5):925-939.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences USA* 95(12):6578-6583.

Woyke, T., D. Tighe, K. Mavromatis, A. Clum, A. Copeland, W. Schackwitz, A. Lapidus, D. Wu, J. P. McCutcheon, B. R. McDonald, N. A. Moran, J. Bristow, J-F. Cheng. 2010. One bacterial cell, one complete genome. *PLoS One* 5(4):e10314.

Wu, D., P. Hugenhotz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J. F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H. P. Klenk, and J. A. Eisen. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056-1060.

Wylie, K. M., R. M. Truty, T. J. Sharpton, K. A. Mihindukulasuriya, Y. Zho, H. Gao, E. Sodergren, G. M. Weinstock, and K. S. Pollard. 2012a. Novel bacterial taxa in the human microbiome. *PLoS One* 7(6):e35294.

Wylie, K. M., K. A. Mihindukulasuriya, E. Sodergren, G. M. Weinstock, and G. A. Storch. 2012b. Sequence analysis of the human virome in febrile and afebrile children. *PLoS ONE* 7(6):e27735.

Zhou, J., L. Wu, Y. Deng, X. Zhi, Y. H. Jiang, Q. Tu, J. Xie, J. D. Van Nostrand, Z. He, and Y. Yang. 2011. Reproducibility and quantitation of amplicon sequencing-based detection. *ISME Journal* 5(8):1303-1313.

Zhou, Y., H. Gao, K. A. Mihindukulasuriya, P. S. Rosa, K. M. Wylie, T. Vishnivetskaya, M. Podar, B. Warner, P. I. Tarr, D. E. Nelson, J. D. Fortenberry, M. J. Holland, S. E. Burr, W. D. Shannon, E. Sodergren, and G. M. Weinstock. 2013. Biogeography of the ecosystems of the healthy human body. *Genome Biology* 14(1):R1.

Appendix A

A1

THE MICROBIAL FORENSICS PATHWAY FOR USE OF MASSIVELY PARALLEL SEQUENCING TECHNOLOGIES

Bruce ***Budowle***,[1,2] *Sarah E. Schmedes*,[1,2] *and Randall S. Murch*[1,3]

The Challenge

Eliminating the threat of terrorist or criminal attacks with microorganisms or toxin weapons is a continual challenge for biodefense and biosecurity programs. The task is difficult for several reasons: (1) the relative ease of access to a variety of effective source materials (Srivatsan et al., 2008) and options for the delivery of a bioweapon, (2) the minute quantities of materials that can be transferred and yet still be effective, (3) the difficulties in detection and analysis of microbiological evidence, and (4) the lack of well-defined approaches regarding credible inferences that can be made from microbial forensic evidence given extant data. At the onset of an event, it may be difficult to distinguish between a deliberate attack and a naturally occurring outbreak of an infectious disease (Morse and

[1] Institute of Applied Genetics, University of North Texas Health Science Center, Fort Worth, TX.
[2] Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, Fort Worth, TX.
[3] Virginia Tech, National Capital Region, Arlington, VA.

Budowle, 2006; Morse and Khan, 2005). Even if evidence strongly supports the hypothesis of a deliberate attack, it may still be very difficult to attribute the attack with certainty to those responsible (i.e., attribution). Attempts to resolve the crime will require advanced methods for characterizing microbial agents, as well as a combination of traditional investigation and intelligence gathering activities.

The Approach

In response to the need to determine the nature of the threat and the source of the weapon and to identify those who perpetrated the crime, the scientific community rose to the occasion beginning in 1996 and developed the field of microbial forensics. Microbial forensics is the scientific discipline dedicated to analyzing evidence from a bioterrorism act, biocrime, hoax, or inadvertent microorganism/toxin release for attribution purposes (Budowle et al., 2003, 2005a; Köser et al., 2012; Morse and Budowle, 2006). Another goal can be to support analysis of potential bioweapons capabilities for counter-proliferation, treaty verification, and/or interdiction. A forensics investigation initially will attempt to determine the identity of the causal agent and/or source of the bioweapon in much the same manner as in an epidemiological investigation. The epidemiological concerns are identification and characterization of specific disease-causing pathogens or their toxins, their modes of transmission, and any manipulations that may have been performed intentionally to increase their effects against human, animal, or plant targets (Morse and Budowle, 2006; Morse and Khan, 2005). A microbial forensics investigation proceeds further in that evidence is characterized to assist in determining the specific source of the sample, as individualizing as possible, and the methods, means, processes, and locations involved to determine the identity of the perpetrator(s) of the attack or to determine that an act is in preparation. A systems analysis may be able to determine the processes used to generate the weapon or how it was delivered, which also can help inform the investigation and attribution decision. The ultimate goal is attribution—to identify the perpetrator(s) or to reduce the potential perpetrator population to as few individuals as possible so investigative and intelligence methods can be effectively and efficiently applied to "build the case" (Figure A1-1).

Forensic Targets

Microbial forensic evidence may include the microbe, toxin, nucleic acids, protein signatures, inadvertent microbial contaminants, stabilizers, additives, dispersal devices, and indications of the methods used in a preparation. In addition, traditional types of forensic evidence may be informative and should be part of the toolbox of potential analyses of evidence from an act of bioterrorism or biocrime. Traditional evidence includes fingerprints, body fluids and tissues, hair, fibers, documents, photos, digital evidence, videos, firearms, glass, metals, plastics, paint, powders, explosives, tool marks, and soil. Other types of relevant

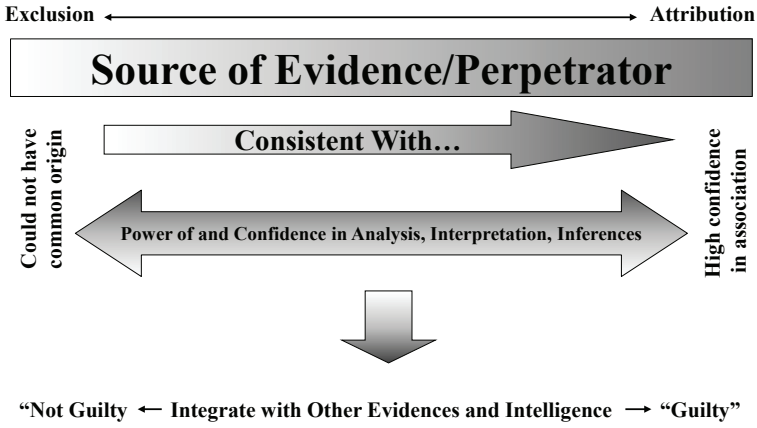


FIGURE A1-1 The microbial forensics attribution continuum.

evidence must be considered to exploit avenues to better achieve attribution, including proteins and chemical signatures. These types of signatures can only be obtained from crimes where the weaponized material or delivery device is found; they have little use in covert attacks where the biological agent is derived from the victims. Many of these methods are based on sound technologies and are complementary. They can be combined to identify signatures of sample growth, processing, and chronometry (Morse and Budowle, 2006). Matching of sample properties can help to establish the relatedness of disparate incidents. Furthermore, mismatches might have exclusionary power or signify a more complex causal relationship between the events under investigation. The results of these analyses can provide information on how, when, and/or where microorganisms were grown and weaponized. While the goal of a microbial forensic analysis is to characterize a sample such that it can be traced to a unique source or at least eliminate other sources, it is unlikely that microbial forensic evidence alone is currently adequate to meet this goal.

Emerging Science and Technology

To enhance attribution capabilities with microbial evidence, considerable attention is being invested in molecular genetics, genomics, and bioinformatics. These fields are essential to microbial species/strain identification, fine genome variation, virulence determination, pathogenicity characterization, possible genetic engineering, and attaining source attribution to the highest degree possible. The various tools that have been, or are being, developed in these areas will help to narrow the potential sources from which the pathogen used in an attack may have originated. Indeed, sequencing of an entire genome has been demonstrated

as feasible in epidemiological investigations, such as the recent studies of outbreaks of *E. coli* O104:H4 in Germany and cholera in Haiti (Brzuszkiewicz et al., 2008; Chin et al., 2011; Grad et al., 2012; Hasan et al., 2012; Hendriksen et al., 2011; Mellmann et al., 2011; Rasko et al., 2011; Rohde et al., 2011). In addition, metagenomics studies may become foundational on describing diversity and endemicity. Endemicity becomes important when the relationship between microbes or their genetic residues in samples collected from a site of interest and microbes in the environmental background need to be defined. While the inferential capacity of microbial forensics genetics has yet to reach its full power, the phenomenal new generations of sequencing technology and the concomitant developments for bioinformatics capabilities to handle and extract the explosion of data offer potentials for enhancing microbial forensic investigations. Indeed, the science and technology supporting microbial forensics are advancing at an inconceivable rate. For example, in 2002 in response to the anthrax letter attack, whole genomes of a few isolates were sequenced using shotgun sequencing by TIGR (Budowle et al., 2005b; NRC, 2009; Ravel et al., 2009; Read et al., 2002, 2003). That seemingly nominal analysis, by today's capabilities, cost approximately $250,000 for one genome, took several weeks, and was unable to characterize but a few samples. Today, such enterprises are a fraction of the cost (and continue to drop dramatically), are becoming more automatable, and provide gigabases and terabytes of data in a matter of days (Bentley et al., 2008; Holt et al., 2008; Loman et al., 2012; MacLean et al., 2009; Margulies et al., 2005).

Given the enhanced capabilities of nucleic acid sequencing of microbes the microbial forensics community will embrace these molecular tools. Although developments are needed, one can envision identification of microbes at the species, strain, and isolate levels being transformed using next- (or better termed "current-") generation sequencing (CGS). Fine genome detail could become available for routine microbial forensic use. Because CGS provides whole genome characterization capabilities with high depths of coverage (100s to 1,000s fold and beyond), the technology will serve a critical role for research, such as genetic diversity and endemicity studies via metagenomics, and become a rapid diagnostic tool initially when viable and culturable microbes are available. Indeed, whole genome sequencing will reduce the need for *a priori* design of assays directed at defined species. The technology should apply at some resolution level to any genome without knowledge of the target. In addition, whole genome sequencing offers the capability to evaluate a sample for indications of genetic engineering.

Current Realities

However, not all microbial forensic evidence will present itself in a manner where copious quantities of target are available. Some samples will be highly degraded and/or contaminated. Thus, there will be challenges to extract the most

information possible from limited materials and non-viable organisms. To meet these challenges, improved sample collection and extraction methods will be needed, nucleic acid repair methods will be sought, target amplification strategies such as whole genome amplification and selective target capture will be sought, and sequencing chemistries will be enhanced. Because of the throughput, CGS technologies can analyze multiple samples and not even begin to exploit the full throughput of the systems (Brzuszkiewicz et al., 2011; Cummings et al., 2012; Eisen, 2007; Hasan et al., 2012; Holt et al., 2008; Howden et al., 2011; Loman et al., 2012; MacLean et al., 2009; Relman, 2011; Rohde et al., 2011). However, the technology still is evolving and currently does not offer the sensitivity of detection to analyze low-quantity and low-quality DNA samples without some amplification approach prior to sequencing. Nonetheless, CGS is sufficiently mature to be considered useful for microbial forensic applications. Alternatively, technologies, such as mass spectrometry analyses of nucleic acids and real-time PCR, will continue to be used because they offer rapid detection (at species and strain levels) at substantially lower costs (Jacob et al., 2012; Kenefic et al., 2008; Sampath et al., 2005, 2009; U'ren et al., 2005; Vogler et al., 2008).

There are a number of CGS instruments and different chemistries. They include Miseq® System and Hiseq™ Sequencing Systems (Illumina, Inc., San Diego, CA), Ion Personal Genome Machine™ (PGM™) Sequencer, Ion Proton™ Sequencer and SOLiD® Systems (Life Technologies, Foster City, CA), and the 454 Genome Sequencer FLX and GS Junior Systems (Roche Diagnostics Corporation, Indianapolis, IN) (Bentley et al., 2008; Cummings et al., 2010; Loman et al., 2012; Margulies et al., 2005). In addition, single molecule detection platforms, such those from Pacific Bioscience (Chin et al., 2011; Eid et al., 2009) and possibly Oxford Nanopore (Branton et al., 2008) are on the horizon. Each system offers some advantages and limitations for sequencing that will need to be defined with considerations of library preparation, read length, and accuracy. The evaluations should be based on the needs of application-oriented laboratories and not necessarily those of a research laboratory. Initially, microbial forensics instruments will be maintained in controlled laboratory environments.

Library preparation is one of the critical limiting factors for transferring CGS technology from a research environment to that of an operational laboratory. Currently, only a few samples can be prepared at any given time. Thus, while the sequencing throughput of the platforms is high, a sufficient number of samples cannot be readily prepared in an appropriate amount of time to meet the full capacity of the system. Library preparation needs to be simplified. Haloplex (Agilent, Santa Clara, CA) is an example of a library preparation process that potentially can reduce the preparation work required (www.halogenomics.com). This library preparation approach is a single-tube target amplification methodology that enables a large number of library samples to be prepared manually. The general process is: (1) restriction digest and denature the sample; (2) hybridize probes to targeted ends of the digested fragments; (3) circularize and ligate the

molecules; and (4) introduce bar codes and amplify the targets by polymerase chain reaction (PCR). Eventually with automation the process might accommodate the number of samples that may be encountered by high-throughput operational laboratories. As many as 96 bar codes are available, which fits well with the 96-well format and reduces the preparation time from 2 weeks or several days to 6 hours. However, currently Haloplex is not available for use with non-human nucleic acids. One constraint is that the Haloplex system employs restriction digestion of the DNA. The restriction enzymes can potentially cleave a target site of interest (either a single nucleotide polymorphism (SNP) site or within a repeat motif) and render the marker untypable. Unfortunately, the enzymes used in Haloplex are proprietary, and one cannot readily scan for the restriction sites that would be incompatible with the designated targets (although palindromes can be sought for potential sites that may be obliterated). Another strategy for simplifying library preparation and decreasing sample input is that of the Nextera XT DNA Sample Preparation Kit (www.illumina.com). Strategies, such as the Haloplex system and the Nextera XT DNA kit, hold promise for simplifying and possibly automating library preparation.

Another factor to consider with CGS technology is sequencing read length and accuracy. Current read lengths for the most widely used CGS instruments typically do not exceed 200 bases, and when they do, the quality of base calling decreases substantially along the length of a read. Longer reads with higher accuracy are necessary. Advances in technology for some platform systems suggest that reads up to 400 bases will be feasible in 2012.

Another consideration of platform selection is for situations where rapid responses are required (such as in military operations, some pandemics, and bioterrorism acts). Initially, platforms will be placed in laboratories with controlled environments. One can envision the technology being taken to the field for immediate response and exigent circumstances. Robustness of the instrumentation, supply lines of reagents, and service support will be part of the decision process for the instrumentation/chemistry of choice. Fortunately, the technology and supporting interpretation tools continue to evolve and likely will become more robust.

Seeking More Power and Depth

For design and selection of systems and diagnostics, different diagnostic-based strategies can be considered. They can be based on the sample type, the sample matrix, the amount of work, or the question that one is attempting to address. The latter may be the best suited for conceiving workflow systems. The different scenarios should be considered where nucleic acid analyses may be applied, because these will help guide the needs for the microbial forensic community. They likely are (1) identification of species/strain (i.e., similar to epidemiological needs), (2) attribution, (3) genetic engineering, (4) sample-to-sample

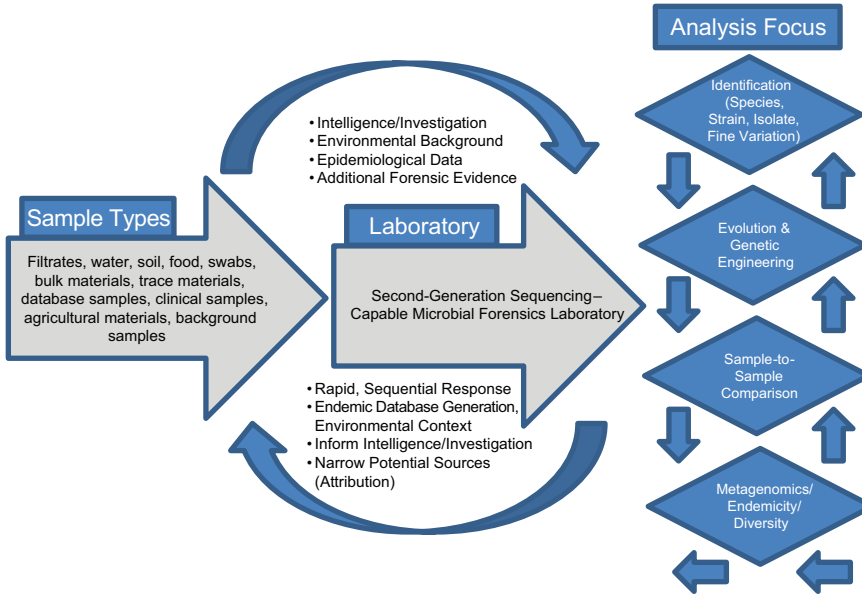


FIGURE A1-2 A general overview of the work and information flow from sample to analysis to information developed based on use of second-generation sequencing technology.

comparisons, and (5) metagenomics for endemicity (or a modified metagenomics for sample characterization) (Figure A1-2).

Sample identification generally would be direct characterization to identify the agent for immediate determination of potential threat and probable cause to investigate further. The process of attribution would drill down to the finest resolution possible and make comparisons to other reference samples, databases, or repositories to reduce the possible sources from which the sample originated or to a recent common ancestor. Genetic engineering could be detected by whole genome sequencing.

Metagenomics studies have been performed on several platforms, and they will likely provide some foundational data on diversity and endemicity (Eisen, 2007; Relman, 2011; Tringe et al., 2005). The value could be searching various niches for select agents. Suppose that in every sample tested certain select agents are identified. Then there can be two consequences: one is that it may be more difficult to elucidate natural outbreaks versus intentional releases (although strain resolution may reduce the uncertainty); the second could be that such high resolution may be less informative at some threshold depth of coverage.

Most metagenomic work to date has been by exploiting a small, single sequence target (16s rRNA), at a very high depth of coverage (Rusch et al., 2007; Venter et al., 2004). These studies often cannot provide resolution beyond family to genus levels. Clearly such broad range definition will not enable

individualization or identify select agents. The anthrax investigation could have benefited from a modified metagenomics characterization. The putative common source of the material (RMR1029) was composed of a population of very similar cells. The colony morphological variants found in the evidence from the 2001 anthrax letter attacks were minority components and because of sample preparation and stochastic effects the minor variants potentially could be difficult to detect with PCR-based assays that were developed for the investigation. Because of the high depth of coverage with CGS, the population of low-level variants may be more readily detected, especially if an amplification enrichment step was included that focused only on the known variant sites that defined the morphology types. Such high depth of coverage would substantially reduce the false-positive rate and improve confidence in the potential relationship of the most similar samples to focus investigative leads (Cummings et al., 2010). Indeed, the depth of coverage could be in the millions. While exquisitely sensitive, platform- and chemistry-specific errors may confound interpretation, and thus thresholds of reliability may be necessarily invoked.

One could envision extending this population depth analysis, which in essence is a simplified metagenomic analysis, and exploiting the concept of using a multi-locus sequence typing (MLST) approach to provide a species-level identification capability (Maiden et al., 1998; Spratt, 1999). A few loci (perhaps the seven typically applied to MLST to 15) could be selected as a standard (e.g., for bacteria). If there is a combination of sufficiently stable sites and evolutionarily rapid sites, the loci could indicate species- to strain-level presence in mixed and metagenomic samples. Using the core seven used for MLST could allow some questions regarding time and place of isolation, host or niche, serotype, and some clinical or drug resistance profiles. This will not be a trivial process because each of the sites will not be physically linked. However, one could determine, if the complete set or a reasonable subset of targets are in a sample, whether there is confidence that a particular species or sets of species are present. In theory this approach could be extended to strain levels. There certainly is enough throughput to consider this capability. The potential already has been established with electrospray ionization mass spectrometry of targeted genes for rapid bacterial species identification (and even for viruses such as influenza). There are sufficient bacterial genomes that have been sequenced to test our hypothesis, and work is under way.

Inferences about the significance of genetic evidence may not reach the ultimate goal of attribution. The most confounding constraint on reaching the full power of attribution is scant data on diversity and endemicity. The vast diversity of the microbial world is unknown and will not be defined substantially with current approaches in the area where a biocrime or bioterrorist attack has occurred. This limitation is not the sole purview of the microbial forensics community; it plagues the epidemiologists as well. Another limitation that evidentiary samples will likely have is an unknown history. Lack of knowledge on how it was manipulated (e.g., number of passages, exposure to mutagenic agents, length of

storage) will complicate providing inferences about the significance or strength of sequencing results, especially because the distance between samples will be determined by the degree of similarity or dissimilarity. Indeed, even defining what is a "match" or "similar" may not be straightforward. Keim (personal communication) has stressed this uncertainty and proffered new terminology—a "member," to the microbial forensic lexicon based on phylogenetics for the relationship of a sample to some reference samples. Regardless of the terminology used, some data will be needed to define the uncertainty of a "membership" or "association." In 2006, the need for reconciliation between microbial genomics and systematics was described; microbial forensics and epidemiology were seen to offer useful, practical venues to frame the gaps and priorities (Buckley and Roberts, 2006). This challenge remains.

Some assessment of the strength or significance of an analytical result and subsequent comparison also is needed (Budowle et al., 2008; Chakraborty and Budowle, 2011). Of course, because of scant supporting data, such an endeavor will be challenging. Qualitative and/or quantitative statements of the significance of the finding will need to be developed. As an example, consider a forensic analysis of whole genome sequence data that compared two or more sequences, such as an evidence sample profile with that of a reference sample that may be considered a possible direct link or have a common ancestor. The evolutionary rates of the variants will need to be known. But perhaps as consequential, sequencing error and other factors could inflate the dissimilarity between samples and add a degree of "uncertainty" to some extent. Thus, efforts in defining and quantifying the error rates associated with each CGS platform and chemistry are critically important.

Beyond comparison of samples for identification purposes are inferences by whole genome sequencing of phenotypic (i.e., functional) properties of a microbe. For example, even with a whole genome sequence whether a microbe phenotypically displays antimicrobial resistance or susceptibility is still limited. Bacteria may contain multiple pathways, and how the different genes interact is far from being completely understood (Eisen, 2007; Köser et al., 2012; Relman, 2011). Substantial research will be needed such that genotype can be used reliably to predict phenotype.

Making Sense of Data

The ever-increasing amount of microbial genomic sequence data presents a variety of challenges related to the handling and storage of data and the development of bioinformatics methods that can accommodate such large numbers of whole genomes. Being able to analyze the vast amounts of data in a timely fashion is a key challenge to leveraging the power of these newer sequencing platforms. Software, hardware, and IT support may be the greatest barrier to use of CGS technology. It is unlikely that dedicated bioinformaticists will reside in every microbial forensics laboratory. Data cannot be sent to web-based clouds

and be analyzed because the results may be classified. Instead, some standardization and standard operating data analysis and interpretation approaches will be needed. Pipeline and interpretation software will need to be evaluated for reliability and seamless diagnostic flow without bioinformatics expert intervention. The output of results must be intelligible to the microbial forensic analyst as well. The ideal software should be a comprehensive tool(s) enabling microbe detection to determination of engineering.

The government should rely heavily on industry and well-established genome centers. The commercial competitive environment is driving down costs and improving informatics pipelines without the need for extensive investment. Leveraging these efforts will help meet the needs of microbial forensics more expeditiously than going it alone. The centers (to include the national laboratories) are evaluating platforms and chemistries and are generating data at unprecedented levels. They are providing solutions to massive data handling, including storage, curation of reference data, annotation, and data analysis.

Collection and databases are needed to house the microbial genomic data and when possible the accompanying meta-data. No standards yet exist for building databases to meet the needs of the microbial forensic community. Requirements for storage and retrieval of raw sequence data in microbial forensics cases and supporting inferential data must be developed. Given the high throughput and anticipated speed of analyses, it is conceivable that meaningful databases can be developed "on the fly" that better reflect the diversity where the crime was committed (to include the preparation laboratory to the crime scene).

The power of microbial forensics techniques, tools, software, and databases that are used need to be understood, and their limitations even more so need to be understood. To achieve this goal methods need to be validated, and validation should be a requisite of any forensic repertoire. Indeed the forensic sciences in general are facing well-deserved criticism for not necessarily having sound foundations and overstating the strength of the evidence (NRC, 2009). Attempts to attribute any attack to a person(s) or group should rely on acurate and credible results. The interpretation of such results might seriously impact the course or focus of an investigation, thus affecting the liberties of individuals or even being used as a justification for a government's military response to an attack or threat of an attack. Therefore, the methods for collection, extraction, and analysis of microbial evidence that could generate key results need to be as scientifically robust as possible, so the methods can be high performing and the results defensible for decision makers and to the legal, international government, law enforcement, and scientific communities, as well as scrutiny by the media.

Validation Is Essential

Validation is frequently used to connote confidence in a test or process, but it may be better thought of as defining the limitation of a method, process, or assay (Budowle et al., 2003, 2006, 2008). It still is common for the term *validation*

to be used vaguely or to remain undefined when applied to process performance evaluation. The degree of validation varies from nominal to rigorous. The consequences of such varied requirements can be catastrophic if methods used in microbial forensic investigations are poorly constructed, under-developed, or generate results that are difficult to interpret. The validation process needs to be defined as to what is expected to be achieved by a validation study.

Validation determines the limits of a test. It does not mean that a test must be 100 percent accurate or have no cross-reactivity, false-positive results, or false-negative results to be considered useful. It is often thought of as a process applied to the analytical portion of a system. This concept is only partly correct. The limits that the methods can provide must be demonstrated and documented for all steps of the process to include sample collection, preservation, extraction, analytical characterization, and data interpretation. Furthermore, it is recognized that as new technologies and capabilities are developed to address the needs of the microbial forensics community, key principles and performance parameters including accuracy, precision, bias, reliability, sensitivity, and robustness will need to be determined. Robust quality assurance and data control systems are required to achieve confidence in results by diverse users of the information. It is imperative that both technical and interpretation limitations (and thus accuracy and error) be defined. Additionally, a key resource for microbial forensic research, validation, and analysis is access to well-defined and curated microbial collections and data sets that are as comprehensive as is possible to the task. This effort includes the structure, content, and quality of the data sets. While some collections have been started for use in research, or created for case-specific use, no comprehensive repository exists to support microbial forensics, and standards are not codified for meta-data and data curation.

The implications of highly technical data, epidemiological data, traditional evidence data, and investigative or intelligence information are complex and need to be appreciated for their strengths and limitations. Because scientific data can affect the decision-making process for retaliation, preemptive actions, and/or courtroom deliberations, it is imperative that those directly involved in microbial forensics or those who may use the results for investigative lead value or more direct associations be properly educated (or at least properly apprised) of the implications of such data. To meet this necessary goal, education and training are critical to disseminate the principles, development, and applications of the evolving field of microbial forensics. Educational strategies and programs need to be constructed and training programs developed on the varied scientific foundations that support microbial forensics.

If validation processes are not defined and not followed and proper training or communication is not provided, then it is possible that a false sense of confidence may be associated with a poor method or process or from a result of limited significance. There are myriad methods, processes, targets, platforms, and applications. Yet some basic requirements transcend individual differences in methods, and these can be reinforced by contextual description (Table A1-1).

Validation needs to be codified. Efforts are under way and should be applied equally across the user space.

Conclusion

Microbial forensics should embrace and validate newly developed and emerging molecular biology technologies and phylogenetics approaches, and pursue potential forensic information and comparative sources, such as might be achieved through metagenomics. Genetic analyses of microorganisms often are a powerful tool for differentiating species, isolates, and strains. Similar to human DNA forensic identification, DNA sequences of microorganisms can be used to identify and differentiate between isolates and strains of a single microbial species; however, nucleic acid–based identification is not as resolving with respect to source attribution in microbial forensics as with human DNA forensic analysis. The basic constituents of nucleic acids essentially are the same for bacteria and humans; however, unlike humans, bacteria, viruses, and fungi multiply rapidly in a clonal fashion and can readily share or exchange genetic material between and among species. These differences and uncertainties due to scant supporting data must be taken into consideration during analysis, interpretation, and reporting related to the findings derived from microbial genetic evidence. For the foreseeable future

TABLE A1-1 Validation Criteria List

- Sensitivity
- Specificity
- Reproducibility
- Precision
- Accuracy
- Resolution
- Reliability
- Robustness
- Specified samples
- Purity
- Input values
- Quantitation
- Dynamic range
- Limit of detection
- Controls
- Window of performance for operational steps of assay
- Critical equipment calibration
- Critical reagents
- Databases

NOTE: It is difficult to prescribe the criteria for validation of the variety of methods that may be considered. The list is provided for consideration and is not meant to be exhaustive.
SOURCE: Derived from Budowle et al. (2008).

the ability of microbial forensics to establish that a sample collected from either a crime scene or a person of interest can be attributed to a known source to a high degree of scientific certainty will be limited. Therefore, the methods must be reliable and robust, and the uncertainty associated with any interpretation should be properly conveyed.

Microbial forensics experts and those who contribute in closely related fields need to work together to advance the science, to validate methods to scientific and legal standards, and to transition interpretation of results and conclusions from such analyses into something that can be used by the criminal justice system, the policy community, and other stakeholders. It is incumbent upon the microbial forensics community to make every effort to interpret and communicate objectively and effectively the advantages and limitations of both microbial forensics and traditional forensic science analyses. Consumers of microbial forensic information who incorporate this evidence into decision making should be provided accurate, reliable, credible, and defensible results, interpretations, and context.

References

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. L. Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53-59.

Branton, D., D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss. 2008. The potential and challenges of nanopore sequencing. *Nature Biotechnology* 26:1146-1153.

Brzuszkiewicz, E., A. Thurmer, J. Schuldes, A. Leimbach, H. Liesegang, F. D. Meyer, J. Boelter, H. Petersen, G. Gottschalk, R. Daniel. 2011. Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-aggregative-haemorrhagic *Escherichia coli* (EAHEC). *Archives of Microbiology* 193:883-891.

Buckley, M., and R. J. Roberts. 2006. Reconciling microbial systematics and genomics. *Report of a Colloquium of the American Academy of Microbiology*, Washington, DC: ASM Press.

Budowle, B., S. E. Schutzer, A. Einseln, L. C. Kelley, A. C. Walsh, J. A. Smith, B. L. Marrone, J. Robertson, and J. Campos. 2003. Building microbial forensics as a response to bio-terrorism. *Science* 301:1852-1853.

Budowle, B., S. E. Schutzer, M. S. Ascher, R. M. Atlas, J. P. Burans, R. Chakraborty, J. J. Dunn, C. M. Fraser, D. R. Franz, T. J. Leighton, S. A. Morse, R. S. Murch, J. Ravel, D. L. Rock, T. R. Slezak, S. P. Velsko, A. C. Walsh, R. A. Walters. 2005a. Toward a system of microbial forensics: From sample collection to interpretation of evidence. *Applied and Environmental Microbiology* 71:2209-2213.

Budowle, B., M. D. Johnson, C. M. Fraser, T. J. Leighton, R. S. Murch, and R. Chakraborty. 2005b. Genetic analysis and attribution of microbial forensics evidence. *Critical Reviews in Microbiology* 31(4):233-254.

Budowle, B., S. E. Schutzer, J. P. Burans, D. J. Beecher, T. A. Cebula, R. Chakraborty, W. T. Cobb, J. Fletcher, M. L. Hale, R. B. Harris, M. A. Heitkamp, F. P. Keller, C. Kuske, J. E. LeClerc, B. L. Marrone, T. S. McKenna, S. A. Morse, L. L. Rodriguez, N. B. Valentine, and J. Yadev. 2006. Quality sample collection, handling, and preservation for an effective microbial forensics program. *Applied and Environmental Microbiology* 72(10):6431-6438.

Budowle, B., S. E. Schutzer, S. A. Morse, K. F. Martinez, R. Chakraborty, B. L. Marrone, S. L. Messenger, R. S. Murch, P. J. Jackson, P. Williamson, R. Harmon, and S. P. Velsko. 2008. Criteria for validation of methods in microbial forensics. *Applied and Environmental Microbiology* 74:5559-5607.

Chakraborty, R., and B. Budowle. 2011. Population genetic considerations in statistical interpretation of microbial forensic data in comparison with the human DNA forensic standard. In: *Microbial Forensics*, 2nd ed., edited by: B. Budowle, S. E. Schutzer, R. Breeze, P. S. Keim, and S. A. Morse. Amsterdam: Academic Press. Pp. 561-580.

Chin, C. S., J. Sorenson, J. B. Harris, W. P. Robins, R. C. Charles, R. R. Jean-Charles, J. Bullard, D. R. Webster, A. Kasarskis, P. Peluso, E. E. Paxinos, Y. Yamaichi, S. B. Calderwood, J. J. Mekalanos, E. E. Schadt, and M. K. Waldor. 2011. The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* 364:33-42.

Cummings, C. A., C. A. Bormann-Chung, R. Fang, M. Barker, P. Brzoska, P. C. Williamson, J. Beaudry, M. Matthews, J. Schupp, D. M. Wagner, D. Birdsell, A. J. Vogler, M. R. Furtado, P. Keim, and B. Budowle. 2010. Accurate, rapid, and high-throughput detection of strain-specific polymorphisms in *Bacillus anthracis* and *Yersinia pestis* by next-generation sequencing. *BMC Investigative Genetics* 1:5.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323:133-138.

Eisen, J. A. 2007. Environmental shotgun sequencing: Its potential and challenges for studying the hidden world of microbes. *PLoS Biology* 5(3): e82.

Grad, Y. H., M. Lipsitch, M. Feldgarden, H. M. Arachchi, G. C. Cerqueira, M. Fitzgerald, P. Godfrey, B. J. Haas, C. I. Murphy, C. Russ, S. Sykes, B. J. Walker, J. R. Wortman, S. Young, Q. Zeng, A. Abouelleil, J. Bochicchio, S. Chauvin, T. DeSmet, S. Gujja, C. McCowan, A. Montmayeur, S. Steelman, J. Frimodt-Møller, A. M. Petersen, C. Struve, K. A. Krogfelt, E. Bingen, F-X. Weill, E. S. Lander, C. Nusbaum, B. W. Birren, D. T. Hung, and W. P. Hanage. 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proceedings of the National Academy of Sciences USA* 109:3065-3070.

Hasan, N. A., S. Y. Choi, M. Eppinger, P. W. Clark, A. Chen, M. Alam, B. J. Haley, E. Taviani, E. Hine, Q. Su, L. J. Tallon, J. B. Prosper, K. Furth, M. M. Hog, H. Li, C. M. Fraser-Liggett, A. Cravioto, A. Hug, J. Ravel, T. A. Cebula, and R. R. Colwell. 2012. Genomic diversity of 2010 Haitian cholera outbreak strains. *Proceedings of the National Academy of Sciences USA* 109(29):E2010-E2017.

Hendriksen, R. S., L. B. Price, J. M. Schupp, J. D. Gillece, R. S. Kaas, D. M. Engelthaler, V. Bortolaia, T. Pearson, A. E. Waters, B. P. Upadhyay, S. D. Shrestha, S. Adhikai, G. Shakya, P. S. Keim, and F. M. Aarestrup. 2011. Population genetics of *Vibrio cholerae* from Nepal in 2010: Evidence on the origin of the Haitian outbreak. *MBio* 2(4):e00157-e00111.

Holt, K. E., J. Parkhill, C. J. Mazzoni, P. Roumagnac, F-X. Weill, I. Goodhead, R. Rance, S. Baker, D. J. Maskell, J. Wain, C. Dolecek, M. Achtman, and G. Dougan. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi*. *Nature Genetics* 40:987-993.

Howden, B. P., C. R. E. McEvoy, D. L. Allen, K. Chua, W. Gao, P. F. Harrison, J. Bell, G. Coombs, V. Bennett-Wood, J. L. Porter, R. Robins-Browne, J. K. Davies, T. Seemann, T. P. Stinear. 2011. Evolution of multidrug resistance during *Staphylococcus aureus* infection involves mutation of the essential two component regulator WalKR. *PLoS Pathogens* 7(11):e1002359.

Jacob, D., U. Sauer, R. Housley, C. Washington, K. Sannes-Lowery, D. J. Ecker, R. Sampath, R. Grunow. 2012. Rapid and high-throughput detection of highly pathogenic bacteria by Ibis PLEX-ID technology. *PLoS One* 7(6):e39928.

Kenefic, L. J., J. Beaudry, C. Trim, R. Daly, R. Parmar, S. Zanecki, L. Huynh, M. N. Van Ert, D. M. Wagner, T. Graham, and P. Keim. 2008. High resolution genotyping of *Bacillus anthracis* outbreak strains using four highly mutable single nucleotide repeat markers. *Letters in Applied Microbiology* 46:600-603.

Köser, C. U., M. J. Ellington, E. J. Cartwright, S. H. Gillespie, N. M. Brown, M. Farrington, M. T. G. Holden, G. Dougan, S. D. Bentley, J. Parkhill, and S. J. Peacock. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathogens* 8(8):e1002824.

Loman, N. J., R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* 30(5):434-439.

MacLean, D., J. D. Jones, and D. J. Studholme. 2009. Application of "next-generation" sequencing technologies to microbial genetics. *Nature Reviews Microbiology* 7(4):287-296.

Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences USA* 95:3140-3145.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K B. Jirage, J-B Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

Mellmann, A., D. Harmsen, C. A. Cummings, E. B. Zentz, S. R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji, W. Zhang, S. F. McLaughlin, J. K. Henkhaus, B. Leopold, M. Bielaszewska, R. Prager, P. M. Brzoska, R. L. Moore, S. Guenther, J. M. Rothberg, and H. Karch. 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next-generation sequencing technology. *PLoS One* 6(7):e22751.

Morse, S. A., and B. Budowle. 2006. Microbial forensics: Application to bioterrorism preparedness and response. *Infectious Disease Clinics of North America* 20:455-473.

Morse, S. A., and A. S. Khan. 2005. Epidemiologic investigation for public health, biodefense, and forensic microbiology. In: *Microbial Forensics*, edited by R. Breeze, B. Budowle, and S. Schutzer. Amsterdam: Academic Press. Pp. 157-171.

NRC (National Research Council). 2009. *Strengthening forensic science in the United States: A path forward.* Washington, DC: The National Academies Press.

Rasko, D. A., P. L. Worshamb, T. G. Abshireb, S. T. Stanley, J. D. Bannand, M. R. Wilson, R. J. Langham, R. S. Decker, L. Jianga, T. D. Reade, A. M. Phillippy, S. L. Salzberg, M. Pop, M. N. Van Ert, L. J. Kenefic, P. S. Keim, C. M. Fraser-Liggett, and J. Ravel. 2011. *Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation. *Proceedings of the National Academy of Sciences USA* 108(12):5027-5032.

Ravel, J., L. Jiang, S. T. Stanley, M. R. Wilson, R. S. Decker, T. D. Read, P. Worsham, P. S. Keim, S. L. Salzberg, C. M. Liggett, and D. A. Rasko. 2009. The complete genome sequence of *Bacillus anthracis* Ames "Ancestor." *Journal of Bacteriology* 191:445-446.

Read, T. D., S. L. Salzberg, M. Pop, M. Shumway, L. Umayam, L. Jiang, E. Holtzapple, J. D. Busch, K. L. Smith, J. M. Schupp, D. Solomon, P. Keim, and C. M. Fraser. 2002. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* 296:2028-2033.

Read, T. D., S. N. Peterson, N. Tourasse, L. W. Baillie, I. T. Paulsen, K. E. Nelson, H. Tettelin, D. E. Fouts, J. A. Eisen, S. R. Gill, E. K. Holtzapple, O. A. Okstad, E. Helgason, J. Rilstone, M. Wu, J. F. Kolonay, M. J. Beanman, R. J. Dodson, L. M. Brinkac, M. Gwinn, R. T. DeBoy, R. Madpu, S. C. Daugherty, A. S. Durkin, D. H. Haft, W. C. Nelson, J. D. Peterson, M. Pop, H. M. Khouri, D. Radune, J. L. Benton, Y. Mahamoud, L. Jiang, I. R. Hance, J. F. Wiedman, K. J. Berry, R. D. Plaut, A. M. Wolf, K. L. Watkins, W. C. Nierman, A. Hazen, R. Cline, C. Redmond, J. E. Thwaite, O. White, S. L. Salzberg, B. Thomason, A. M. Friedlander, T. M. Koehler, P. C. Hanna, A. B. Kolstø, and C. M. Fraser. 2003. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* 423:81-86.

Relman, D. A. 2011. Microbial genomics and infectious diseases. *New England Journal of Medicine* 365:347-357.

Rohde, H., J. Qin, Y. Cui, D. Li, N. J. Loman, M. Hentschke, W. Chen, F. Pu, Y. Peng, J. Li, F. Xi, S. Li, Y. Li, Z. Zhang, X. Yang, M. Zhao, P. Wang, Y. Guan, Z. Cen, X. Zhao, M. Christner, R. Kobbe, S. Loos, J. Oh, L. Yang, A. Danchin, G. F. Gao, Y. Song, Y. Li, H. Yang, J. Wang, J. Xu, M. J. Pallen, J. Wang, M. Aepfelbacher, and R. Yang. 2011. *E. coli* O104:H4 Genome Analysis Crowd-Sourcing Consortium 2011. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *New England Journal of Medicine* 365(8):718-724.

Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y-H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter. 2007. The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* 5:e77.

Sampath, R., N. Mulholland, L. B. Blyn, M. W. Eshoo, T. A. Hall, C. Massire, H. M. Levene, J. C. Hannis, P. M. Harrell, B. Neuman, M. J. Buchmeier, Y. Jiang, R. Ranken, J. J. Drader, V. Samant, R. H. Griffey, J. A. McNeil, S. T. Crooke, and D. J. Ecker. 2005. Rapid identification of emerging pathogens: Coronavirus. *Emerging Infectious Diseases* 11:373-379.

Sampath, R., N. Mulholland, L. B. Blyn, C. Massire, C. A. Whitehouse, N. Waybright, C. Harter, J. Bogan, M. S. Miranda, D. Smith, C. Baldwin, M. Wolcott, D. Norwood, R. Kreft, M. Frinder, R. Lovari, I. Yasuda, H. Matthews, D. Toleno, R. Housley, D. Duncan, F. Li, R. Warren, M. W. Eshoo, T. A. Hall, S. A. Hofstadler, and D. J. Ecker. 2009. Comprehensive biothreat cluster identification by PCR/electrospray-ionization mass spectrometry. *Nature Reviews Microbiology* 7(4):287-296.

Spratt, B. G. 1999. Multilocus sequence typing: Molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the Internet. *Current Opinion in Microbiology* 2:312-316.

Srivatsan, A., Y. Han, J. Peng, A. K. Tehranchi, R. Gibbs, J. D. Wang, and R. Chen. 2008. High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genetics* 4(8):e1000139.

Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. 2005. Comparative metagenomics of microbial communities. *Science* 308:554-557.

U'ren, J. M., M. N. Vant, J. M. Schupp, W. R. Easterday, T. S. Simonson, R. T. Okinaka, T. Pearson, and P. Keim. 2005. Use of a real-time PCR TaqMan assay for rapid identification and differentiation of *Burkholderia pseudomallei* and *Burkholderia mallei*. *Journal of Clinical Microbiology* 43:5771-5774.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y-H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66-74.

Vogler, A. J., E. M. Driebe, J. Lee, R. K. Auerbach, C. J. Allender, M. Stanley, K. Kubota, G. L. Andersen, L. Radnedge, P. L. Worsham, P. Keim, and D. M. Wagner. 2008. Assays for the rapid and specific identification of North American *Yersinia pestis* and the common laboratory strain CO92. *BioTechniques* 44:201-207.

A2

MICROBIAL VIRULENCE AS AN EMERGENT PROPERTY: CONSEQUENCES AND OPPORTUNITIES[4]

*Arturo **Casadevall***,[5,*] *Ferric C. Fang*,[6] *Liise-anne Pirofski*[5]

Although an existential threat from the microbial world might seem like science fiction, a catastrophic decline in amphibian populations with the extinction of dozens of species has been attributed to a chytrid fungus (Daszak et al., 1999; Pound et al., 2006), and North American bats are being decimated by *Geomyces destructans*, a new fungal pathogen (Blehert et al., 2009). Hence, individual microbes can cause the extinction of a species. In the foregoing instances, neither fungus had a known relationship with the threatened species; there was neither selection pressure for pathogen attenuation nor effective host defense. Humans are also constantly confronted by new microbial threats as witnessed by the appearance of HIV, SARS coronavirus, and the latest influenza pandemic. While some microbial threats seem to be frequently emerging or re-emerging, others seem to wane or attenuate with time, as exemplified by the decline of rheumatic heart disease (Quinn, 1989), the evolution of syphilis from a fulminant to a chronic disease (Tognotti, 2009), and the disappearance of "English sweating sickness" (Beeson, 1980). A defining feature of infectious diseases is changeability, with change being a function of microbial, host, environmental, and societal changes that together translate into changes in the outcome of a host–microbe interaction. Given that species as varied as amphibians and bats can be threatened with extinction by microbes, the development of predictive tools for identifying microbial threats is both desirable and important.

---

[4] Reprinted from *PLoS Pathogens*. Originally published as Casadevall A, Fang FC, Pirofski L-a (2011) Microbial Virulence as an Emergent Property: Consequences and Opportunities. *PLoS Pathogens* 7(7): e1002136. doi:10.1371/journal.ppat.1002136.

Editor: Glenn F. Rall, The Fox Chase Cancer Center, United States of America.

[5] Department of Microbiology & Immunology and Medicine, Albert Einstein College of Medicine, Bronx, New York, United States of America.

[6] Departments of Laboratory Medicine and Microbiology, University of Washington School of Medicine, Seattle, Washington, United States of America.

Published: July 21, 2011

Funding: The authors are supported by NIH grants AI-45459 (LP), AI44374 (LP), AI39557 (FCF), AI44486 (FCF), AI77629 (FCF), AI91966 (FCF), HL059842 (AC), AI033774 (AC), AI033142 (AC), and AI052733 (AC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: arturo.casadevall@einstein.yu.edu

Virulence as an Emergent Property

To those familiar with the concept of emergence (Box A2-1), it probably comes as no surprise that microbial virulence is an emerging property. However, the traditional view of microbial pathogenesis has been reductionist (Fang and Casadevall, 2011), namely, assigning responsibility for virulence to either the microbe or the host. Such pathogen- and host-centric views, and in turn the scientific approaches fostered by these viewpoints, differ significantly in their historical underpinnings and philosophy (Biron and Casadevall, 2010). In fact, neither alone can account for how new infectious diseases arise. The conclusion that virulence is an emergent property is obvious when one considers that microbial virulence can only be expressed in a susceptible host (Casadevall and Pirofski, 2001). Consequently, the very same microbe can be virulent in one host but avirulent in another (Casadevall and Pirofski, 1999). Furthermore, host immunity can negate virulence, as evidenced by the effectiveness of immunization that renders a microbe as deadly as the variola virus completely avirulent in individuals inoculated with the vaccinia virus. Infection with a microbe can result in diametrically opposed outcomes, ranging from the death of a host to

BOX A2-1
The Concept of Emergent Properties

Emergent properties are properties that cannot be entirely explained by their individual components (Ponge, 2005). An element of novelty is also considered to be an essential attribute of "emergent," a term that contrasts with "resultant" with the latter denoting an outcome that is predicted from the combination of the two components, such that resultant properties are additive whereas emergent properties are non-additive (Ablowitz, 1939). Another facet of emergent properties is that they are irreducible to their constituent components. Most treatises on emergence have emphasized that emergent properties have two components: an outcome that is greater than the sum of the parts and some form of novelty (Ablowitz, 1939; Baylis, 1929; Henpel and Oppenheim, 2011). Although the concept of emergence dates back to antiquity when Aristotle stated that the "whole is not just the sum of its parts," there is increasing interest in emergent properties as it becomes increasingly evident that reductionistic approaches cannot explain many phenomena in our world (Parrish et al., 2011). Examples of emergent properties in liquids are surface tension and viscosity, neither of which can be explained by analysis of individual molecules, as the properties pertain to the macroscopic world, and these phenomena have no corresponding analogs in the molecular realm. Biological systems have been described as characterized by emergent properties that exist at the edge of chaos, such that small fluctuations in their conditions can lead to sudden major changes (Mazzocchi, 2008). Similarly, self-organized movements of individuals, as in schools of fish, can result in a variety of forms that are thought to protect against predators (Parrish et al., 2011).

elimination of the microbe. Hence, virulence is inherently novel, unpredictable, and irreducible to first principles.

Critical to our understanding of virulence as a property that can only be expressed in a susceptible host is that both the microbe and the host bring their own emergent properties to their interaction. Host and microbial cells receive and process information by signaling cascades that manifest emergent properties (Bhalla and Iyengar, 1999); e.g., gene expression studies reveal heterogeneous or bi-stable expression in clonal cell populations with important implications for phenotypic variability and fitness (Dubnau and Losick, 2006; Veening et al., 2008). Other emergent properties that have been identified in microbial and cellular systems could influence pathogenesis. Intracellular parasitism is associated with genome reduction, a phenomenon that could confer emergent properties, given that deliberate genome reduction in *E. coli* has led to unexpected emergent properties, such as ease of electroporation and increased stability of cloned DNA and plasmids (Posfai et al., 2006).

On the host side, many aspects of the immune system have the potential to spawn emergent properties. The antigenic determinants of a microbe are defined by antibodies and processing by host cells, consequently existing only in the context of an immune system (Van Regenmortel, 2004). Microbial determinants can elicit host-damaging immune responses. Such deleterious responses exemplify a detrimental emergent property of the same host defense mechanisms that mediate antimicrobial effects. The outcome of a viral infection can depend on prior infection with related or unrelated viruses that express related antigens; hence, the infection history of a host affects the outcome of subsequent infections (Welsh et al., 2010).

For those accustomed to viewing host–microbe interactions from an evolutionary perspective (Dethlefsen et al., 2007), the emergent nature of virulence is also no surprise, for the evolution of life itself can be viewed as an emergent process (Corning, 2002). Even in relatively well-circumscribed systems such as Darwin's finches on the Galápagos Islands, evolutionary trends over time became increasingly unpredictable as a consequence of environmental fluctuations (Grant and Grant, 2002).

Consequences of the Emergent Nature of Microbial Virulence

The fact that virulence is an emergent property of host, microbe, and their interaction has profound consequences for the field of microbial pathogenesis, for it implies that the outcome of host–microbe interaction is inherently unpredictable. Even with complete knowledge of microbes and hosts, the outcome of all possible interactions cannot be predicted for all microbes and all hosts. Lack of predictability should not be unduly discouraging. Even in systems in which emergent properties reveal novel functions, such as fluid surface tension and viscosity, recognition of these properties can be useful. For example, molecular structure

might not predict the hydrodynamics of a fluid, but the empirical acquisition of information can be exploited to optimize pipeline diameter and flow rates. Novelty is unpredictable but novel events can be interpreted and comprehended once they have occurred (Ablowitz, 1939). A pessimist might argue that living systems are significantly more complex than flowing liquids. However, such pessimism may be unwarranted. The appearance of new influenza virus strains every year is an emergent property resulting from high rates of viral mutation and host selection of variants (Lofgren et al., 2007). Hence, the time or place in which new pandemics will arise or the relative proportion of strains that will circulate each year cannot be predicted with certainty. Nevertheless, the likely appearance of new strains can be estimated from the history of population exposure to given strains and knowledge of recently circulating strains, and this information can be used to formulate the next year's vaccine.

A Probabilistic Framework

Although the field of infectious diseases may never achieve the predictive certainty achieved in other branches of medicine, it may be possible to develop a probabilistic framework for the identification of microbial threats. Although all known pathogenic host–microbe interactions have unique aspects, and it is challenging to extrapolate from experiences with one microbe to another, a probabilistic framework can incorporate extant information and attempt to estimate risks. For example, the paucity of invasive fungal diseases in mammalian populations with intact immunity has been attributed to the combination of endothermy and adaptive immunity (Robert and Casadevall, 2009). This notion could be extrapolated to other environmental microbes, i.e., those that cannot survive at mammalian temperatures have a low probability of emerging as new human pathogens. On the other hand, the identification of known virulence determinants in new bacterial strains may raise concern. In this regard, the expression of anthrax toxin components in *Bacillus cereus* produces an anthrax-like disease that is not caused by *Bacillus anthracis* (Hoffmaster et al., 2004).

Given the experience of recent decades, we can predict with confidence that new infectious diseases are likely to continue to emerge and make some general predictions about the nature of the microbes that could constitute these threats. One possibility is that an emergent pathogen could come from elsewhere in the animal kingdom. A comprehensive survey revealed that three-fourths of emerging pathogens are zoonotic (Taylor et al., 2001). Crossing the species barrier can result in particularly severe pathology, as pathogen and host have not had the opportunity to co-evolve toward equilibrium. Another good bet is that an RNA virus could emerge as a pathogen. The high mutation rate and generally broad host range of RNA viruses may favor species jumps (Woolhouse et al., 2005), and many emergent human pathogens belong to this group, e.g., HIV, H5N1 influenza, SARS coronavirus, Nipah virus, and hemorrhagic fever viruses. On

the other hand, global warming could hasten the emergence of new mammalian pathogenic fungi through thermal adaptation (Garcia-Solache and Casadevall, 2010), given that the relative resistance of mammals to fungal diseases has been attributed to a combination of higher body temperatures and adaptive immunity (Bergman and Casadevall, 2010; Robert and Casadevall, 2009).

Despite abandoning hopes for certainty and determinism in predicting microbial pathogenic interactions, we can attempt to develop a probabilistic framework that endeavors to estimate the pathogenic potential of a microbe based on lessons from known host–microbe interactions. A variety of mathematical models based on game theory or quantitative genetics have been developed in attempts to understand the evolution of virulence (Boots et al., 2009; Day and Proulx, 2004). These have provided interesting new insights into host–pathogen interactions, including the tendency for evolutionary dynamics to produce oscillations and chaos rather than stable fitness-maximizing equilibria, the unpredictability that results when multiple games are played simultaneously, and the tendency for three-way co-evolution of virulence with host tolerance or resistance to select for greater virulence and variability (Carval and Ferriere, 2010; Hashimoto, 2006; Nowak and Sigmund, 2004).

Preparing for the Unpredictable

Emerging infections seem to be becoming more frequent, and it is not difficult to understand why. An interesting experimental system examining a viral pathogen of moth larvae demonstrated that host dispersal promotes the evolution of greater virulence (Boots and Mealor, 2007). When hosts remain local, this encourages more "prudent" behavior by pathogens, but host movement encourages more infections and greater disease severity (Buckling, 2007). Global travel in the modern world can rapidly spread pathogenic microbes, but what is less obvious is that travel may also enhance virulence. Other factors contributing to the emergence and re-emergence of new pathogens include changes in land use, human migration, poverty, urbanization, antibiotics, modern agricultural practices, and other human behaviors (Cleaveland et al., 2007; IOM, 1992). Microbial evolution and environmental change, anthropogenic or otherwise, will continue to drive this process. Another implication of the emergent nature of virulence is recognition of the hubris and futility of thinking that we can simply target resources to the human pathogens that we already know well. The discovery of HIV as the cause of AIDS (Barre-Sinoussi et al., 1983) was greatly facilitated by research on avian and murine retroviruses that had taken place decades before (Hsiung, 1987), at a time when the significance of retroviruses as agents of human disease was unknown.

We share the view that sentinel capabilities are more important than predictive models at the present time (Barre-Sinoussi et al., 1983; Hsiung, 1987), but are optimistic that it will be possible to develop general analytical tools that can

be applied to provide probabilistic assessments of threats from future unspecified agents. Comparative analysis of microbes with differing pathogenic potential and their hosts could provide insight into those interactions that are most likely to result in virulence. Hence, the best preparation for the unexpected and unpredictable nature of microbial threats will be the combination of enhanced surveillance with a broad exploration of the natural world to ascertain the range of microbial diversity from which new threats are likely to emerge.

References

Ablowitz R (1939) The theory of emergence. Phil Sci 6: 1–16.

Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, et al. (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220: 868–871.

Baylis CA (1929) The philosophic functions of emergence. Philos Rev 38: 372–384.

Beeson PB (1980) Some diseases that have disappeared. Am J Med 68: 806–811.

Bergman A, Casadevall A (2010) Mammalian endothermy optimally restricts fungi and metabolic costs. MBio 1: 00212–10.

Bhalla US, Iyengar R (1999) Emergent properties of networks of biological signaling pathways. Science 283: 381–387.

Biron CA, Casadevall A (2010) On immunologists and microbiologists: ground zero in the battle for interdisciplinary knowledge. MBio 1: e00280–10.

Blehert DS, Hicks AC, Behr M, Meteyer CU, Berlowski-Zier BM, et al. (2009) Bat white-nose syndrome: an emerging fungal pathogen? Science 323: 227.

Boots M, Best A, Miller MR, White A (2009) The role of ecological feedbacks in the evolution of host defence: what does theory tell us? Philos Trans R Soc Lond B Biol Sci 364: 27–36.

Boots M, Mealor M (2007) Local interactions select for lower pathogen infectivity. Science 315: 1284–1286.

Buckling A (2007) Epidemiology. Keep it local. Science 315: 1227–1228.

Carval D, Ferriere R (2010) A unified model for the coevolution of resistance, tolerance, and virulence. Evolution 64: 2988–3009.

Casadevall A, Pirofski L (1999) Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. Infect Immun 67: 3703–3713.

Casadevall A, Pirofski L (2001) Host-pathogen interactions: the attributes of virulence. J Infect Dis 184: 337–344.

Cleaveland S, Haydon DT, Taylor L (2007) Overviews of pathogen emergence: which pathogens emerge, when and why? Curr Top Microbiol Immunol 315: 85–111.

Corning PA (2002) The re-emergence of 'emergence': a venerable concept in search for a theory. Complexity 7: 18–30.

Daszak P, Berger L, Cunningham AA, Hyatt AD, Green DE, et al. (1999) Emerging infectious diseases and amphibian population declines. Emerg Infect Dis 5: 735–748.

Day T, Proulx SR (2004) A general theory for the evolutionary dynamics of virulence. Am Nat 163: E40–E63.

Dethlefsen L, McFall-Ngai M, Relman DA (2007) An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature 449: 811–818.

Dubnau D, Losick R (2006) Bistability in bacteria. Mol Microbiol 61: 564–572.

Fang FC, Casadevall A (2011) Reductionistic and holistic science. Infect Immun 79: 1401–1414.

Garcia-Solache MA, Casadevall A (2010) Global warming will bring new fungal diseases for mammals. MBio 1: e00061–10.

Grant PR, Grant BR (2002) Unpredictable evolution in a 30-year study of Darwin's finches. Science 296: 707–711.

Hashimoto K (2006) Unpredictability induced by unfocused games in evolutionary game dynamics. J Theor Biol 241: 669–675.

Henpel CG, Oppenheim P (2011) Studies in the logic of explanation. Phil Sci 15: 135–175.

Hoffmaster AR, Ravel J, Rasko DA, Chapman GD, Chute MD, et al. (2004) Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. Proc Natl Acad Sci U S A 101: 8449–8454. Taylor LH, Latham SM, Woolhouse ME (2001) Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci 356: 983–989.

Hsiung GD (1987) Perspectives on retroviruses and the etiologic agent of AIDS. Yale J Biol Med 60: 505–514.

IOM (1992) Emerging infections: microbial threats to the United States. Washington (D.C.): Institute of Medicine.

Lofgren E, Fefferman NH, Naumov YN, Gorski J, Naumova EN (2007) Influenza seasonality: underlying causes and modeling theories. J Virol 81: 5429–5436.

Mazzocchi F (2008) Complexity in biology. Exceeding the limits of reductionism and determinism using complexity theory. EMBO Rep 9: 10–14.

Nowak MA, Sigmund K (2004) Evolutionary dynamics of biological games. Science 303: 793–799.

Parrish JK, Viscido SV, Grumbaum D (2011) Self organized fish schools: an example of emergent properties. Biol Bull 202: 296–305

Ponge JF (2005) Emergent properties from organisms to ecosystems: towards a realistic approach. Biol Rev Camb Philos Soc 80: 403–411.

Posfai G, Plunkett G, III, Feher T, Frisch D, Keil GM, et al. (2006) Emergent properties of reduced-genome *Escherichia coli*. Science 312: 1044–1046.

Pounds JA, Bustamante MR, Coloma LA, Consuegra JA, Fogden MP, et al. (2006) Widespread amphibian extinctions from epidemic disease driven by global warming. Nature 439: 161–167.

Quinn RW (1989) Comprehensive review of morbidity and mortality trends for rheumatic fever, streptococcal disease, and scarlet fever: the decline of rheumatic fever. Rev Infect Dis 11: 928–953.

Robert VA, Casadevall A (2009) Vertebrate endothermy restricts most fungi as potential pathogens. J Infect Dis 200: 1623–1626.

Tognotti E (2009) The rise and fall of syphilis in Renaissance Europe. J Med Humanit 30: 99–113.

Van Regenmortel MH (2004) Reductionism and complexity in molecular biology. Scientists now have the tools to unravel biological and overcome the limitations of reductionism. EMBO Rep 5: 1016–1020.

Veening JW, Smits WK, Kuipers OP (2008) Bistability, epigenetics, and bet-hedging in bacteria. Annu Rev Microbiol 62: 193–210.

Welsh RM, Che JW, Brehm MA, Selin LK (2010) Heterologous immunity between viruses. Immunol Rev 235: 244–266.

Woolhouse ME, Haydon DT, Antia R (2005) Emerging pathogens: the epidemiology and evolution of species jumps. Trends Ecol Evol 20: 238–244.

A3

MICROBIAL GENOME SEQUENCING TO UNDERSTAND PATHOGEN TRANSMISSION

Jennifer L. ***Gardy***[7]

Outbreak Investigation: A Brief Primer

In public health, we are often confronted with the task of "solving" an infectious disease outbreak—identifying all the cases, determining a source of the illness, and deploying an intervention to prevent further cases. A typical scenario unfolds as follows. A potential outbreak alert is issued when routine laboratory or population-based surveillance methods detect a statistically significant increase in case counts relative to historical norms for a particular disease, or when an astute clinician or public health official notes an unusual clustering of cases. This alert triggers an initial investigation combining descriptive epidemiology with laboratory work. Epidemiologists use interviews and questionnaires to review case data, such as travel history, food exposures, and attendance at social events, with the goal of revealing common behaviours across cases—eating the same food items, visiting the same locations, or shared contact with a particular individual.

At the same time, microbiologists carry out their own epidemiological investigation using genotyping techniques. Similar to the genetic fingerprinting methods used in paternity testing or in forensic crime scene analysis, these "molecular epidemiology" tools, including pulsed-field gel electrophoresis (PFGE) and multi-locus sequencing typing (MLST), can quickly reveal whether a collection of bacterial specimens share a common genetic fingerprint and likely represent a true outbreak, or whether they display a range of genotypes and simply reflect an unusual excess of cases of that particular illness, none of which are related to each other.

The results of the descriptive epidemiology and molecular epidemiology investigations are then compared, and a determination is made as to whether the cluster of cases is truly an outbreak meriting further investigation. If this is indeed the case, then a more robust field epidemiological investigation is typically undertaken. This includes enhanced case-finding using more detailed survey instruments as well as case-control studies in which behaviours of cases are compared to those of controls in order to quantify risk factors strongly associated with illness. Through these analyses, investigators are able to form and test a specific hypothesis regarding the source of the outbreak. Laboratory work is also critical at this stage—new cases are genotyped to determine whether they are part of the outbreak, while genotyping of isolates collected from food, water,

[7] Senior Scientist, Molecular Epidemiology, British Columbia Centre for Disease Control.

and other non-human sources can confirm or rule out these entities as potential sources of the outbreak.

Once a source of the outbreak has been confirmed, intervention measures can be put in place. For food- or water-borne outbreaks, these typically involve issuing a recall for the food item in question, eliminating access to the water source until it has been declared safe, and issuing extensive media alerts warning consumers of the risks associated with the entity in question. For outbreaks involving personal contact or attendance at a shared location, such as a specific hospital ward, active case-finding is used to find and treat all infected patients or potential carriers of an illness, while infection control approaches such as patient decolonization or enhanced cleaning are deployed to prevent further infections.

Unfortunately, not every outbreak can be neatly resolved. A number of factors greatly limit public health's ability to investigate an outbreak from both the field epidemiology and molecular epidemiology perspectives (Figure A3-1). Field investigations are typically limited by resources—not having enough personnel, time, or money to be able to effect a complete investigation—and patients' inability to recall specific events that might be relevant to the investigation. Molecular epidemiology approaches are also limited in their utility. For some pathogens, such as *Salmonella* Enteritidis, unrelated isolates from multiple outbreaks may show identical genetic fingerprints. For others, such as *Campylobacter jejeuni*, one outbreak may comprise multiple distinct genetic fingerprints due to frequent rearrangement of the pathogen's genome. Genotyping typically requires the

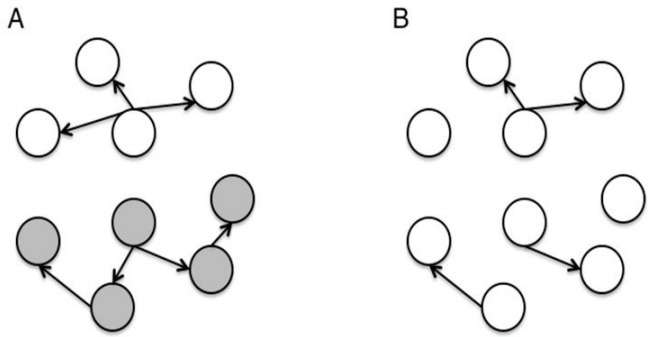


FIGURE A3-1 An example demonstrating how the limitations of field and molecular epidemiology complicate outbreak reconstructions. Panel A shows the "true" outbreak scenario—two different genotypes of pathogen are found in the hosts (white and grey circles), and arrows indicate person-to-person transmission events. Panel B shows what a reconstruction of that outbreak might look like using current tools. In this situation, the molecular epidemiology technique applied was not able to identify the different genotypes, and all isolates were grouped into a single cluster, as often happens in the case of clonal pathogens such as *Salmonella* Enteritidis. Patients' inability to recall specific contacts means that several transmission events are missed.

organism in question to be cultured, which may add several weeks to an investigation in the case of slow-growing organisms such as *Mycobacterium tuberculosis*, and the costs of many molecular epidemiology assays are not insignificant, meaning they are often not routinely deployed.

One of the biggest limitations of current molecular epidemiology methods is the low level of resolution they provide. At best, such tools are only capable of determining whether or not an isolate belongs to an outbreak cluster. Further detail, such as the order of person-to-person transmission, the underlying pattern of spread—superspreader or ongoing chains of transmission—is beyond the scope of current laboratory methods.

An Illustrative Example

In May 2006, a case of pleural tuberculosis (TB) was diagnosed in an adult female in a medium-sized community in British Columbia, Canada. Although pleural TB is suggestive of recently acquired disease, inquiring after the case's contacts did not suggest a potential source for her illness. Molecular analysis using a TB-specific technique called mycobacterial interspersed repetitive unit variable number tandem repeat (MIRU-VNTR) was performed. In MIRU-VNTR, 24 variable number tandem repeat loci around the *M. tuberculosis* genome are amplified using polymerase chain reaction (PCR), followed by capillary electrophoresis to enumerate the number of repeats present at each locus. The patient's MIRU-VNTR genotype indicated she harboured the same strain of TB that had been circulating in her community for several years. She was assumed to represent one of the few annual cases of TB that community regularly observed in a year and was treated for her illness.

Some months later, a second case of TB was reported in the community, this time in an infant female with no epidemiological link to the May case. TB in children is considered to be a marker for recent community transmission, and when MIRU-VNTR revealed the infant to be infected with the same strain of TB as the earlier case, the local public health nurses began an intensive case-finding effort. Using an approach called reverse contact tracing, they identified individuals who had been in contact with the infant and screened them using a tuberculin skin test in an attempt to find out who had been the source of the child's infection. This investigation led to the diagnosis of nine more cases of active TB in the community, and an outbreak was declared.

Extensive investigation soon followed. Each case was interviewed using a detailed questionnaire, and the resulting data—connections between individuals who reported social relationships with each other, links between people and the places they regularly spent time at, and links between people and specific behaviours associated with an increased risk of TB infection, such as smoking, alcohol use, or drug use—was visualized as a network. The network suggested a potential source for the outbreak—an individual who had been symptomatic

and undiagnosed for many months prior to detection of the first case, who had a number of risk factors, and who was a high-degree node in the network—they reported contact with many of the cases.

Although the investigation revealed the likely source case, who was immediately put on treatment, the outbreak eventually grew to include 41 individuals over the 2006–2008 period, with a handful of subsequent diagnoses from 2009 onward.

Attempting to reconstruct the path the organism took through the community proved to be impossible. Despite the rich epidemiological and clinical data available, the social network structure in the community was too dense to interpret—each individual case had an average of six contacts with other cases, and most everyone in the community reported spending time at the same locations, including a series of hotel pubs and crack houses, meaning there were many potential sources for each person's infection (Figure A3-2). All cases had identical 24-loci MIRU-VNTR patterns, but the low resolution of this technique was incapable of identifying smaller subclusters within the outbreak.

While the outbreak eventually abated, our inability to reconstruct individual transmission events meant that an important learning opportunity was missed. We could not describe the underlying pattern of disease spread in the community, we could not compare it to other TB outbreaks to determine whether this organism behaves in a similar way across different outbreaks in different communities with

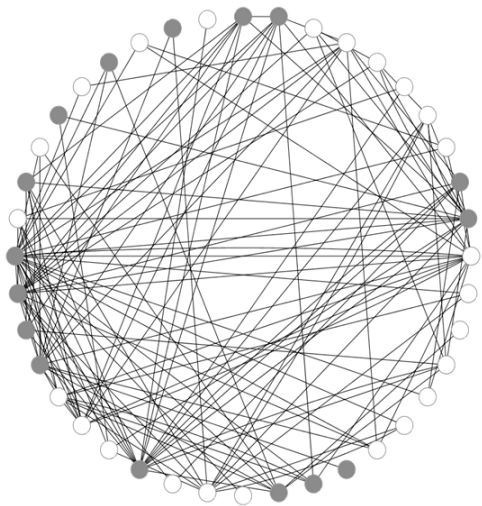


FIGURE A3-2 The dense social network in the outbreak community complicated outbreak reconstruction attempts. Circular nodes represent outbreak cases—grey nodes are individuals with smear-positive tuberculosis; white nodes are individuals with smear-negative disease. Solid lines connect two cases that reported a social contact with each other.

different social network structures, and we could not use our experience to guide future TB outbreak investigations.

This uncertainty about how an outbreak unfolds is not unique to tuberculosis. For the majority of communicable diseases, our understanding of how they behave "in the wild" is limited. Unfortunately, this lack of understanding of pathogens' natural transmission tendencies precludes developing any sort of proactive evidence-based interventions. We do not know whether there are "one-size-fits-all" interventions for a given pathogen or a family of disease, or whether each outbreak is unique and will require a specifically tailored intervention.

The Rise of the Next Generation

We will return to the tuberculosis story in time; for now, we must climb into our microbial genomics time machine and rewind several years. . . . The complete genome has sometimes been described as "the ultimate genotype"—examining the total genetic content of an organism reveals the unique fingerprint that sets each of us apart from the other members of our species. Until recently, however, interrogating the complete genome of anything larger than a virus required a significant investment of time, money, and analytical resources. Sequencing of the first bacterial genome, *Haemophilus influenzae*, in 1995 took more than a year, cost nearly USD$1 million, and involved a large team of researchers running a not insignificant number of DNA sequencers. Subsequent microbial genomics efforts targeted individual bacteria selected to represent a range of common laboratory strains and interesting clinical isolates; experiments reporting the sequencing of more than one isolate were relatively rare and certainly outside the scope of most research groups' technological abilities.

Tracking the number of bacterial genome projects recorded in the Genomes Online Database (GOLD) reveals that after approximately a decade of steady progress in microbial genomics, a sudden and dramatic upswing in the number of sequenced genomes began around 2006 (Figure A3-3). This sea change coincides with the commercial release of the so-called "next-generation" DNA-sequencing technologies. Previously, DNA sequencing was performed using the Sanger method, originally developed by Frederick Sanger in 1977, although subsequently modified to improve throughput. Next-generation sequencing methods, including the pyrosequencing platform commercialized by Roche, a reversible terminator platform marketed by Illumina, and the newer ion semiconductor-based approach available through Life Technologies, all take a fundamentally different approach to sequencing. They are based on the concept of "sequencing by synthesis," in which DNA synthesis is essentially observed in real-time, with the sequencing instrument using one of the above technologies to extend a template base by base, and record which base was added at each step. Although the reads produced by these technologies are much shorter than those resulting from Sanger sequencing, the sheer magnitude of the parallel sequencing made possible by these approaches

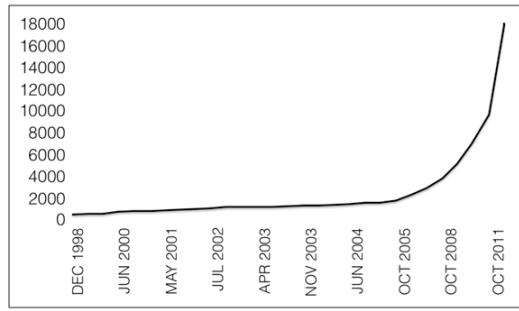


FIGURE A3-3 The number of bacterial genome projects recorded in the Genomes Online Database (GOLD) increased exponentially with the introduction of next-generation sequencing methods in the mid-2000s.

means that next-generation sequencing runs generate orders of magnitude more data in a single run than Sanger sequencers are capable of.

With these new platforms, the cost of sequencing a complete bacterial genome dropped dramatically. Now, many large genome centres operating optimized pipelines and high-volume sequencers are able to offer their clients full bacterial genome sequences for between USD$50-250 per genome. Run times can be as little as a few hours for certain platforms, meaning it is now possible to sequence tens or even hundreds of bacterial genomes within a week for only a few thousand dollars.

A New Tool: Genomic Epidemiology

Soon after the commercialization and adoption of next-generation sequencing technologies, a few astute clinicians and infectious disease researchers recognized the technology's potential for transforming molecular epidemiology. In a wonderful example of convergent evolution, several independent research groups around the world embarked upon proof-of-concept projects in the new area of "genomic epidemiology," with the first few papers in the field appearing in 2010 and 2011.

The basic premise behind genomic epidemiology is that the microevolutionary events occurring within a pathogen's genome over the course of an outbreak can be used as markers of transmission. For example, consider an outbreak in which the first patient is colonized with a bacterium having the genome sequence AAAAA. Any individuals infected by that person would then be colonized with bacteria having the same genome sequence. As a result of the natural process of mutation, many of these second-generation organisms will accrue a small number of nucleotide changes (the number depends on the duration of infection

and the natural mutation rate of the pathogen in question). If we have three second-generation patients, one might accrue no mutations and continue to display the AAAAA genome sequence, one might show the sequence ACAAA, and one might contain the sequence AAAAG. The third generation of cases, those infected by these individuals, will then show genome sequences identical to or descended from these second-generation cases. By sequencing the genomes of all the outbreak organisms and identifying positions that vary over the course of the outbreak, one should, in theory, be able to infer the individual transmission events that gave rise to the outbreak (Figure A3-4).

The words "in theory" are very important in this case. The first two studies to use genomics to identify person-to-person transmission events both revealed that the answers are not so readily forthcoming.

In the first study (Lewis et al., 2010), genome sequence was obtained for six multidrug-resistant *Acinetobacter baumannii* isolates from a hospital outbreak occurring over a seven-week period—four from military patients and two from

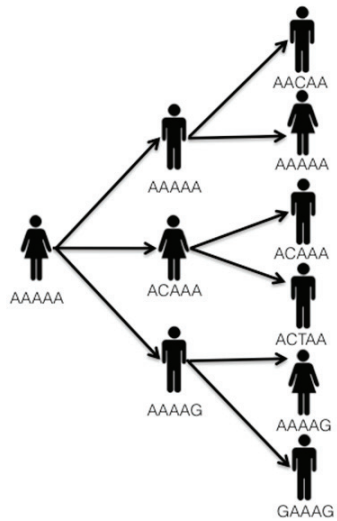


FIGURE A3-4 Using microevolutionary events to track person-to-person spread of a pathogen over a social network. As a pathogen spreads over a contact network, the accrual of mutations can be used to trace person-to-person transmission. When a mutation arises in one person, such as the C in position two of the second-generation female case in this example, it will be transmitted onwards to those individuals that case infects. Isolates may transmit without accruing mutations, as in the AAAAA sequence maintained here across three generations of illness. This can complicate reconstructions if the underlying contact data are unknown—if no contact information was available for this network, the genomic data would suggest that the third-generation AAAAA female case could have been infected by the first- or second-generation AAAAA cases.

civilian patients. The hope was that the study would reveal how the bacterium was transferred from the military patients—who were presumed to have been infected in the field prior to hospital admission—to the civilian patients. Three positions across the approximately 4-megabase pair genome were found to vary between isolates, with patients showing four genotypes at these positions: CAG, TAG, TAT, and TTG. One of the civilian patients shared a genotype with two of the military patients, suggesting that one of these military patients was the source of the civilian's infection. Examination of this hypothesis in the context of the available epidemiological information revealed that one of the military patients was housed in the bed next to the civilian, making him or her the most likely source. The other civilian patient displayed a unique genotype, and the study authors were not able to infer the source of that individual's infection.

In the second study (Gardy et al., 2011), genomics was used to reconstruct the tuberculosis outbreak described earlier in this paper. Genome sequence was obtained for 36 *M. tuberculosis* isolates—32 from outbreak cases and 4 from patients diagnosed in the same community in the decade prior to the outbreak, all of which had the same 24-loci MIRU-VNTR fingerprint as the outbreak cases. More than 200 single nucleotide polymorphisms (SNPs) were found among the isolates, and the authors realized the nature of TB infection meant it would be impossible to trace the outbreak's path SNP by SNP, as was done by Lewis et al. In TB, an individual may be infectious for a period of many months, during which time the colonizing organism is continuing to accrue mutations. If the individual transmits the disease to an individual on day 1 of his or her illness, to another individual on day 180, and to a third individual on day 270, and is diagnosed and his or her organism sampled at day 300, his or her isolate might show similarity to the isolate in patient 3, but could be very different from the isolate in patient 1, making it difficult to ascribe patient 1's disease to the source case. The variable periods of latency associated with TB further complicate SNP-by-SNP reconstruction of transmission.

Instead, the authors used a phylogenetic tree of the data to demonstrate that two separate lineages of *M. tuberculosis*—labeled A and B—could be resolved within the single MIRU-VNTR genotype. Thus the genomic data acted as a sort of enhanced genotyping method, able to break one MIRU-VNTR cluster containing all the isolates into two distinct genome-based clusters, A and B. Although the original social network describing the relationships between all the outbreak cases was too complex to resolve, when it was broken down into two networks—one showing connections between A cases and one showing connections between B cases—the data became much more interpretable and several person-to-person transmission events could be identified. This revealed that several key individuals acting as superspreaders were associated with the majority of transmission events and that factors including delays in diagnosis, clinical presentation, and risk behaviours contributed toward these individuals' role as sources of infection. Not every transmission event could be identified, however, and the genomic data also

suggested that some individuals in the network might have exhibited coinfection with both an A and a B strain.

Best Practices and Future Directions

The earliest genomic epidemiology studies suggested that when combined with epidemiological and clinical data, whole genome sequencing has the potential to inform reconstructions of communicable disease outbreaks. Since the publication of the first few studies, several other papers have described using whole genome sequencing to solve outbreaks of other organisms, including *Clostridium difficile*, methicillin-resistant *Staphylococcus aureus*, and *Klebsiella pneumoniae*. Projects are becoming larger and more ambitious, sequencing hundreds of isolates collected across large regions over many years, and the number of outbreak reconstructions available for an individual pathogen is growing as well.

As this emerging field continues to find its place in the realm of public health microbiology, it is important to note several "best practices" that must be considered when doing such a study.

1. Genomic data alone cannot reliably identify individual transmission events. The genomic data must be combined with epidemiological and clinical information if a plausible reconstruction of an outbreak is to be achieved.
2. The bioinformatics methods for identifying positions of variation across a series of isolates are not perfect. The results of any analysis must be carefully examined to ensure that errors in alignment or inappropriate scoring thresholds are not causing variants to be erroneously called or missed.
3. The data must be considered in terms of biological plausibility. An expected level of variation over an outbreak can be inferred from organisms' mutation rates. If the observed variation is much less or much greater than the expected variation, then the analysis used to generate that data must be reexamined.
4. For others to evaluate a study's accuracy and reproduce the results, the raw sequencing data for each isolate should be made freely available in a public repository. Manuscripts describing genomic epidemiology studies should include, as appendices, the analysis commands used to generate the data and a detailed description of how the data were filtered and processed after genome assembly and SNP calling.
5. There is a significant amount of interesting biology that can be mined from outbreak-derived genome sequences, particularly in the area of population genetics. To maximize the value of a sequencing data set, it is well worth identifying academic partners who can use a study data set for further analyses.

As more and more genomic epidemiology projects are undertaken, the natural behaviour of pathogens in the wild will slowly be revealed. We will know much more about their spatial and temporal patterns of spread, whether superspreading is as common as early reports are indicating, and the factors that influence an individual's tendency to spread disease. It will then be up to public health agencies to use this valuable information to develop evidence-based interventions—for example, directing case-finding efforts around contacts of potential superspreaders, or designing prevention programs targeted to specific high-risk communities or individuals. In our own work, we are sequencing 20 years worth of TB in a single Canadian province to identify province-wide transmission routes, community-level transmission events, and socioeconomic and clinical risk factors for acting as a source or sink community for disease. It is our hope that the resulting data will allow us to reshape our current TB prevention and control programs, enabling us to use our limited resources for maximum effect.

As sequencing technologies improve—generating longer reads at lower costs—and as bioinformatics methods become more reliable at identifying variation, we anticipate even more accurate and detailed outbreak reconstructions. The coming decade will be an exciting time for genomic epidemiology as it moves from proof of concept to a routine component of clinical practice.

References

Gardy, J. L., J. C. Johnston, S. J. Ho Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. M. Jones, F. S. L. Brinkman, R. C. Brunham, and P. Tang. 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New England Journal of Medicine* 364:730-739.

Lewis, T., N. J. Loman, L. Bingle, P. Jumaa, G. M. Weinstock, D. Mortiboy, and M. J. Pallen. 2010. High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. *Journal of Hospital Infection* 75(1):37-41.

A4

PRESENCE OF OSELTAMIVIR-RESISTANT PANDEMIC A/H1N1 MINOR VARIANTS BEFORE DRUG THERAPY WITH SUBSEQUENT SELECTION AND TRANSMISSION[8]

Elodie ***Ghedin***,[9,*] *Edward C. Holmes*,[10,11]
Jay V. DePasse,[8] *Lady Tatiana Pinilla*,[12] *Adam Fitch*,[8]
Marie-Eve Hamelin,[13] *Jesse Papenburg*,[12] *Guy Boivin*[11,*]

Abstract

A small proportion (1-1.5%) of 2009 pandemic A/H1N1 influenza viruses (A(H1N1)pdm09) are oseltamivir-resistant, due almost exclusively to a H275Y mutation in the neuraminidase protein. However, many individuals infected with resistant strains had not received antivirals. Whether drug-resistant viruses are initially present as minor variants in untreated subjects before they emerge as the dominant strain in a virus population is of great importance for predicting the speed at which resistance will arise. To address this issue, we employed ultra-deep sequencing of viral populations from serial nasopharyngeal specimens from an immunocompromised child and from two individuals in a household outbreak. We observed that the Y275

[8] Reprinted with permission by permission of Oxford University Press. Originally published as Ghedin, E, et al. (2012) Presence of oseltamivir-resistant pandemic A/H1N1 minor variants before drug therapy with subsequent selection and transmission. *Journal of Infectious Diseases* 206(10), 1504-1511. doi: 10.1093/infdis/jis571.

[9] Department of Computational & Systems Biology, Center for Vaccine Research, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA.

[10] Center for Infectious Disease Dynamics, Department of Biology,
The Pennsylvania State University, University Park, PA 16802, USA.

[11] Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA.

[12] Centre de Recherche du Centre Hospitalier Universitaire de Québec and Université Laval, Québec City, Québec, Canada.

[13] McGill University Health Centre, Montréal, Québec, Canada.

Conflict of interest statement: E.G., E.C.H., J.V.D., T.P., A.F., M.-E.H., J.P.: No conflicts
G.B.: Research grant from GlaxoSmithKline

Funding statement: This work was supported in part by National Institute of Allergy and Infectious Diseases at the National Institutes of Health [grant number U54 GM088491]; National Institute of General Medical Sciences at the National Institutes of Health [grant number 2R01 GM080533-06 to E.C.H.] and the Canadian Institutes of Health Research [to G.B.].

* Corresponding authors: Dr. Elodie Ghedin, Center for Vaccine Research, University of Pittsburgh School of Medicine, 3501 5th Avenue, BST3 Room 9043b, Pittsburgh, PA 15261. Phone: (412) 383-5850. E-mail: elg21@pitt.edu. Dr. Guy Boivin, Centre de Recherche en Infectiologie, CHUL, room RC-709, 2705 Laurier, Québec City, QC, Canada. E-mail: guy.boivin@crchul.ulaval.ca

Key words: Influenza; A(H1N1)pdm09; drug resistance; deep sequencing; oseltamivir.

mutation was present as a minor variant in infected hosts prior to onset of therapy. We also found evidence for the transmission of this drug-resistant variant alongside drug-sensitive viruses. These observations provide important information on the relative fitness of the Y275 mutation in the absence of oseltamivir.

The 2009 pandemic A/H1N1 influenza virus (A(H1N1)pdm09) emerged following reassortment between two swine viruses circulating in North America and Eurasia (Garten et al., 2009). Between 1 and 1.5% of A(H1N1)pdm09 strains analyzed to date have been found to be resistant to oseltamivir, a neuraminidase (NA) inhibitor that constitutes the current standard of care (Pizzorno et al., 2011a). Virtually all oseltamivir-resistant A(H1N1)pdm09 viruses contain an H275Y amino acid substitution in the viral NA gene (Pizzorno et al., 2011b). Among the drug-resistant strains recovered from immunocompetent patients, approximately one-third have been recovered from untreated individuals (WHO, 2011). Whether drug-resistant variants are initially present as minor variants in untreated subjects due to transmission from a host harboring a minority drug-resistant population, or whether they emerge following *de novo* replication, is of great importance for predicting the speed at which resistance will arise: the selection of resistant mutations will occur more rapidly if they are already present within hosts as pre-existing minor variants (Bonhoeffer and Nowak, 1997). In addition, the presence (or not) of the H275Y mutation in pre-treatment samples provides important information on the relative fitness of drug resistance mutations in the absence of oseltamivir.

To determine whether the H275Y mutation is present as a minor variant within hosts infected with influenza A virus, we performed ultra-deep sequencing of viral populations from nasopharyngeal specimens of two sets of individuals infected with A(H1N1)pdm09 viruses. First, we examined longitudinal samples collected from an immunocompromised child who remained infected for more than 6 weeks, during which time a drug-resistant strain came to dominate the virus population. Second, we analyzed the emergence of oseltamivir-resistant viruses in an household outbreak of A(H1N1)pdm09 infections in which the contact case developed influenza symptoms 24 hours after starting post-exposure oseltamivir prophylaxis (Baz et al., 2009).

Materials and Methods

Study 1: Immunocompromised Child

A 31-month-old boy weighing 13.4 kg, diagnosed three months earlier with medulloblastoma, was admitted on January 5, 2011, for consolidation chemotherapy in preparation for the first of 3 consecutive autologous bone marrow transplants (ABMT). On admission, the child presented rhinorrhea and mild cough

but was afebrile. Members of his immediate family, including his older sister and his father, had cold-like symptoms 1-2 weeks prior; none of the family members, including the patient, had received the 2010-11 influenza vaccine, the monovalent A(H1N1)pdm09 vaccine or any antiviral drug. A nasopharyngeal aspirate (NPA) collected on admission was positive for the A(H1N1)pdm09 virus by real-time RT-PCR (Semret M et al., 2009) and by viral culture on A549 and Mink lung cells. Treatment with oseltamivir (30 mg, twice daily) was started on January 6. The following day, the patient developed fever (max. 39.2°C), coincident with dropping neutrophil counts. The child received his first ABMT on January 10. NPA specimens collected throughout admission remained positive for A(H1N1) pdm09 influenza virus by RT-PCR (Table A4-1). Oseltamivir therapy was continued during the hospitalisation and after discharge on January 22. The patient was readmitted from January 27 to February 14, 2011 for his second ABMT. A NPA specimen collected on January 28 was positive for A(H1N1)pdm09 by RT-PCR. Because of persistent viral excretion, oseltamivir was replaced by zanamivir (25 mg inhaled four times daily) on February 1 and continued until negative RT-PCR results on February 17. The patient received a third ABMT on February 18 and he recovered from his influenza infection without complications.

Study 2: Transmission in Household

A detailed description of the familial cluster of infections with A(H1N1) pdm09 virus has been reported elsewhere (Baz et al., 2009). Briefly, a 13-year-old asthmatic male developed infection with A(H1N1)pdm09 confirmed by RT-PCR testing of a NPA. The child was started on oseltamivir (60 mg twice daily for 5 days) and discharged home the same day. Simultaneously to treatment of the index case, post-exposure oseltamivir prophylaxis (75 mg once daily for 10 days) was prescribed to the 59-year-old father with chronic obstructive pulmonary disease. Approximately 24 hours after beginning oseltamivir prophylaxis, the father developed influenza-like symptoms. On day 8 of oseltamivir prophylaxis, he consulted his general practitioner for persistent cough. An NPA collected at that time was positive by RT-PCR and by culture for A(H1N1)pdm09. The father had an uneventful clinical course, and an NPA sampled at the end of his illness was negative. The son's A(H1N1)pdm09 isolate collected before oseltamivir therapy was susceptible to oseltamivir (50% inhibitory concentration or IC_{50}: 0.27 nM), whereas the father's A(H1N1)pdm09 isolate was highly resistant to oseltamivir ($IC_{50} > 400$ nM). The complete (consensus sequence) virus genomes of the father (GenBank accession FN434454) differed by one amino substitution (H275Y) in the NA protein compared to the virus present in the son (GenBank FN434445).

TABLE A4-1 Virological testing of nasopharyngeal aspirates sampled from a young boy undergoing autologous bone marrow transplantation and infected with A(H1N1)pdm09 influenza. n.e. = not evaluated. CM1 and CM2 = Culture passages 1 and 2. S = Primary specimen (nasopharyngeal aspirates).

Sample	Antiviral therapy (43 days)	Multiplex real time diagnostic PCR (pH1N1)	Discriminatory real-time RT-PCR		Deep Sequencing				
			H275 copies/mL ± stdev (%)	Y275 copies/mL ± stdev (%)	H275 %	Y275 %	C.I.	Coverage #reads	Phenotypic drug susceptibility
05-01-2011—S1	None	Positive	$1.78\times10^{7} \pm 0.71\times10^{7}$ (99.91)	$1.44\times10^{4} \pm 0.29\times10^{4}$ (0.08)	95.7	3.7	1.7	488	n.e.
05-01-2011—CM2	None	Positive	$3.15\times10^{10} \pm 2.05\times10^{10}$ (99.99)	$3.16\times10^{5} \pm 3.41\times10^{5}$ (0.001)	96.9	2.7	0.7	1914	Susceptible to oseltamivir, zanamivir, and peramivir
10-01-2011—S2	Oseltamivir	Positive	$5.99\times10^{8} \pm 3.15\times10^{8}$ (99.75)	$1.53\times10^{6} \pm 1.26\times10^{6}$ (0.25)	94.6	4.4	1.0	1552	n.e.
17-01-2011—S3	Oseltamivir	Positive	$4.21\times10^{5} \pm 4.09\times10^{5}$ (3.13)	$1.30\times10^{7} \pm 6.57\times10^{6}$ (96.87)	2.3	97	0.8	1341	n.e.
20-01-2011—S4	Oseltamivir	Positive	$2.24\times10^{6} \pm 1.5610^{6}$ (4.08)	$5.26\times10^{7} \pm 3.67\times10^{7}$ (95.92)	3.3	96	1.0	1170	n.e.
20-01-2011—CM1	Oseltamivir	Positive	$6.25\times10^{5} \pm 5.73\times10^{5}$ (4.62)	$1.29\times10^{7} \pm 0.83\times10^{7}$ (95.38)	3.3	96.5	0.8	1838	n.e.
20-01-2011—CM2	Oseltamivir	n.e.	0	$9.40\times10^{9} \pm 3.24\times10^{9}$ (100.00)	n.e.	n.e.	n.e.	n.e.	Resistant to peramivir and oseltamivir, susceptible to zanamivir
28-01-2011—S5	Zanamivir	Positive	$2.85\times10^{4} \pm 1.15\times10^{4}$ (16.54)	$1.44\times10^{5} \pm 0.53\times10^{5}$ (83.46)	9.9	90	2.4	131	n.e.
08-02-2011—S6	Zanamivir	Positive	$8.23\times10^{5} \pm 4.98\times10^{5}$ (11.53)	$6.32\times10^{6} \pm 3.18\times10^{6}$ (88.47)	6.1	93.9	6.1*	66	n.e.
17-02-2011—S7	Zanamivir	Negative	0	0	n.e.	n.e.	n.e.	n.e.	n.e.

*Not significant with 95% confidence.

Informed Consent

Written consent was obtained for report of the case described in Study 1. Samples used in Study 2 were obtained as part of an investigation of the Public Health Department of the Ministry of Health, Quebec, Canada.

Clinical Specimens and Viral Culture

In Study 1 (immunocompromised child), 7 NPAs were collected between January 5 and February 17, 2011, for RT-PCR testing (Table A4-1 and Figure A4-1). Viral isolates were also obtained by culture from NPAs sampled on January 5 and January 20. In Study 2 (household transmission), the NPA from the index case (son) was collected prior to oseltamivir treatment whereas the NPA from his father was obtained on day 8 of oseltamivir prophylaxis (Figure A4-1).

NA Inhibition Assay

The drug resistance phenotype to NA inhibitors was determined by NA inhibition assays (Potier et al., 1979). The IC_{50} values were determined from the dose response curve. A virus was considered resistant to a drug if its IC_{50} value was 10-fold greater than that of the wild-type (WT) virus (Mishin et al., 2005).

RNA Extraction

Total RNA was extracted from 200 µL of thawed specimen or culture using the MagNA Pure instrument and the MagNA Pure LC total nucleic acid isolation kit (Roche Applied Science) according to the manufacturer's instructions and stored at –80°C.

Discriminative Real-time PCR Assay

To discriminate between WT and H275Y oseltamivir-resistant strains of A(H1N1)pdm09, a modified version of a previously reported real-time RT-PCR method (van der Vries et al., 2010) was used to test samples. This technique requires a reverse (panN1-H275-sense 5′–cagtcgaaatgaatgcccctaa-3′) and a forward (panN1-H275-antisense 5′–tgcacacacatgtgatttcactag-3′) primer for both the WT and the H275Y viruses and two labelled allele-specific probes: panN1-275H-probe (5′–ttaTCActAtgAggaatga-6-FAM/BHQ-1) and panN1-275Y-probe (5′–ttaTTActAtgAggaatga-HEX/BHQ-1). In the aforementioned probe sequences, locked nucleic acid (LNA) nucleotides are denoted in upper case, DNA nucleotides are denoted in lower case, and the single nucleotide polymorphism (SNP) is underlined. The limits of detection for the assay are 50 copies for the H275Y target and between 10 and 50 copies for the WT target. RT-PCR conditions are available upon request. Data acquisition was performed in both FAM and HEX filters

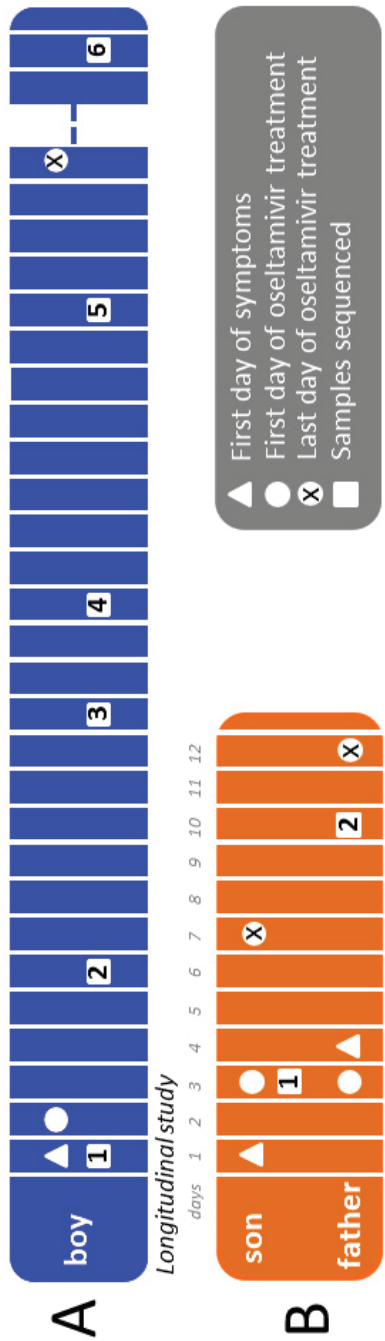


FIGURE A4-1 Outline of studies indicating day of onset, day when oseltamivir treatment was started, and sampling timeline. A) Study 1: Immunocomprised 31-month-old boy. B) Study 2: Son–father transmission.

during the annealing/extension step. Standard curves were constructed using 10-fold serial dilutions of pJET1.2-NA-Y275 and pJET1.2-NA-H275 plasmids.

Sequencing and analysis RNA isolated from two cultured isolates and seven primary specimens collected for Study 1 (Figure A4-1A), and two primary specimens for Study 2 (Figure A4-1B), was subjected to a multisegment RT-PCR (M-RT-PCR) step (Zhou et al., 2009) and random priming with barcoding using the SISPA (sequence independent single primer amplification) protocol (Djikeng et al., 2008). For each RNA sample, we performed two M-RT-PCR reactions using the One Step Superscript III RT kit (Invitrogen). Reactions were purified independently using the Qiagen MinElute kit and quantitated on a Nanodrop spectrophotometer; 100-200 ng of each purified M-RT-PCR reaction was used in two separate SISPA reactions with two different barcode tags for a total of 4 tagged reactions per original RNA sample. Products were then separated on a 1% agarose gel and fragments from 200–400bp purified with the Qiagen MinElute kit. Pooled samples were sent for paired end (PE) library preparation and 100 base sequencing on the Illumina Hi-Seq2000 platform.

The barcoded amplification products were sequenced on one lane of the sequence run. Analyses were performed to reduce the distortion caused by SISPA amplification, account for both PCR and sequencing errors, and provide a "clean" comparison between the mapped reads of the experimental samples. The trimmed reads were mapped to A/Quebec/144147/2009(H1N1) (GenBank accession FN434457-FN434464) using the bowtie short-read aligner (Langmead et al., 2009).

The frequency of each codon observed in the set of mapped reads from each amplification replicate was tabulated across each of the 10 influenza genes. To account for sequence-specific errors (Minoche et al., 2011; Nakamura et al., 2011), the variant counts for the forward and reverse direction reads were calculated separately, and only those variants for which counts were within 50% of each other in both directions were retained. For these summaries, the unique reads from all amplification replicates were pooled and total coverage is reported for each codon site. The proportion of codons expected to differ from the consensus due to background mutation and technical error was estimated from a separate cell culture of the PR8 strain that was otherwise processed in exactly the same manner as the specimens in this study. This proportion, found to be 0.00392, lies well outside of the 95% confidence interval for any variant codon in our study that is (a) represented by more than 4 sequence reads, and (b) found in at least 2% of all sequence reads mapped to that position. The lower limit of the 95% confidence interval determined by computing the inverse of the appropriate cumulative Beta distribution is 0.00813.

Results

Presence of Drug-Resistant Viruses Before Drug Treatment in an Immunocompromised Child (Study 1)

The results of the NA gene H275Y discriminatory real-time RT-PCR assay performed on the seven primary specimens and the two viral isolates (January 5 and 20) are presented in Table A4-1. In the first NPA collected on January 5 (day 1), prior to antiviral therapy (initiated January 6), 99.9% of the viral population was WT at NA position 275 by our discriminatory assay. Nevertheless, a very small sub-population of H275Y mutant was also detectable (0.08%). The corresponding viral isolate (05-01-2011—CM2 in Table 1) contained 99.9% of WT virus and was susceptible to oseltamivir (IC_{50}=0.77 nM ± 0.02), zanamivir (IC_{50}=0.15 nM ± 0.02), and peramivir (IC_{50}=0.05 nM ± 0.01). Notably, the H275Y mutation could not be detected by conventional RT-PCR and Sanger sequencing in the original sample. A second NPA collected on January 10 (day 6) also demonstrated a predominance of the WT population (99.8%). However, the proportion of the H275Y mutant detected in NPAs collected on January 17, 20, and 28 increased to 96.9, 95.9, and 83.5%, respectively, during continuous oseltamivir treatment. Furthermore, the second passage on Madin Darby canine kidney (MDCK) cells of the January 20 viral isolate (20-01-2011-CM2 in Table A4-1) resulted in 100% H275Y mutant population compared with 95.4% from the primary culture recovered from A549 and Mink lung cells. This viral isolate exhibited an IC_{50} value of 556.75 nM ± 61.32 for oseltamivir, 0.22 nM ± 0.01 for zanamivir, and 34.81 nM ± 5.77 for peramivir, which indicates a resistance phenotype to oseltamivir and peramivir. Antiviral therapy was changed to zanamivir on February 1st. The February 8 sample contained a predominance of 88.5% of H275Y mutant virus, whereas the last NPA collected on February 17 was negative for A(H1N1)pdm09 by RT-PCR.

A number of the primary specimens (January 5, 10, 17, 20, 28, and February 8; corresponding to samples 1-6 in Figure A4-1A) for which M-RTPCR product could be generated, as well as the viral isolates, were subjected to deep sequencing to better evaluate the genetic diversity of the viral population, including the presence of drug-resistant mutants. Based on the average depth of coverage across each of the virus segments, we highlighted codons represented by at least 2% of the sequence reads covering each position (Table S1). This percentage is conservative enough that, even in low-coverage areas, it excludes potential sequence and PCR errors.

The positions on the NA and NS1 proteins that display evidence for the presence of minor variants at a frequency of 2% or above in more than one of the samples are shown in Figure A4-2. Similar patterns are observed for all other proteins (Table S1). Over time, the ratios of the minor variants to the dominant codon remain relatively stable, except for NA position 275 where a shift of H to Y is apparent on 17 January 2011. The ratios are similar to the ones observed in

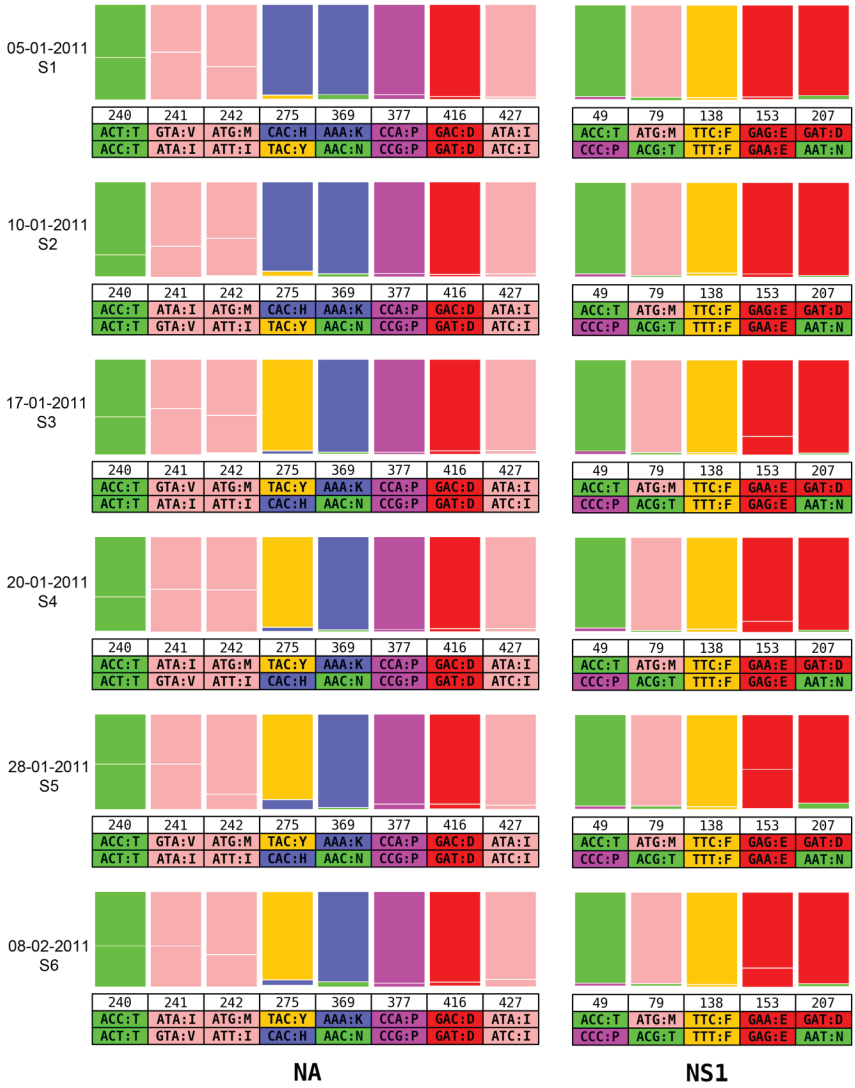


FIGURE A4-2 Longitudinal study of variant codon prevalence across multiple time-points in an infected immunocompromised child. Ratios of major and minor codons are represented at each position where the variant codons appear in more than 2% of the deep sequence data reads in at least two of the time-points. Data collection dates are represented on the left hand side and each group corresponds to the positions where variant residues are observed in the NA and NS1. Codons and single letter amino acid codes are indicated below the position number.

the real-time discriminatory RT-PCR assays for each of the samples tested (Table A4-1), although values across both assays are not identical. No other position on the NA protein appears to co-vary with the 275Y variant. The same pattern is observed in the culture isolates (05-01-2011-CM2 and 20-01-2011-CM1 in Table S2). However, position 153 in NS1 displays a similar switch, although involving a synonymous mutation (from codon GAG to GAA, for E (glutamic acid)). Hence, the sample from the original infection contained a drug-associated minor variant *prior* to the onset of drug treatment, and this minor variant differed from the dominant strain by only two nucleotide positions. Due to drug-associated selection pressure, this minor variant eventually became dominant in the host. The variant codons observed at the other positions are also possibly representative of other minor variants in the original virus population but, as they remained minor members of the viral population, they are unlikely to have a selective advantage.

Evidence for Transmission of Drug-Resistant Viruses in Household (Study 2)

In a separate study, we observed a similar phenomenon where oseltamivir resistance emerged quickly in the household contact (father) of an index case (son). Both family members were started on oseltamivir on the same day (Figure A4-1B) i.e., twice a day treatment for the son and once a day prophylaxis for the father. The latter developed influenza-like symptoms 24 hours after drug treatment was begun. Such a rapid clinical presentation suggests that he was already infected at the time prophylaxis was initiated, and that drug-resistant viruses were most likely already present.

We characterized the genetic diversity of the virus populations in both individuals by deep sequencing. An example for the HA and NA genes where most of the variants seen in the son are also observed in the father is shown in Figure A4-3. While the dominant viruses are drug-sensitive in the son and drug-resistant in the father, apparent by the switch from H275 to 275Y, it is striking that a minor population of viruses in the son already carries the drug resistance mutation; minor drug resistant variant residue 275Y is present in more than 2.4% of the reads in the son (which was not detected by conventional RT-PCR and Sanger sequencing). Hence, it is likely that viruses carrying this mutation were transmitted to the father along with drug sensitive viruses, and became dominant in that individual following selection associated with a subtherapeutic (prophylactic) dose of oseltamivir.

Also of note was that the same minor variants were found in both the father and the son at 60 residue positions across all 10 viral proteins (Table S3). We estimate that there were 8 days of replication in the father from the time he was possibly infected by the son (assuming it occurred 24 hours before any symptoms) to the time the specimen was collected. Over that time, variant representation could have fluctuated such that the set of 60 variants seen in both samples is likely to underestimate the true number. While the number of conserved variants points

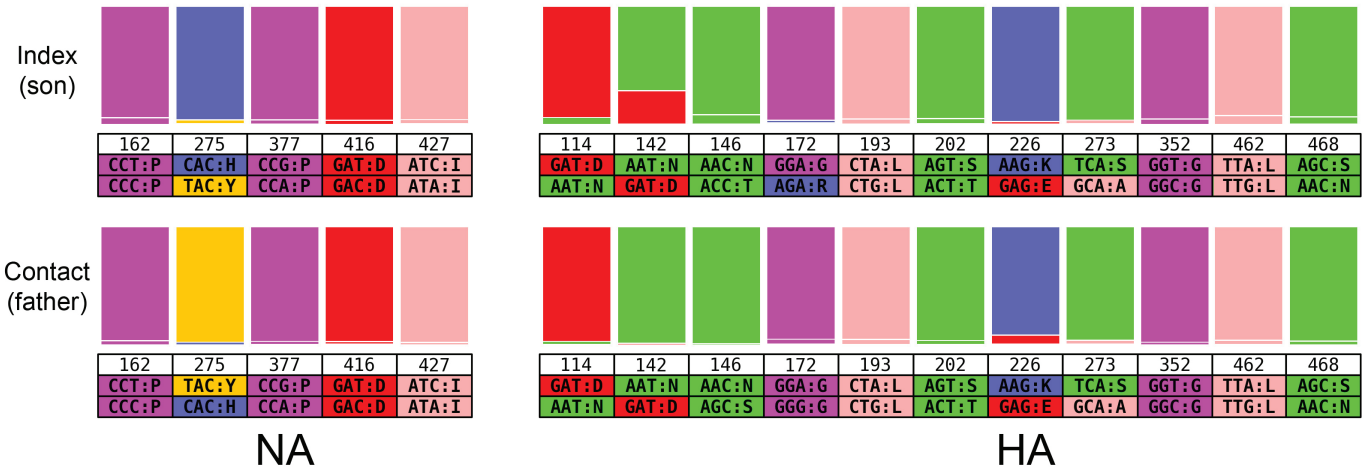


FIGURE A4-3 Transmission study of variant codon prevalence compared between son and father specimens. Ratios of major and minor codons are represented at each position of the neuraminidase (NA) and hemagglutinin (HA) where the variant codons appear in more than 2% of the deep sequence data reads in at least one of the samples. Codons and single letter amino acid codes are indicated below the position number.

to possible transmission, and the probability that the same variants could appear in both the son and the father by chance alone is extremely low, we do not have other potential contacts or index cases to test in order to confirm this observation.

Discussion

The most striking observation from both of these studies is that the mutation most commonly associated with resistance to oseltamivir (H275Y) is present in the viral population of some individuals *prior* to the onset of drug treatment. In addition, this minor drug-resistant population could not be revealed by conventional methods such as phenotypic resistance tests and Sanger sequencing. This observation is important for a number of reasons. First, the prior existence of Y275 means that the selection for drug resistance will proceed much more rapidly following the onset of drug selection pressure than if only wild-type viruses are present in the population, as there is no waiting time for the correct mutation to appear (Bonhoeffer and Nowak, 1997). Further, that the Y275 mutation is present in untreated hosts indicates that this mutation is not strongly deleterious in the absence of oseltamivir, and likely does not need compensatory mutations to enable its fixation (Hamelin et al., 2010; Memoli et al., 2010; Seibert et al., 2010). Indeed, in both cases studied here, we observed no amino acid changes that were fixed concordant with Y275, and only a single synonymous mutation (in NS1) in the case of the immunocomprised child. In these circumstances, the pre-existence of Y275 means that oseltamivir resistance will likely spread rapidly as soon as there is drug selection pressure, especially in immunocompromised individuals and when suboptimal antiviral dosage is used.

If the Y275 mutation is present in individual hosts prior to the onset of drug treatment then it is also likely to have been transmitted between individuals as a minor variant. This in turn suggests that there may not often be a severe population bottleneck during the inter-host transmission of influenza virus. Indeed, mixed infections of multiple variants of influenza virus have been observed in both natural human infections (Ghedin et al., 2009; Ghedin et al., 2011; Pajak et al., 2011) and experimental animal infections (Murcia et al., 2010; Murcia et al., 2012), and hence may be commonplace. Co-infection with major and minor variants, captured by deep sequencing, was also observed during the course of human rhinovirus infections (Cordey et al., 2010), indicating that this phenomenon is not unique to influenza. In contrast, sequencing studies of HIV suggest that a small number of viral particles initiate infection, such that most variants are produced following replication within the newly infected host (Keele et al., 2008).

Such transmission of multiple variants is most clearly documented in the son-father case, where perhaps 60 mutational variants are passed between these two individuals, one of which confers oseltamivir resistance. However, the availability of only short sequence reads makes it impossible to determine the exact number of distinct viral haplotypes these correspond to. In addition, our sampling

protocol in the son-father transmission case dictates that we cannot exclude that there was rapid selection of oseltamivir resistance in the son after we sampled his virus population, such that a *majority* Y275 population was in fact transmitted to the father. However, this would entail extremely rapid selection for resistance and does not change the central observation that multiple variants are transmitted between hosts as both H275 and Y275 are found in the father.

That the Y275 mutation is present in the son prior to oseltamivir treatment and so soon after symptom onset suggests that this resistance mutation was also present in the viral population initially transmitted to the son. Similarly, the presence of Y275 in the immunocomprised child suggests that this mutation may have been transmitted to the child in a mixed infection containing both drug-sensitive and -resistant mutations, although it cannot be excluded that the variant appeared *de novo*. If Y275 is indeed present in the founding population in both individuals then it is possible that this mutation is present as a low frequency variant in many individuals infected with A(H1N1)pdm09, and that its presence reflects the combined action of past selection for drug resistance in patients receiving oseltamivir, incomplete reversion to the wild-type H275 mutation in patients that are not on the drug, and a lack of strongly deleterious fitness effects in the absence of drug. The large-scale ultra-deep sequencing of additional A(H1N1)pdm09 patients who have not received oseltamivir will clearly be central to answering this question.

Next generation ultra-deep sequencing of intra-host viral populations such as that undertaken here promises to transform our understanding of the evolution of drug resistance in acute viral infections, allowing the dissection of the mutational spectrum at an unprecedented level of precision. Indeed, it is striking that in the two cases conventional RT-PCR failed to detect the presence of oseltamivir resistance even though Y275 was present in the viral population. However, despite its undoubted potential, ultra-deep sequencing also comes with a number of inherent analytical difficulties. First, because the sequencing protocol leads to the generation of short sequence reads, nucleotide positions cannot be linked either within or among individual genes except if they are close enough to appear on the same sequence read, or if they have the same pattern of prevalence. More fundamentally, it is critical to ensure that minor genetic variants are not the result of PCR/sequencing artefacts. Amplification leads to the well-known problem of "PCR duplicates," sometimes resulting in severe distortion to the observed proportions of true variant subpopulations and the possible creation of false variant sequences through PCR errors. To address these problems, each specimen from our study was amplified in four independent reactions using different barcodes, allowing us to track amplification products and their respective sequence reads. Future work will employ a simpler and more cost-effective approach using modified primers that include unique tags for each template (Jabara et al., 2011).

References

Baz, M., Y. Abed, J. Papenburg, X. Bouhy, M. E. Hamelin, and G. Boivin. 2009. Emergence of oseltamivir-resistant pandemic H1N1 virus during prophylaxis. *New England Journal of Medicine* 361(23):2296-2297.

Bonhoeffer, S., and M. A. Nowak. 1997. Pre-existence and emergence of drug resistance in HIV-1 infection. *Proceedings: Biological Sciences* 264(1382):631-637.

Cordey, S., T. Junier, D. Gerlach, F. Gobbini, L. Farinelli, E. M. Zdobnov, B. Winther, C. Tapparel, and L. Kaiser. 2010. Rhinovirus genome evolution during experimental human infection. *PloS One* 5(5):e10588.

Djikeng, A., R. Halpin, R. Kuzmickas, J. Depasse, J. Feldblyum, N. Sengamalay, C. Afonso, X. Zhang, N. G. Anderson, E. Ghedin, and D. J. Spiro. 2008. Viral genome sequencing by random priming methods. *BMC Genomics* 9:5.

Garten, R. J., C. T. Davis, C. A. Russell, B. Shu, S. Lindstrom, A. Balish, W. M. Sessions, X. Xu, E. Skepner, V. Deyde, M. Okomo-Adhiambo, L. Gubareva, J. Barnes, C. B. Smith, S. L. Emery, M. J. Hillman, P. Rivailler, J. Smagala, M. de Graaf, D. F. Burke, R. A. Fouchier, C. Pappas, C. M. Alpuche-Aranda, H. Lopez-Gatell, H. Olivera, I. Lopez, C. A. Myers, D. Faix, P. J. Blair, C. Yu, K. M. Keene, P. D. Dotson, Jr., D. Boxrud, A. R. Sambol, S. H. Abid, K. St George, T. Bannerman, A. L. Moore, D. J. Stringer, P. Blevins, G. J. Demmler-Harrison, M. Ginsberg, P. Kriner, S. Waterman, S. Smole, H. F. Guevara, E. A. Belongia, P. A. Clark, S. T. Beatrice, R. Donis, J. Katz, L. Finelli, C. B. Bridges, M. Shaw, D. B. Jernigan, T. M. Uyeki, D. J. Smith, A. I. Klimov, and N. J. Cox. 2009. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* 325(5937):197-201.

Ghedin, E., A. Fitch, A. Boyne, S. Griesemer, J. DePasse, J. Bera, X. Zhang, R. A. Halpin, M. Smit, L. Jennings, K. St George, E. C. Holmes, and D. J. Spiro. 2009. Mixed infection and the genesis of influenza virus diversity. *Journal of Virology* 83(17):8832-8841.

Ghedin, E., J. Laplante, J. DePasse, D. E. Wentworth, R. P. Santos, M. L. Lepow, J. Porter, K. Stellrecht, X. Lin, D. Operario, S. Griesemer, A. Fitch, R. A. Halpin, T. B. Stockwell, D. J. Spiro, E. C. Holmes, and K. St George. 2011. Deep sequencing reveals mixed infection with 2009 pandemic influenza A (H1N1) virus strains and the emergence of oseltamivir resistance. *Journal of Infectious Diseases* 203(2):168-174.

Hamelin, M. E., M. Baz, Y. Abed, C. Couture, P. Joubert, E. Beaulieu, N. Bellerose, M. Plante, C. Mallett, G. Schumer, G. P. Kobinger, and G. Boivin. 2010. Oseltamivir-resistant pandemic A/H1N1 virus is as virulent as its wild-type counterpart in mice and ferrets. *PLoS Pathogens* 6(7):e1001015.

Jabara, C. B., C. D. Jones, J. Roach, J. A. Anderson, and R. Swanstrom. 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer id. *Proceedings of the National Academy of Sciences of the United States of America* 108(50):20166-20171.

Keele, B. F., E. E. Giorgi, J. F. Salazar-Gonzalez, J. M. Decker, K. T. Pham, M. G. Salazar, C. Sun, T. Grayson, S. Wang, H. Li, X. Wei, C. Jiang, J. L. Kirchherr, F. Gao, J. A. Anderson, L. H. Ping, R. Swanstrom, G. D. Tomaras, W. A. Blattner, P. A. Goepfert, J. M. Kilby, M. S. Saag, E. L. Delwart, M. P. Busch, M. S. Cohen, D. C. Montefiori, B. F. Haynes, B. Gaschen, G. S. Athreya, H. Y. Lee, N. Wood, C. Seoighe, A. S. Perelson, T. Bhattacharya, B. T. Korber, B. H. Hahn, and G. M. Shaw. 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* 105(21):7552-7557.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10(3):R25.

Memoli, M. J., R. J. Hrabal, A. Hassantoufighi, B. W. Jagger, Z. M. Sheng, M. C. Eichelberger, and J. K. Taubenberger. 2010. Rapid selection of a transmissible multidrug-resistant influenza a/h3n2 virus in an immunocompromised host. *Journal of Infectious Diseases* 201(9):1397-1403.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer. 2011. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome Biology* 12(11):R112.

Mishin, V. P., F. G. Hayden, and L. V. Gubareva. 2005. Susceptibilities of antiviral-resistant influenza viruses to novel neuraminidase inhibitors. *Antimicrobial Agents and Chemotherapy* 49(11):4515-4520.

Murcia, P. R., G. J. Baillie, J. Daly, D. Elton, C. Jervis, J. A. Mumford, R. Newton, C. R. Parrish, K. Hoelzer, G. Dougan, J. Parkhill, N. Lennard, D. Ormond, S. Moule, A. Whitwham, J. W. McCauley, T. J. McKinley, E. C. Holmes, B. T. Grenfell, and J. L. Wood. 2010. Intra- and interhost evolutionary dynamics of equine influenza virus. *Journal of Virology* 84(14):6943-6954.

Murcia, P. R., J. Hughes, P. Battista, L. Lloyd, G. J. Baillie, R. H. Ramirez-Gonzalez, D. Ormond, K. Oliver, D. Elton, J. A. Mumford, M. Caccamo, P. Kellam, B. T. Grenfell, E. C. Holmes, and J. L. N. Wood. 2012. Evolution of an eurasian avian-like influenza virus in native and vaccinated pigs. *PLoS Pathogens* 8(5):e1002730.

Nakamura, K., T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara, and S. Kanaya. 2011. Sequence-specific error profile of illumina sequencers. *Nucleic Acids Research* 39(13):e90.

Pajak, B., I. Stefanska, K. Lepek, S. Donevski, M. Romanowska, M. Szeliga, L. B. Brydak, B. Szewczyk, and K. Kucharczyk. 2011. Rapid differentiation of mixed influenza A/H1N1 virus infections with seasonal and pandemic variants by multitemperature single-stranded conformational polymorphism analysis. *Journal of Clinical Microbiology* 49(6):2216-2221.

Pizzorno, A., Y. Abed, and G. Boivin. 2011a. Influenza drug resistance. *Seminars in Respiratory and Critical Care Medicine* 32(4):409-422.

Pizzorno, A., X. Bouhy, Y. Abed, and G. Boivin. 2011b. Generation and characterization of recombinant pandemic influenza A(H1N1) viruses resistant to neuraminidase inhibitors. *Journal of Infectious Diseases* 203(1):25-31.

Potier, M., L. Mameli, M. Belisle, L. Dallaire, and S. B. Melancon. 1979. Fluorometric assay of neuraminidase with a sodium (4-methylumbelliferyl-alpha-D-N-acetylneuraminate) substrate. *Analytical Biochemistry* 94(2):287-296.

Seibert, C. W., M. Kaminski, J. Philipp, D. Rubbenstroth, R. A. Albrecht, F. Schwalm, S. Stertz, R. A. Medina, G. Kochs, A. Garcia-Sastre, P. Staeheli, and P. Palese. 2010. Oseltamivir-resistant variants of the 2009 pandemic H1N1 influenza a virus are not attenuated in the guinea pig and ferret transmission models. *Journal of Virology* 84(21):11219-11226.

Semret, M., S. Fenn, H. Charest, J. McDonald, C. Frenette, and V. Loo. 2009 (18-21 June). *A real-time RT-PCR assay for detection of influenza H1N1 (swine-type) and other respiratory viruses.* Paper presented at 26th International Congress of Chemotherapy and Infection, Toronto, ON.

van der Vries, E., M. Jonges, S. Herfst, J. Maaskant, A. Van der Linden, J. Guldemeester, G. I. Aron, T. M. Bestebroer, M. Koopmans, A. Meijer, R. A. Fouchier, A. D. Osterhaus, C. A. Boucher, and M. Schutten. 2010. Evaluation of a rapid molecular algorithm for detection of pandemic influenza A (H1N1) 2009 virus and screening for a key oseltamivir resistance (H275Y) substitution in neuraminidase. *Journal of Clinical Virology* 47(1):34-37.

WHO. 2011. Weekly update on oseltamivir resistance in influenza A(H1N1)2009 viruses.

Zhou, B., M. E. Donnelly, D. T. Scholes, K. St George, M. Hatta, Y. Kawaoka, and D. E. Wentworth. 2009. Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and swine origin human influenza a viruses. *Journal of Virology* 83(19):10309-10313.

A5

DESIGN CONSIDERATIONS FOR HOME AND HOSPITAL MICROBIOME STUDIES

Daniel P. Smith,[14] *John C. Alverdy*,[15]
Jeffrey A. Siegel,[16,17] *and Jack A.* ***Gilbert***[14,18]

Milestones in Home and Hospital Microbiome Research

The populations of developed nations spend approximately 90 percent of their time indoors (Moschandreas, 1981), leading scientists and the public alike to take an interest in the microbial communities that share these spaces with us. This is especially true in healthcare environments, where hospital-acquired infections (HAIs) have long been among the leading causes of patient deaths (Anderson and Smith, 2005; Groseclose et al., 2004; Hall-Baker et al., 2010; Klevens et al., 2007). The first study of airborne pathogens in a hospital can be attributed to Bourdillon and Colebrook, who, in 1946, investigated the concentration of bacteria present in the air of a surgical changing room (Bourdillon and Colebrook, 1946). Their findings, and the findings of similar studies published over the following two decades, revealed levels of airborne bacteria that were cause for concern (Blowers and Wallace, 1960; Colebrook and Cawston, 1948; Cvjetanović, 1957; Greene et al., 1962a,b; Warner and Glassco, 1963) and prompted a rethinking of ventilation designs for hospitals.

The air-sampling techniques developed for hospitals were soon applied to other indoor spaces including subway trains (Williams and Hirch, 1950), classrooms (Williams et al., 1956), movie theaters (Cvjetanović, 1957), and apartments (Simard et al., 1983). Articles by Finch and colleagues in 1978 and Scott and colleagues in 1982 complemented these air-based studies with the first characterizations of bacteria living on bathroom and kitchen surfaces (Finch et al., 1978; Scott et al., 1982). The larger of the two studies, conducted by Scott and colleagues, examined 60 locations in 251 homes, and agreed well with the conclusions from an earlier study by Finch and colleagues of 21 homes that the dominant species on the studied surfaces were enterobacteria, Pseudomonads, micrococci, *Bacillus*, and *Aeromonas hydrophila*, with a lower incidence of *Salmonella*, *Staphylococcus aureus*, and *Bacillus cereus*.

[14] Argonne National Laboratory, Institute for Genomic and Systems Biology, Argonne, IL, USA.
[15] Department of Surgery, The University of Chicago Medical Center, Chicago, IL, USA.
[16] Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX, USA.
[17] Department of Civil Engineering, The University of Toronto, Toronto, ON, Canada.
[18] Department of Ecology and Evolution, The University of Chicago, Chicago, IL, USA.

Current literature regarding the relationship between the indoor environment and humans primarily explores the development of fungal contamination with damp surfaces (Hyvarinen et al., 2002; Jaffal et al., 1997; Lignell et al., 2008; Nevalainen and Seuri, 2005), the role of hygiene in removing microbial communities (Bright et al., 2009; Grice and Segre, 2011), and the length of time microbes can survive on surfaces (Kramer et al., 2006). There have been a number of studies to explore the microbial diversity of communities associated with dust (Pitkäranta et al., 2008; Rintala et al., 2008; Sebastian and Larsson, 2003) and air (Huttunen et al., 2008; Tringe et al., 2008). One study that investigated temporal succession of microbial communities performed on indoor dust found seasonal patterns, and these were building specific, probably as a result of skin cells shed from inhabitants within the buildings (Rintala et al., 2008). These existing studies demonstrate fundamental principles regarding experimental design, explicitly regarding the types of environmental conditions that need to be monitored (e.g., surface material, moisture, HVAC system), and the observation that the architectural design of an indoor space influences the potential community structure and hence human health (Guenther and Vittori, 2008). The influence of air ventilation and the number of people in a space must be explored with regard to the impact on microbial community structure (Hospodsky et al., 2012; Kembel et al., 2012; Qian and Li, 2010; Qian et al., 2012). Additionally, the variability associated with body sites (Costello et al., 2009; Fierer et al., 2010; Grice et al., 2009) will have a major impact on the interpretation of the analyses, because different body sites interact with different surfaces differently. The most diverse skin sites are the driest areas and hence are less likely to be transferred to a surface with sebaceous exudates. This will affect the time that the microbial community maintains structural cohesion with reference to relative abundance of members on a surface (Kramer et al., 2006).

Studies of indoor microbiology are highly relevant in today's age, because concern for protecting ourselves from microbial pathogens is ever-present. Studies in this field do much to put this threat in perspective by identifying incorrect preconceptions. For instance, using cleaning products containing the antibacterial agent triclosan over the course of a year does not result in an increase in antimicrobial drug-resistant bacteria in homes (Aiello et al., 2005; Cole et al., 2003). Several research groups have studied the beneficial effect of childhood exposure to dirty environments, which is particularly pronounced in the inverse correlation between children who live on farms and their reduced likelihood of later developing asthma and other respiratory problems (Adler et al., 2005; Alfvén et al., 2006; Klintberg et al., 2001; Leynaert et al., 2001; Merchant et al., 2005; Remes et al., 2005; Riedler et al., 2001; Schram et al., 2005). Hospitals have found that by opening the windows in patient rooms, the percentage of potentially pathogenic microbes in the air is significantly reduced (Escombe et al., 2007; Kembel et al., 2012). These are just a few examples of a larger trend toward questioning the culture of cleanliness, or as its come to be called—the hygiene hypothesis

(Bloomfield et al., 2006; Martinez, 2001; Rook, 2009; Rook and Stanford, 1998; Yazdanbakhsh and Matricardi, 2004). The study of pathogens in indoor environments is valuable, but perhaps even more valuable is gaining a better understanding of the competition between pathogenic and non-pathogenic microorganisms and how we might shift that balance in our favor.

A New Scientific Community

The expansion of indoor microbiome research from healthcare environments into homes and offices has been driven in large part by funding initiatives by the Alfred P. Sloan Foundation. This private agency stipulates a high degree of collaboration between its grantees and as a result has brought about the development of Internet portals designed to enable any indoor microbiome researcher to exchange raw data and results with other scientists as well as with reporters and the general public. One such nexus, the Microbiology of the Built Environment Network website,[19] tracks investigators, projects, publications, computational resources, protocols, standards, press releases, social media, conferences, and workshops relating to indoor microbiology. A community data archive is hosted by the Microbiome of the Built Environment Data Analysis Core (MoBEDAC), which integrates with data analysis tools including visualization and analysis of microbial population structures (VAMPS), quantitative insights into microbial ecology (QIIME), meta-genome rapid annotation using system technology (MG-RAST), and FungiDB (Caporaso et al., 2010; Meyer et al., 2008; Stajich et al., 2011).

While microBE.net provides general information applicable to a wide range of indoor microbiology projects, working groups for specialists in this field have also emerged. The Berkeley Indoor Microbial Ecology Research Consortium (BIMERC) focuses on identifying the source populations and human influences on the microbial components of indoor air. Discovering the mechanisms and rates with which microbial communities spread throughout healthcare facilities is the goal of the Hospital Microbiome Consortium.[20] The Biology and the Built Environment Center at the University of Oregon[21] has begun training students in the investigation of how architecture can influence the structure of indoor microbiomes.

Indoor Microbiology Without Culturing?

Early studies of the indoor microbiome relied heavily on culture-based methods: agar plates were commonly pressed directly against the surface of interest,

[19] http://www.microBE.net.
[20] www.hospitalmicrobiome.com.
[21] BioBE; http://biobe.uoregon.edu.

and the resultant colonies were counted and microscopically characterized in order to establish quantitative and taxonomic abundance metrics. However, with the introduction of next-generation sequencing technologies, rapid, high-throughput characterization of taxonomic marker genes (e.g., 16S/18S rRNA) and whole genomic DNA from environmental samples is now financially viable, altering the landscape of tools available to the researcher interested in the indoor microbiome.

The Pros and Cons of High-Throughput Sequencing

Since 2007, high-throughput sequencing technologies offered by Illumina, Roche, and ABI have enabled researchers to quickly and inexpensively profile the relative abundances of taxonomic groups in samples of environmental microbial communities. This approach to microbial ecology offers three advantages over culture-based methods. First, uncultivable species can be identified, thereby providing a more complete characterization of the microbial community. Second, organisms can be systematically classified by computer-aided alignment of DNA sequences to reference genes and genomes. And lastly, this entire process is relatively easily scalable from tens of samples to tens of thousands of samples.

Classical culturing, however, still yields important information that cannot be attained through high-throughput DNA sequencing of ribosomal genes. Measuring the absolute abundance of colony-forming units is very straightforward when working with agar press-plates, but it is a difficult metric to attain from nanogram quantities of DNA extracted from environmental samples. Growth of colonies on plates also provides concrete evidence that the cell taken from the environment was viable. In addition, sequencing ribosomal genes does not provide information on whether the detected microbial species harbor genetic cassettes encoding antibiotic-resistance genes, an important factor in evaluating the pathogenicity of microorganisms. The ability to retain bacterial colonies for further studies is a third advantage of plate-based culturing, allowing one to subject any detected species to thorough examination. In light of these considerations, a study seeking to characterize both currently uncultivable microorganisms and antibiotic-resistant human pathogens would need to draw on both classical and next-generation microbial community analysis techniques.

Microarrays for Rapid Identification of Antibiotic Resistance

A new approach to rapidly screening microbial communities for multiple antibiotic resistance markers has recently been developed by Taitt et al. (2012). Their Antimicrobial Resistance Determinant Microarray chips are designed to detect more than 250 resistance genes covering 12 classes of antibiotics and have been shown to be compatible with low concentrations of DNA extracted from swabs (personal communication). Advances such as this are quickly closing the

gap between the high-throughput sequencing technologies and classical culture-based phenotyping.

Sample and Metadata Collection

The selection of sampling locations and environmental parameters to monitor is fundamental to any microbiome project. Balancing the comprehensiveness of an investigation against logistical constraints has led to studies that examine highly specific aspects of the indoor microbiome (Hilton and Austin, 2000; Kelley et al., 2004; Kembel et al., 2012; Kopperud et al., 2004; Krogulski and Szczotko, 2011; Tang, 2009; Wiener-Well et al., 2011). As might be expected, the sampling locations that these studies chose are those that humans most come into contact with on a daily basis. A list of these locations for homes (Table A5-1) and hospitals (Table A5-2) are provided below.

Air

The air within built environments is arguably one of the most important mediums to consider when investigating the interaction between humans and the indoor microbiome. Air inside of buildings is biologically distinct from outdoor air, containing a greater proportion of human-associated microflora shed by its occupants (Bouillard et al., 2005; Clark, 2009; Fox et al., 2010; Hospodsky et al., 2012; Korves et al., 2012; Noble et al., 1976; Noris et al., 2011; Qian et al., 2012; Rintala et al., 2008; Täubel et al., 2009). Low air exchange rates and recycling of air for conditioning purposes can exacerbate the negative effects of aerosolized microorganisms that have effects on human health due to pathogenic, toxic, and/or allergic properties (D'Amato et al., 2005; Monto, 2002; Peccia et al., 2008; Pope et al., 1993).

Airborne particles are commonly collected using one of four methods: settle plates, impactors, impingers, and filtration, each of which offer differing advantages and efficiencies (Fahlgren et al., 2010; Griffin et al., 2010; Morrow et al.,

TABLE A5-1 Home High-Touch Surfaces and Bacterial Reservoirs

All rooms	Light switches, air, dust, floor, rugs, door knobs
Kitchen	Countertop, sink, faucet handles, drain, u-pipe, refrigerator handle, refrigerator shelves, microwave buttons, dish sponge, drying towels, drying rack
Bathroom	Countertop, sink, u-pipes, shower floor, shower curtain, showerhead, shower poufs, bar soap, toilet bowl, toilet water, toilet seat, toilet flush handle, hand towels
Bedroom	Pillows, sheets
Living room	Seats, arm rests, head rests, pillows, blankets, remote controls
Office, etc.	Keyboard, mouse, water from water heater, mop head, HVAC filters

TABLE A5-2 Hospital High-Touch Surfaces and Bacterial Reservoirs

Patient area	Bed rails, tray table, call boxes, telephone, bedside tables, patient chair, IV pole, floor, light switches, glove box, air, air exhaust filter
Patient restroom	Sink, faucet handles, inside faucet head, hot tap water, cold tap water, light switches, door knob, handrails, toilet seats, flush lever, bed pan cleaning equipment, floor, air, air exhaust filter
Additional equipment	IV Pump control panel, monitor control panel, monitor touch screen, monitor cables, ventilator control panel, blood pressure cuff, janitorial equipment
Water	Cold tap water, hot tap water, water used to clean floors
Patient	Stool sample, nasal swab, hand
Staff	Nasal swab, bottom of shoe, dominant hand, cell phone, pager, iPad, computer mouse, work phone, shirt cuff, stethoscope
Travel areas	Corridor floor, corridor wall, steps, stairwell door knobs, stairwell door kick plates, elevator buttons, elevator floor, handrails, air
Lobby	Front desk surface, chairs, coffee tables, floor, air
Public restroom	Floor, door handles, sink controls, sink bowl, soap dispenser, towel dispenser, toilet seats, toilet lever, stall door lock, stall door handle, urinal flush lever, air, air exhaust filter

2012). Settle plates offer a silent and inexpensive option for enumerating colony-forming units deposited by gravity onto agar petri dishes. Settle plates will preferentially sample larger particles, because they are more likely to be deposited by gravitational settling. Impactors increase the rate and control of particle deposition by accelerating air in an arc relative to the agar surface, utilizing centrifugal forces to select for a specific range of particle masses. The same design principle is used by impingers, where the deposition media is liquid rather than solid state. However, the mechanical stresses introduced by impactors and impingers can rupture cellular membranes, thereby reducing culturing viability. Filtration of air through a porous membrane is less mechanically stressful to cells, but may result in desiccation. Although active samplers—impactors, impingers, and filtering devices—offer the added benefit of providing a quantitative accounting of the volume of air sampled, the noise generated by the unit's pump or fan may preclude their use in occupied buildings. Filters from central HVAC systems have also been used in lieu of portable sampling units (Bonetta et al., 2009; Drudge et al., 2011; Farnsworth et al., 2006; Hospodsky et al., 2012; Korves et al., 2012; Noris et al., 2009, 2011; Stanley et al., 2008).

Water

Municipal water supplies have long been known to contain biofilm-forming and planktonic microorganisms, including *Mycobacterium* and *Legionella*

(Angenent, 2005; Boe-Hansen et al., 2002; du Moulin et al., 1988; Embil et al., 1997; Falkinham et al., 2001; Le Dantec et al., 2002; Lee et al., 1988; Leoni et al., 1999, 2001; Thomas et al., 2006; Vaerewijck et al., 2001, 2005). Tap water therefore may be an important source of microbes in built environments. Cell counts can be attained visually by microscopy using non-specific DNA stains such as SYBR Gold, or in an automated fashion with flow cytometry. Both methods can be adapted to fluorescent in situ hybridization (FISH) analysis, in which taxon-specific fluorescent probes replace or complement non-specific DNA stains. The particle size of aggregated cells can also be used to determine if cells are biofilm-originating or planktonic. Because biofilms may form on faucet heads, sampling strategies may opt to collect water samples as soon as water begins flowing from the tap (to favor collection of the tap's biofilms) and/or wait until water has been flowing through the tap for several minutes (to measure systemic contaminants). Collection of both hot and cold tap water samples is crucial, because the different water temperatures can have an effect on the microorganisms that are able to persist in these water systems.

External Factors Influencing the Indoor Microbiome

Microbial community structure is highly habitat dependent. Therefore, the collection of metadata is fundamental to any project seeking to characterize the microbial community composition and structure. For constantly fluctuating parameters such as temperature, relative humidity, and brightness, one might consider recording not only the value at the time of sampling, but also the recent highs, lows, and averages for the sampled location. Table A5-3 lists parameters that may have an influence on microbial populations and communities. Many of these factors, such as temperature and humidity, are constantly

TABLE A5-3 Environmental Parameters

Building	Room	Surface
Latitude, longitude, altitude	Window closed vs. open	Material (carpet, granite, etc.)
Foundation type	Window direction (N, S, E, W)	Water activity
HVAC sterility	Light bulb type	Time since last cleaning
Surrounding flora	Hours per day occupied	Type of cleaning
Construction materials	Barefoot vs. shoe traffic	Light exposure
	Exposure to pets, vermin, etc.	Surface temperature
	Plants or water features	History of moisture events
	Number of occupants	Occupant proximity and interaction
	Connections to other rooms	
	Air temperature	
	Relative humidity	
	Percentage recirculated air	
	Air exchange rate	

changing, therefore measuring these variables continuously in order to observe maximums, minimums, and averages can be useful in generating a comprehensive site analysis.

To facilitate the adoption of a consistent metadata ontology for these types of measurements, a Minimum Information about any Sequence (MIxS) standard specific to the built environment (MIxS-BE) was presented to the Genomics Standards Consortium (GSC) on March 7, 2012 (Gilbert et al., 2012). At the time of this writing, the MIxS-BE standard is available as a working draft on the Microbiology of the Built Environment website (microBE.net) in order to solicit feedback and provide direction to early adopters. The MIxS GSC standards have been integral to generating comparable results among different research centers, with the existing GSC standards for genomics (MIGS), metagenomics (MIMS), and genetic markers (MIMARKS) (Field et al., 2008; Kottmann et al., 2008; Yilmaz et al., 2011) having enabled tens of thousands of environmental samples from dozens of laboratories and hundreds of geographic locations to be included in the Earth Microbiome Project (Gilbert et al., 2010a,b, 2011).

Automated Monitoring

Many of the parameters in Table A5-3 can be automatically measured and recorded at regular intervals by specialized data loggers designed for this purpose. Temperature and humidity monitors are relatively inexpensive, whereas devices for assessing air exchange rate, fraction of recirculated air, HVAC system flows, and occupancy and activities to a high level of precision can be a significant portion of a research budget and can require substantial post-processing. In selecting monitoring equipment, it is also important to consider the options of using battery versus electrical outlet power, and whether to store data locally on memory cards versus transmitting readings off site. In hospital settings, it is often necessary to obtain permission from technical administrators before installing wireless transmitters, because such devices may interfere with sensitive medical equipment.

Personnel- and asset-tracking infrastructure can also be of immense value to a study of microbial communities in an environment where human movement is hypothesized to be a driving factor in the introduction of new microbial species to surfaces and airborne particles. Through a combination of uniquely identifiable radio-frequency identification (RFID) tags worn by hospital occupants and RFID sensors placed throughout the building, this system is able to continuously monitor the location of personnel, thereby providing time-stamped information on person-to-person and person-to-room interactions. From this data, one can examine the connection between movement of staff between rooms and the movement of bacterial populations between rooms. Furthermore, by combining the observed number of occupants per room with the air exchange rate, it is possible to estimate the CO_2 and airborne microbe concentrations at each point in time. RFID systems are commonly used to track equipment and can therefore

also provide information regarding which equipment is shared among patients, such as IVs and dialysis machines. These data enable researchers to observe not only microbial communities over time, but also the influence of human interaction with those communities.

Seasonal changes in outdoor humidity and temperature have previously been found to influence the composition of microbial communities in indoor environments (Augustowska and Dutkiewicz, 2006; Eber et al., 2011; Kaarakainen et al., 2009; Park et al., 2000; Pitkäranta et al., 2008; Rintala et al., 2008, 2012; Yamada, 2007). Therefore, the recording of meteorological conditions is an important aspect of indoor microbiome studies. This can be accomplished by retrieving publicly available National Oceanic and Atmospheric Administration records through the National Climatic Data Center website at www.ncdc.noaa.gov. This collection offers hourly measurements from a network of temperature, humidity, pressure, wind velocity, and precipitation sensors gathered from automated weather monitoring stations throughout the United States.

Special Considerations

Effect of Cleaning

Cleaning practices in indoor environments are inherently designed to affect the resident microbiota; therefore, it is important to take note of the strategies used to disinfect the environments under study and the time-points at which cleaning regimens were conducted. In previous studies, the effects of cleaning were found to be highly dependent upon the cleaning products used (Barker et al., 2004; Exner et al., 2004; Josephson et al., 1997; Marshall et al., 2012; Rusin et al., 1998; Rutala et al., 2000; Scott et al., 1984): antimicrobial agents, bleach, ethanol, peroxide, and Lysol were much more effective at sterilizing a surface than surfactants, detergents, vinegar, ammonia, or baking soda. Although there has been speculation that households using antimicrobial cleaning products may select for antibiotic-resistant bacteria, randomized studies investigating this hypothesis did not observe differences in bacterial population structure or antibiotic resistance in response to antimicrobial cleaning products (Aiello et al., 2005; Cole et al., 2003).

Healthcare Facility Sampling

Prior studies have identified numerous hospital-associated pathogens (HAPs) as well as routes of transmission between patients, staff, equipment, surfaces, and recycled air. HAPs that are of particular relevance are coagulase-negative staphylococci, *Staphylococcus aureus*, *Enterococcus* species, *Candida* species, *Escherichia coli*, *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, *Enterobacter* species, *Acinetobacter baumannii*, and *Klebsiella oxytoca*, which have been

previously found to collectively account for 84 percent of HAIs over a 21-month period in 463 hospitals (Hidron et al., 2008). These bacteria have been found on physician's and nursing staff's clothing (Babb et al., 1983; Biljan et al., 1993; Loh et al., 2000; Lopez et al., 2009; Perry et al., 2001; Snyder et al., 2008; Treakle et al., 2009; Wiener-Well et al., 2011; Wong et al., 1991; Zachary et al., 2001), cell phones (Akinyemi et al., 2009; Brady et al., 2006, 2009; Datta et al., 2009; Hassoun et al., 2004; Kilic et al., 2009; Ulger et al., 2009), stethoscopes (Marinella et al., 1997; Zachary et al., 2001), computer keyboards (Bures et al., 2000; Doğan et al., 2008), faucet handles (Bures et al., 2000), telemetry leads (Safdar et al., 2012), electronic thermometers (Livornese et al., 1992), blood-pressure cuffs (Myers, 1978), X-ray cassettes (Kim et al., 2012), gels for ultrasound probes (Schabrun et al., 2006), and in the air of patient rooms (Berardi and Leoni, 1993; Fleischer, 2006; Genet et al., 2011; Huang et al., 2006; Sudharsanam et al., 2008).

Patient microflora are one of the most significant drivers of microbial ecology within a hospital room (Bhalla et al., 2004; Drees et al., 2008b). Microorganisms are readily transferred from patients to hospital staff (Bhalla et al., 2004) and to the next occupant of the room after it has been cleaned (Drees et al., 2008a; Huang et al., 2006). Human traffic that enters and leaves multi-specialty medical centers includes patients with active diseases and infections as well as healthy patients undergoing invasive medical and surgical procedures. The inherent risk of cross-contamination from interactions between healthcare workers, patients, and their families presents a major obstacle to protecting patient health using only the practices of isolation and containment. Several studies have found that regular washing of patients' skin with the bactericide chlorhexidine can reduce the likelihood of the patient acquiring a nosocomial antibiotic-resistant infection (Bleasdale, 2007; Climo et al., 2009; Kassakian et al., 2011; O'Horo et al., 2012; Paulson, 1993; Popovich et al., 2010; Vernon et al., 2006). Diagnostic testing performed by hospital laboratories in the course of patient treatment produces a detailed accounting of specific patient-associated microorganisms that, with institutional review board approval, can be included in study metadata to identify point sources of microbial populations within the larger hospital environment under observation.

Characterization of the Microbial Community

Ribosomal RNA sequencing is a common method to identify the microbial community structure in environmental samples. This approach involves PCR amplifying a variable region of the 16S (bacterial), 18S (eukaryotic), or ITS (fungal) rRNA gene using multifunctional DNA oligos that contain not only a complementary nucleotide sequence for priming the PCR reaction, but also a multiplex barcode for marking amplified sequences with a unique sample-specific DNA sequence and 5′ region encoding base pairs needed by the sequence technology.

The amplified, barcoded sequences from multiple samples are pooled together in equimolar concentrations for sequencing and then demultiplexed with computer algorithms based on their barcode sequence. Several software suites are freely available for processing high-throughput sequencing data including QIIME (Caporaso et al., 2010), MG-RAST (Meyer et al., 2008), mother (Schloss et al., 2009), Galaxy (Goecks et al., 2010), HUMAnN (Abubucker et al., 2012), and MEGAN (Huson et al., 2011).

The influence of environmental parameters on microbial populations can be quantified using multivariate non-parametric algorithms (e.g., principal coordinate analysis, principal component analysis, non-metric multidimensional scaling) for community composition and univariate analysis of variance (ANOVA) tests for diversity measures. These statistical tools calculate the percentages of variation that can be explained by individual parameters such as treatment, temperature, building material, adjacent microbiome, and any other environmental characteristic measured in concert with sample collections. Multivariate-crossed analyses are particularly useful in determining if specific combinations of environmental parameters (interactions) have a synergistic effect on population structure or composition. Univariate tests of diversity indices use higher-way ANOVA and are calculated with distribution-free, permutation-based (PERMANOVA) routines (Anderson, 2001). Additionally, following taxonomic characterization of the communities, using the QIIME pipeline (Caporaso et al., 2010), and production of an abundance matrix of operational taxonomic units against experimental condition, community similarity between samples can be represented by calculating a Bray-Curtis similarity matrix and UniFrac distances (Lozupone and Knight, 2005). Non-metric multidimensional scaling can be used to visualize the relationship between the experimental factors and formally tested using a combination of permutation-based PERMANOVA and fully non-parametric ANOSIM tests (Clarke, 1993). The QIIME, MoBEDAC, and VAMPS web servers calculate these metrics as well as facilitate the public release of such data sets. State-of-the-art artificial neural network software developed Larsen and colleagues can be employed to generate models for predicting the development of microbial communities based on the bacterial abundances observed in the study (Larsen et al., 2012). Source-tracking algorithms developed by Knights and colleagues identifies transference of communities from one sampling site to another (Knights et al., 2011). Taken together, these analyses provide insights into the driving factors behind microbial community development.

In addition to examining the relationships between microbial community structure and environmental variables, it is often desirable to compare population structures directly to one another. Such metrics include the diversity of species in a sample (alpha diversity) and the closely related measures of richness and evenness that describe the quantity of species and the range of population sizes (Whittaker, 1960, 1972). Beta diversity takes into account the average alpha diversity and the combined diversity of species across all samples (gamma

diversity) to evaluate the presence or absence of a core microbiome commonly shared across a significant subset of samples. This is particularly relevant in the study of disease-causing microorganisms, where it is important to differentiate between systemic populations of opportunistic pathogens that become disease-causing under specific conditions and microbes that are only associated with infections.

When describing the presence of potentially pathogenic microorganisms discovered in surveys of ribosomal sequences, care should be taken to place these results in context. Taxonomic assignment of reads is reliant on reference databases that contain a disproportionate number of sequences from disease-causing microbes, leading many novel operational taxonomic units to phylogenetically ordinate closest to a pathogen with which they may or may not share specific infectivity characteristics. The RDP Classifier (Wang et al., 2007) used in QIIME (Caporaso et al., 2010) and other taxonomic assignment software partially alleviates this issue by providing a confidence score for the assignment of a read into each taxonomic level—domain, kingdom, phylum, class, order, family, genus, and species. However, even correct species-level assignments fail to provide information on the presence or absence of genetic elements that are responsible for many pathogenic and antibiotic-resistant phenotypes.

Acknowledgments

The ideas presented in this article have been shaped in large part by Hospital Microbiome Consortium (hospitalmicrobiome.com/consortium) discussions held at the University of Chicago on June 7, 2012. We also thank the Arthur P. Sloan Foundation for funding the Home Microbiome Project (2011-6-05) and the Hospital Microbiome Consortium Workshop grant (2012-3-25).

References

Abubucker, S., N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Methé, P. D. Schloss, D. Gevers, M. Mitreva, and C. Huttenhower. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Computational Biology* 8(6):e1002358. doi:10.1371/journal.pcbi.1002358.

Adler, A., I. Tager, and D. R. Quintero. 2005. Decreased prevalence of asthma among farm-reared children compared with those who are rural but not farm-reared. *Journal of Allergy and Clinical Immunology* 115(1):67-73. doi:10.1016/j.jaci.2004.10.008.

Aiello, A. E., B. Marshall, S. B. Levy, P. Della-Latta, S. X. Lin, and E. Larson. 2005. Antibacterial cleaning products and drug resistance. *Emerging Infectious Diseases* 11(10):1565-1570. doi:10.3201/eid1110.041276.

Akinyemi, K. O., A. D. Atapu, O. O. Adetona, and A. O. Coker. 2009. The potential role of mobile phones in the spread of bacterial infections. *Journal of Infection in Developing Countries* 3(8):628-632.

Alfvén, T., C. Braun-Fahrländer, B. Brunekreef, E. von Mutius, J. Riedler, A. Scheynius, M. van Hage, M. Wickman, M. R. Benz, J. Budde, K. B. Michels, D. Schram, E. Ublagger, M. Wasser, G. Pershagen, and PARSIFAL study group. 2006. Allergic diseases and atopic sensitization in children related to farming and anthroposophic lifestyle—the PARSIFAL Study. *Allergy* 61(4):414-421. doi:10.1111/j.1398-9995.2005.00939.x.

Anderson, M. J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26(1):32-46. doi:10.1111/j.1442-9993.2001.01070.pp.x.

Anderson, R. N., and B. L. Smith. 2005. Deaths: Leading causes for 2002. *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* 53(17):1-89.

Angenent, L. T. 2005. Molecular identification of potential pathogens in water and air of a hospital therapy pool. *Proceedings of the National Academy of Sciences of the USA* 102(13):4860-4865. doi:10.1073/pnas.0501235102.

Augustowska, M., and J. Dutkiewicz. 2006. Variability of airborne microflora in a hospital ward within a period of one year. *Annals of Agricultural and Environmental Medicine: AAEM* 13(1):99-106.

Babb, J. R., J. G. Davies, and G. A. J. Ayliffe. 1983. Contamination of protective clothing and nurses' uniforms in an isolation ward. *Journal of Hospital Infection* 4(2):149-157. doi:10.1016/0195-6701(83)90044-0.

Barker, J., I. B. Vipond, and S. F. Bloomfield. 2004. Effects of cleaning and disinfection in reducing the spread of norovirus contamination via environmental surfaces. *Journal of Hospital Infection* 58(1):42-49. doi:10.1016/j.jhin.2004.04.021.

Berardi, B. M., and E. Leoni. 1993. Indoor air climate and microbiological airborne: Contamination in various hospital areas. *Zentralblatt Für Hygiene Und Umweltmedizin = International Journal of Hygiene and Environmental Medicine* 194(4):405-418.

Bhalla, A., N. J. Pultz, D. M. Gries, A. J. Ray, E. C. Eckstein, D. C. Aron, and C. J. Donskey. 2004. Acquisition of nosocomial pathogens on hands after contact with environmental surfaces near hospitalized patients. *Infection Control and Hospital Epidemiology* 25(2):164-167. doi:10.1086/502369.

Biljan, M. M., C. A. Hart, D. Sunderland, P. R. Manasse, and C. R. Kingsland. 1993. Multicentre randomised double bind crossover trial on contamination of conventional ties and bow ties in routine obstetric and gynaecological practice. *BMJ* 307(6919):1582-1584. doi:10.1136/bmj.307.6919.1582.

Bleasdale, S. C. 2007. Effectiveness of chlorhexidine bathing to reduce catheter-associated bloodstream infections in medical intensive care unit patients. *Archives of Internal Medicine* 167(19):2073. doi:10.1001/archinte.167.19.2073.

Bloomfield, S. F., R. Stanwell-Smith, R. W. R. Crevel, and J. Pickup. 2006. Too clean, or not too clean: The hygiene hypothesis and home hygiene. *Clinical & Experimental Allergy* 36(4):402-425. doi:10.1111/j.1365-2222.2006.02463.x.

Blowers, R., and K. R. Wallace. 1960. Environmental aspects of staphylococcal infections acquired in hospitals. III. Ventilation of operating rooms—bacteriological investigations. *American Journal of Public Health and the Nation's Health* 50:484-490.

Boe-Hansen, R., H-J. Albrechtsen, E. Arvin, and C. Jørgensen. 2002. Bulk water phase and biofilm growth in drinking water at low nutrient conditions. *Water Research* 36(18):4477-4486.

Bonetta, S., S. Bonetta, S. Mosso, S. Sampò, and E. Carraro. 2009. Assessment of microbiological indoor air quality in an Italian office building equipped with an HVAC system. *Environmental Monitoring and Assessment* 161(1-4):473-483. doi:10.1007/s10661-009-0761-8.

Bouillard, L., O. Michel, M. Dramaix, and M. Devleeschouwer. 2005. Bacterial contamination of indoor air, surfaces, and settled dust, and related dust endotoxin concentrations in healthy office buildings. *Annals of Agricultural and Environmental Medicine: AAEM* 12(2):187-192.

Bourdillon, R. B., and L. Colebrook. 1946. Air hygiene in dressing-rooms for burns or major wounds. *Lancet* 1(6400):601.

Brady, R. R. W., A. Wasson, I. Stirling, C. McAllister, and N. N. Damani. 2006. Is your phone bugged? The incidence of bacteria known to cause nosocomial infection on healthcare workers' mobile phones. *Journal of Hospital Infection* 62(1):123-125. doi:10.1016/j.jhin.2005.05.005.

Brady, R. R. W., J. Verran, N. N. Damani, and A. P. Gibb. 2009. Review of mobile communication devices as potential reservoirs of nosocomial pathogens. *Journal of Hospital Infection* 71(4):295-300. doi:10.1016/j.jhin.2008.12.009.

Bright, K. R., S. A. Boone, and C. P. Gerba. 2009. Occurrence of bacteria and viruses on elementary classroom surfaces and the potential role of classroom hygiene in the spread of infectious diseases. *Journal of School Nursing* 26(1):33-41. doi:10.1177/1059840509354383.

Bures, S., J. T. Fishbain, C. F. Uyehara, J. M. Parker, and B. W. Berg. 2000. Computer keyboards and faucet handles as reservoirs of nosocomial pathogens in the intensive care unit. *American Journal of Infection Control* 28(6):465-471. doi:10.1067/mic.2000.107267.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7(5):335-336. doi:10.1038/nmeth.f.303.

Clark, R. P. 2009. Skin scales among airborne particles. *Journal of Hygiene* 72(01):47. doi:10.1017/S0022172400023196.

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. *Austral Ecology* 18(1):117-143. doi:10.1111/j.1442-9993.1993.tb00438.x.

Climo, M. W., K. A. Sepkowitz, G. Zuccotti, V. J. Fraser, D. K. Warren, T. M. Perl, K. Speck, J. A. Jernigan, J. R. Robles, and E. S. Wong. 2009. The effect of daily bathing with chlorhexidine on the acquisition of methicillin-resistant *Staphylococcus aureus*, vancomycin-resistant *Enterococcus*, and healthcare-associated bloodstream infections: Results of a quasi-experimental multicenter trial. *Critical Care Medicine* 37(6):1858-1865. doi:10.1097/CCM.0b013e31819ffe6d.

Cole, E. C., R. M. Addison, J. R. Rubino, K. E. Leese, P. D. Dulaney, M. S. Newell, J. Wilkins, D. J. Gaber, T. Wineinger, and D. A. Criger. 2003. Investigation of antibiotic and antibacterial agent cross-resistance in target bacteria from homes of antibacterial product users and nonusers. *Journal of Applied Microbiology* 95(4):664-676. doi:10.1046/j.1365-2672.2003.02022.x.

Colebrook, L., and W. C. Cawston. 1948. Microbic content of air on roof of city hospital, at street level, and in wards. *Medical Research Council, Special Report* 262:233-241.

Costello, E. K., C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight. 2009. Bacterial community variation in human body habitats across space and time. *Science* 326(5960):1694-1697. doi:10.1126/science.1177486.

Cvjetanović, B. 1957. Determination of bacterial air pollution in various premises. *Journal of Hygiene* 56(02):163. doi:10.1017/S0022172400037657.

D'Amato, G., G. Liccardi, M. D'Amato, and S. Holgate. 2005. Environmental risk factors and allergic bronchial asthma. *Clinical & Experimental Allergy* 35(9):1113-1124. doi:10.1111/j.1365-2222.2005.02328.x.

Datta, P., H. Rani, J. Chander, and V. Gupta. 2009. Bacterial contamination of mobile phones of health care workers. *Indian Journal of Medical Microbiology* 27(3):279-281. doi:10.4103/0255-0857.53222.

Doğan, M., B. Feyzioğlu, M. Ozdemir, and B. Baysal. 2008. Investigation of microbial colonization of computer keyboards used inside and outside hospital environments. *Mikrobiyoloji Bülteni* 42(2):331-336.

Drees, M., D. R. Snydman, C. H. Schmid, L. Barefoot, K. Hansjosten, P. M. Vue, M. Cronin, S. A. Nasraway, and Y. Golan. 2008a. Prior environmental contamination increases the risk of acquisition of vancomycin-resistant enterococci. *Clinical Infectious Diseases* 46(5):678-685. doi:10.1086/527394.

Drees, M., D. R. Snydman, C. H. Schmid, L. Barefoot, K. Hansjosten, P. M. Vue, M. Cronin, S. A. Nasraway, and Y. Golan. 2008b. Antibiotic exposure and room contamination among patients colonized with vancomycin-resistant enterococci. *Infection Control and Hospital Epidemiology* 29(8):709-715. doi:10.1086/589582.

Drudge, C. N., S. Krajden, R. C. Summerbell, and J. A. Scott. 2011. Detection of antibiotic resistance genes associated with methicillin-resistant *Staphylococcus aureus* (MRSA) and coagulase-negative staphylococci in hospital air filter dust by PCR. *Aerobiologia* 28(2):285-289. doi:10.1007/s10453-011-9219-x.

du Moulin, G. C., K. D. Stottmeier, P. A. Pelletier, A. Y. Tsang, and J. Hedley-Whyte. 1988. Concentration of *Mycobacterium avium* by hospital hot water systems. *Journal of the American Medical Association* 260(11):1599-1601.

Eber, M. R., M. Shardell, M. L. Schweizer, R. Laxminarayan, and E. N. Perencevich. 2011. Seasonal and temperature-associated increases in gram-negative bacterial bloodstream infections among hospitalized patients. *PLoS ONE* 6(9):e25298. doi:10.1371/journal.pone.0025298.

Embil, J., P. Warren, M. Yakrus, R. Stark, S. Corne, D. Forrest, and E. Hershfield. 1997. Pulmonary illness associated with exposure to mycobacterium-avium complex in hot tub water. Hypersensitivity pneumonitis or infection? *Chest* 111(3):813-816.

Escombe, A. R., C. C. Oeser, R. H. Gilman, M. Navincopa, E. Ticona, W. Pan, C. Martínez, J. Chacaltana, R. Rodríguez, D. A. Moore, J. S. Friedland, and C. A. Evans. 2007. Natural ventilation for the prevention of airborne contagion. *PLoS Medicine* 4(2):e68. doi:10.1371/journal.pmed.0040068.

Exner, M., V. Vacata, B. Hornei, E. Dietlein, and J. Gebel. 2004. Household cleaning and surface disinfection: New insights and strategies. *Journal of Hospital Infection* 56(Suppl 2):S70-S75. doi:10.1016/j.jhin.2003.12.037.

Fahlgren, C., G. Bratbak, R.-A. Sandaa, R. Thyrhaug, and U. Li Zweifel. 2010. Diversity of airborne bacteria in samples collected using different devices for aerosol collection. *Aerobiologia* 27(2):107-120. doi:10.1007/s10453-010-9181-z.

Falkinham, J. O., C. D. Norton, and M. W. LeChevallier. 2001. Factors influencing numbers of *Mycobacterium avium*, *Mycobacterium intracellulare*, and other mycobacteria in drinking water distribution systems. *Applied and Environmental Microbiology* 67(3):1225-1231. doi:10.1128/AEM.67.3.1225-1231.2001.

Farnsworth, J. E., S. M. Goyal, S. W. Kim, T. H. Kuehn, P. C. Raynor, M. A. Ramakrishnan, S. Anantharaman, and W. Tang. 2006. Development of a method for bacteria and virus recovery from heating, ventilation, and air conditioning (HVAC) filters. *Journal of Environmental Monitoring* 8(10):1006. doi:10.1039/b606132j.

Field, D., G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, M. Ashburner, N. Axelrod, S. Baldauf, S. Ballard, J. Boore, G. Cochrane, J. Cole, P. Dawyndt, P. De Vos, C. dePamphilis, R. Edwards, N. Faruque, R. Feldman, J. Gilbert, P. Gilna, F. O. Glöckner, P. Goldstein, R. Guralnick, D. Haft, D. Hancock, H. Hermjakob, C. Hertz-Fowler, P. Hugenholtz, I. Joint, L. Kagan, M. Kane, J. Kennedy, G. Kowalchuk, R. Kottmann, E. Kolker, S. Kravitz, N. Kyrpides, J. Leebens-Mack, S. E. Lewis, K. Li, A. L. Lister, P. Lord, N. Maltsev, V. Markowitz, J. Martiny, B. Methe, I. Mizrachi, R. Moxon, K. Nelson, J. Parkhill, L. Proctor, O. White, S. Assunta-Sansone, A. Spiers, R. Stevens, P. Swift, C. Taylor, Y. Tateno, A. Tett, S. Turner, D. Ussery, B. Vaughan, N. Ward, T. Whetzel, I. San Gil, G. Wilson, and A. Wipat. 2008. The Minimum Information About a Genome Sequence (MIGS) Specification. *Nature Biotechnology* 26(5):541-547. doi:10.1038/nbt1360.

Fierer, N., C. L. Lauber, N. Zhou, D. McDonald, E. K. Costello, and R. Knight. 2010. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences of the USA* 107(14):6477-6481. doi:10.1073/pnas.1000162107.

Finch, J. E., J. Prince, and M.Hawksworth. 1978. A bacteriological survey of the domestic environment. *Journal of Applied Microbiology* 45(3):357-364. doi:10.1111/j.1365-2672.1978.tb04236.x.

Fleischer, M. 2006. Microbiological control of airborne contamination in hospitals. *Indoor and Built Environment* 15(1):53-56. doi:10.1177/1420326X06062230.

Fox, K., A. Fox, T. Elssner, C. Feigley, and D. Salzberg. 2010. MALDI-TOF mass spectrometry speciation of staphylococci and their discrimination from micrococci isolated from indoor air of schoolrooms. *Journal of Environmental Monitoring* 12(4):917-923. doi:10.1039/b925250a.

Genet, C., G. Kibru, and W. Tsegaye. 2011. Indoor air bacterial load and antibiotic susceptibility pattern of isolates in operating rooms and surgical wards at Jimma University Specialized Hospital, southwest Ethiopia. *Ethiopian Journal of Health Sciences* 21(1):9-17.

Gilbert, J. A., F. Meyer, D. Antonopoulos, P. Balaji, C. T. Brown, N. Desai, J. A. Eisen, D. Evers, D. Field, W. Feng, D. Huson, J. Jansson, R. Knight, J. Knight, E. Kolker, K. Konstantindis, J. Kostka, N. Kyrpides, R. Mackelprang, A. McHardy, C. Quince, J. Raes, A. Sczyrba, A. Shade, and R. Stevens. 2010a. Meeting report: The Terabase Metagenomics Workshop and the Vision of an Earth Microbiome Project. *Standards in Genomic Sciences* 3(3):243-248. doi:10.4056/sigs.1433550.

Gilbert, J. A., F. Meyer, J. Jansson, J. Gordon, N. Pace, J. Tiedje, R. Ley, N. Fierer, D. Field, N. Kyrpides, F-O. Glöckner, H-P. Klenk, K. E. Wommack, E. Glass, K. Docherty, R. Gallery, Rick Stevens, and R. Knight. 2010b. The Earth Microbiome Project: Meeting report of the 1st EMP Meeting on Sample Selection and Acquisition at Argonne National Laboratory, October 6, 2010. *Standards in Genomic Sciences* 3(3):249-253. doi:10.4056/aigs.1443528.

Gilbert, J. A., M. Bailey, D. Field, N. Fierer, J. A. Fuhrman, B. Hu, J. Jansson, R. Knight, G. A. Kowalchuk, N. C. Kyrpides, F. Meyer, and R. Stevens. 2011. The Earth Microbiome Project: The Meeting Report for the 1st International Earth Microbiome Project Conference, Shenzhen, China, June 13-15, 2011. *Standards in Genomic Sciences* 5(2):243-247. doi:10.4056/sigs.2134923.

Gilbert, J. A., Y. Bao, H. Wang, S-A. Sansone, S. C. Edmunds, N. Morrison, F. Meyer, L. M. Schriml, N. Davies, P. Sterk, J. Wilkening, G. M. Garrity, D. Field, R. Robbins, D. P. Smith, I. Mizrachi, and C. Moreau. 2012. Report of the 13th Genomic Standards Consortium Meeting, Shenzhen, China, March 4-7, 2012. *Standards in Genomic Sciences* 6(2):276-286. doi:10.4056/sigs.2876184.

Goecks, J., A. Nekrutenko, J. Taylor, and The Galaxy Team. 2010. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11(8):R86. doi:10.1186/gb-2010-11-8-r86.

Greene, V. W., D. Vesley, R. G. Bond, and G. S. Michaelsen. 1962a. Microbiological contamination of hospital air. I. Quantitative studies. *Applied Environmental Microbiology* 10:561-566.

Greene, V. W., D. Vesley, R. G. Bond, and G. S. Michaelsen. 1962b. Microbiological contamination of hospital air. II. Qualitative studies. *Applied Environmental Microbiology* 10:567-571.

Grice, E. A., H. H. Kong, S. Conlan, C. B. Deming, J. Davis, A. C. Young, NISC Comparative Sequencing Program, G. G. Bouffard, R. W. Blakesley, P. R. Murray, E. D. Green, M. L. Turner, and J. A. Segre. 2009. Topographical and temporal diversity of the human skin microbiome. *Science* 324(5931):1190-1192. doi:10.1126/science.1171700.

Grice, E. A., and J. A. Segre. 2011. The skin microbiome. *Nature Reviews Microbiology* 9(4):244-253. doi:10.1038/nrmicro2537.

Griffin, D. W., C. Gonzalez, N. Teigell, T. Petrosky, D. E. Northup, and M. Lyles. 2010. Observations on the use of membrane filtration and liquid impingement to collect airborne microorganisms in various atmospheric environments. *Aerobiologia* 27(1):25-35. doi:10.1007/s10453-010-9173-z.

Groseclose, S. L., W. S. Brathwaite, P. A. Hall, F. J. Connor, P. Sharp, W. J. Anderson, R. F. Fagan, J. J. Aponte, G. F. Jones, D. A. Nitschke, C. A. Worsham, N. Adekoya, M-H. Chang, T. Doyle, R. Dhara, and R. A. Jajosky. 2004. Summary of notifiable diseases—United States, 2002. *Morbidity and Mortality Weekly Report* 51(53):1-84.

Guenther, R., and G. Vittori. 2008. *Sustainable healthcare architecture*. Hoboken, NJ: John Wiley & Sons.

Hall-Baker, P. A., E. Nieves, R. A. Jajosky, D. A. Adams, P. Sharp, W. J. Anderson, J. J. Aponte, A. E. Aranas, S. B. Katz, M. Mayes, M. S. Wodajo, D. H. Onweh, J. Baillie, M. Park. 2010. Summary of notifiable diseases—United States, 2008. *Morbidity and Mortality Weekly Report* 57(54):1-100.

Hassoun, A., E. M. Vellozzi, and M. A. Smith. 2004. Colonization of personal digital assistants carried by healthcare professionals. *Infection Control and Hospital Epidemiology* 25(11):1000-1001. doi:10.1086/502334.

Hidron, A. I., J. R. Edwards, J. Patel, T. C. Horan, D. M. Sievert, D. A. Pollock, S. K. Fridkin, National Healthcare Safety Network Team, and Participating National Healthcare Safety Network Facilities. 2008. NHSN annual update: Antimicrobial-resistant pathogens associated with healthcare-associated infections: Annual summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2006–2007. *Infection Control and Hospital Epidemiology* 29(11):996-1011. doi:10.1086/591861.

Hilton, A. C., and E. Austin. 2000. The kitchen dishcloth as a source of and vehicle for foodborne pathogens in a domestic setting. *International Journal of Environmental Health Research* 10(3):257-261. doi:10.1080/09603120050127202.

Hospodsky, D., J. Qian, W. W. Nazaroff, N. Yamamoto, K. Bibby, H. Rismani-Yazdi, and J. Peccia. 2012. Human occupancy as a source of indoor airborne bacteria. *PLoS ONE* 7(4):e34867. doi:10.1371/journal.pone.0034867.

Huang, L. L., I. F. Mao, M. L. Chen, and C. T. Huang. 2006. The microorganisms of indoor air in a teaching hospital. *Taiwan Journal of Public Health* 25(4):315-322.

Huang, S. S., R. Datta, and R. Platt. 2006. Risk of acquiring antibiotic-resistant bacteria from prior room occupants. *Archives of Internal Medicine* 166(18):1945-1951. doi:10.1001/archinte.166.18.1945.

Huson, D. H., S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Research* 21(9):1552-1560. doi:10.1101/gr.120618.111.

Huttunen, K., H. Rintala, M-R. Hirvonen, A. Vepsäläinen, A. Hyvärinen, T. Meklin, M. Toivola, and A. Nevalainen. 2008. Indoor air particles and bioaerosols before and after renovation of moisture-damaged buildings: The effect on biological activity and microbial flora. *Environmental Research* 107(3):291-298. doi:10.1016/j.envres.2008.02.008.

Hyvarinen, A., T. Meklin, A. Vepsäläinen, and A. Nevalainen. 2002. Fungi and actinobacteria in moisture-damaged building materials—concentrations and diversity. *International Biodeterioration & Biodegradation* 49(1):27-37. doi:10.1016/S0964-8305(01)00103-2.

Jaffal, A. A., I. M. Banat, A. A. El Mogheth, H. Nsanze, A. Bener, and A. S. Ameen. 1997. Residential indoor airborne microbial populations in the United Arab Emirates. *Environment International* 23(4):529-533. doi:10.1016/S0160-4120(97)00055-X.

Josephson, K. L., J. R. Rubino, and I. L. Pepper. 1997. Characterization and quantification of bacterial pathogens and indicator organisms in household kitchens with and without the use of a disinfectant cleaner. *Journal of Applied Microbiology* 83(6):737-750.

Kaarakainen, P., H. Rintala, A. Vepsäläinen, A. Hyvärinen, A. Nevalainen, and T. Meklin. 2009. Microbial content of house dust samples determined with qPCR. *Science of the Total Environment* 407(16):4673-4680. doi:10.1016/j.scitotenv.2009.04.046.

Kassakian, S. Z., L. A. Mermel, J. A. Jefferson, S. L. Parenteau, and J. T. Machan. 2011. Impact of chlorhexidine bathing on hospital-acquired infections among general medical patients. *Infection Control and Hospital Epidemiology* 32(3):238-243. doi:10.1086/658334.

Kelley, S. T., U. Theisen, L. T. Angenent, A. St Amand, and N. R. Pace. 2004. Molecular analysis of shower curtain biofilm microbes. *Applied and Environmental Microbiology* 70(7):4187-4192. doi:10.1128/AEM.70.7.4187-4192.

Kembel, S. W., E. Jones, J. Kline, D. Northcutt, J. Stenson, A. M. Womack, B. J. M. Bohannan, G. Z. Brown, and J. L. Green. 2012. Architectural design influences the diversity and structure of the built environment microbiome. *ISME Journal.* doi:10.1038/ismej.2011.211.

Kilic, I. H., M. Ozaslan, I. D. Karagoz, Y. Zer, and V. Davutoglu. 2009. The microbial colonisation of mobile phone used by healthcare staffs. *Pakistan Journal of Biological Sciences* 12(11):882-884. doi:10.3923/pjbs.2009.882.884.

Kim, J.-S., H.-S. Kim, J.-Y. Park, H.-S. Koo, C.-S. Choi, W. Song, H. C. Cho, and K. M. Lee. 2012. Contamination of X-ray cassettes with methicillin-resistant *Staphylococcus aureus* and methicillin-resistant *Staphylococcus haemolyticus* in a radiology department. *Annals of Laboratory Medicine* 32(3):206. doi:10.3343/alm.2012.32.3.206.

Klevens, R. M., J. R. Edwards, C. L. Richards Jr., T. C. Horan, R. P. Gaynes, D. A. Pollock, and D. M. Cardo. 2007. Estimating health care-associated infections and deaths in U.S. hospitals, 2002. *Public Health Reports (Washington, D.C.: 1974)* 122(2):160-166.

Klintberg, B., N. Berglund, G. Lilja, M. Wickman, and M. van Hage-Hamsten. 2001. Fewer allergic respiratory disorders among farmers' children in a closed birth cohort from Sweden. *European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology* 17(6):1151-1157.

Knights, D., J. Kuczynski, E. S. Charlson, J. Zaneveld, M. C. Mozer, R. G. Collman, F. D. Bushman, R. Knight, and S. T. Kelley. 2011. Bayesian community-wide culture-independent microbial source tracking. *Nature Methods* 8(9):761-763. doi:10.1038/nmeth.1650.

Kopperud, R. J., A. R. Ferro, and L. M. Hildemann. 2004. Outdoor versus indoor contributions to indoor particulate matter (PM) determined by mass balance methods. *Journal of the Air & Waste Management Association (1995)* 54(9):1188-1196.

Korves, T. M., Y. M. Piceno, L. M. Tom, T. Z. DeSantis, B. W. Jones, G. L. Andersen, and G. M. Hwang. 2012. Bacterial communities in commercial aircraft high-efficiency particulate air (HEPA) filters assessed by PhyloChip analysis. *Indoor Air.* doi:10.1111/j.1600-0668.2012.00787.x. http://doi.wiley.com/10.1111/j.1600-0668.2012.00787.x.

Kottmann, R., T. Gray, S. Murphy, L. Kagan, S. Kravitz, T. Lombardot, D. Field, and F. O. Glöckner. 2008. A standard MIGS/MIMS compliant XML schema: Toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS: A Journal of Integrative Biology* 12(2):115-121. doi:10.1089/omi.2008.0A10.

Kramer, A., I. Schwebke, and G. Kampf. 2006. How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. *BMC Infectious Diseases* 6:130. doi:10.1186/1471-2334-6-130.

Krogulski, A., and M. Szczotko. 2011. Microbiological quality of hospital indoor air. Determinant factors for microbial concentration in air of operating theatres. *Roczniki Państwowego Zakładu Higieny* 62(1):109-113.

Larsen, P. E., D. Field, and J. A. Gilbert. 2012. Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods.* doi:10.1038/nmeth.1975.

Le Dantec, C., J.-P. Duguet, A. Montiel, N. Dumoutier, S. Dubrou, and V. Vincent. 2002. Occurrence of mycobacteria in water treatment lines and in water distribution systems. *Applied and Environmental Microbiology* 68(11):5318-5325. doi:10.1128/AEM.68.11.5318-5325.2002.

Lee, T. C., J. E. Stout, and V. L. Yu. 1988. Factors predisposing to *Legionella pneumophila* colonization in residential water systems. *Archives of Environmental Health: An International Journal* 43(1):59-62. doi:10.1080/00039896.1988.9934375.

Leoni, E., P. Legnani, M. T. Mucci, and R. Pirani. 1999. Prevalence of mycobacteria in a swimming pool environment. *Journal of Applied Microbiology* 87(5):683-688. doi:10.1046/j.1365-2672.1999.00909.x.

Leoni, E., P. Legnani, M. A. Bucci Sabattini, and F. Righi. 2001. Prevalence of *Legionella* spp. in swimming pool environment. *Water Research* 35(15):3749-3753. doi:10.1016/S0043-1354(01)00075-6.

Leynaert, B., C. Neukirch, D. Jarvis, S. Chinn, P. Burney, and F. Neukirch. 2001. Does living on a farm during childhood protect against asthma, allergic rhinitis, and atopy in adulthood? *American Journal of Respiratory and Critical Care Medicine* 164(10 Pt 1):1829-1834.

Lignell, U., T. Meklin, H. Rintala, A. Hyvärinen, A. Vepsäläinen, J. Pekkanen, and A. Nevalainen. 2008. Evaluation of quantitative PCR and culture methods for detection of house dust fungi and streptomycetes in relation to moisture damage of the house. *Letters in Applied Microbiology* 47(4):303-308. doi:10.1111/j.1472-765X.2008.02431.x.

Livornese, L. L., Jr., S. Dias, C. Samel, B. Romanowski, S. Taylor, P. May, P. Pitsakis, G. Woods, D. Kaye, and M. E. Levison. 1992. Hospital-acquired infection with vancomycin-resistant *Enterococcus faecium* transmitted by electronic thermometers. *Annals of Internal Medicine* 117(2):112-116.

Loh, W., V. V. Ng, and J. Holton. 2000. Bacterial flora on the white coats of medical students. *Journal of Hospital Infection* 45(1):65-68. doi:10.1053/jhin.1999.0702.

Lopez, P.-J., O. Ron, P. Parthasarathy, J. Soothill, and L. Spitz. 2009. Bacterial counts from hospital doctors' ties are higher than those from shirts. *American Journal of Infection Control* 37 (1):79-80. doi:10.1016/j.ajic.2008.09.018.

Lozupone, C., and R. Knight. 2005. UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* 71(12):8228-8235. doi:10.1128/AEM.71.12.8228-8235.2005.

Marinella, M. A., C. Pierson, and C. Chenoweth. 1997. The stethoscope: A potential source of nosocomial infection? *Archives of Internal Medicine* 157(7):786-790. doi:10.1001/archinte.1997.00440280114010.

Marshall, B. M., E. Robleto, T. Dumont, and S. B. Levy. 2012. The frequency of antibiotic-resistant bacteria in homes differing in their use of surface antibacterial agents. *Current Microbiology* 65(4):407-415. doi:10.1007/s00284-012-0172-x.

Martinez, F. D. 2001. The coming-of-age of the hygiene hypothesis. *Respiratory Research* 2(3):129-132. doi:10.1186/rr48.

Merchant, J. A., A. L. Naleway, E. R. Svendsen, K. M. Kelly, L. F. Burmeister, A. M. Stromquist, C. D. Taylor, P. S. Thorne, S. J. Reynolds, W. T. Sanderson, and E. A. Chrischilles. 2005. Asthma and farm exposures in a cohort of rural Iowa children. *Environmental Health Perspectives* 113(3):350-356.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. 2008. The metagenomics RAST Server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9(1):386. doi:10.1186/1471-2105-9-386.

Monto, A. S. 2002. Epidemiology of viral respiratory infections. *American Journal of Medicine* 112(6):4-12. doi:10.1016/S0002-9343(01)01058-0.

Morrow, J. B., A. S. Downey, and J. Peccia. 2012. Challenges in microbial sampling in the indoor environment. National Institutes of Standards and Technology. http://www.nist.gov/customcf/get_pdf.cfm?pub_id=910577.

Moschandreas, D. J. 1981. Exposure to pollutants and daily time budgets of people. *Bulletin of the New York Academy of Medicine* 57(10):845-859.

Myers, M. G. 1978. Longitudinal evaluation of neonatal nosocomial infections: Association of infection with a blood pressure cuff. *Pediatrics* 61(1):42-45.

Nevalainen, A., and M. Seuri. 2005. Of microbes and men. *Indoor Air* 15(s9):58-64. doi:10.1111/j.1600-0668.2005.00344.x.

Noble, W. C., J. D. F. Habbema, R. Van Furth, I. Smith, and C. De Raay. 1976. Quantitative studies on the dispersal of skin bacteria into the air. *Journal of Medical Microbiology* 9(1):53-61. doi:10.1099/00222615-9-1-53.

Noris, F., J. A. Siegel, and K. A. Kinney. 2009. Biological and chemical contaminants in HVAC filter dust. *ASHRAE Transactions* 115(2):484-491.

Noris, F., J. A. Siegel, and K. A. Kinney. 2011. Evaluation of HVAC filters as a sampling mechanism for indoor microbial communities. *Atmospheric Environment* 45(2):338-346. doi:10.1016/j.atmosenv.2010.10.017.

O'Horo, J. C., G. L. M. Silva, L. S. Munoz-Price, and N. Safdar. 2012. The efficacy of daily bathing with chlorhexidine for reducing healthcare-associated bloodstream infections: A meta-analysis. *Infection Control and Hospital Epidemiology* 33(3):257-267. doi:10.1086/664496.

Park, J. H., D. L. Spiegelman, H. A. Burge, D. R. Gold, G. L. Chew, and D. K. Milton. 2000. Longitudinal study of dust and airborne endotoxin in the home. *Environmental Health Perspectives* 108(11):1023-1028.

Paulson, D. S. 1993. Efficacy evaluaton of a 4% chlorhexidine gluconate as a full-body shower wash. *American Journal of Infection Control* 21(4):205-209. doi:10.1016/0196-6553(93)90033-Z.

Peccia, J., D. K. Milton, T. Reponen, and J. Hill. 2008. A role for environmental engineering and science in preventing bioaerosol-related disease. *Environmental Science & Technology* 42(13):4631-4637. doi:10.1021/es087179e.

Perry, C., R. Marshall, and E. Jones. 2001. Bacterial contamination of uniforms. *Journal of Hospital Infection* 48(3):238-241. doi:10.1053/jhin.2001.0962.

Pitkäranta, M., T. Meklin, A. Hyvärinen, L. Paulin, P. Auvinen, A. Nevalainen, and H. Rintala. 2008. Analysis of fungal flora in indoor dust by ribosomal DNA sequence analysis, quantitative PCR, and culture. *Applied and Environmental Microbiology* 74(1):233-244. doi:10.1128/AEM.00692-07.

Pope, A. M., R. Patterson, H. Burge, and Institute of Medicine (U.S.). Committee on the Health Effects of Indoor Allergens. 1993. *Indoor allergens assessing and controlling adverse health effects.* Washington, DC: National Academy Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1100.

Popovich, K. J., B. Hota, R. Hayes, R. A. Weinstein, and M. K. Hayden. 2010. Daily skin cleansing with chlorhexidine did not reduce the rate of central-line associated bloodstream infection in a surgical intensive care unit. *Intensive Care Medicine* 36(5):854-858. doi:10.1007/s00134-010-1783-y.

Qian, H., and Y. Li. 2010. Removal of exhaled particles by ventilation and deposition in a multibed airborne infection isolation room. *Indoor Air* 20(4):284-297. doi:10.1111/j.1600-0668.2010.00653.x.

Qian, J., D. Hospodsky, N. Yamamoto, W. W. Nazaroff, and J. Peccia. 2012. Size-resolved emission rates of airborne bacteria and fungi in an occupied classroom. *Indoor Air.* doi:10.1111/j.1600-0668.2012.00769.x.

Remes, S. T., H. O. Koskela, K. Iivanainen, and J. Pekkanen. 2005. Allergen-specific sensitization in asthma and allergic diseases in children: The Study on Farmers' and Non-farmers' Children. *Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology* 35(2):160-166. doi:10.1111/j.1365-2222.2005.02172.x.

Riedler, J., C. Braun-Fahrländer, W. Eder, M. Schreuer, M. Waser, S. Maisch, D. Carr, R. Schierl, D. Nowak, and E. von Mutius. 2001. Exposure to farming in early life and development of asthma and allergy: A cross-sectional survey. *Lancet* 358(9288):1129-1133. doi:10.1016/S0140-6736(01)06252-3.

Rintala, H., M. Pitkaranta, M. Toivola, L. Paulin, and A. Nevalainen. 2008. Diversity and seasonal dynamics of bacterial community in indoor environment. *BMC Microbiology* 8(1):56. doi:10.1186/1471-2180-8-56.

Rintala, H., M. Pitkäranta, and M. Täubel. 2012. Microbial communities associated with house dust. *Advances in Applied Microbiology* 78:75-120. Elsevier. http://linkinghub.elsevier.com/retrieve/pii/B978012394805200004X.

Rook, G. A. W. 2009. Review series on helminths, immune modulation and the hygiene hypothesis: The broader implications of the hygiene hypothesis. *Immunology* 126(1):3-11. doi:10.1111/j.1365-2567.2008.03007.x.

Rook, G. A. W., and J. L. Stanford. 1998. Give us this day our daily germs. *Immunology Today* 19(3):113-116. doi:10.1016/S0167-5699(98)80008-X.

Rusin, P., P. Orosz-Coughlin, and C. Gerba. 1998. Reduction of faecal coliform, coliform and heterotrophic plate count bacteria in the household kitchen and bathroom by disinfection with hypochlorite cleaners. *Journal of Applied Microbiology* 85(5):819-828.

Rutala, W. A., S. L. Barbee, N. C. Aguiar, M. D. Sobsey, and D. J. Weber. 2000. Antimicrobial activity of home disinfectants and natural products against potential human pathogens. *Infection Control and Hospital Epidemiology* 21(1):33-38. doi:10.1086/501694.

Safdar, N., J. Drayton, J. Dern, S. Warrack, M. Duster, and M. Schmitz. 2012. Telemetry leads harbor nosocomial pathogens. *International Journal of Infection Control* 8(2). doi:10.3396/ijic.v8i2.012.12.

Schabrun, S., L. Chipchase, and H. Rickard. 2006. Are therapeutic ultrasound units a potential vector for nosocomial infection? *Physiotherapy Research International* 11(2):61-71. doi:10.1002/pri.329.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. 2009. Introducing Mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75(23):7537-7541. doi:10.1128/AEM.01541-09.

Schram, D., G. Doekes, M. Boeve, J. Douwes, J. Riedler, E. Ublagger, E. Mutius, J. Budde, G. Pershagen, F. Nyberg, J. Alm, C. Braun-Fahrländer, M. Waser, B. Brunekreef, and the PARSIFAL Study Group. 2005. Bacterial and fungal components in house dust of farm children, Rudolf Steiner school children and reference children—the PARSIFAL Study. *Allergy* 60(5):611-618. doi:10.1111/j.1398-9995.2005.00748.x.

Scott, E., S. F. Bloomfield, and C. G. Barlow. 1982. An investigation of microbial contamination in the home. *Journal of Hygiene* 89(2):279-293.

Scott, E., S. F. Bloomfield, and C. G. Barlow. 1984. Evaluation of disinfectants in the domestic environment under 'in use' conditions. *Journal of Hygiene* 92(2):193-203.

Sebastian, A., and L. Larsson. 2003. Characterization of the microbial community in indoor environments: A chemical-analytical approach. *Applied and Environmental Microbiology* 69(6):3103-3109. doi:10.1128/AEM.69.6.3103-3109.2003.

Simard, C., M. Trudel, G. Paquette, and P. Payment. 1983. Microbial investigation of the air in an apartment building. *Journal of Hygiene* 91(2):277-286.

Snyder, G. M., K. A. Thom, J. P. Furuno, E. N. Perencevich, M.-C. Roghmann, S. M. Strauss, G. Netzer, and A. D. Harris. 2008. Detection of methicillin-resistant *Staphylococcus aureus* and vancomycin-resistant enterococci on the gowns and gloves of healthcare workers. *Infection Control and Hospital Epidemiology* 29(7):583-589. doi:10.1086/588701.

Stajich, J. E., T. Harris, B. P. Brunk, J. Brestelli, S. Fischer, O. S. Harb, J. C. Kissinger, W. Li, V. Nayak, D. F. Pinney, C. J. Stoekert Jr., and D. S. Roos. 2011. FungiDB: An integrated functional genomics database for fungi. *Nucleic Acids Research* 40(D1):D675-D681. doi:10.1093/nar/gkr918.

Stanley, N. J., T. H. Kuehn, S. W. Kim, P. C. Raynor, S. Anantharaman, M. A. Ramakrishnan, and Sagar M. Goyal. 2008. Background culturable bacteria aerosol in two large public buildings using HVAC filters as long term, passive, high-volume air samplers. *Journal of Environmental Monitoring* 10(4):474. doi:10.1039/b719316e.

Sudharsanam, S., P. Srikanth, M. Sheela, and R. Steinberg. 2008. Study of the indoor air quality in hospitals in South Chennai, India—microbial profile. *Indoor and Built Environment* 17(5):435-441. doi:10.1177/1420326X08095568.

Taitt, C. R., T. Leski, D. Stenger, G. J. Vora, B. House, M. Nicklasson, G. Pimentel, D. V. Zurawski, B. C. Kirkup, D. Craft, P. E. Waterman, E. P. Lesho, U. Bangurae, and R. Ansumana. 2012. Antimicrobial resistance determinant microarray for analysis of multi-drug resistant isolates. *SPIE* 8371: 83710X-83710X-10.

Tang, J. W. 2009. The effect of environmental parameters on the survival of airborne infectious agents. *Journal of The Royal Society Interface* 6(Suppl_6):S737-S746. doi:10.1098/rsif.2009.0227.focus.

Täubel, M., H. Rintala, M. Pitkäranta, L. Paulin, S. Laitinen, J. Pekkanen, A. Hyvärinen, and A. Nevalainen. 2009. The occupant as a source of house dust bacteria. *Journal of Allergy and Clinical Immunology* 124(4):834-840.e47. doi:10.1016/j.jaci.2009.07.045.

Thomas, V., K. Herrera-Rimann, D. S. Blanc, and G. Greub. 2006. Biodiversity of amoebae and amoeba-resisting bacteria in a hospital water network. *Applied and Environmental Microbiology* 72(4):2428-2438. doi:10.1128/AEM.72.4.2428-2438.2006.

Treakle, A. M., K. A. Thom, J. P. Furuno, S. M. Strauss, A. D. Harris, and E. N. Perencevich. 2009. Bacterial contamination of health care workers' white coats. *American Journal of Infection Control* 37(2):101-105. doi:10.1016/j.ajic.2008.03.009.

Tringe, S. G., T. Zhang, X. Liu, Y. Yu, W. H. Lee, J. Yap, F. Yao, S. T. Suan, S. K. Ing, M. Haynes, F. Rohwer, C. L. Wei, P. Tan, J. Bristow, E. M. Rubin, and Y. Ruan. 2008. The airborne metagenome in an indoor urban environment. *PLoS ONE* 3(4):e1862. doi:10.1371/journal.pone.0001862.

Ulger, F., S. Esen, A. Dilek, K. Yanik, M. Gunaydin, and H. Leblebicioglu. 2009. Are we aware how contaminated our mobile phones with nosocomial pathogens? *Annals of Clinical Microbiology and Antimicrobials* 8(1):7. doi:10.1186/1476-0711-8-7.

Vaerewijck, M. J. M., G. Huys, J. Carlos Palomino, J. Swings, and F. Portaels. 2005. Mycobacteria in drinking water distribution systems: Ecology and significance for human health. *FEMS Microbiology Reviews* 29(5):911-934. doi:10.1016/j.femsre.2005.02.001.

Vernon, M. O., M. K. Hayden, W. E. Trick, R. A. Hayes, D. W. Blom, and R. A. Weinstein. 2006. Chlorhexidine gluconate to cleanse patients in a medical intensive care unit: The effectiveness of source control to reduce the bioburden of vancomycin-resistant Enterococci. *Archives of Internal Medicine* 166(3):306-312. doi:10.1001/archinte.166.3.306.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 73(16):5261-5267. doi:10.1128/AEM.00062-07.

Warner, P., and A. Glassco. 1963. Enumeration of air-borne bacteria in hospital. *Canadian Medical Association Journal* 88:1280-1283.

Whittaker, R. H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs* 30(3):279. doi:10.2307/1943563.

Whittaker, R. H. 1972. Evolution and measurement of species diversity. *Taxon* 21(2/3):213. doi:10.2307/1218190.

Wiener-Well, Y., M. Galuty, B. Rudensky, Y. Schlesinger, D. Attias, and A. M. Yinnon. 2011. Nursing and physician attire as possible source of nosocomial infections. *American Journal of Infection Control* 39(7):555-559. doi:10.1016/j.ajic.2010.12.016.

Williams, R. E. O., and A. Hirch. 1950. Bacterial contamination of air in underground trains. *Lancet* 1(6595):128-131. doi:10.1016/S0140-6736(50)90081-X.

Williams, R. E. O., O. M. Lidwell, and A. Hirch. 1956. The bacterial flora of the air of occupied rooms. *Journal of Hygiene* 54(04):512. doi:10.1017/S002217240004479X.

Wong, D., K. Nye, and P. Hollis. 1991. Microbial flora on doctors' white coats. *BMJ* 303(6817):1602-1604. doi:10.1136/bmj.303.6817.1602.

Yamada, K. 2007. A study on the behavior and control of indoor airborne microbe in a clinic. *Journal of the National Institute of Public Health* 56(3):300-302.

Yazdanbakhsh, M., and P. M. Matricardi. 2004. Parasites and the hygiene hypothesis: Regulating the immune system? *Clinical Reviews in Allergy & Immunology* 26(1):15-24. doi:10.1385/CRIAI:26:1:15.

Yilmaz, P., R. Kottmann, D. Field, R. Knight, J. R. Cole, L. Amaral-Zettler, J. A. Gilbert, I. Karsch-Mizrachi, A. Johnston, G. Cochrane, R. Vaughan, C. Hunter, J. Park, N. Morrison, P. Rocca-Serra, P. Sterk, M. Arumugam, M. Bailey, L. Baumgartner, B. W. Birren, M. J. Blaser, V. Bonazzi, T. Booth, P. Bork, F. D. Bushman, P. L. Buttigieg, P. S. G. Chain, E. Charlson, E. K. Costello, H. Huot-Creasy, P. Dawyndt, T. DeSantis, N. Fierer, J. A. Fuhrman, R. E. Gallery, D. Gevers, R. A. Gibbs, I. San Gil, A. Gonzalez, J. I. Gordon, R. Guralnick, W. Hankeln, S. Highlander, P. Hugenholtz, J. Jansson, A. L. Kau, S. T. Kelley, J. Kennedy, D. Knights, O. Koren, J. Kuczynski, N. Kyrpides, R. Larsen, C. L. Lauber, T. Legg, R. E. Ley, C. A. Lozupone, W. Ludwig, D. Lyons, E. Maguire, B. A. Methé, F. Meyer, B. Muegge, S. Nakielny, K. E. Nelson, D. Nemergut, J. D. Neufeld, L. K. Newbold, A. E. Oliver, N. R. Pace, G. Palanisamy, J. Peplies, J. Petrosino, L. Proctor, E. Pruesse, C. Quast, J. Raes, S. Ratnasingham, J. Ravel, D. A. Relman, S. Assunta-Sansone, P. D. Schloss, L. Schriml, R. Sinha, M. I. Smith, E. Sodergren, A. Spor, J. Stombaugh, J. M. Tiedje, D. V. Ward, G. M. Weinstock, D. Wendel, O. White, A. Whiteley, A. Wilke, J. R. Wortman, T. Yatsunenko, and F. O. Glöckner. 2011. Minimum Information about a Marker Gene Sequence (MIMARKS) and Minimum Information About Any (x) Sequence (MIxS) Specifications. 2011. *Nature Biotechnology* 29(5):415-420. doi:10.1038/nbt.1823.

Zachary, K. C., P. S. Bayne, V. J. Morrison, D. S. Ford, L. C. Silver, and D. C. Hooper. 2001. Contamination of gowns, gloves, and stethoscopes with vancomycin-resistant Enterococci. *Infection Control and Hospital Epidemiology* 22(9):560-564. doi:10.1086/501952.

A6

SEQUENCING ERRORS, DIVERSITY ESTIMATES AND THE RARE BIOSPHERE

*Susan M. **Huse**,[22] David B. Mark Welch, and Mitchell L. Sogin*

Introduction

Our understanding of microbial communities is in a time of rapid change. The application of polymerase chain reaction (PCR), cloning, and DNA sequencing to microbial diversity research has rapidly expanded our appreciation of the extent of the microbial world. In particular, analysis of PCR amplicons from various regions of the small subunit ribosomal RNA (SSU rRNA or 16S) gene generated from culture-independent samples is now the accepted standard for cataloguing microbial communities. As sequencing technologies improved, it became feasible to assess community membership from more than 1,000 individual SSU rRNA amplicons. With the advent of next-generation sequencing (NGS) that did not require the cloning of individual amplicons, researchers transitioned from generating thousands of ~800–1,100 nt reads to hundreds of thousands of 100–500 nt sequencing reads (454 technology). Illumina technology can now

[22] Brown University.

produce millions of 100 nt reads from hundreds of samples in a single run, potentially providing a nearly exhaustive survey of microbes present in a sample.

While the NGS technologies provide deeper sampling, the trade-off for depth has been shorter read lengths. The ~800–1,100 nt reads produced by the late 1990s using Sanger sequencing on ABI or LICOR platforms could be used to reconstruct the entire SSU rRNA gene through multiple sequencing of the same clone. To make use of the shorter reads produced by NGS technology, Sogin et al. capitalized on the structure of the SSU rRNA gene. The gene includes a series of regions that are highly conserved across the bacterial domain, interspersed with a series of nine hypervariable regions. This structure lends itself conveniently to NGS because oligonucleotide primers that target conserved regions on either side of the hypervariable regions can amplify DNA from across the bacterial domain. The more rapidly evolving hypervariable regions in contrast are unique for most microbial genera and in many cases can differentiate below the genus level. Sogin et al. used primers to conserved flanking regions to amplify the V6 hypervariable region, which at 60–80 nt in length could reliably be completely sequenced on a 454 GS20. As with conventionally sequenced clone libraries, each read in principle represents an SSU rRNA operon and is a proxy for a microbe from the sample. By comparing the hypervariable region sequences against databases of SSU rRNA gene sequences from known taxonomy, such as RDP (Wang et al., 2007), SILVA (Pruesse et al., 2007), or Greengenes (DeSantis et al., 2006), the reads become tags for cataloging the taxonomy of the community being studied.

As the technology for microbial community research has evolved, so has our understanding of the communities we study. In the first published study implementing NGS in environmental samples, Sogin et al. (2006) examined several marine environments and discovered a richness and diversity in microbial community structures previously unknown. Each community exhibited a relatively small number of highly abundant taxa and a large number of low abundance taxa, a pattern often described as a long-tail distribution (Figure A6-1). Because of the unevenness of this community structure, previous studies using hundreds or thousands of sequencing reads were able to identify only the most abundant members and a small fraction of the taxa in the long tail. The greater sequencing depth of NGS methods revealed the breadth of the low abundance taxa—the "rare biosphere."

Impact of Sequencing Errors and Clustering Methods

As NGS provided a means to explore ever deeper into microbial community structures, the gap between the number of named species and the number of sequence phylotypes increased. Unfortunately, the short read lengths of NGS technology, especially when applied to hypervariable regions with few stable phylogenetically informative positions, are poorly suited for the traditional phylogenetic analyses required for registering new taxa. Researchers turned to

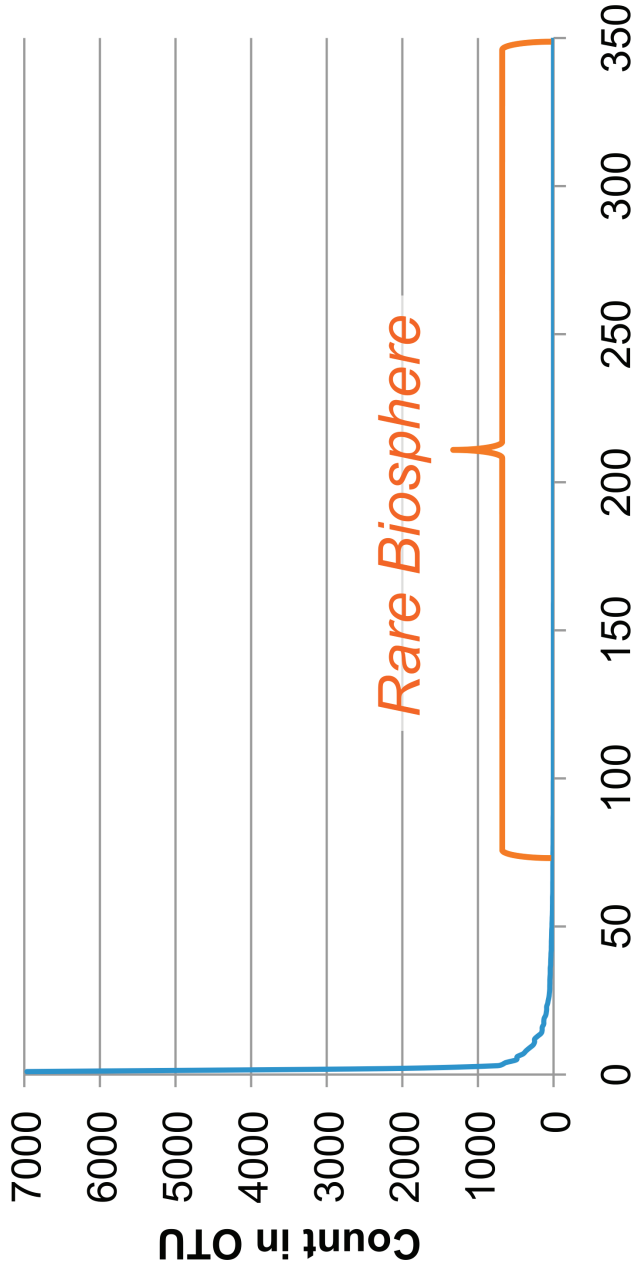


FIGURE A6-1 An example rank abundance curve with a long-tail distribution. The OTUs are ordered from most abundant on the left to the least abundant on the right. The y-axis plots the abundance of each OTU. There are a very small number of highly abundant OTUs on the extreme left. Then there are a small number of moderately sized OTUs, and then much of the graph and all OTUs after the first 40 or so are low abundance or rare. The OTUs ranked below 75 fall within the general category of "rare biosphere."

taxonomic-independent sequence clustering methods for characterizing microbial communities. By assuming that very similar sequences represent closely related organisms, and that more divergent sequences represent more distantly related organisms, the sequences can be clustered into groups of similar organisms, each cluster or "operational taxonomic unit" (OTU) presumed to represent a phylotype (Schloss and Handelsman, 2005). The width of the clustering, meaning the percent identity threshold for sequence tags to be placed in the same OTU, represents the similarity of the microbes in each OTU.

A critical element in taxonomy-independent analyses of diversity is sequencing error. Random errors can be tolerated in an assembly project where the goal is a consensus sequence. In OTU clustering, however, each read is assumed to represent an individual organism, and if a read has sufficient errors, then it will not cluster with its template, instead forming a new, spurious OTU. Thus, if not filtered out or unaccounted for, sequencing error can lead to inflation in the number of OTUs attributed to a community.

To address the issue of sequence quality, several authors developed quality filtering (Huse et al., 2007) and data de-noising (Quince et al., 2009) techniques for processing raw 454 sequencing reads to reduce sequencing errors and thereby reduce OTU inflation. In 2009, in a paper titled *Wrinkles in the Rare Biosphere*, Kunin et al. (2010) highlighted the impact of error rates on OTU analyses by sequencing a single strain of *E. coli* and generating more than 600 OTUs. Reeder and Knight followed this with a "News and Views" piece: *The 'rare biosphere': A reality check* (Reeder and Knight, 2009). The combination of these two publications spotlighted the very real concerns about the impact of sequencing error on microbial community diversity estimates. They posed a critical question: *Is the rare biosphere real or simply an artifact of sequencing errors?*

Even with very high-quality sequencing and stringent quality filtering, the depth of sampling afforded by NGS technology leads to more absolute OTU inflation than the earlier Sanger sequencing. This is for two reasons: First, when processing data from sequencing hundreds to thousands of Sanger capillary reads, individual chromatograms are often read by hand and confirmed by forward and reverse reads resulting in very high-quality sequence assemblies. While it is in principle possible to develop a similar skill reading 454 flowgrams, it is not conceivable to hand-edit hundreds of thousands of reads. In this regard it is worth noting that the generally accepted error rate in automated high-throughput capillary sequencing is 1 percent (i.e., an average Phred score of 20). The second reason is that the sheer number of reads produced by NGS will result in more spurious OTUs even if the error rate is much lower. Most OTU clustering methods are based only on the percent identity between sequences. If a read has sufficient errors that the difference between it and its template is greater than the clustering threshold, the clustering algorithm will place it in a new OTU. If a sequencing error rate leads to 1 read per thousand that fails to cluster with its template, a traditional clone library Sanger project with 1,000 reads will have, on average,

one spurious OTU. With NGS, *with the same error rate*, a data set of 100,000 reads will have, on average, 100 spurious OTUs. In practice, quality-controlled NGS reads tend to have a lower overall error rate (Schloss et al., 2011) than automated ABI capillary sequencing. So while, the relative rate of OTU inflation per read may have dropped with NGS, the absolute number of spurious OTUs has increased considerably because of dramatically increased depths of sequencing.

As it turns out, though, much of the OTU inflation observed in NGS projects was not due to sequencing errors. *Ironing out the Wrinkles in the Rare Biosphere* (Huse et al., 2010) demonstrated that the most commonly used method for OTU generation dramatically compounded the problem of sequencing error. Following on the Kunin technique, Huse et al. clustered DNA amplified from a single *E. coli* gene, and using only sequences with an error rate below the clustering threshold, showed that switching from a single multiple sequence (MS) alignment to multiple pairwise (PW) alignments, and from complete linkage (CL) clustering to average linkage (AL) clustering reduced the OTU inflation from 599 OTUs to 24. They introduced a single-linkage preclustering (SLP) to smooth errors prior to clustering. Using SLP-PWAL for clustering brought the OTU count to 1.

The effect of reads with more errors than the clustering threshold still needs to be taken into consideration. With established, simple sequence quality filtering, SLP clustering of reads generated on the Roche Genome Sequencer FLX platform from amplicons of the V6 SSU rRNA hypervariable region, errant reads produce spurious singleton OTUs at a rate of ~1 spurious OTU per 1,000 reads. When applied to analysis of control communities of limited diversity this rate can sound alarming. Processing 50,000 reads of a control community with 40 known members would generate 40 + 50,000/1,000 = 90 OTUs. However it is important to recognize that the number of spurious OTUs produced scales with the sequencing depth, not the complexity of the community. If a biological community sampled to a depth of 50,000 reads were to have 1,000 observed OTUs, then 50 of these would be due to sequencing error. As shown in Table A6-1, algorithm choice has a much greater effect on OTU inflation than sequencing error.

As NGS technologies produce longer reads (> 500 nt with Roche/454 and 300 nt with the Illumina MiSeq at the time of writing) the same error rate results in fewer errant reads generating new OTUs, because more errors per read are required for a sequence to fail to cluster with its template. Instead, a second type of error becomes of increasing importance: chimeric sequences from two or more templates during amplification. Chimeras are generated when a template is incompletely replicated during the elongation step of PCR. This truncated sequence can hybridize to other targets in subsequent PCR cycles and act as a primer, generating a single sequence from multiple templates. The frequency of chimeras scales with SSU rRNA amplicon length (Huber et al., 2009). The amplification of sample DNA is largely independent of sequencing platform (though the effect of platform-specific primer adapters has not been thoroughly explored) and several modifications of standard PCR result in reduced chimera formation (Acinas et al., 2004; Lahr and Katz, 2009; Qiu et al., 2001). However, chimeras remain

TABLE A6-1 OTU Inflation Due to Clustering Algorithm and Sequencing Error

Sample	Number of Reads	MS-CL OTUs	SLP-PWAL OTUs	Estimated Spurious OTUs	Estimated Sequencing Inflation	Estimated Algorithm Inflation
Deep-sea vent Archaea	63,133	709	470	63	15%	59%
English Channel	12,851	1,154	859	13	1.5%	35%
Human gut	15,239	803	566	15	2.7%	43%
Sewage	33,082	2,383	1,831	33	1.8%	31%
North Atlantic deep water	15,497	1,713	1,339	15	1.1%	28%

NOTE: OTUs found using multiple sequence alignment + complete linkage (MS-CL) and single linkage preclustering followed by pairwise average linkage (SLP-PWAL) in four example V6 data sets sequenced using a 454 GSFLX. The number of spurious SLP-PWAL OTUs is estimated to be 1 for every 1,000 sequence reads. The number of true OTUs is estimated to be the number of SLP-PWAL OTUs minus the estimated spurious OTUs. The estimated sequencing inflation is the ratio of estimated spurious OTUs to the number of estimated true OTUs. The estimated algorithm inflation is the ratio of the number of OTUs generated using MS-CL minus the number of spurious OTUs due to sequencing to the number of estimated true OTUs.
SOURCE: Adapted from Table 1 in Huse et al. (2010).

in most heterogeneous amplicon pools. Several methods have been developed for identifying and removing chimeric reads. Haas et al. (2011) developed Chimera Slayer to remove chimeras by comparing each sequence read against a curated database of non-chimeric, SSU rRNA genes. Quince et al. (2011) and Edgar et al. (2011) have developed chimera checkers (Perseus and UChime, respectively) using reference comparison, but optimized for shorter NGS reads. UChime also performs a *de novo* check by comparing triplets of reads in a data set to see if any reads appear to be a combination of two other reads from the same amplicon pool. Because chimeras have large sections of sequence substitutions, we can conservatively presume that each unique chimera will likely create a new OTU during clustering. Chimera checking, therefore, is just as important for improving OTU richness estimates as basic sequence quality filtering.

Sequencing and Clustering Best Practices

DNA Amplification

Given the known limitations of both NGS technologies and of clustering methods, it is particularly important to exercise best practices in both generation and use of NGS data. The first way to reduce the impact of sequencing errors on microbial ecology investigations is to reduce the rates of base incorporation errors and chimera formation in the DNA amplification step. This includes the use

of a high-fidelity polymerase such as Platinum Taq, reducing contaminants that interfere with polymerase processivity or proofreading, minimizing the number of PCR cycles, and optimizing the amount of input DNA (Acinas et al., 2004; Lahr and Katz, 2009; Qiu et al., 2001). Technical replicates of the amplification step facilitate discrimination of novel sequences from high-frequency errors or chimeras.

Quality Filtering

Several authors have provided extensive analyses of quality filtering methods for both 454 and Illumina sequencing technologies (Huse et al., 2007; Meacham et al., 2011; Minoche et al., 2011; Quince et al., 2011; Schloss et al., 2011). In brief, removing reads with ambiguous bases (*Ns*), with low-quality scores, that are truncated, and that have known mismatches in the primer region are computationally simple means of decreasing the error rate several-fold. More computationally intensive algorithms such as AmpliconNoise (Quince et al., 2011) (implemented either directly or as implemented in QIIME [Caporaso et al., 2010] or mothur [Schloss et al., 2009]) can be used as the first step for quality filtering pyrosequencing data. Illumina paired-end technology allows another, more traditional approach to quality filtering: each amplicon is sequenced in both directions, so if the amplicon length is less than twice the read length, overlap of the complementary sequences can be used to assess accuracy (Bartram et al., 2011; Gloor et al., 2010). If the two reads overlap completely, meaning the amplicon length is less than or equal to the read length, requiring no mismatches could lead to data sets with little or no sequencing error, although systematic errors that are the same in both directions, could still exist.

Sequence quality filtering should be followed by chimera checking. Using a reference database for chimera detection is the standard method for identifying and removing chimeras and is very effective. Unfortunately, sequencing of novel environments is fast outstripping curated database growth, and novel genes that are parents of chimeras will be missed by reference comparison methods. A combination of both reference comparison and *de novo* chimera checking should always be used.

Smoothing Imperfect Data by Aggregation

Even with the best quality filtering and chimera checking, sequencing data will still contain reads with base incorporation and base calling errors, chimeric reads, reads from contaminating DNA, and reads from amplification of non-target areas of sample DNA. The methods chosen for downstream analysis of the data will determine the degree of impact these errors will have on research results.

Assigning reads to their closest match in a database of sequences annotated with defined taxonomy is a simple, straightforward way to minimize the impact of small sequence differences, segregate chimeras, identify contaminants, and

eliminate reads from non-target amplification. For instance, many of the reference SSU rRNA databases provide taxonomy primarily to the genus level. While this may not be sufficient resolution for some analyses, the use of genus-level taxonomy for analysis will assign similar sequences to the same genus, so that sequences with even a moderate number of errors are likely to still be classified together with their template. Chimeras of sequences within the same genus will remain in that genus, and chimeras of sequences from different genera within a family will tend to not be classified at the genus level; these cross-genera chimeras will aggregate as sequences classified to the family level but no further. The presence of unexpected genera, such as *Ralstonia* in deep-sea sediment samples, can indicate contamination.

Routinely assigning taxonomy to all sequence reads has the additional advantage of quickly identifying non-target reads. Occasionally, the SSU rRNA primers can amplify DNA from a section of the genome other than the SSU rRNA gene. The resulting amplicons will be quite divergent from any 16S reference gene. Finally, PCR primers designed to be specific for domains or other groups often amplify the rRNA SSU gene from a subset of species outside that group, generally in a non-quantitative manner. Reads that map to taxa outside the group to which the primers were designed can easily be eliminated from downstream analyses.

The other common way to aggregate similar sequences is to cluster them into OTUs based on percent similarity, as described previously. This is often done after assigning taxonomy to reads so that reads from non-template amplicons or from taxonomic groups outside the target range of the primers can be eliminated. Clustering reads based on a similarity score of 97 percent has become a common way of approximating "species" or phylotype in the absence of taxonomic resolution. However, it is important to note that different algorithms create very different 97 percent OTUs, and some of these methods lead to OTU inflation, as discussed previously. Among those that do not inflate OTU counts are SLP-PWAL clustering, which uses a nearest-neighbor approach to link sequences likely to be derived from base incorporation error with average neighbor linkage to form OTUs, and methods known as greedy clustering algorithms. One of the more popular of these is UClust (Edgar, 2010). Briefly, the UClust ranks sequences in order of abundance and seeds the first OTU with the most abundant sequence. The next sequence is compared to the first, and if it is within the clustering threshold then it is added to the OTU, if not then it becomes the seed of a second OTU. Each sequence is compared to the OTU seeds, in order, until all reads have been assigned to an existing OTU or used to create a new OTU.

Diversity in an Imperfect World

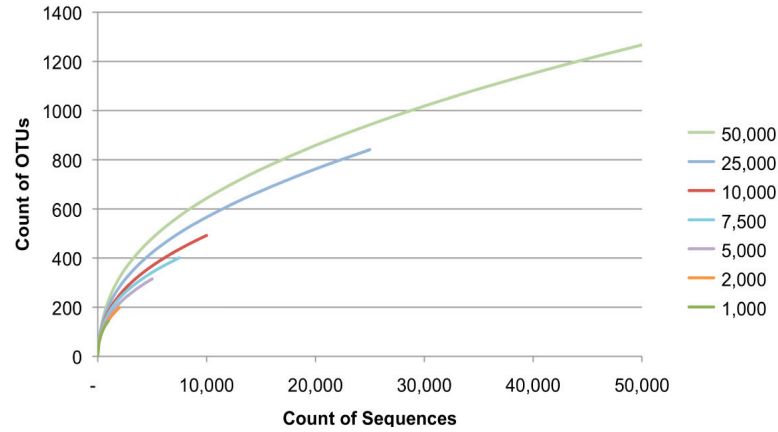
Ecologists measure the diversity of a microbial community in a variety of ways. Richness is the number of different members (OTUs, species, genera, phylotypes, etc.) in a community. Evenness describes the distribution of relative

abundances of the members, whether they have similar abundances or are skewed with some highly abundant and others rare. A community's diversity combines richness and evenness (although richness is often called diversity as well). The richness of a single community is often referred to as alpha diversity. The degree of similarity or difference between two or more communities in richness or evenness is beta diversity. One important conceptual difference between these two measures is that alpha diversity is nearly always used to designate an estimate of the true richness of the community from which a data sample was taken, while beta diversity is generally a metric describing the similarity between the observed richness or evenness of two samples (though there are methods for estimating community similarity from sample data [Chao, 2004], they are rarely used in molecular microbial ecology).

Algorithms used to calculate diversity differ in the stability of their results in the presence of residual sequencing errors and chimeras (as reflected in OTU inflation) and in the depth of sampling. Estimates of sample richness (alpha diversity) are particularly susceptible to the impacts of both. Rarefaction curves, while not true estimators of alpha diversity, are often used to illustrate the relationship between the observed number of community members (i.e., OTUs) and sampling depth. The number of OTUs observed for a range of subsampling sizes are plotted against the average number of OTUs observed for each subsample size, describing the number of new members discovered for an incremental increase in sampling effort. With very small subsamples, small increases in sampling depth lead to the discovery of many new members. As sampling depth increases, the number of new members found decreases and begins to asymptote as the sampling depth provides a more complete picture of the underlying community. If OTUs are created using a clustering method that inflates with sampling depth, the rarefaction curves cannot reach an asymptote and instead will increase linearly as a direct function of sampling depth. Superimposing multiple rarefaction curves from complete-linkage OTUs demonstrates the dependence of the slope of the rarefaction curve on the sample depth (Figure A6-2, Panel A). Rarefaction curves based on depth-independent OTUs (e.g., SLP-PWAL or UClust) have essentially identical slopes for sample depths ranging over two orders of magnitude (Figure A6-2, Panel B).

Some nonparametric methods of estimating alpha diversity remain sensitive to sampling depth independent of OTU inflation or other forms of error. Two of the most common estimators, ACE (Chao and Lee, 1992) and Chao1 (Chao, 1984), are well known to underestimate richness when used for populations where many of the members are "unseen" in the sample, and are more accurately considered lower bounds rather than true estimates. In other words, they do not perform well when a community is drastically undersampled, as is the case for most samples of microbial communities. Panel A of Figure A6-3 illustrates the impact of both sampling depth and clustering method on the estimated richness for a human gut microbiome sample. With increasing depth, the sample more

A



B

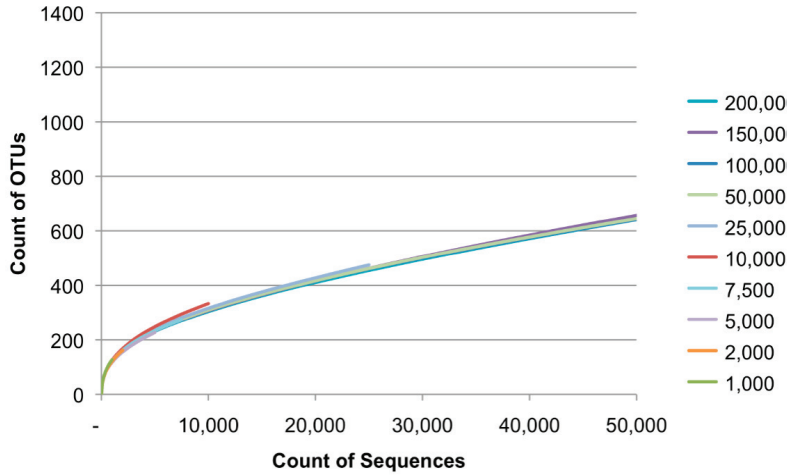


FIGURE A6-2 Panel A, Rarefaction curve for OTUs generated from Human Microbiome Project stool samples using the V3-V5 region. We subsampled the 50,000 reads into smaller samples ranging from 1,000 to the full 50,000 and then performed complete linkage clustering independently for each sample size. A rarefaction curve shows the average number of OTUs (y) found for a given number of sequences (x) in the clustering set. The larger the sample size in the clustering, the most OTUs found for a given number of sequences, demonstrating that the complete linkage algorithm is depth dependent. Panel B, Rarefaction curves for OTUs generated by average linkage clustering. The curves overlap or decrease negligibly, demonstrating the depth independence of the average linkage algorithm.

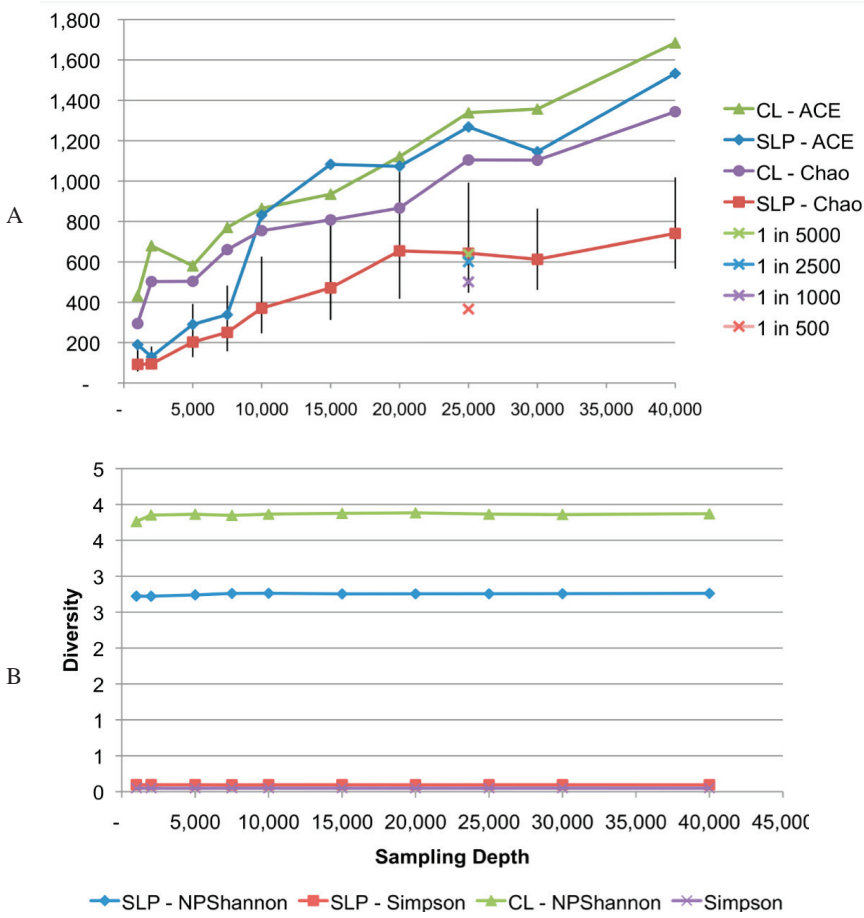


FIGURE A6-3 Panel A, We calculated alpha diversity (richness) using both ACE and Chao estimators with clusters based on complete linkage and average linkage algorithms. Chao values are consistently smaller than ACE values, and SLP-PWAL are lower than complete linkage. All Chao values for sampling depths from 20,000 to 40,000 are within the error bars for all of those estimates, demonstrating a stability in the estimate for this data set with subsamples of 20,000 or more. We recalculated the Chao values for SLP-PWAL clusters adjusting for spurious OTUs at rates from 1 in 5,000 reads to 1 in 500 reads. The Chao value assuming 1 spurious OTU in 1,000 reads is still within the error bars of the value without removing spurious OTUs. Panel B, Simpson and Shannon diversity estimators are independent of the depth of sampling, providing consistent values from subsample sizes ranging from 2,000 to 40,000.

adequately reflects the underlying community, and the richness estimate can stabilize. In this example, Chao1 estimates rise rapidly with sampling depth from 1,000 to 20,000, but then the estimates level. Doubling the sample depth from 20,000 to 40,000 does not change the estimated richness beyond the error bars. Plots of richness against subsample depth may serve as a reality check on the stability (if not accuracy) of calculated richness for a given sample. Parametric estimators such as CatchAll (Bunge, 2011) are less sensitive to sample size and undersampling and also make use of a wider range of OTU sizes in extrapolating total richness.

Even with depth-independent OTU clustering methods, community richness estimates can be vulnerable to OTU inflation. Most algorithms for estimating richness heavily weight the number of OTUs with one or two reads (singleton and doubleton OTUs). We can assume that most spurious OTUs are singletons (this will not always be the case, for instance early-round chimeras can be amplified in a sample, but it is a conservative assumption). By removing the estimated number of spurious OTUs (e.g., 1 in 1,000) from the count of singletons in the species abundance data used to calculate richness, we can compensate for OTU inflation in estimating alpha diversity. We do not need to know which of our OTUs are spurious and which are true; we only need an estimated number. Panel A of Figure A6-3 shows the impact of several OTU inflation rates on richness estimates for a subsample of 25,000.

At any sampling depth, microbial communities consistently display a long-tail distribution, and therefore evenness will be low at all sampling depths. Both Simpson's and Shannon's diversity estimates show very little direct susceptibility to the sampling depth (Figure A6-3, Panel B), but they are still affected by the OTU clustering method. Clearly, choosing a clustering algorithm that minimizes OTU inflation and that is stable to sample size is critical at all times.

The list of beta diversity metrics that compare the degree of similarity or difference between two communities is very long. We highlight the importance of using metrics that are robust to both differences in sampling depth and undersampling using three distance metrics: Jaccard presence/absence, Bray-Curtis, and Morisita-Horn. Results using the Yue-Clayton distance were consistently similar to Morisita-Horn (results not shown). We subsampled a large data set to provide pseudo-replicates that we expect to be similar. If we compare a subsample with itself, then the distance by any metric will be zero (or 1), and the expectation is that a comparison of two different random subsamples should give a very similar value. In Figure A6-4, Panel A, we take multiple subsamples of 5,000 reads from a sample of 50,000 reads and calculate the beta diversity of pairs of replicates. Although we would like the distances to approach zero, we know that microbial diversity is so great that there will still be differences between the replicates due to incomplete sampling. Morisita-Horn, Bray-Curtis, and Yue-Clayton (not shown) all return distances of about 5 percent or less. The Jaccard presence/absence metric, on the other hand, returns a community distance of 50-60 percent.

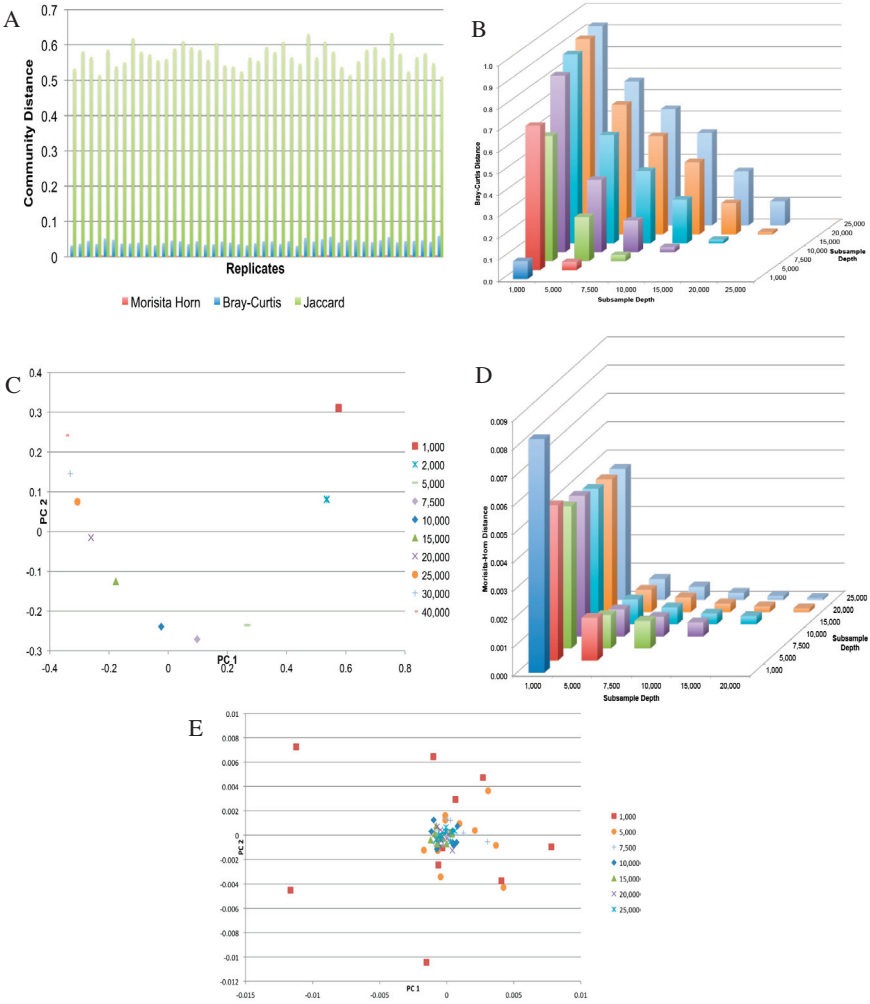


FIGURE A6-4 Panel A, Selecting multiple random subsamples of 5,000 reads from a larger data set of 50,000 reads, we created a set of pseudo-replicate samples. Because they all represent the same larger sample, the pairwise distances should be very small. The use of Jaccard presence/absence is highly affected by the membership within the rare biosphere. Bray-Curtis includes abundance and returns much smaller values. Morisita-Horn was specifically designed for smaller samples and to adjust for different sample sizes and returns values approaching zero. Panel B, Bray-Curtis uses absolute counts, rather than relative abundances, and displays increasing community distance values with increasing differences in sampling depth. Panel C, Even when considering only the Bray-Curtis distance between samples of the same depth, the values returned are still affected

The test of presence or absence in a community, where fewer members have been detected than not, will always show large differences between the communities. The small number of abundant members will be consistent across replicates, but rare members detected in any given replicate will vary. It is not surprising that upwards of half the members of a replicate pair can be different.

In the case of similar subsample size, both Bray-Curtis and Morisita-Horn returned suitably low beta diversity values. Unfortunately, NGS data sets vary greatly in the number of reads depending on the amount of the amplicon library loaded on the sequencer and the quality of the particular sequencing run. Bray-Curtis and Morisita-Horn compare not only which members are present, but also the abundance of each. To adjust for undersampling, Morisita-Horn emphasizes the abundant members, assuming that if a member is abundant in one sample and not detected in the other sample, then this reflects a true difference in the communities, while if a rare member is detected in one and not the other, this may be an artifact of undersampling. Bray-Curtis includes but does not differentially weight the abundance information.

In Panels B through E in Figure A6-4, we illustrate the effect of different sample sizes on beta diversity values. Bray-Curtis returns small diversity values for subsamples of the same size and increasingly larger values when comparing samples of different sizes; the average distance between subsamples of 1,000 and 25,000 is about 90 percent (Figure A6-4, Panel B). This disparity comes in part because the Bray-Curtis method uses absolute rather than relative abundance. Even if data are subsampled to the same depth, Bray-Curtis can still cause misinterpretations of results when combined with one of the most common visualization tools for illustrating community similarity, principal coordinate analysis (PCoA). Panel C of Figure A6-4, shows that Bray-Curtis measures of subsample similarity pairs subsamples based on read depth, *even though they are subsamples from the very same community*. Even though the absolute distances were low for pairs from the same subsampling depth, a PCoA plot does not report absolute differences but scales according to the set of distances used. For subtle changes in

---

by the sampling depth. In a classic principal coordinate analysis plot, commonly used in microbial community studies, the clustering is affected by the sampling depth used. Panel D, The Morisita-Horn metric, which uses relative abundances and places more emphasis on the more abundant community members, is not appreciably affected by comparing samples of different sizes. The larger distances consistently reported for samples of 1,000 reads likely reflects the lower bound of sample depth required to be representative for this data set. Even so, the values returned are below 1 percent. Panel E, The distance values returned by Morisita-Horn cluster together for samples with more than 7,500 reads. The increasing spread of points for sampling depths less than 7,500 presumably reflects the lower bound of representative sampling for this data set, rather than an inherent limitation of the Morisita-Horn metric.

the community, a method such as Bray-Curtis could lead to erroneous interpretations of community shifts.

Morisita-Horn effectively compensates for sampling depth, returning beta diversity values less than 1 percent for all subsample comparisons (Figure A6-4, Panel D). Interestingly, the Morisita-Horn distances for pairs where one subsample was at 1,000 reads were noticeably higher than the other distances. With increasing depth this divergence disappears. In the PCoA plot, the 1,000 read depth pairs do not cluster either with each other or with any of the other data (Figure A6-4, Panel E). Depth pairs with 5,000 are still divergent but much less so than 1,000. In this particular data set, it appears that a minimum sampling depth of 10,000 is necessary to adequately reflect the community in the subsample.

Continued Evidence for the Rare Biosphere

In evaluating both the extent of the rare biosphere and our ability to meaningfully sample it, it is helpful to put the word "rare" into perspective. Estimates vary, but let us assume that there are at least 1×10^{11} bacterial cells in a single gram dry weight of human stool (Franks et al., 1998). If we sequence DNA from a single gram of stool and analyze a relatively large data set of 50,000 (5×10^4) reads, we are sampling a tiny fraction of the census population. An OTU found as a singleton may be present at a frequency of only 1/50,000, but that is 2×10^6 cells/g, which may not be an insignificant number.

But is the long-tail distribution, while consistent across bacterial communities sampled from human and other hosts, marine, freshwater, soil, sand, leaves, sewage, and any number of other environments, merely an artifact of the known phenomenon of OTU inflation caused by deep sequencing? Returning to the empirically derived estimate of 1 spurious OTU per 1,000 reads, we can remove a fraction of singleton OTUs equal to those attributed to OTU inflation (Figure A6-5, Panel A). What we see is that even if we remove 1 out of every 500 singleton OTUs, the distribution retains its characteristic shape, because the fraction of singletons removed compared to the number observed is relatively small. Any spurious OTUs are simply extending the end of the tail incrementally; they are not fundamentally altering the shape of the species abundance curve.

Everything May or May Not Be Everywhere, but Everything Is Rare Somewhere

One of the stronger pieces of evidence supporting the existence of the rare biosphere comes from comparing the sequences found in different microbial communities. The human gut microbiome, for example, varies greatly between subjects. In practice, this leads to a wide range of relative abundances of even the most common OTUs. Panels B and C in Figures A6-5 show the 100 most abundant OTUs across 208 Human Microbiome Project subjects in rank order (Huse et al., 2012). The maximum abundance in a single subject for each OTU is

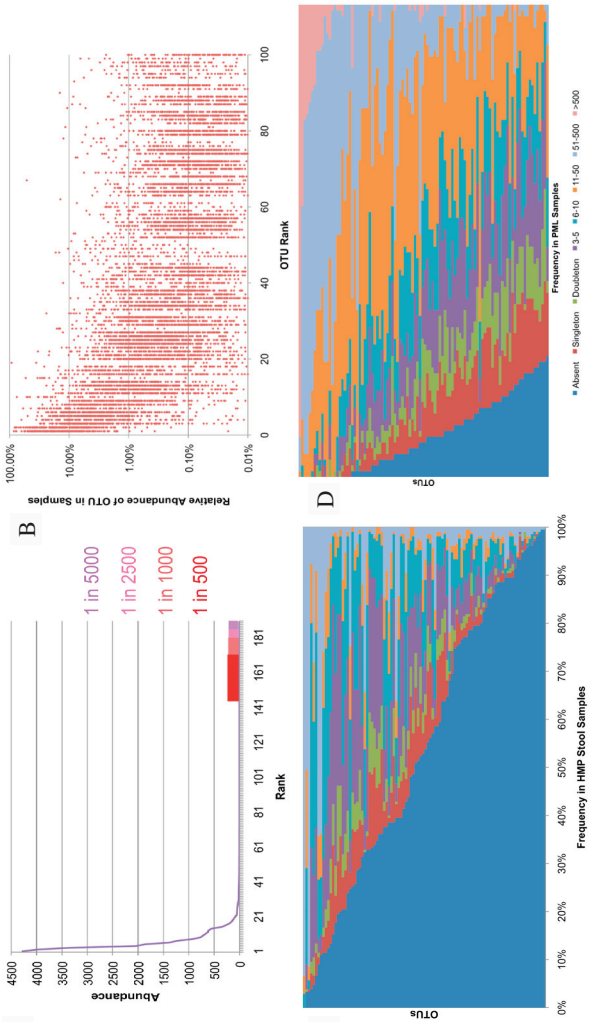


FIGURE A6-5 Panel A, The rank abundance curve shows only minor reductions in the long tail even assuming that as many as 1 in 500 reads generates a spurious OTU. Panel B, The relative abundance of the 100 most abundant OTUs in the Human Microbiome Project stool samples. Each dot represents a separate subject. Even the most abundant OTUs that dominant some samples, are also rare in other samples, indicating that an OTU that is rare in an individual is not necessarily either spurious or consistently rare. Low abundance taxa should not be dismissed based on only a few samples. Panel C, The same stool samples as plotted in Panel D, but here with absolute abundance in a sample. The blue expanse on the left represents the percentage of samples that do not contain the OTU. Even OTUs that are more absent across samples than present appear consistently in multiple samples with an abundance greater than 100. Panel D, An another example of the abundance graph in Panel C, with data from aquatic samples of the English Channel. Here the intersample variation is much lower.

1-100 percent (Figure A6-5, Panel B). The minimum abundance for each of these (except the first most abundant OTU) is within the rare biosphere for at least one subject. In other words, essentially all of the most abundant gut OTUs are highly abundant in some subjects and rare or not detected in others. Panel C of Figure A6-5 uses absolute rather than relative abundance and frequencies to portray the same data. Looking only at a single subject, we might be tempted to discount rare OTUs as noise in the data rather than a true rare biosphere signal. In the greater context of many samples, however, we realize that true rare members are prevalent across subjects, and these same members can dominate the microbiome in other subjects. This same pattern can be seen in other environments including the waters of the English Channel (Gilbert et al., 2009) (Figure A6-5, Panel D).

Conclusions

The use of NGS methods has revolutionized microbial ecology. But, as with any new technology, new challenges must be met. For accurate results, great care must be taken to reduce the rates of sequencing errors and to remove DNA amplification chimeras, using high-quality de-noising or paired-end overlap filtering, and chimera detection. These initial steps, however, are not enough. Researchers must also select bioinformatics tools that avoid artificially inflating the number of OTU clusters, alpha diversity estimates, and beta diversity estimates. OTU clustering methods such as SLP-PWAL and UClust reduce inflation, whereas methods employing multiple sequence alignments and complete linkage clustering overestimate the appropriate number of OTUs.

The selection of diversity metrics affects the research results. Simple richness estimates are sample size dependent. Because the most common estimators are known to be affected by undersampling, larger sample sizes (in the absence of depth-dependent OTU inflation) may provide the most accurate richness estimates. One means to enhance the interpretation of richness estimates is to plot the richness at subsample depths for a given sample to see whether the sample depth is sufficient for the estimate to be stable or whether the sample depth is still within the zone of distinct depth-dependence. Fortuitously, both Simpson's and Shannon's diversity estimates show independence of sample depth.

Beta diversity should only be calculated with an appropriately robust metric that can accommodate sample depth. Even in cases where multiple samples are subsampled to the same depth before calculating intercommunity distance, the use of a depth-dependent metric such as Bray-Curtis will still be affected by depth and can in cases of more subtle shifts in community structure skew PCoA plots, resulting in possible misinterpretation of results. The practice of subsampling to the minimum can introduce artifacts of undersampling as demonstrated in Figure A6-5. A much more robust method is to select a beta diversity algorithm such as Morisita-Horn or Yue-Clayton that does not require subsampling. For all alpha and beta diversity calculations there are thresholds of undersampling

that no metric selection can overcome. Therefore, in the absence of other depth-dependent overestimates (such as poor selection of clustering method), it is best to use full sample sizes rather than subselecting to a minimal and therefore less representative sample size.

Even in the best of all research worlds, errors, OTU inflation, chimeras, contamination, and other inaccuracies will still exist. In this light, the use of multiple samples for determining when low abundance "errare" OTUs or taxa are errors and when they are true rare members is necessary. One straightforward means for deciding to trust the validity is if an OTU occurs abundantly in any sample. By clustering OTUs or using taxonomy and performing bioinformatics analyses across multiple samples at once, it is easy to detect abundant members in the set of samples, validating those members in communities where they are rare. Given our current techniques, context is the best method for discerning truth from fiction in the rare biosphere.

References

Acinas, S. G., V. Klepac-Ceraj, D. E. Hunt, C. Pharino, I. Ceraj, D. L. Distel, and M. F. Polz. 2004. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* 430:551-554.

Bartram, A. K., M. D. J. Lynch, J. C. Stearns, G. Moreno-Hagelsieb, and J. D. Neufeld. 2011. Generation of multimillion-sequence 16s rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Applied and Environmental Microbiology* 77(11):3846-3852.

Bunge, J., 2011. Estimating the number of species with CatchAll. Proceedings of the 2011 Pacific Symposium on Biocomputing.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7(5):335-336.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* 11:265-270.

Chao, A., and S.-M. Lee. 1992. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* 87(417):210-217.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with ARB. *Applied and Enviromental. Microbiology* 72(7):5069-5072.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. 2011. Uchime improves sensitivity and speed of chimera detection. *Bioinformatics* 27(16):2194-2200.

Franks, A. H., H. J. M. Harmsen, G. C. Raangs, G. J. Jansen, F. Schut, and G. W. Welling. 1998. Variations of bacterial populations in human feces measured by fluorescent in situ hybridization with group-specific 16s rRNA-targeted oligonucleotide probes. *Applied and Environmental Microbiology* 64(9):3336-3345.

Gilbert, J. A., F. Dawn, S. Paul, N. Lindsay, O. Anna, S. Tim, J. S. Paul, H. Sue, and J. Ian. 2009. The seasonal structure of microbial communities in the western English Channel. *Environmental Microbiology* 11(12):3132-3139.

Gloor, G. B., R. Hummelen, J. M. Macklaim, R. J. Dickson, A. D. Fernandes, R. MacPhee, and G. Reid. 2010. Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS ONE* 5(10):e15406.

Haas, B. J., D. Gevers, A. M. Earl, M. Feldgarden, D. V. Ward, G. Giannoukos, D. Ciulla, D. Tabbaa, S. K. Highlander, E. Sodergren, B. Methé, T. Z. DeSantis, C. The Human Microbiome, J. F. Petrosino, R. Knight, and B. W. Birren. 2011. Chimeric 16s rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* 21(3):494-504.

Huber, J. A., H. G. Morrison, S. M. Huse, P. R. Neal, M. L. Sogin, and D. B. Mark Welch. 2009. Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environmental Microbiology* 11(5):1292-1302.

Huse, S., J. Huber, H. Morrison, M. Sogin, and D. Mark Welch. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* 8(7):R143.

Huse, S. M., D. Mark Welch, H. G. Morrison, and M. L. Sogin. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* 12(7):1889-1898.

Huse, S. M., Y. Ye, Y. Zhou, and A. A. Fodor. 2012. A core human microbiome as viewed through 16s rRNA sequence clusters. *PLoS ONE* 7(6):e34242. doi:10.1371/journal.pone.0034242.

Kunin, V., A. Engelbrektson, H. Ochman, and P. Hugenholtz. 2010. Wrinkles in the rare biosphere: Pyrosequencing errors lead to artificial inflation of diversity estimates. *Environmental Microbiology* 12(1):118-123.

Lahr, D. J. G., and L. A. Katz. 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques* 47:857-866.

Meacham, F., D. Boffelli, J. Dhahbi, D. Martin, M. Singer, and L. Pachter. 2011. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 12(1):451.

Minoche, A., J. Dohm, and H. Himmelbauer. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology* 12(11):R112.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glockner. 2007. Silva: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* 35(21):7188-7196.

Qiu, X., L. Wu, H. Huang, P. E. McDonel, A. V. Palumbo, J. M. Tiedje, and J. Zhou. 2001. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16s rRNA gene-based cloning *Applied and Environmental Microbiology* 67:880-887.

Quince, C., A. Lanzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan. 2009. Noise and the accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods* 6(9):639-641.

Quince, C., A. Lanzen, R. Davenport, and P. Turnbaugh. 2011. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12(1):38.

Reeder, J., and R. Knight. 2009. The "rare biosphere": A reality check. *Nature Methods* 6(9):636-637.

Schloss, P. D., D. Gevers, and S. L. Westcott. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16s rRNA-based studies. *PLoS ONE* 6(12):e27310.

Schloss, P. D., and J. Handelsman. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology* 71(3):1501-1506.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. 2009. Introducing mothur: Open source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75(23):7537-7541.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proceedings of the National Academy of Sciences* 103(32):12115-12120.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. A naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261-5267.

A7

PHYLOGEOGRAPHY AND MOLECULAR EPIDEMIOLOGY OF *YERSINIA PESTIS* IN MADAGASCAR[23]

Amy J. Vogler,[24] *Fabien Chan*,[25] *David M. Wagner*,[24] *Philippe Roumagnac*,[26,¤a] *Judy Lee*,[24] *Roxanne Nera*,[24] *Mark Eppinger*,[27] *Jacques Ravel*,[26] *Lila Rahalison*,[24,¤b] *Bruno W. Rasoamanana*,[24,¤c] *Stephen M. Beckstrom-Sternberg*,[24,28] *Mark Achtman*,[24,29] *Suzanne Chanteau*,[24,¤d] *and Paul* ***Keim***[24,27,*]

Abstract

Background Plague was introduced to Madagascar in 1898 and continues to be a significant human health problem. It exists mainly in the central highlands, but

[23] Reprinted from *PLoS Neglected Tropical Diseases*. Originally published as: Vogler AJ, Chan F, Wagner DM, Roumagnac P, Lee J, et al. (2011) Phylogeography and Molecular Epidemiology of *Yersinia pestis* in Madagascar. *PLoS Negl Trop Dis* 5(9): e1319. doi:10.1371/journal.pntd.0001319

Editor: Mathieu Picardeau, Institut Pasteur, France.

[24] Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, Arizona, USA.

[25] Institut Pasteur de Madagascar, Antananarivo, Madagascar.

[26] Max Planck Institut für Infektionsbiologie, Berlin, Germany.

[27] Institute for Genomic Sciences (IGS), School of Medicine, University of Maryland, Baltimore, Maryland, USA.

[28] Translational Genomics Research Institute, Phoenix, Arizona, United States of America.

[29] Environmental Research Institute, University College Cork, Cork, Ireland.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Paul.Keim@nau.edu

¤a Current address: Unite Mixte de Recherche 6191, Centre National de la Recherche Scientifique-Commissariat à l'Energie Atomique-Aix-Marseille Université, Commissariat à l'Energie Atomique Cadarache, Saint Paul Lez Durance, France.

¤b Current address: Centers for Disease Control and Prevention, Atlanta, Georgia, USA.

¤c Current address: Laboratoire de Biologie Médicale du Tampon, Le Tampon, Reunion Island, France.

¤d Current address: Institut Pasteur de Nouvelle-Calédonie, Nouméa, New Caledonia.

in the 1990s it was reintroduced to the port city of Mahajanga, where it caused extensive human outbreaks. Despite its prevalence, the phylogeography and molecular epidemiology of *Y. pestis* in Madagascar has been difficult to study due to the great genetic similarity among isolates. We examine island-wide geographic-genetic patterns based upon whole-genome discovery of SNPs, SNP genotyping, and hypervariable variable-number tandem repeat (VNTR) loci to gain insight into the maintenance and spread of *Y. pestis* in Madagascar.

Methodology and principal findings We analyzed a set of 262 Malagasy isolates using a set of 56 SNPs and a 43-locus multi-locus VNTR analysis (MLVA) system. We then analyzed the geographic distribution of the subclades and identified patterns related to the maintenance and spread of plague in Madagascar. We find relatively high levels of VNTR diversity in addition to several SNP differences. We identify two major groups, Groups I and II, which are subsequently divided into 11 and 4 subclades, respectively. *Y. pestis* appears to be maintained in several geographically separate subpopulations. There is also evidence for multiple long distance transfers of *Y. pestis*, likely human mediated. Such transfers have resulted in the reintroduction and establishment of plague in the port city of Mahajanga, where there is evidence for multiple transfers both from and to the central highlands.

Conclusions and Significance The maintenance and spread of *Y. pestis* in Madagascar is a dynamic and highly active process that relies on the natural cycle between the primary host, the black rat, and its flea vectors as well as human activity.

Author Summary

Plague, caused by the bacterium *Yersinia pestis*, has been a problem in Madagascar since it was introduced in 1898. It mainly affects the central highlands, but also has caused several large outbreaks in the port city of Mahajanga, after it was reintroduced there in the 1990s. Despite its prevalence, the genetic diversity and related geographic distribution of different genetic groups of *Y. pestis* in Madagascar has been difficult to study due to the great genetic similarity among isolates. We subtyped a set of Malagasy isolates and identified two major genetic groups that were subsequently divided into 11 and 4 subgroups, respectively. *Y. pestis* appears to be maintained in several geographically separate subpopulations. There is also evidence for multiple long distance transfers of *Y. pestis*, likely human mediated. Such transfers have resulted in the reintroduction and establishment of plague in the port city of Mahajanga where there is evidence for multiple transfers both from and to the central highlands. The maintenance and spread of *Y. pestis* in Madagascar is a dynamic and highly active process that

relies on the natural cycle between the primary host, the black rat, and its flea vectors as well as human activity.

Introduction

Throughout recorded history, *Yersinia pestis*, etiologic agent of plague, has spread multiple times from foci in central Asia in greatly widening swaths as human-mediated transport became more efficient (Morelli et al., 2010). Plague attained its current global distribution during the current "third" pandemic, which began in 1855 in the Chinese province of Yünnan, when it was introduced to many previously unaffected countries via infected rats on steam ships (Perry and Fetherston, 1997). Plague caused widespread outbreaks during this introduction period (~1900 A.D.), and though disease incidence has since largely decreased, plague remains a significant human health threat due to the severe and often fatal nature of the disease, the many natural plague foci (Perry and Fetherston, 1997), and its potential as a bioterror agent (it is currently classified as a Class A Select Agent [Rotz et al., 2002]). Plague is of particular significance in Madagascar, which has reported some of the highest human plague case numbers (18%–60% of the world total each year between 1995 and 2009) (WHO, 2010) and was the origin of a natural multi-drug resistant strain of *Y. pestis* (Galimand et al., 1997; Welch et al., 2007).

Plague has been a problem in Madagascar since its introduction during the current pandemic. It was first introduced to Toamasina in 1898 (Brygoo, 1966), likely via India (Morelli et al., 2010), with outbreaks in other coastal cities soon after. In 1921, plague reached the capital, Antananarivo, likely via infected rats transported on the railroad linking Toamasina and Antananarivo. Subsequent rat epizootics signaled the establishment of plague in the central highlands (Brygoo, 1966). Plague then disappeared from the coast and now exists within two large areas in the central and northern highlands above 800 m in elevation (Chanteau et al., 1998). This elevational distribution of plague is linked to the presence of the flea vectors *Xenopsylla cheopis* and *Synopsyllus fonquerniei*, which are less abundant and absent, respectively, below 800 m (Duplantier, 2001; Duplantier et al., 1999). Plague has never disappeared from this region, and although it was relatively controlled in the 1950s due to public hygiene improvements and the introduction of antibiotics and insecticides, disease incidence began increasing in 1989 (Chanteau et al., 1998, 2000; Migliani et al., 2006). Human plague cases peaked in 1997 but continue to occur at high frequencies, making Madagascar among the top three countries for human plague cases during the past 15 years (WHO, 2010).

A third, newly emerged plague focus outside the central and northern highlands is the port city of Mahajanga, located ~400 km by air from Antananarivo (Chanteau et al., 1998). Plague first appeared in Mahajanga during an outbreak in 1902. Subsequent outbreaks occurred in 1907 and between 1924 and 1928

(Brygoo, 1966). Plague then disappeared from Mahajanga for a period of 62 years before reappearing during a large outbreak in 1991 (Laventure et al., 1991). Subsequent outbreaks occurred from 1995–1999 (Boisier et al., 1997, 2000; Rasolomaharo et al., 1995). During this time, the Mahajanga focus was responsible for ~30% of the reported human plague cases in Madagascar (Boisier et al., 2002). Interestingly, this focus likely represents one of the only examples of plague being reintroduced to an area where it had gone extinct, rather than emergence from a silently cycling rodent reservoir without telltale human cases (Duplantier et al., 2005).

Molecular subtyping of *Y. pestis* for epidemiological tracking has been difficult due to a lack of genetic diversity (Achtman et al., 1999). SNP genotyping (Achtman et al., 2004; Eppinger et al., 2010; Morelli et al., 2010), ribotyping (Guiyoule et al., 1994), IS*100* insertion element restriction fragment length polymorphism (RFLP) analysis (Achtman et al., 1999), PCR-based IS*100* genotyping (Achtman et al., 2004; Motin et al., 2002) and pulsed-field gel electrophoresis (PFGE) (Lucier and Brubaker, 1992) have been used to differentiate global isolate collections; however, SNP genotyping provides the most robust phylogenetic reconstructions. SNP genotyping (Morelli et al., 2010), ribotyping (Guiyoule et al., 1997), IS*100* insertion element RFLP analysis (Huang et al., 2002), different region (DFR) analysis (Li et al., 2008), clustered regularly interspaced short palindromic repeats (CRISPR) analysis (Cui et al., 2008), ERIC-PCR (Kingston et al., 2009), ERIC-BOX-PCR (Kingston et al., 2009), and PFGE (Huang et al., 2002; Zhang et al., 2009) have shown limited to moderate ability in differentiating isolates on a regional scale. Of these, ribotyping has been applied to a set of 187 Malagasy isolates, but only revealed four ribotypes, three of which were unique to Madagascar (Guiyoule et al., 1997). SNP genotyping of 82 Malagasy isolates provided greater and more phylogenetically informative resolution, revealing two major groups and an additional 10 subgroups derived from these two major groups that were mostly isolate-specific (Morelli et al., 2010). In contrast to these other molecular subtyping methods, multi-locus variable-number tandem repeat (VNTR) analysis (MLVA) has shown high discriminatory power at global (Achtman et al., 2004; Klevytska et al., 2001; Pourcel et al., 2004), regional (Girard et al., 2004; Klevytska et al., 2001; Li et al., 2009; Lowell et al., 2007; Zhang et al., 2009), and local scales (Girard et al., 2004), indicating its likely usefulness for further differentiation among *Y. pestis* isolates from Madagascar.

The use of SNPs and MLVA together, in a hierarchical approach, has been successfully applied to clonal, recently emerged pathogens (Keim et al., 2004; Van Ert et al., 2007; Vogler et al., 2009). Point mutations that result in SNPs occur at very low rates, making SNPs relatively rare in the genome, but discoverable through intensive sampling (i.e., whole genome sequencing). In addition, since each SNP likely occurred only once in the evolutionary history of an organism, SNPs represent highly stable phylogenetic markers that can be used for identifying key phylogenetic positions (Keim et al., 2004). However, SNPs discovered

from a limited number of whole genome sequences will have limited resolving power (Keim et al., 2004) since they will only be able to identify phylogenetic groups along the evolutionary path(s) linking the sequenced genomes (Pearson et al., 2004). In contrast, VNTRs possess very high mutation rates and multiple allele states, allowing them to provide a high level of resolution among isolates. Unfortunately, these high mutation rates can lead to mutational saturation and homoplasy, which can obscure deeper phylogenetic relationships, leading to inaccurate phylogenies. Using these two marker types together, in a nested hierarchical approach, with SNPs used to identify major genetic groups followed by VNTRs to provide resolution within those groups, allows for both a deeply rooted phylogenetic hypothesis and high resolution discrimination among closely related isolates (Keim et al., 2004).

We investigated the phylogeography and molecular epidemiology of *Y. pestis* in Madagascar through extensive genotyping and mapping of genetic groups. We genotyped 262 Malagasy isolates from 25 districts from 1939–2005 using 56 SNPs and a 43-marker MLVA system to identify island specific subclades. We then spatially mapped the subclades to examine island-wide geographic-genetic patterns and potential transmission routes.

Methods

Ethics Statement

The DNAs analyzed in this study (Table S1) were extracted from *Y. pestis* cultures that were previously isolated by the Malagasy Central Laboratory for plague and Institut Pasteur de Madagascar as part of Madagascar's national plague surveillance plan. The Malagasy Ministry of Health, as part of this national plague surveillance plan, requires declaration of all suspected human plague cases and collection of biological samples from those cases. These biological samples are analyzed by the Malagasy Central Laboratory for plague and Institut Pasteur de Madagascar, which also maintains any cultures derived from these samples. These cultures are all de-linked from the patients from whom they originated and analyzed anonymously if used in any research study. Thus, for purposes of this study, all of the DNAs derived from *Y. pestis* cultures from human patients were analyzed anonymously. No Malagasy review board existed during the collection period of the cultures (1939–2001) from which the DNAs used in this study were derived. In addition, the Institutional Review Board of Northern Arizona University, where the DNA genotyping was done, did not require review of the research due to the anonymous nature of the samples.

DNAs

DNA was obtained from 262 isolates from 25 different districts from 1939–2005 (Figure S1, Table S1). DNAs consisted of simple heat lysis preparations or whole genome amplification (WGA) (QIAGEN, Valencia, CA) products generated from the heat lysis preps. Most of the isolates were collected by the Malagasy Central Laboratory for plague supervised by the Institut Pasteur de Madagascar and were primarily isolated from human cases with a few isolated from other mammals or fleas. A handful of other isolates were from other institutions (still originally collected by the Malagasy Central Laboratory for plague) or represent publically available whole genome sequences (Table S1).

SNP Genotyping

A total of 56 SNPs were chosen to genotype the Malagasy isolates because they either marked the branches leading to or from the Madagascar clades in a worldwide analysis (Morelli et al., 2010) or were polymorphic among Malagasy isolates (Table S2). These SNPs were either previously identified in a worldwide SNP study on *Y. pestis* using a combination of denaturing high performance liquid chromatography (dHPLC) and whole genome sequence comparisons (Morelli et al., 2010) or identified here through whole genome sequence comparisons among 2 Malagasy whole genome sequences (MG05-1020 [GenBank:AAYS00000000] and IP275 [GenBank:AAOS00000000] [Morelli et al., 2010]) and 14 other *Y. pestis* strain sequences (CO92 [GenBank:AL590842] (Parkhill et al., 2001), FV-1 [GenBank:AAUB00000000] (Touchman et al., 2007), CA88-4125 [GenBank:ABCD00000000] (Auerbach et al., 2007), Antiqua [GenBank:CP000308], Nepal 516 [GenBank:CP000305] (Chain et al., 2006), UG05-0454 [GenBank:AAYR00000000] (Morelli et al., 2010), KIM 10 [GenBank:AE009952] (Deng et al., 2002), F1991016 [GenBank:ABAT00000000], E1979001 [GenBank:AAYV00000000], K1973002 [GenBank:AAYT00000000], B42003004 [GenBank:AAYU00000000] (Eppinger et al., 2009), Pestoides F [GenBank:CP000668] (Garcia et al., 2007), Angola [GenBank:CP000901] (Eppinger et al., 2010) and 91001 [GenBank:AE017042] [Song et al., 2004]). These whole genome sequence comparisons involved comparing the predicted gene sequences of the closed genome of *Y. pestis* strain CO92 (Parkhill et al., 2001) to the completed and draft genomes of all other strains using MUMmer and in-house Perl scripts (Delcher et al., 2002). For genomes with deposited underlying Sanger sequencing read information, a polymorphic site was considered of high quality when its underlying sequence in the query comprised at least three sequencing reads with an average Phred quality score >30 (Eppinger et al., 2010; Ewing et al., 1998).

A TaqMan-minor groove binding (MGB) assay or a melt mismatch amplification mutation assay (Melt-MAMA) was developed for each SNP for use in

genotyping the Malagasy DNAs. A TaqMan-MGB assay was designed around one SNP known to divide Malagasy isolates into two major groups (Mad-43, Table S2). Melt-MAMA assays were designed around the other 55 SNPs as previously described (Vogler et al., 2009). SNP locations, primer sequences, primer concentrations and other information for these assays are presented in Table S2. Primers and probes were designed using Primer Express 3.0 software (Applied Biosystems, Foster City, CA). Each 5 µl TaqMan-MGB reaction contained primers and probes (for concentrations see Table S2), 1× Platinum Quantitative PCR SuperMix-UDG with ROX (Invitrogen, Carlsbad, CA), water and 1 µl of template. Each 5 µl Melt-MAMA reaction contained 1× SYBR Green PCR Master Mix (Applied Biosystems) or 1× EXPRESS SYBR GreenER qPCR Supermix with Premixed ROX (Invitrogen) (for assay-specific master mix see Table S2), derived and ancestral allele-specific MAMA primers, a common reverse primer (for primer concentrations see Table S2), water and 1 µl of diluted DNA template. DNA templates were diluted 1/10 for heat lysis preparations or 1/50 for WGA products. All assays were performed on an Applied Biosystems 7900HT Fast Real-Time PCR System with SDS software v2.3. Thermal cycling conditions for the TaqMan-MGB assay were as follows: 50°C for 2 min, 95°C for 2 min and 50 cycles of 95°C for 15 s and 66°C for 1 min. Thermal cycling conditions for the Melt-MAMA assays were as follows: 50°C for 2 min, 95°C for 10 min and 40 cycles of 95°C for 15 s and 55–65°C for 1 min (see Table S2 for assay-specific annealing temperatures). Melt-MAMA results were interpreted as previously described (Vogler et al., 2009).

MLVA

All 262 Malagasy isolates were also genotyped using a 43-marker MLVA system as previously described (Girard et al., 2004).

Node Assignment

In general, missing SNP data (<0.5% of dataset) were not a factor in node assignment (see SNP phylogenetic analysis below) since data were usually available for an equivalent SNP, thus leading to unambiguous node assignments for most isolates. However, there were four cases where the node assignment was potentially ambiguous. For three isolates missing data for SNP Mad-21 (branch 1.ORI3.k-1.ORI3.o, Table S2), the ancestral allele state was assumed for that SNP for those isolates, since in this and in a previous worldwide analysis (Morelli et al., 2010), only a single isolate, not included among these three, belonged to node "o." For a single isolate missing data for SNP Mad-46 (branch 1.ORI3.d-1.ORI3.h1, Table S2) the derived state was assumed, due to the placement of that isolate in MLVA subclade II.B in a neighbor-joining analysis and the observed

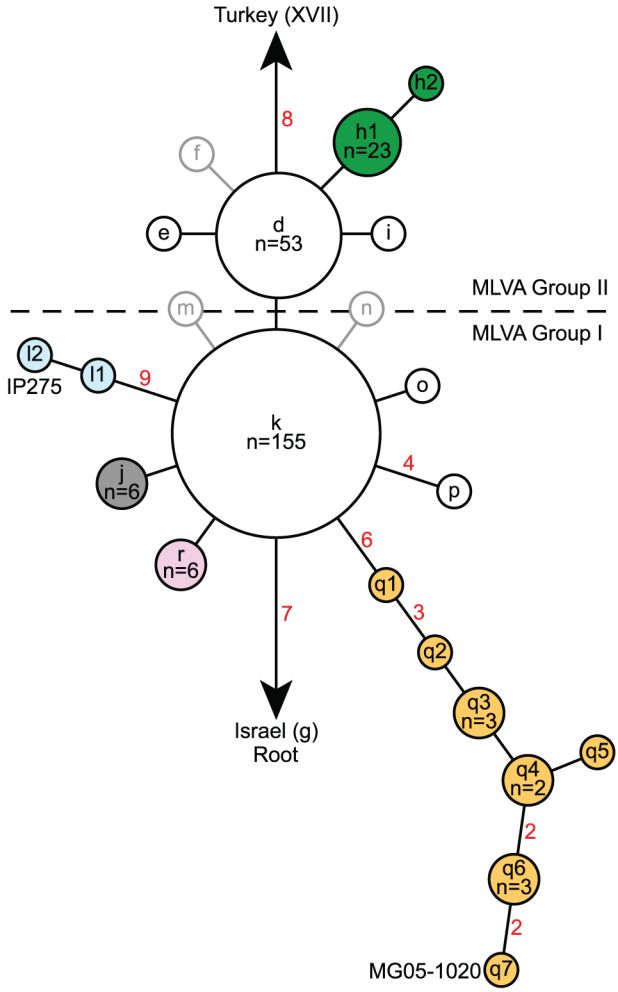


FIGURE A7-1 SNP phylogeny of 262 Malagasy isolates. Nodes were named as in Morelli et al. (2010) (lower case letters) and belong to the 1.ORI3 group described there (Morelli et al., 2010). Previously identified nodes (Morelli et al., 2010) that were expanded in this analysis (h, l and q) have additional number designations (e.g., q1) given to each new node in the expansions. The one entirely new node was assigned a new letter, "r." Previously identified nodes (Morelli et al., 2010) that were not represented by any isolates in this study are represented by gray outlines. Colored nodes correspond to MLVA-identified subclades and are colored the same as their matching MLVA subclades in Figure 2A–B. The number of isolates in nodes with >1 isolate are indicated as are the number of SNPs on branches (red numbers) with >1 SNP. The nodes containing the two sequenced Malagasy strains, MG05-1020 and IP275, are labeled with the strain names.

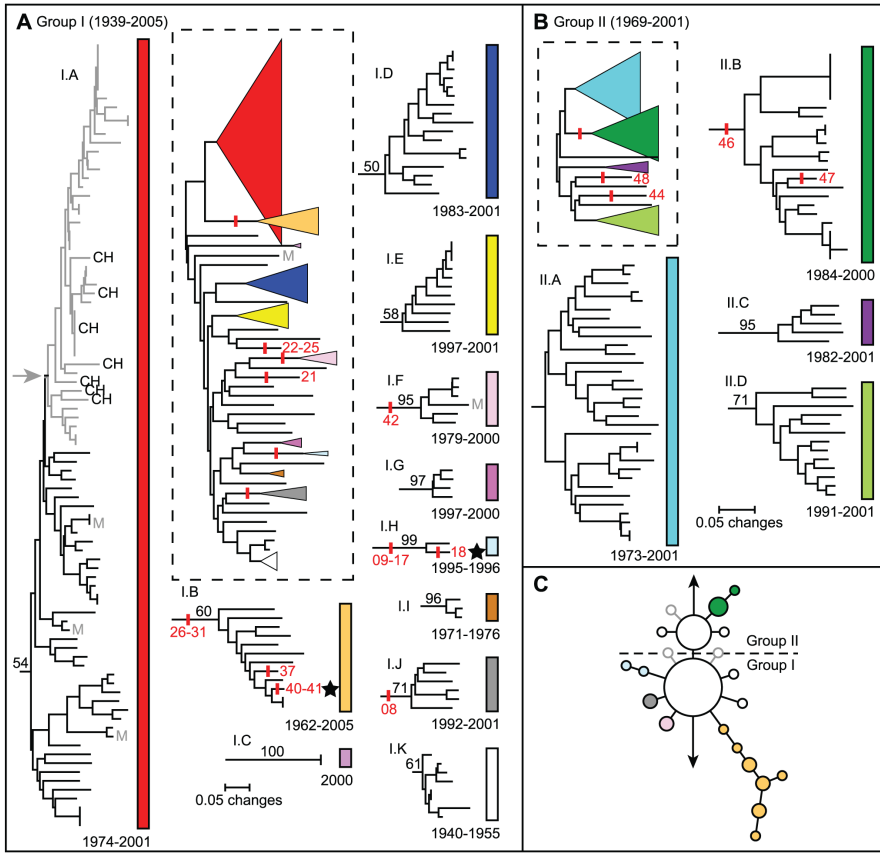


FIGURE A7-2 Neighbor-joining dendrograms based upon MLVA data. Dendrograms for Group I (A) and Group II (B) are indicated. The SNP phylogeny from Figure 1 is also indicated (C) for comparison. Subclades within Groups I and II are collapsed in the full phylogenies (dotted boxes) for those groups (colored triangles) and are then individually expanded to show the structure within each subclade. The expanded subclades are labeled based upon their membership in Group I or II and by a capital letter (e.g., I.A) and are indicated by colored bars. Bootstrap values ≥50 supporting individual subclades are indicated on the expanded subclade phylogenies. SNP locations are indicated by vertical red lines. These red lines are labeled with the SNP ID numbers presented in Table S2 on the full phylogenies for unaffiliated isolate-specific SNPs and on the expanded phylogenies for all other SNPs. The years of isolation for isolates within each full and expanded phylogeny are indicated beside the panel label and underneath the individual phylogeny, respectively. The gray subcluster marked by the gray arrow in subclade I.A represents the "Mahajanga I.A subcluster," a subcluster containing most of the isolates from the Mahajanga plague focus. Seven isolates from the central highlands that also fell within this subcluster are labeled with a "CH." Five Mahajanga isolates that did not belong in this subcluster are labeled with a gray "M" (A). Black stars indicate the locations of the two sequenced Malagasy strains, MG05-1020 in subclade I.B and IP275 in subclade I.H.

congruence between the "h" nodes and MLVA subclade II.B (see phylogenetic analyses below, Table S1).

Phylogenetic Analyses

A hierarchical approach was applied to the phylogenetic analysis of the Malagasy isolates. First, a SNP phylogeny was generated using data from all 56 SNPs (Figure A7-1). Second, neighbor-joining dendrograms based upon MLVA data were constructed using MEGA 3.1 (Kumar et al., 2001) for the two main groups in the SNP phylogeny, Groups I and II (Figure A7-2A–B). These groups corresponded to the two major Malagasy groups in a previous worldwide analysis (Morelli et al., 2010) and so were separated prior to analyzing with MLVA. The remaining SNPs showing variation among the Malagasy isolates mostly defined subclades observed in the MLVA phylogenies or were specific to single isolates, and so were not used to further separate the isolates prior to applying MLVA. The locations of these additional SNPs are marked on the two MLVA phylogenies where applicable (Figure A7-2A–B). A small set of SNPs provided very fine-scale resolution of the lineage leading to the whole genome sequenced MG05-1020 strain and are not marked on the MLVA phylogeny due to disagreement between the SNP and MLVA phylogenies on this small scale. Distance matrices for the two MLVA phylogenies were based upon mean character differences. Bootstrap values were based upon 1,000 simulations and were generated using PAUP 4.0b10 (D. Swofford, Sinauer Associates, Inc., Sunderland, MA). Branches with ≥50% bootstrap support and/or supported by one or more SNPs were identified as subclades. One other cluster (II.A) was also considered a subclade despite a lack of bootstrap support because of the proximity of a SNP-defined subclade (Figure A7-2B).

Geographic Distribution of Subclades

We mapped the geographic distributions of the Group I and II subclades we identified to determine their phylogeographic patterns (Figure A7-3).

Statistical Analyses

Analysis of similarity (ANOSIM) (Clarke, 1993) tests were performed using PRIMER software version 5 to test the hypotheses that 1) Groups I and II form distinct geographic groups and 2) the identified subclades form distinct geographic groups. These tests were performed on all subclades with ≥5 members (N = 221 isolates), thus excluding the unaffiliated isolates and subclades I.C, I.H, I.I and I.G (Table S1). The results of all 55 pairwise comparisons among the subgroups were evaluated at $\alpha = 0.000909$ (global α of 0.05 divided by 55). To determine if there was a rank relationship between genetic distance and geographic

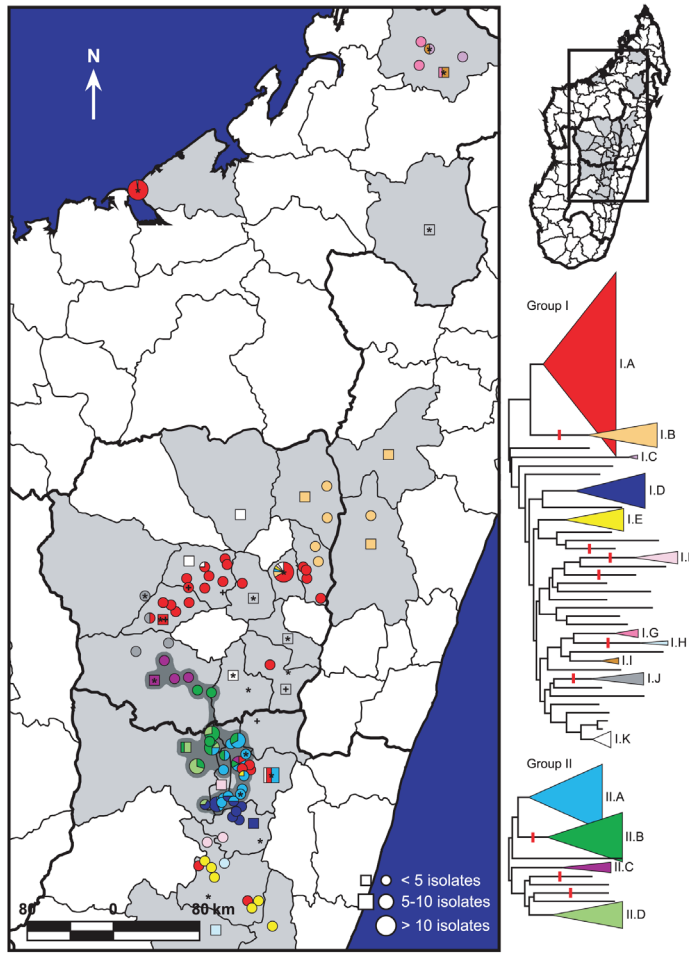


FIGURE A7-3 Geographic distribution of MLVA subclades in Madagascar. The MLVA phylogenies for Groups I and II from Figure A7-2A–B are presented with labeled subclades. Light gray shaded districts indicate Madagascar districts where *Y. pestis* isolates used in this study were obtained. Colors within the mapped circles and squares correspond to the subclade color designations in the MLVA phylogenies. Divisions within those circles and squares indicate that multiple subclades were found at that location. Circles represent isolates where the city/commune of origin is known. Squares represent isolates where only the district of origin is known and are placed within their corresponding districts near to cities/communes containing the same subclade(s) where possible. Six isolates had unknown districts of origin and were not mapped. Unaffiliated Group I and II isolates are indicated by an "*" and a "+," respectively; these symbols surrounded by a square indicate unaffiliated isolates where only the district of origin is known. The dark gray-shaded area indicates the geographic area where Group II subclades are found. Note that some Group I subclades are also found in this area.

distance, a Spearman correlation coefficient was generated using the RELATE function in PRIMER software with significance of the resulting statistics determined using 10,000 random permutations of the data. This analysis utilized all isolates with any geographic data (N = 256), with district centroids used as the geographic location for isolates for which only district level geographic information was available (N = 33); city/commune point geographic data were used for the remaining 223 isolates. Six isolates lacking any geographic information were excluded from both statistical analyses (Table S1).

Results

Genetic Diversity of Y. Pestis *in Madagascar*

Our hypervariable-locus and genome-based approaches identified a relatively high level of genetic diversity among the 262 Malagasy isolates from 25 districts from 1939–2005. We confirmed the presence of two major genetic groups, Groups I and II, differentiated by a single SNP, Mad-43 (Figure A7-1, Table S2), and many VNTR mutational steps. Groups I and II were further differentiated into eleven (I.A–I.K, Figure A7-2A, Table S1) and four (II.A–II.D, Figure A7-2B, Table S1) subclades, respectively, based upon MLVA and/or SNPs. All but one of these subclades was at least weakly supported by bootstrap values ≥50 and/or one or more SNPs (Figure A7-2A–B). The high mutation rates at VNTR loci can lead to homoplasy and, consequently, to low bootstrap support for deeper phylogenetic relationships when analyzing isolates from regional or worldwide collections (Achtman et al., 2004; Johansson et al., 2004; Keim et al., 2004; Lowell et al., 2007). Nevertheless, subsequent analyses using more phylogenetically stable molecular markers (i.e., SNPs) have confirmed MLVA-determined clades with weak or even no bootstrap support (Achtman et al., 2004; Vogler et al., 2009), leading us to use even weak bootstrap support to validate subclades in this analysis. Of the two MLVA identified subclades without bootstrap support, II.A and II.B, subclade II.B was supported by SNP Mad-46 (Table S2) and subclade II.A was designated due to its proximity to and clear separation from the SNP-identified subclade II.B (Figure A7-2B). Subclades I.B, I.F, and I.H were supported by SNPs Mad-26 to 31, Mad-42, and Mad-09 to 17 (Table S2), respectively, and bootstrap analysis (Figure A7-2A). MLVA also identified 23 and 5 isolates in Groups I and II, respectively, that did not belong to any of the identified subclades within those groups (hereafter referred to as unaffiliated isolates) (Figure A7-2A–B, I.NONE and II.NONE isolates in Table S1). Four of these unaffiliated isolates and isolates in subclades I.B, I.H and II.B were also identified by apparently isolate-specific SNPs (Figure A7-2A–B). Overall, MLVA identified 226 genotypes among the 262 isolates, constituting far better resolution than that achieved using ribotyping (Guiyoule et al., 1997).

The SNP and MLVA analyses showed remarkable congruence. Nearly all of the nodes in the SNP phylogeny either corresponded to MLVA subclades or were specific to individual isolates, allowing the combined analysis of SNP and MLVA data discussed above. Three nodes (f, m and n, Figure A7-1) did not have representatives in this study, but appeared to be specific for individual isolates in a previous analysis (Morelli et al., 2010). The only exception to this congruence was within the lineage leading to the whole genome sequenced strain, MG05-1020 (q nodes in Figure A7-1 and subclade I.B in Figure A7-2A). In this case, the SNP phylogeny (q nodes, Figure A7-1) was more accurate than and provided nearly as much resolution as the corresponding MLVA phylogeny (I.B, Figure A7-2A). This fine-scale phylogenetic resolution was due to the use of a high resolution SNP discovery method, whole genome sequence comparisons, to discover SNPs along this lineage as opposed to the lower resolution dHPLC method used to discover most of the other Malagasy SNPs (Morelli et al., 2010). Interestingly, comparable resolution was not seen in the lineage leading to the other whole genome sequenced strain, IP275 (l nodes in Figure A7-1 and subclade I.H in Figure A7-2A), likely due to the very low number of isolates (N = 2) within that lineage in this analysis.

Missing data for two SNP assays suggested a potential genomic rearrangement (e.g., deletion) in some of the Malagasy strains. Twenty-five of the 262 isolates were missing data for two SNP assays despite repeated attempts at amplification (Table S1). The two SNPs, Mad-28 and Mad-41, were located <850 bp apart at CO92 positions 2,208,345 and 2,207,531, respectively (Table S2), suggesting that there may have been a genomic rearrangement affecting this region in these strains. Intriguingly, IS*100* elements were located flanking these SNPs at CO92 positions 2,135,459-2,137,412 and 2,236,265-2,238,215. IS elements are important facilitators of genomic rearrangements in *Y. pestis* (Auerbach et al., 2007; Chain et al., 2006) and may have played a role in this result. If so, the same or a similar genomic rearrangement must have occurred multiple times since the 25 isolates were members of six different nodes in the SNP phylogeny (Table S1). This hypothesis is supported by the fact that IS*100* elements are known potential hotspots for genomic rearrangements and excisions in *Y. pestis* (Achtman, 2004; Auerbach et al., 2007).

Geographic Distribution of Isolates

Significant geographic separation was observed among the identified subclades. Overall, there was a small, but highly significant relationship between genetic and geographic distance (Spearman correlation coefficient $\rho = 0.226$, $p<0.0001$). In addition, the two main genetic groups, Groups I and II, formed distinct geographic groups based upon an ANOSIM ($R = 0.091$, $p = 0.0007$). Group II isolates, which possessed the derived state for SNP Mad-43 (Table S2), were essentially restricted to three of the most active plague districts in the central

highlands, Betafo, Manandriana and Ambositra (Chanteau et al., 2000), and an adjacent district, Ambatofinandrahana (Figure A7-3, S1). The only exceptions to this were the five unaffiliated Group II isolates, which were scattered in districts to the east and north (+ symbols, Figure A7-3). In contrast, Group I isolates were found in all three foci, both the central and northern highlands and Mahajanga. Geographic separation among the individual Group I and II subclades was also apparent (Figure A7-3) and statistically supported in an ANOSIM ($R = 0.232$, $p<0.0001$). Post-hoc analyses of the pairwise comparisons among subclades indicated that most of the eleven tested subclades formed distinct geographic groups (data not shown). Indeed, several interesting geographic patterns were apparent for the different subclades, only some of which are described below. Separate Group I subclades were found in the northern (I.C, I.G, and I.I, Figure A7-3, Table S1) versus the central (I.A, I.B, I.D, I.E, I.F, I.H, I.J, and I.K, Figure A7-3, Table S1) highlands. Subclade I.A, the largest single subclade, was the dominant subclade found in the capital, Antananarivo, and the surrounding area (Figure A7-3, S1). With the exception of two isolates, it was also the only subclade found in Mahajanga (Figure A7-3, S1, Table S1), indicating a central highlands origin for the *Y. pestis* responsible for the series of Mahajanga plague outbreaks from 1991–1999 (Boisier et al., 1997, 2002; Laventure et al., 1991; Rasolomaharo et al., 1995). Subclade I.B was the only subclade found in the northeastern portion of the central highlands (Figure A7-3). Geographic analysis of the corresponding SNP phylogeny (q nodes, Figure A7-1) for this subclade revealed some additional geographic-genetic patterns. Isolates with the same SNP genotype tended to be clustered geographically, although no distinct spreading pattern could be discerned, possibly due to the limited number of isolates (Figure A7-4). Subclade I.E was predominantly found in the southern central highlands, in district Fianarantsoa, and also appears to be the subclade responsible for the reemergence of plague in the Ikongo district (Migliani et al., 2001), adjacent to Fianarantsoa on the southeast (Figure A7-3, S1).

Three subclades, I.F, I.H and I.K, did not show distinct geographic patterns (Figure A7-3). In the cases of subclades I.F and I.H, this may be due to the limited numbers of isolates within those subclades (Figure A7-2A, Table S1). The geographically widespread nature of subclade I.K isolates, however, may be related to their older dates of isolation. All of the subclade I.K isolates were isolated between 1940 and 1955 (Figure A7-2A, Table S1), just 19–34 years after plague was introduced to the central highlands. Therefore, these isolates may represent a subclade that was formerly spread throughout much of the central highlands but that currently does not exist in nature in Madagascar. Similarly, subclade I.I, although it was not geographically widespread (Figure A7-3), only contained isolates isolated from 1971–1976 (Figure A7-2A, Table S1) and may represent a former, now extinct subclade from the northern highlands. However, the limited number of isolates makes this difficult to determine. Alternatively,

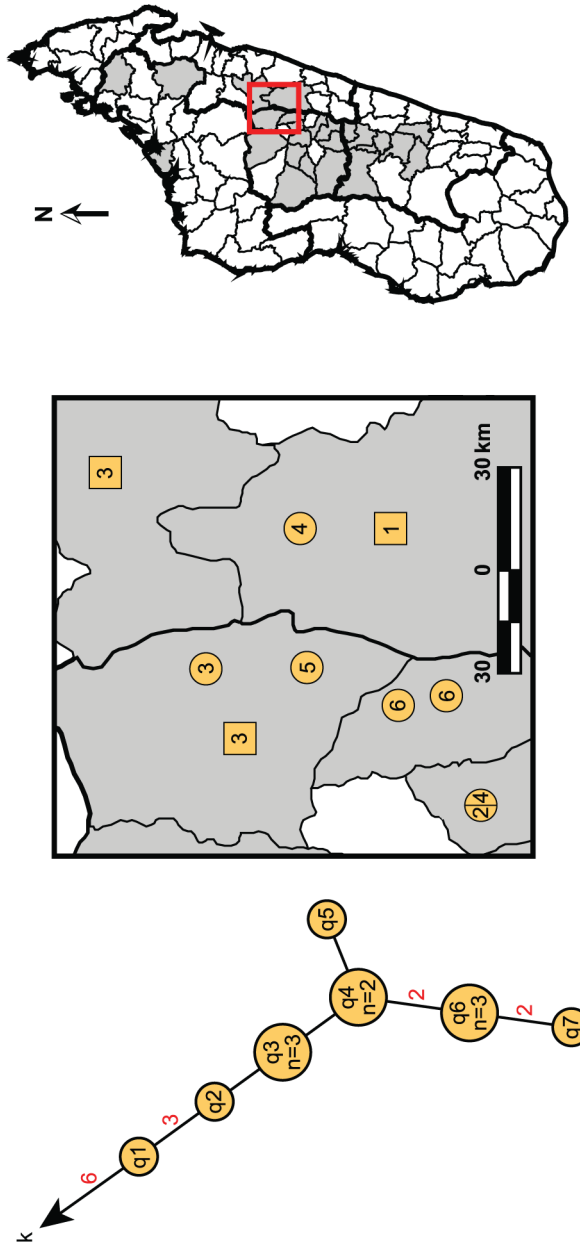


FIGURE A7-4 Geographic distribution of SNP-defined nodes in the strain MG05-1020 lineage. The strain MG05-1020 lineage portion of the SNP phylogeny from Figure A7-1 is indicated as well as an enlarged cutout of the map from Figure A7-3 showing the geographic distribution of isolates from this lineage. For an explanation of the mapped circles and squares see the figure legend for Figure A7-3. Circles, squares and pie chart slices in the map are numbered based upon the node number in the SNP phylogeny for the isolates represented by those shapes. The isolate in node "q7" is not mapped due to its geographic origin being unknown.

these subclades may still exist, but may have decreased in frequency and/or be very rare in nature.

Interestingly, the other older isolates tended to be the unaffiliated isolates. Eighteen of the 28 unaffiliated isolates were isolated between 1939 and 1978. Another 3 had unknown dates of isolation (Table S1). Given their older dates of isolation, these unaffiliated isolates may also be representatives of older, now extinct subclades from Madagascar. The lack of comparable isolates to these unaffiliated isolates among the rest of the isolate collection could be due to the limited sampling from earlier years (Table S1). Alternatively, the unaffiliated isolates may simply be representatives of very rare subclades. A final possibility could involve the accumulation of VNTR mutations due to repeated passages associated with prolonged storage in the laboratory, which could lead to the older isolates being inaccurate representatives of the original isolates. This is unlikely, however, as the rate of VNTR evolution in the laboratory, even with passaging, should be much slower than in nature. Thus, while these isolates may not be exactly the same as when they were first isolated, they should be close. Also, multiple copies of a subset of the Malagasy isolates in this study that were stored at different temperatures showed identical MLVA genotypes (data not shown), indicating that these VNTR loci are relatively stable in these isolates under the storage conditions used. Regardless, the unaffiliated nature of many of the older isolates is consistent with and most likely related to their older dates of isolation.

Several cities and communes yielded isolates of subclades predominantly found elsewhere, suggesting importation from other locations. Antananarivo, in particular, contained isolates from five subclades in addition to the dominant subclade (Figure A7-3, S1). Commune Andina Firaisana in the Ambositra district is another example, containing representatives of four different subclades (Figure 3, S1). One of these, subclade I.A, was also found in the nearby surrounding area. However, this area is considerably south of the area where the majority of subclade I.A isolates were found, suggesting that this subclade may have been imported to this area from further north or vice versa (Figure A7-3). Of the other three subclades found in Andina Firaisana, subclades II.A and II.B are also found in nearby areas and so may be naturally occurring in Andina Firaisana rather than due to transfer events. Subclade II.C, in contrast, appears to have been transferred to Andina Firaisana from the Betafo district in the northwest or vice versa (Figure A7-3, S1). Another nearby commune, Ivato, contained a single subclade I.E isolate, suggesting a transfer event from district Fianarantsoa in the south (Figure A7-3, S1).

Plague in Mahajanga

Our data suggest that *Y. pestis* was reintroduced to Mahajanga from the central highlands. The majority of the Mahajanga isolates (39 of 44) belonged to a single subcluster within subclade I.A (hereafter referred to as the Mahajanga

I.A subcluster) (Figure A7-2A), suggesting that there was an introduction to Mahajanga from the central highlands that became established in Mahajanga and then underwent local cycling. Though this Mahajanga I.A subcluster did not have either SNP or MLVA support (Figure A7-2A), close examination of the isolates within this subcluster revealed very close genetic relationships, with most differences involving only a single repeat change at a single VNTR locus (data not shown). This is consistent with an outbreak scenario originating from a single introduction and strengthens the identification of this subcluster as a genetic group. In contrast, subclade I.A isolates outside of the Mahajanga I.A subcluster exhibited much greater variation both in the number of VNTR loci displaying polymorphisms and the number of alleles observed at those loci (data not shown), consistent with an older, more geographically dispersed and more differentiated set of isolates.

Our data also suggest that there have been multiple transfers of *Y. pestis* between Mahajanga and the central highlands. Specifically, seven isolates within the Mahajanga I.A subcluster were isolated from central highland locations rather than from Mahajanga (Figure A7-2A), suggesting that *Y. pestis* was also transferred back from Mahajanga to the central highlands. Two other Mahajanga isolates belonged to subclade I.F and were unaffiliated, respectively (Figure A7-2A), suggesting that there has been more than one introduction of *Y. pestis* to Mahajanga as well. The final three Mahajanga isolates, although they belonged to subclade I.A, were not part of the Mahajanga I.A subcluster and were instead more closely related to subclade I.A isolates from the central highlands (Figure A7-2A), again suggesting multiple introductions. However, it is unclear as to whether any of these other introductions became established in Mahajanga due to the lack of other Mahajanga isolates similar to these five outliers. Finally, although our data suggest that there have been multiple transfers of *Y. pestis* between Mahajanga and the central highlands, there is no evidence in these data for an introduction to Mahajanga from the northern highlands, as was previously suggested by PFGE analyses (Boisier et al., 2002; Duplantier et al., 2005).

Discussion

Madagascar is one of the most active plague regions in the world. However, few studies have investigated the molecular epidemiology of *Y. pestis* from Madagascar and none have done so using very high resolution genomic methodologies. Here, we investigated the phylogeography and molecular epidemiology of *Y. pestis* in Madagascar by using a combination of SNPs and MLVA to analyze 262 Malagasy isolates from 25 districts from 1939–2005. In contrast with previous analyses that utilized ribotyping or SNPs alone (Guiyoule et al., 1997; Morelli et al., 2010), we identified a very high level of genetic diversity with 226 MLVA genotypes among the 262 isolates. These genotypes were distributed amongst 15

subclades that displayed significant geographic separation (Figure A7-3), leading to insights into the maintenance and spread of plague in Madagascar.

The use of MLVA was particularly effective at identifying genetic groups in Madagascar. SNPs, though useful, mostly provided confidence in genetic groups that were already apparent via MLVA. This is somewhat counter to the conventional hierarchical approach wherein SNPs are used first to identify major genetic groups followed by MLVA to provide resolution within those groups, thus minimizing the problems of mutational saturation/homoplasy that can occur with highly variable markers such as VNTRs (Keim et al., 2004). In this study, only SNP Mad-43 (Table S2), which differentiated Groups I and II, was useful in this conventional sense to identify "major genetic groups" that were obscured in the MLVA phylogeny (data not shown). All of the other subclades identified by SNPs were also identified by MLVA, suggesting that at this regional scale, MLVA alone may be effective at identifying robust genetic groups. Importantly, though MLVA was excellent at identifying these genetic groups, the relationships among those groups, such as the division between Groups I and II, remained unclear using MLVA alone (data not shown) whereas they were very clearly depicted as a star phylogeny in the SNP phylogeny (Figure A7-1). Where knowledge of deeper genetic relationships or fine-scale phylogenetic analysis of specific lineages (e.g., the strain MG05-1020 lineage here) is desired, SNPs will remain the preferred methodology for clonal pathogens such as *Y. pestis*. However, until whole genome sequencing for entire isolate collections becomes feasible, MLVA will continue to be a useful tool for examining genetic diversity whether used in conjunction with SNPs or alone.

Our analyses suggest that plague is being maintained in Madagascar in multiple geographically separated subpopulations. We revealed significant geographic separation among the identified subclades (Figure A7-3), suggesting that these subclades are undergoing local cycling with limited gene flow from other subclades. This is consistent with the population genetics and ecology of the black rat (*Rattus rattus*), the primary plague host in rural Madagascar (Brygoo, 1966; Duplantier, 2001). The black rat in Madagascar exhibits limited gene flow between subpopulations (Gilabert et al., 2007) as well as limited geographic ranges (Rahelinirina et al., 2010). This limited mobility, a high reproduction rate (Duplantier and Rakotondravony, 1999), and the development of some resistance to plague (Tollenaere et al., 2010) are all likely important factors that allow the black rat to maintain plague in these genetically distinct, geographically separated subpopulations. The two flea vectors, *X. cheopis* and *S. fonquerniei* (Duplantier, 2001; Duplantier and Rakotondravony, 1999), may also play a role in maintaining genetically distinct subpopulations (i.e., Groups I and II), though more data would be needed to confirm this hypothesis.

In contrast, transport of *Y. pestis* across longer distances in Madagascar is likely human-mediated. Historically, there is ample evidence for the influence of human traffic on the spread of plague, including transport along trade routes

such as the Silk Road in the early pandemics and transport via steam ship to numerous new locations during the "third" pandemic (Morelli et al., 2010; Perry and Fetherston, 1997). The SNP phylogeny determined by Morelli et al. (2010) suggests the progression of plague from Israel to Madagascar to Turkey (Figure A7-1), a series of transfer events that were almost certainly human-mediated, though the details remain unknown. In Madagascar, plague was most likely transported from its introduction point on the coast to the central highlands, where it became permanently established, via the railroad linking Toamasina and Antananarivo (Brygoo, 1966). More recently, plague was most likely reintroduced to Mahajanga via the transport of infected rats and fleas together with foodstuffs from the central highlands. Indeed, our data suggest multiple transfers between Mahajanga and the central highlands, all likely human-mediated. Additional long distance transfers of *Y. pestis* in Madagascar are suggested by the multiple subclades identified in cities/communes such as Antananarivo and Andina Firaisana (Figure A7-3, S1, Table S1).

Though long distance transfers of *Y. pestis* undoubtedly occur, it is unclear how often such transfers result in the successful establishment of the transferred genotypes in new locations. At least one transfer to Mahajanga became successfully established and underwent local cycling as evidenced by the Mahajanga I.A subcluster described here (Figure A7-2A). However, many of the other examples of long distance transfers where multiple subclades were found in a single location are not as clear regarding the establishment of the transferred subclade(s). Antananarivo, for example, is clearly dominated by subclade I.A with only 1–2 representatives of each of the other five subclades identified there (Figure A7-3, S1, Table S1), suggesting that the presence of these alternative subclades may have been only transitory.

Successful establishment of subclades in new locations following a long distance transfer may be related to adaptive advantages possessed by some genotypes (Keim and Wagner, 2009). For instance, subclade I.A appears to be particularly successful in our analysis. The earliest subclade I.A isolate in our dataset was collected in 1974 from the Ambositra district (Table S1), one of the most active plague districts in Madagascar (Chanteau et al., 2000). Subsequent isolates indicate that this subclade continued to exist in a small area of the Ambositra district but also became well established over a large geographic area including and surrounding the capital, Antananarivo. This subclade was also successfully introduced to and established in Mahajanga and appears to have been transferred to the Fianarantsoa district, though it is unclear whether or not it became established there (Figure A7-3, S1, Table S1). This widespread geographical success may indicate that this subclade possesses an adaptive advantage that enhances its ability to be transferred long distances and become established in new locations (Keim and Wagner, 2009). Alternatively, the particular success of this subclade may simply be due to chance.

The central highlands focus remains the most active plague focus in Madagascar (Chanteau et al., 2000) and is, consequently, a likely place for new genotypes to emerge. This is particularly true for those central highlands districts with the highest plague activity. For instance, the three unique ribotypes identified in a previous study belonged to isolates from two highly active districts, Ambositra and Ambohimahasoa (Guiyoule et al., 1997). Here, isolates belonging to Group II and its subclades were found in three highly active districts, Betafo, Manandriana, and Ambositra (Figure A7-3, S1). As discussed above, Ambositra may also have been the district of origin for the highly successful subclade I.A. Overall, the Ambositra district was one of the two most diverse districts in our analysis, containing representatives from six different subclades (Figure A7-3, Table S1). This diversity is consistent with the Ambositra district's status as one of the three most important plague districts in Madagascar (Chanteau et al., 1998; 2000).

The maintenance and spread of *Y. pestis* in Madagascar is a dynamic and highly active process, depending on the natural cycle between the black rat and its flea vectors as well as human activity. *Y. pestis* in Madagascar is maintained in multiple, genetically distinct, geographically separated subpopulations, likely via the black rat. The exact geographic landscape of these subpopulations is probably ever changing, with some subclades going extinct or decreasing in frequency (e.g., subclade I.K), new subclades emerging and becoming established, and some subclades being transferred to new locations, where they may become established either temporarily or more long-term. Much of the long distance spread of *Y. pestis* in Madagascar is likely due to human activities that allow for the transport of plague infected rats and fleas from one location to another.

Acknowledgments

We would like to thank Dr. Kimothy Smith for initially suggesting the collaboration that led to this work. Note that the use of products/names does not constitute endorsement by the DHS of the United States.

Author Contributions

Conceived and designed the experiments: AJV DMW SC PK. Performed the experiments: AJV FC JL RN. Analyzed the data: AJV DMW. Contributed reagents/materials/analysis tools: FC PR ME JR LR BWR SMB-S MA SC. Wrote the paper: AJV PK.

Funding

This work was funded by the Department of Homeland Security Science and Technology Directorate (award numbers NBCH2070001 and HSHQDC-08-C-00158), the Cowden Endowment in Microbiology at Northern Arizona

University, and the National Institute of Allergy and Infectious Diseases (NIAID), US National Institutes of Health (NIH), Department of Health and Human Services (HHS) (award number AI065359). This work was also supported by the Science Foundation of Ireland (award number 05/FE1/B882) (MA), the NIAID NIH HHS (award number N01 AI-30071) (ME JR), the Malagasy Ministry of Health (contract Nu01/95 IDA 2252-MAG) (FC LR BWR SC), and the French Cooperation (FAC Nu 94008 300) (FC LR BWR SC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, et al. (2004) Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl Acad Sci U S A 101: 17837-17842.

Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, et al. (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc Natl Acad Sci U S A 96: 14043-14048.

Auerbach RK, Tuanyok A, Probert WS, Kenefic L, Vogler AJ, et al. (2007) *Yersinia pestis* evolution on a small timescale: comparison of whole genome sequences from North America. PLoS One 2: e770.

Boisier P, Rahalison L, Rasolomaharo M, Ratsitorahina M, Mahafaly M, et al. (2002) Epidemiologic features of four successive annual outbreaks of bubonic plague in Mahajanga, Madagascar. Emerg Infect Dis 8: 311-316.

Boisier P, Rasolomaharo M, Ranaivoson G, Rasoamanana B, Rakoto L, et al. (1997) Urban epidemic of bubonic plague in Majunga, Madagascar: epidemiological aspects. Trop Med Int Health 2: 422-427.

Brygoo ER (1966) Epidémiologie de la peste à Madagascar. Arch Inst Pasteur Madagascar 35: 9-147.

Chain PS, Hu P, Malfatti SA, Radnedge L, Larimer F, et al. (2006) Complete genome sequence of *Yersinia pestis* strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. J Bacteriol 188: 4453-4463.

Chanteau S, Ratsifasoamanana L, Rasoamanana B, Rahalison L, Randriambelosoa J, et al. (1998) Plague, a reemerging disease in Madagascar. Emerg Infect Dis 4: 101-104.

Chanteau S, Ratsitorahina M, Rahalison L, Rasoamanana B, Chan F, et al. (2000) Current epidemiology of human plague in Madagascar. Microbes Infect 2: 25-31.

Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. Aust J Ecol 18: 117-143.

Cui Y, Li Y, Gorgé O, Platonov ME, Yan Y, et al. (2008) Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. PLoS ONE 3: e2652.

Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30: 2478-2483.

Deng W, Burland V, Plunkett G, 3rd, Boutin A, Mayhew GF, et al. (2002) Genome sequence of *Yersinia pestis* KIM. J Bacteriol 184: 4601-4611.

Duplantier JM (2001) The black rat's role in spreading human plague in Madagascar. L'Institut de recherche pour le développement Scientific Bulletin 131: 1-3.

Duplantier JM, Duchemin JB, Chanteau S, Carniel E (2005) From the recent lessons of the Malagasy foci towards a global understanding of the factors involved in plague reemergence. Vet Res 36: 437-453.

Duplantier JM, Rakotondravony D (1999) The rodent problem in Madagascar: agricultural pest and threat to human health. In: Singleton G, Hinds L, Leirs H, Zhang Z, editors. Ecologically-based management of rodent pests. Canberra: Australian Centre for International Agricultural Research. pp. 441-459.

Eppinger M, Guo Z, Sebastian Y, Song Y, Lindler LE, et al. (2009) Draft genome sequences of *Yersinia pestis* isolates from natural foci of endemic plague in China. J Bacteriol 191: 7628-7629.

Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, et al. (2010) Genome sequence of the deep-rooted *Yersinia pestis* strain Angola reveals new insights into the evolution and pangenome of the plague bacterium. J Bacteriol 192: 1685-1699.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res 8: 175-185.

Galimand M, Guiyoule A, Gerbaud G, Rasoamanana B, Chanteau S, et al. (1997) Multidrug resistance in *Yersinia pestis* mediated by a transferable plasmid. N Engl J Med 337: 677-680.

Garcia E, Worsham P, Bearden S, Malfatti S, Lang D, et al. (2007) Pestoides F, an atypical *Yersinia pestis* strain from the former Soviet Union. Adv Exp Med Biol 603: 17-22.

Gilabert A, Loiseau A, Duplantier JM, Rahelinirina S, Rahalison L, et al. (2007) Genetic structure of black rat populations in a rural plague focus in Madagascar. Can J Zool 85: 965-972.

Girard JM, Wagner DM, Vogler AJ, Keys C, Allender CJ, et al. (2004) Differential plague-transmission dynamics determine *Yersinia pestis* population genetic structure on local, regional, and global scales. Proc Natl Acad Sci U S A 101: 8408-8413.

Guiyoule A, Grimont F, Iteman I, Grimont PA, Lefévre M, et al. (1994) Plague pandemics investigated by ribotyping of *Yersinia pestis* strains. J Clin Microbiol 32: 634-641.

Guiyoule A, Rasoamanana B, Buchrieser C, Michel P, Chanteau S, et al. (1997) Recent emergence of new variants of *Yersinia pestis* in Madagascar. J Clin Microbiol 35: 2826-2833.

Huang XZ, Chu MC, Engelthaler DM, Lindler LE (2002) Genotyping of a homogeneous group of *Yersinia pestis* strains isolated in the United States. J Clin Microbiol 40: 1164-1173.

Johansson A, Farlow J, Larsson P, Dukerich M, Chambers E, et al. (2004) Worldwide genetic relationships among *Francisella tularensis* isolates determined by multiple-locus variable-number tandem repeat analysis. J Bacteriol 186: 5808-5818.

Keim P, Van Ert MN, Pearson T, Vogler AJ, Huynh LY, et al. (2004) Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. Infect Genet Evol 4: 205-213.

Keim PS, Wagner DM (2009) Humans and evolutionary and ecological forces shaped the phylogeography of recently emerged diseases. Nat Rev Microbiol 7: 813-821.

Kingston JJ, Tuteja U, Kapil M, Murali HS, Batra HV (2009) Genotyping of Indian *Yersinia pestis* strains by MLVA and repetitive DNA sequence based PCRs. Antonie Van Leeuwenhoek 96: 303-312.

Klevytska AM, Price LB, Schupp JM, Worsham PL, Wong J, et al. (2001) Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome. J Clin Microbiol 39: 3179-3185.

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17: 1244-1245.

Laventure S, Andrianaja V, Rasoamanana B (1991) Epidémie de peste à Majunga en 1991. Rapport de mission de l'Institut Pasteur de Madagascar 1991: 1-26. 1-26.

Li Y, Cui Y, Hauck Y, Platonov ME, Dai E, et al. (2009) Genotyping and phylogenetic analysis of *Yersinia pestis* by MLVA: insights into the worldwide expansion of Central Asia plague foci. PLoS One 4: e6000.

Li Y, Dai E, Cui Y, Li M, Zhang Y, et al. (2008) Different region analysis for genotyping *Yersinia pestis* isolates from China. PLoS ONE 3: e2166.

Lowell JL, Zhansarina A, Yockey B, Meka-Mechenko T, Stybayeva G, et al. (2007) Phenotypic and molecular characterizations of *Yersinia pestis* isolates from Kazakhstan and adjacent regions. Microbiology 153: 169-177.

Lucier TS, Brubaker RR (1992) Determination of genome size, macrorestriction pattern polymorphism, and nonpigmentation-specific deletion in *Yersinia pestis* by pulsed-field gel electrophoresis. J Bacteriol 174: 2078-2086.

Migliani R, Chanteau S, Rahalison L, Ratsitorahina M, Boutin JP, et al. (2006) Epidemiological trends for human plague in Madagascar during the second half of the 20th century: a survey of 20,900 notified cases. Trop Med Int Health 11: 1228-1237.

Migliani R, Ratsitorahina M, Rahalison L, Rakotoarivony I, Duchemin JB, et al. (2001) [Resurgence of the plague in the Ikongo district of Madagascar in 1998. 1. Epidemiological aspects in the human population]. Bull Soc Pathol Exot 94: 115-118.

Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, et al. (2010) *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet 42: 1140-1143.

Motin VL, Georgescu AM, Elliott JM, Hu P, Worsham PL, et al. (2002) Genetic variability of *Yersinia pestis* isolates as predicted by PCR-based IS*100* genotyping and analysis of structural genes encoding glycerol-3-phosphate dehydrogenase (*glpD*). J Bacteriol 184: 1019-1027.

Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, et al. (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413: 523-527.

Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, et al. (2004) Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. Proc Natl Acad Sci U S A 101: 13536-13541.

Perry RD, Fetherston JD (1997) *Yersinia pestis*-etiologic agent of plague. Clin Microbiol Rev 10: 35-66.

Pourcel C, André-Mazeaud F, Neubauer H, Ramisse F, Vergnaud G (2004) Tandem repeats analysis for the high resolution phylogenetic analysis of *Yersinia pestis*. BMC Microbiol 4: 22.

Rahelinirina S, Duplantier JM, Ratovonjato J, Ramilijaona O, Ratsimba M, et al. (2010) Study on the movement of *Rattus rattus* and evaluation of the plague dispersion in Madagascar. Vector Borne Zoonotic Dis 10: 77-84.

Rasolomaharo M, Rasoamanana B, Andrianirina Z, Buchy P, Rakotoarimanana N, et al. (1995) Plague in Majunga, Madagascar. Lancet 346: 1234.

Rotz LD, Khan AS, Lillibridge SR, Ostroff SM, Hughes JM (2002) Public health assessment of potential biological terrorism agents. Emerging infectious diseases 8: 225-230.

Song Y, Tong Z, Wang J, Wang L, Guo Z, et al. (2004) Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. DNA Res 11: 179-197.

Tollenaere C, Rahalison L, Ranjalahy M, Duplantier JM, Rahelinirina S, et al. (2010) Susceptibility to *Yersinia pestis* experimental infection in wild *Rattus rattus*, reservoir of plague in Madagascar. Ecohealth 7: 242-247.

Touchman JW, Wagner DM, Hao J, Mastrian SD, Shah MK, et al. (2007) A North American *Yersinia pestis* draft genome sequence: SNPs and phylogenetic analysis. PLoS One 2: e220.

Van Ert MN, Easterday WR, Huynh LY, Okinaka RT, Hugh-Jones ME, et al. (2007) Global genetic population structure of *Bacillus anthracis*. PLoS One 2: e461.

Vogler AJ, Birdsell D, Price LB, Bowers JR, Beckstrom-Sternberg SM, et al. (2009) Phylogeography of *Francisella tularensis*: global expansion of a highly fit clone. J Bacteriol 191: 2474-2484.

Welch TJ, Fricke WF, McDermott PF, White DG, Rosso ML, et al. (2007) Multiple antimicrobial resistance in plague: an emerging public health risk. PLoS One 2: e309.

World Health Organization (WHO) (2010) Human plague: review of regional morbidity and mortality, 2004–2009. Wkly Epidemiol Rec 85: 40-45.

Zhang X, Hai R, Wei J, Cui Z, Zhang E, et al. (2009) MLVA distribution characteristics of *Yersinia pestis* in China and the correlation analysis. BMC Microbiol 9: 205.

Zhang Z, Hai R, Song Z, Xia L, Liang Y, et al. (2009) Spatial variation of *Yersinia pestis* from Yunnan Province of China. Am J Trop Med Hyg 81: 714-717.

A8

BIG DATA IN BIOLOGY: PITFALLS WHEN USING SHOTGUN METAGENOMICS TO DEFINE HYPOTHESES ABOUT MICROBIAL COMMUNITIES

*Folker **Meyer***[30] *and Elizabeth M. Glass*[30]

Introduction

Next-generation sequencing (NGS) has opened up access to genomic data from diverse microbial communities, and studies are emerging that cover a wide variety of systems (Gilbert et al., 2010; Human Microbiome Project, 2012; Nealson and Venter, 2007; Tara Expeditions, 2012; Terragenome Consortium, 2012). A number of techniques are used to extract genome-based information either using single reference genes (usually 16s rDNA) or random shotgun metagenomics using entire genomes. There is an abundance of reviews of the subject (Desai et al., 2012; Thomas et al., 2012). Systems such as MG-RAST (Meyer et al., 2008) now provide access to thousands of metagenomic data sets (see Figure A8-1).

However this newly found data richness is not without its challenges. The main problem stems from dramatic changes that converted an ecosystem that was until recently data poor, to one that is now overflowing with data. Environmental biology and molecular ecology went from being overwhelmed by several hundred megabytes of data generated in 2005 by the Global Ocean Survey (Nealson and Venter, 2007) to generating many terabytes of data in 2012. Biology and medicine, however, lack the tradition and experience to handle big data—only a few areas such as cancer diagnosis have established stable pipelines and data formats that allow exchange.

Shotgun Metagenomics as an Example

Metagenomic shotgun sequencing using NGS technology can serve as a blueprint for how biology is and will be impacted by big data. With sequence data already significantly cheaper than the corresponding analysis (Wilkening et al., 2009), and as the cost of sequencing drops by a factor of 10 annually, it seems clear that a paradigm shift will be required to handle data analysis and storage.

There is significant value in the comparative analysis of metagenomic data sets, yet comparison requires data sets to have undergone more or less the same

[30] Argonne National Laboratory, Institute for Genomics and Systems Biology, 9700 S. Cass Ave, Lemont, IL 60439, U.S.A.

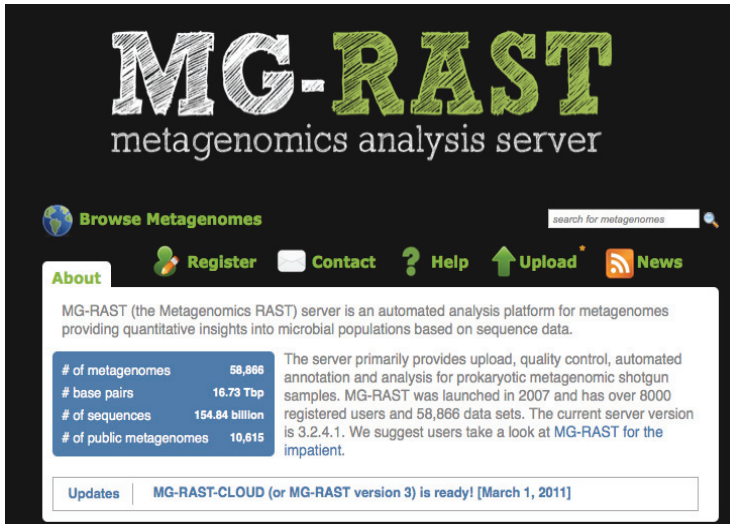


FIGURE A8-1 The MG-RAST system has more than 58,000 metagenomic data sets totaling over 16.5 terabase pairs of information.

analytical processes. The existing paradigm of publishing the raw data and summary statistics in tabular form as auxiliary material does not allow any third parties to benefit from the work already done; instead it requires any future authors to re-analyze data. Consider the Human Microbiome Project (Turnbaugh et al., 2007) that has recently published more than 5 terabases of sequence data from 172 human subjects. Any researcher attempting to compare their finding to the data will discover that they need to re-analyze all of the data.

As an interesting side note, the value and need for comparative analyses also necessitates asking questions of how the reviewing process was handled. Did the reviewers in fact take a look at any of the analysis performed, or did they take the results produced by a complex sequence analysis pipeline at face value? Often the information that was derived from the data is a product of a complex pipeline that was not described in sufficient detail. Further, the reviewer has no way to know whether the same results could be obtained from the same data. A rigorous review process cannot possibly be maintained under current mechanisms and requirements. The prohibitive cost renders such analyses effectively irreproducible. Thus, it is all the more critical to require detailed documentation of data handling in analyses.

One of the key missing concepts is the notion of rigorous analysis of data quality prior to deriving any statements about biology from the data. Many factors contribute to data quality; in DNA sequencing, noise can be added at various steps in the pipeline. While some vendor-specific schemes exist to determine sequence

noise, in the past there were no generally accepted vendor-neutral ways to characterize the noise in sequence data. For 16s rDNA-based amplicon metagenomic data, this has already led to a major debate over the amount of microbial diversity (Reeder and Knight, 2010), leading to a number of approaches that will de-noise sequence data prior to analysis (Quince et al., 2009; Reeder and Knight, 2010). For shotgun metagenomics the DRISEE approach (Keegan et al., 2012) provides a vendor-neutral estimate of sequence error. Interestingly, results show that the errors found are not specific to sequencing platforms; variations in quality within the platforms are significant, as highlighted by Figure A8-2.

Taking the data sets underlying Figure A8-2 as an example, the sequence analysis pipelines will be required to use different approaches for data with less than 1 percent error and with more than 45 percent error. The MG-RAST analysis pipeline was the first to include the examination of sequence quality systematically and to highlight sequence quality issues. A surprising amount of data sets submitted to the system are rejected initially because they are, for example, too low in quality or contain contamination.

In addition to "low-level" sequence error, a number of other significant sources of problems exist. Tom Schmidt's group described the existence of artificially duplicated reads in 454 data (Gomez-Alvarez et al., 2009). (These artifacts also exist in Ion Torrent and Illumina data.) If left uncorrected, such duplicates

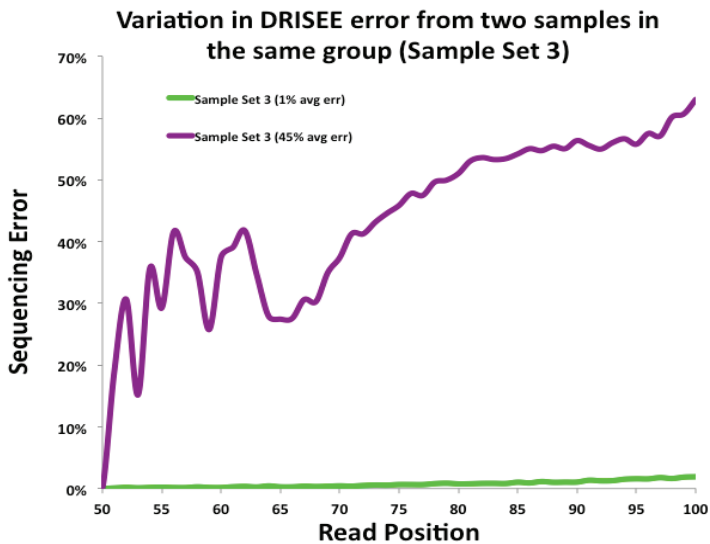


FIGURE A8-2 The DRISEE error profiles for two anonymous projects with three shotgun libraries. The predicted cumulative error per position is plotted showing dramatic variation with the green data set near perfect and the purple data exceeding 40 percent error after 70 bases.

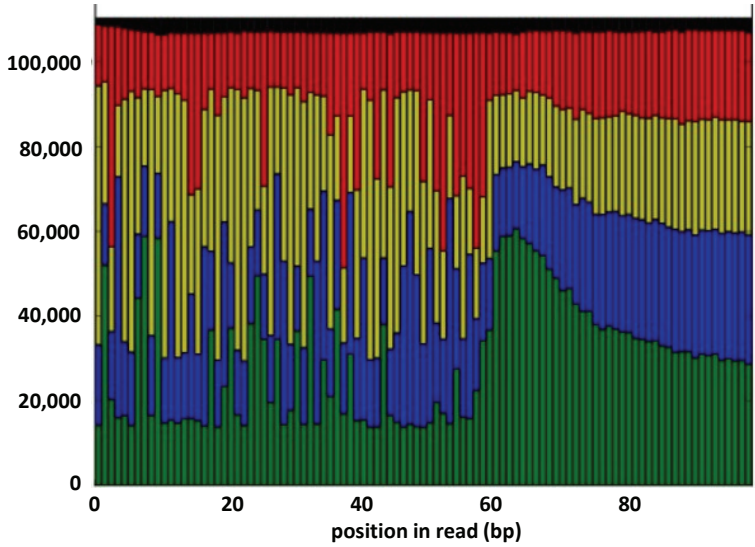


FIGURE A8-3 A simple representation of average base abundance per base demonstrates that data are not distributed randomly. The four bases are represented green (A), blue (C), yellow (G), and red (T); black indicates a missing base call. In this (anonymous) data set, the first 53 base pairs contain an adapter. Note the peak of A's (green) at positions 54–73, likely indicating the presence of poly-A artifacts in addition to the adapters.

lead to significant biases in the interpretation of sequence data, as some areas of sequence are misrepresented.

Other frequently found artifacts include leftover adapter sequences or primer di-mers (see Figure A8-3 for an example of both problems). Most researchers will agree that for a shotgun sample, a more or less even distribution of each base on each position of all reads is to be expected. Unfortunately, in cases that deviate from the expected distribution, as in Figure A8-3, the only information that can be derived from the sequence data is that the sequencing run has failed. Yet, the data shown in Figure A8-3 have been interpreted biologically and were accepted for publication.

Once the potential obstacles with data quality have been eliminated, a number of bioinformatics tools can be used to predict genes of interest for downstream analysis. While significant progress has been made in recent years for the prediction of genes in more or less complete microbial genomes (Overbeek et al., 2007), the same cannot be said for the state of the art for predicting genes in noisy metagenomic data. Trimble et al. (2012) show that only one of the existing tools accounts for the possibility that sequences might contain sequencing error, despite that, as we have mentioned above, the presence of noise (or imperfect

data) is a reality in shotgun metagenomics where data are frequently not or only partially assembled.

A side effect of assuming all data to be perfect is the significant impairment of tool performance when tools are used on data with realistic error properties. This mismatch of assumption with reality leads to performance reductions of 10–20 percent accuracy in the presence of 3 percent error (see Trimble et al., 2012).

Following gene prediction, functional assignments are computed by mapping the prediction features against a database of known proteins in some form. Again different pipelines apply different approaches; however, they all rely on some flavor of sequence similarity searching (e.g., BLAST [Altschul et al., 2009], BLAT [Kent, 2002], HMMer [Eddy, 1998]). What all these approaches have in common is their failure to identify novelty. Any unknown gene will remain uncharacterized, and a non-homologous replacement for a known protein function (no matter how important for the sample) will be represented as an unknown protein in the metagenomic analysis presented.

While this might present a problem in some cases, the majority of the protein databases contain a variety of annotations for proteins of the same function. Often, multiple annotations for proteins that are 100 percent sequence identical can be found (e.g., the alcohol dehydrogenase gene from various *Streptococcus* strains has 20 different annotations). While some annotations would be informative for a human reader, in the majority of cases, a computer would not be able to recognize the fact that the two functions described are identical. When comparing the relative abundance of alcohol dehydrogenase genes, results would be affected by the fact that several alcohol dehydrogenase genes would have slightly different names. As a result, annotations derived from similarity searches are less then useful for quantitative studies in microbial ecology unless carried out on a higher functional level as for example KEGG (Kanehisa, 2002) pathways or the very successful SEED subsystems (Overbeek et al., 2005). These higher-level aggregations subsume a significant number of genes in categories (e.g., KEGG pathways).

Using those categories to represent gene function abundance, we can represent the genetic material in environmental samples as an abundance vector, allowing cross-sample comparison of gene abundance. With the newly minted BIOM (McDonald et al., 2012) format, various existing tools, such as MG-RAST (Meyer et al., 2008) and QIIME (Caporaso et al., 2010) can be used together to analyze functional abundance data.

An Attempt to Estimate the Scale of the Problem

"Raw" sequence data can be transformed into abundance vectors that describe the abundance of specific gene categories in environmental samples. However even slight variations in one of the transforming steps will introduce

significant variations in the outcome, as would be the case for two data sets that have used different gene prediction algorithms with different levels of tolerance for sequence error.

Because there is no culture of sharing data beyond raw sequences in the INSDC archives, consumers of data from any specific study can either rely on tables published as auxiliary material with a study, or they can re-analyze the data. As mentioned before, the computational cost makes that undesirable (Wilkening et al., 2009), Imagine yourself in the position of comparing your data to the Human Microbiome Project jump-start project (described above). Instead of simply analyzing their own data, researchers would find themselves re-analyzing various other data sets they are using in comparison. While this type of approach was common in the early days of genomics and still is used for single microbial genomes—such as SEED (Aziz et al., 2008), IMG (Markowitz et al., 2006), and GenDB (Meyer et al., 2003)—using the same approach for more computationally demanding data types like metagenomes will lead to a situation in which groups are no longer limited by their ability to acquire sequence data, but by their ability to analyze it. While it is likely that metagenomic data analysis will undergo significant improvements (as compared to the improvements for individual microbial genomes discussed in Overbeek et al. [2007]), that will not suffice. With sequencing costs continuing to drop, data acquisition costs will soon be a small fraction of the data analysis cost.

Sharing of computational results (e.g., gene calling results, computed similarities, and other intermediate data types) would alleviate this problem. But this only works if the community can agree on a small number of standard formats to represent the data, and only if the majority of the tools support those standards.

Ways Out of the Current Dilemma

While for the first time biology now has access to abundant data, the challenges in handling the data and using the data in a robust way to define new research hypotheses seem insurmountable. We have described the challenge of using big data in the context of metagenomics above.

A number of aspects to this challenge need to be addressed separately. Perhaps the most important aspect is a change of culture recognizing that in the presence of abundant data, the standard operating procedures from before are no longer sufficient. Among the things that need to change are data archives that now can no longer attempt to capture all data, computational approaches that need to take computational costs into account, and last but not least the individual researchers that need to learn that data analysis must be planned and budgeted for appropriately.

One suggestion likely acceptable to most readers is that, in the presence of big data, computational data analysis needs to take a more prominent role in the training of the next generation of bio-scientists.

There are also some technical steps that can be taken to improve the current situation. Standards for describing sequence data have been established by the Genomics Standards Consortium (GSC) (Field et al., 2011). These standards are currently enabling data exchange at a hitherto unprecedented scale, enabling data consumption by many third parties for many purposes.

Based on these positive experiences, the GSC now has initiated a long-term project to define standards for sharing processed data. The results of this M5 project will enable researchers to consume data from a published study for their analysis without having to re-analyze everything from scratch and yet allowing them to both change and understand all the fine details of these studies.

In addition, the M5 project aims to define encodings for data that will allow existing and new analysis service providers to exchange analyzed data. This ability will allow comparison of different analytical approaches and will help with the evolution of analysis approaches. We predict that this new openness will lead to more acceptance for the established analysis approaches and reduce the number of ad hoc analysis pipelines that reinvent analysis processes. By embracing the needed cultural changes, adhering to standards, and promoting openness, many third parties will be liberated to innovate like never before.

References

Altschul, S. F., E. M. Gertz, R. Agarwala, A. A. Schaffer, and Y. K. Yu. 2009. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Research* 37(3):815-824.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko. 2008. The RAST server: Rapid annotations using subsystems technology. *BMC Genomics* 9:75.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7(5):335-336.

Desai, N., D. Antonopoulos, J. A. Gilbert, E. M. Glass, and F. Meyer. 2012. From genomics to metagenomics. *Current Opinion in Biotechnology* 23(1):72-76.

Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics* 14(9):755-763.

Field, D., L. Amaral-Zettler, G. Cochrane, J. R. Cole, P. Dawyndt, G. M. Garrity, J. Gilbert, F. O. Glöckner, L. Hirschman, and I. Karsch-Mizrachi. 2011. The Genomic Standards Consortium. *PLoS Biology* 9(6):e1001088.

Gilbert, J. A., F. Meyer, J. Jansson, J. Gordon, N. Pace, J. Tiedje, R. Ley, N. Fierer, D. Field, N. Kyrpides, F. O. Glöckner, H. P. Klenk, K. E. Wommack, E. Glass, K. Docherty, R. Stevens, and R. Knight. 2010. The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory, October 6, 2010. *Standards in Genomic Sciences* 3(3):249-253.

Gomez-Alvarez, V., T. K. Teal, and T. M. Schmidt. 2009. Systematic artifacts in metagenomes from complex microbial communities. *ISME Journal* 3(11):1314-1317.

HMP (Human Microbiome Project). 2012. Human Microbiome Project. http://nihroadmap.nih.gov/hmp (accessed October 2, 2012).

Kanehisa, M. 2002. The KEGG database. *Novartis Foundation symposium* 247:91-101; discussion 101-103, 119-128, 244-252.

Keegan, K. P., W. L. Trimble, J. Wilkening, A. Wilke, T. Harrison, M. D'Souza, and F. Meyer. 2012. A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. *PLoS Computational Biology* 8(6):e1002541.

Kent, W. J. 2002. BLAT—the BLAST-like alignment tool. *Genome Research* 12(4):656-664.

Markowitz, V. M., F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X. Zhao, I. Dubchak, P. Hugenholtz, I. Anderson, A. Lykidis, K. Mavromatis, N. Ivanova, and N. C. Kyrpides. 2006. The integrated microbial genomes (IMG) system. *Nucleic Acids Research* 34 (Database issue):D344-D348.

McDonald, D., J. C. Clemente, J. Kuczynski, J. Rideout, J. Stombaugh, D. Wendel, A. Wilke, S. Huse, J. Hufnagle, F. Meyer, R. Knight, and J. Caporaso. 2012. The Biological Observation Matrix (BIOM) format or: How I learned to stop worrying and love the ome-ome. *Gigascience* 1:7.

Meyer, F., A. Goesmann, A. C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich, and A. Pühler. 2003. GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Research* 31(8):2187-2195.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. 2008. The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics [electronic resource]* 9:386.

Nealson, K. H., and J. C. Venter. 2007. Metagenomics and the global ocean survey: What's in it for us, and why should we care? *ISME Journal* 1(3):185-187.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, N. Diaz, H-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Hamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. 2005. The subsystems approach to genome annotation and its use in the Project to Annotate 1000 Genomes. *Nucleic Acids Research* 33(17).

Overbeek, R., D. Bartels, V. Vonstein, and F. Meyer. 2007. Annotation of bacterial and archaeal genomes: Improving accuracy and consistency. *Chemical Reviews* 107(8):3431-3447.

Quince, C., A. Lanzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods* 6(9):639-641.

Reeder, J., and R. Knight. 2010. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nature Methods* 7(9):668-669.

Tara Expeditions. 2012. http://oceans.taraexpeditions.org (accessed October 2, 2012).

Terragenome Consortium. 2012. http://www.terragenome.org (accessed October 2, 2012).

Thomas, T., J. Gilbert, and F. Meyer. 2012. Metagenomics—a guide from sampling to data analysis. *Microbial Informatics and Experimentation* 2(1):3.

Trimble, W. L., K. P. Keegan, M. D'Souza, A. Wilke, J. Wilkening, J. Gilbert, and F. Meyer. 2012. Short-read reading-frame predictors are not created equal: Sequence error causes loss of signal. *BMC Bioinformatics [electronic resource]* 13(1):183.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. 2007. The human microbiome project. *Nature* 449(7164):804-810.

Wilkening, J., A. Wilke, N. Desai, and F. Meyer. 2009. Using clouds for metagenomics: A case study. In: *IEEE Cluster 2009.*

A9

HIGH-THROUGHPUT BACTERIAL GENOME SEQUENCING: AN EMBARRASSMENT OF CHOICE, A WORLD OF OPPORTUNITY[31]

Nicholas J. Loman,[32] *Chrystala Constantinidou*,[32]
Jacqueline Z. M. Chan,[32] *Mihail Halachev*,[32] *Martin Sergeant*,[32]
Charles W. Penn,[32] *Esther R. Robinson*,[33] *and Mark J.* ***Pallen***[32]

Abstract

Here, we take a snapshot of the high-throughput sequencing platforms, together with the relevant analytical tools, that are available to microbiologists in 2012, and evaluate the strengths and weaknesses of these platforms in obtaining bacterial genome sequences. We also scan the horizon of future possibilities, speculating on how the availability of sequencing that is 'too cheap to metre' might change the face of microbiology forever.

In bacteriology, the genomic era began in 1995, when the first bacterial genome was sequenced using conventional Sanger sequencing (Fleischmann et al., 1995). Back then, sequencing projects required six-figure budgets and years of effort. A decade later, in 2005, the advent of the first high-throughput (or "next-generation") sequencing technologies signalled a significant advance in the ease and cost of sequencing (Metzker et al., 2005), delivering bacterial genome sequences in hours or days rather than months or years. High-throughput sequencing now delivers sequence data thousands of times more cheaply than is possible with Sanger sequencing. The availability of a growing abundance of platforms and instruments presents the user with an embarrassment of choice. Better still, vigorous competition between manufacturers has resulted in sustained

[31] Reprinted with kind permission from Nature Publishing Group.

[32] Nicholas J. Loman, Chrystala Constantinidou, Jacqueline Z. M. Chan, Mihail Halachev, Martin Sergeant, Charles W. Penn and Mark J. Pallen are at the Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK.

[33] Esther R. Robinson is at the Nuffield Department of Clinical Laboratory Sciences, University of Oxford, Oxford OX3 9DU, UK.

Correspondence to: Mark J. Pallen—Email: m.pallen@bham.ac.uk

technical improvements on almost all platforms. This means that in recent years our sequencing capability has been doubling every 6–9 months—much faster than Moore's law.

Here, we describe the sequencing technologies themselves, examine the practicalities of producing a sequence-ready template from bacterial cultures and clinical samples, and weigh up the costs of labour and kits. We look at the types of data that are delivered by each instrument, and describe the approaches, programs and pipelines that can be used to analyse these data and thus move from draft to complete genomes.

Several high-throughput sequencing platforms are now chasing the US$1,000 human genome (Venter, 2010). Given that the average bacterial genome is less than one-thousandth the size of the human genome, a back-of-the-envelope calculation suggests that a $1 bacterial genome sequence is an imminent possibility. In closing, we assess how close to reality the $1 bacterial genome actually is and explore the ways in which high-throughput sequencing might change the way that all microbiologists work.

A Variety of Approaches

High-throughput sequencing platforms can be divided into two broad groups depending on the kind of template used for the sequencing reactions. The earliest, and currently most widely used, platforms depend on the production of libraries of clonally amplified templates. These are produced through amplification of immobilized libraries made from a single DNA molecule in the initial sample. More recently, we have seen the arrival of single-molecule sequencing platforms, which determine the sequence of single molecules without amplification. Within these broad categories, there is considerable variation in performance—including in throughput, read length and error rate—as well as in factors affecting usability, such as cost and run time.

Template amplification technologies In general terms, all of the platforms that are currently on the market rely on a three-stage workflow of library preparation, template amplification and sequencing (Figure A9-1). Library preparation begins with the extraction and purification of genomic DNA. Depending on the protocol, the amount of DNA required can vary from a few nanograms to tens of micrograms, meaning that success in this step depends on the ability to grow sufficient biomass. For some microorganisms, obtaining suitable DNA—in terms of quantity and quality—can prove difficult. Therefore, before using expensive reagents for library preparation and sequencing, it is advisable to confirm, by fluorometry, that DNA of sufficient quantity and quality has been obtained. However, purchasing a suitable instrument to do this adds to the costs of establishing a sequencing capability (Box A9-1).

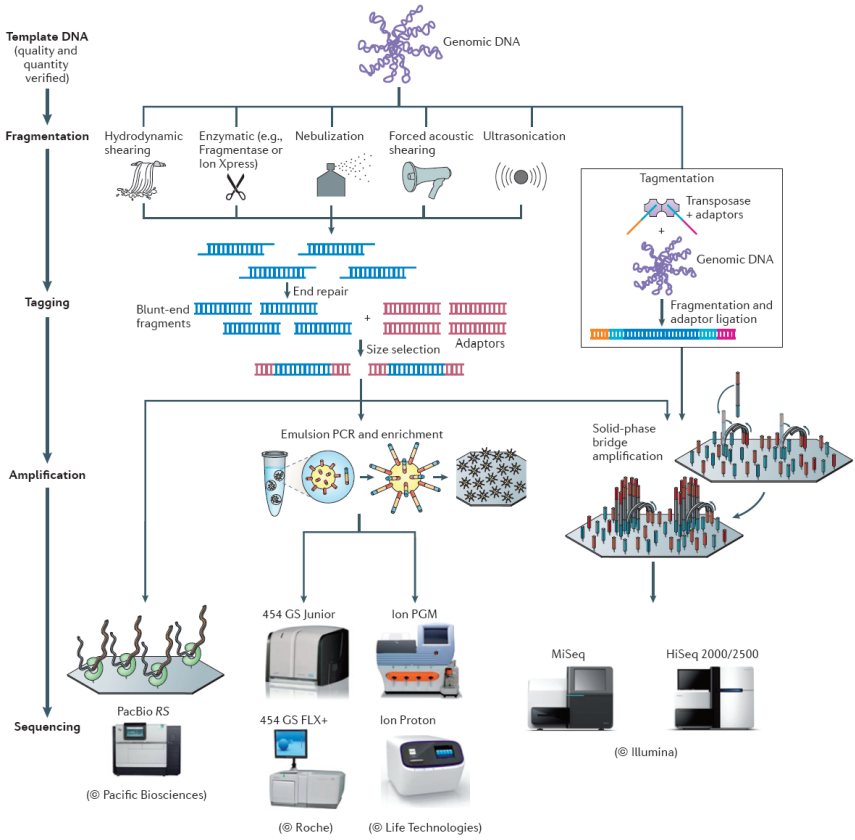


FIGURE A9-1 High-throughput sequencing platforms. The schematic shows the main high-throughput sequencing platforms available to microbiologists today, and the associated sample preparation and template amplification procedures. For full details, see main text. PGM, Personal Genome Machine. The tagmentation schematic is modified, with permission, from Adey et al. © (2010) BioMed Central.

For shotgun sequencing, an initial fragmentation step is required to generate random, overlapping DNA fragments. Depending on the platform and application, these fragments can range from 150 bp to 800 bp in length; size selection either involves harvesting from agarose gels or exploits paramagnetic-bead-based technology. The selected fragments must also be sufficiently abundant to provide comprehensive and even coverage of the target genome. Two types of fragmentation are widely used: mechanical and enzymatic. Early protocols relied on mechanical methods such as nebulization or ultrasonication. Nebulization is an

BOX A9-1
The Add-on Cost of Sequencing

The costs of sequencing instruments and reagents are not the only issues that need to be taken into account when setting up a sequencing facility for microbial applications. So, what else do you need? Well, first you have to buy a high-end fluorometer such as a Life Technologies Qubit (around US$2,000) and/or an Agilent Technologies 2100 Bioanalyzer (around $18,000). Then, if you want to save time by parallel processing, you should consider investing in an ultrasonicator (for example, from Covaris, at around $45,000) and a liquid-handling robot (for example, the Biomek FX^P, at around $310,000, or one of the SPRIworks systems, at around $45,000; both from Beckman Coulter). To carry out sequencing on the 454 GS FLX+ instrument from Roche, you need a bead counter for emulsion PCR (up to $20,000), and for the Genome Analyzer IIx or HiSeq machines from Illumina, you need to buy an Illumina cBot (~$55,000). For some platforms, you may have to buy additional centrifuges and/or rotors; for example, the ULTRA-TURRAX Tube Drive system from IKA ($1,000) is required by the Ion Torrent platform (from Life Technologies) if the OneTouch system is not used. You also need to buy a server to take receipt of the data coming off your instrument (for example, a $5,000 desktop), and then a cluster of servers for analysing and storing the data (ranging from $20,000 upwards). In addition, you may have to update your laboratory infrastructure by investing in a dedicated electrical connection and appropriate air-conditioning units for your sequencing instrument, and uninterruptible power supplies for your sequencer and servers. Most laboratories also want to invest in a backup solution that is both fast and available. This may be a mirrored set of hard drives, or even a shelf full of disconnected USB drives. Illumina offers a cloud-based backup and basic-analysis solution called BaseSpace which can store sequence results as they are generated on the Illumina MiSeq. Currently, this is a free solution, but users are likely to have to pay a subscription in the future.

inexpensive method that can be easily adopted by any laboratory, but it results in large losses of input material and a broad range of fragment sizes, runs the risk of cross-contamination and cannot handle parallel processing. By contrast, ultrasonication instruments such as systems from Covaris or the Bioruptor systems from Diagenode allow parallel sample processing and minimize hands-on time and sample loss but come at a price that could be prohibitive for small laboratories. Mechanically generated fragments require repair and end-polishing before platform-specific adaptors can be ligated to the ends of the target molecules. These adaptors act as primer-binding sites for the subsequent template amplification reaction.

More recently, enzymatic methods have provided an alternative approach to producing random fragments of the desired length. These require less input DNA and offer easier, faster sample processing. Fragmentase (from New England Biolabs) is a mixture of a nuclease, which randomly nicks double-stranded

DNA, and a T7 endonuclease, which cleaves the DNA. Together, these enzymes generate random double-strand DNA breaks in a time-dependent manner, allowing the user to tailor protocols in order to obtain products of the required length. Adaptors can then be ligated to these fragments in the usual way. Tagmentation (Caruccio, 2011) is a promising transposase-based approach that, in a single step, fragments DNA and incorporates sequence tags, which then take the place of adaptors. Currently, the only available implementation of tagmentation is within the Nextera system, which is only available for the Illumina platform. Several companies have produced automated liquid-handling machines that greatly reduce the hands-on time required for fragmentation approaches but significantly increase costs (Box A9-1).

In addition to supporting fragment-based sequencing, all template amplification platforms support mate pair sequencing, in which the ends of DNA fragments of a certain size (typical sizes are 3 kb, 6 kb, 8 kb or 20 kb) are joined together to form circular molecules. These molecules are then fragmented a second time. Fragments flanking the joins are then selected and end adaptors added. Sequencing through the joins provides valuable information about the location of sequences dispersed across the genome, facilitating assembly.

Paired-end sequencing has similarities to mate pair sequencing, but DNA fragments are sequenced from each end without the need for additional library preparation steps. The Illumina platform has direct support for paired-end sequencing. Short fragments that are less than the read length from the forward and reverse ends (for example, 180 bp fragments combined with 2 × 100 base sequencing) permits overlapping pseudo long reads to be generated. Alternatively, fragments of up to ~800 bp can be used. Longer fragments may result in a loss of amplification efficiency. The Ion Personal Genome Machine (PGM) (using the Ion Torrent platform, from Life Technologies) also has a bidirectional sequencing protocol that requires the removal of the chip after the initial run, a digestion step and a second sequencing run using a different sequencing primer. All platforms can handle PCR products, allowing adaptor sequences to be incorporated into the 5′ ends of primers.

For all platforms, it is highly advisable to assess the quality and quantity of the sequence library before subjecting it to amplification. Different instruments for quality assessment are recommended by different manufacturers. Examples include the 2100 Bioanalyzer (from Agilent Technologies), fluorometers such as the NanoDrop 3300 (from Thermo Scientific) or the Qubit (from Life Technologies), and quantitative PCR using any of a number of available quantitative PCR machines along with either own-design or commercially available assays. Purchasing a suitable instrument for this step can add several thousand dollars to the costs of establishing a sequencing capability (Box A9-1).

In preparation for amplification, template molecules are immobilized on a solid surface, which is a flow cell for sequencing with the Illumina platform and solid beads or ion sphere particles for other approaches. Simultaneous solid-phase

amplification of millions or billions of spatially separated template fragments prepares the way for massively parallel sequencing. For the Illumina platform, template amplification is automated and is performed either directly on the instrument (for the MiSeq, and the HiSeq 2500 sequencer in rapid-run mode) or using the cBot, a separate instrument that is dedicated to this task (used in conjunction with the Genome Analyzer IIx and the HiSeq 2000 machine). Clusters are generated by bridge amplification on the surface of the flow cell. For platforms that use bead-based immobilization (the SOLiD [from Life Technologies], 454 and Ion Torrent platforms), amplified template sequence libraries are prepared off-instrument, relying on an emulsion PCR, in which the beads are enclosed in aqueous-phase microreactors and are kept separated from each other in a water-in-oil emulsion.

Sequencing chemistry Although these platforms rely on a sequencing-by-synthesis design, they differ in the details of the sequencing chemistry and the approach used to read the sequence. The Illumina sequencing platform depends on Solexa chemistry (Bentley et al., 2008), which includes reversible termination of sequencing products. In each sequencing cycle, a mixture of fluorescently labelled 'reversible terminator' nucleotides with protected 3′-OH groups (and a different emission wavelength for each nucleotide) is perfused across the flow cell. Wherever a complementary nucleotide is present on the template strand, the terminator is incorporated and imaged, and then the signal is quenched and the terminator nucleotide is chemically deprotected at the 3′-OH group.

The 454 and Ion Torrent sequencing platforms avoid the use of terminators. Instead, in each cycle a single kind of dNTP is flowed across the template. When there is base complementarity between the dNTP and the next available position in the template, the DNA polymerase incorporates the base onto the extending strand, liberating pyrophosphate and hydrogen ions. When there is no complementarity, DNA synthesis is halted temporarily; each type of dNTP is flowed across the template in turn according to the dispensing cycle, and DNA synthesis is thus re-initiated when the next complementary dNTP is added. The 454 platform exploits a pyrosequencing approach (Margulies et al., 2005; Ronaghi et al., 1998) whereby the presence of pyrophosphate is signalled by visible light as the result of an enzyme cascade. The order and intensity of the light peaks are recorded as "flowgrams." The Ion Torrent platform relies on a modified silicon chip to detect hydrogen ions that are released during base incorporation; the resulting lack of reliance on imaging makes this platform the first "post-light" sequencing instrument (Rothberg et al., 2011).

The SOLiD platform (Valouev et al., 2008) and the platform from Complete Genomics (Drmanac et al., 2010) depend on sequencing by ligation. In this approach, fluorescent probes undergo iterative steps of hybridization and ligation to complementary positions in the template strand at the 5′ end of the extending strand, followed by fluorescence imaging to identify the ligated probe.

Single-molecule sequencing Single-molecule sequencing brings the promise of freedom from amplification artefacts as well as from onerous sample and library preparations. The HeliScope Single-Molecule Sequencer (from Helicos BioSciences) was the first platform for single-molecule sequencing to hit the market place in 2009 (Bowers et al., 2009). This technology applies one-colour reversible-terminator sequencing to unamplified single-molecule templates. However, this platform has been hampered by its high price and poor instrument sales and, following the delisting of the company from the stock market, there are significant doubts over the future of the platform.

More recently, Pacific Biosciences has delivered "real-time sequencing," in which dye-labelled nucleotides are continuously incorporated into a growing DNA strand by a highly processive, strand-displacing ϕ29-derived DNA polymerase (Eid et al., 2009). Each DNA polymerase molecule is tethered within a zero-mode waveguide detector, which allows continuous imaging of the labelled nucleotides as they enter the strand (Levene et al., 2003).

Choosing a Platform

High-end instruments The high-throughput sequencing market presents the user with a challenging choice between bulky, expensive high-end instruments and the new generation of bench-top instruments (Tables A9-1, A9-2). The high-end machines include PacBio*RS* (from Pacific Bioseciences), the HiSeq instruments, Genome Analyzer IIx, the SOLiD 5500 series and the 454 GS FLX+ system. These deliver a high throughput and/or long read lengths but come with set-up costs of hundreds of thousands of dollars, placing them beyond the reach of the average research laboratory or even department. These machines are thus only suitable for large sequencing centres or core facilities. This raises the important question of where an 'average' microbiologist should source sequencing from.

These instruments can deliver dozens to thousands of bacterial genomes per run, as illustrated by several high-impact publications on bacterial genomes and metagenomes (Harris et al., 2012; Hess et al., 2011; Mutreja et al., 2011; Qin et al., 2010). However, to achieve efficiencies in time and cost, optimum sequencing of microbial samples on such instruments requires onerous and expensive bar-coding and multiplexing of samples and/or subdivision of runs (for example, through gaskets or the use of single channels on the Illumina platform), as well as a sophisticated scheduling system. Compare sequencing a single human genome with the equivalent sequencing throughput for 1,000 average-sized bacterial genomes: although the sequencing run itself may be comparable in both scenarios, >1,000 samples and libraries need to be prepared for the bacterial run, compared with just one for the human genome. The costs and effort involved in sequencing 1,000 bacterial genomes therefore vastly outweigh the requirements for sequencing a single human genome, so the hasty calculation that one human

TABLE A9-1 Comparison of Next-Generation Sequencing Platforms

Machine (manufacturer)	Chemistry	Modal read length* (bases)	Run time	Gb per run	Current approximate cost (US$)†	Advantages	Disadvantages
High-end instruments							
454 GS FLX+ (Roche)	Pyrosequencing	700–800	23 hours	0.7	500,000	• Long read lengths	• Appreciable hands-on time • High reagent costs • High error rate in homopolymers
HiSeq 2000/2500 (Illumina)	Reversible terminator	2 × 100	11 days (regular mode) or 2 days (rapid run mode)§	600 (regular mode) or 120 (rapid run mode)§	750,000	• Cost-effectiveness • Steadily improving read lengths • Massive throughput • Minimal hands-on time	• Long run time • Short read lengths • HiSeq 2500 instrument upgrade not available at time of writing (available end 2012)
5500xl SOLiD (Life Technologies)	Ligation	75+35	8 days	150	350,000	• Low error rate • Massive throughput	• Very short read lengths • Long run times
PacBio RS (Pacific Biosciences)	Real-time sequencing	3,000 (maximum 15,000)	20 minutes	3 per day	750,000	• Simple sample preparation • Low reagent costs • Very long read lengths	• High error rate • Expensive system • Difficult installation

continued

TABLE A9-1 Continued

Machine (manufacturer)	Chemistry	Modal read length* (bases)	Run time	Gb per run	Current approximate cost (US$)†	Advantages	Disadvantages
Bench-top instruments							
454 GS Junior (Roche)	Pyrosequencing	500	8 hours	0.035	100,000	• Long read lengths	• Appreciable hands-on time • High reagent costs • High error rate in homopolymers
Ion Personal Genome Machine (Life Technologies)	Proton detection	100 or 200	3 hours	0.01–0.1 (314 chip), 0.1–0.5 (316 chip) or up to 1 (318 chip)	80,000 (including OneTouch and server)	• Short run times • Appropriate throughput for microbial applications	• Appreciable hands-on time • High error rate in homopolymers
Ion Proton (Life Technologies)	Proton detection	Up to 200	2 hours	Up to 10 (Proton I chip) or up to 100 (Proton II chip)	145,000 + 75,000 for compulsory server	• Short run times • Flexible chip reagents	• Instrument not available at time of writing
MiSeq (Illumina)	Reversible terminator	2 × 150	27 hours	1.5	125,000	• Cost-effectiveness • Short run times • Appropriate throughput for microbial applications • Minimal hands-on time	• Read lengths too short for efficient assembly

* Average read length for a fragment-based run

† Approximate cost per machine plus additional instrumentation and service contract. See Glenn (2011).

§ Available only on the HiSeq 2500.

TABLE A9-2 The Applicability of the Major High-Throughput Sequencing Platforms

		Machine*						
Example application in bacteriology	Desirable characteristics	454 GS Junior‡	454 GS FLX+‡	Ion Personal Genome Machine (318 chip)§	MiSeq‖	HiSeq 2000‖	5500x1 SOLiD§	PacBio *RS*¶
De novo sequencing of novel strains to generate a single-scaffold reference genome	• Long reads • Paired-end protocol and/or long mate-pair protocol • Even coverage of genome	✓	✓✓	✓	✓	✓	✗	✓✓
Rapid characterization of a novel pathogen (draft *de novo* assembly of a genome for a single strain)	• Total run time (library preparation plus sequencing) of under 48 hours • Sufficient coverage of a bacterial genome in a single run	✓	✓✓	✓✓	✓✓	✗	✗	✓✓
Rough-draft *de novo* sequencing of small numbers of strains (<20) for comparative analysis of gene content	• Long or paired-end reads • High throughput • Ease of library and sequencing workflow • Cost-effective	✗	✓	✓	✓✓	✓✓	✓	✓
Re-sequencing of many similar strains (>50) for the discovery of single nucleotide polymorphisms and for phylogenetics	• Very high throughput • Low-cost, high-throughput sequence library construction • High accuracy	✗	✗	✓	✓	✓✓	✓	✓
Small-scale transcriptomics-by-sequencing experiments (for example, two strains under four growth conditions with two biological replicates, so 16 strains)	• High per-isolate coverage	✗	✓	✓	✓	✓✓	✓✓	✓✓

continued

TABLE A9-2 Continued

Example application in bacteriology	Desirable characteristics	Machine*						
		454 GS Junior‡	454 GS FLX+‡	Ion Personal Genome Machine (318 chip)§	MiSeq‖	HiSeq 2000‖	5500xl SOLiD§	PacBio *RS*¶
Phylogenetic profiling to genus-level using partial 16S rRNA gene amplicon sequencing	• High coverage • Long amplicon input (≥500 bp) • Long reads • High single-read accuracy (error rate <1%)	✓	✓✓	✓	✓✓	✓	✓	✗
Whole-genome metagenomics for the reconstruction of multiple genomes in a single sample	• Long reads or paired-end reads • Very high throughput • Low error rate	✗	✓	✓	✓	✓✓	✓	✓

* ✓✓, particularly well suited; ✓, suitable; ✗, not suitable.
‡ From Roche; § From Life Technologies; ‖ From Illumina; ¶ From Pacific Biosciences.

genome-sequencing project equates to 1,000 bacterial genome-sequencing projects starts to look rather optimistic.

Bench-top instruments Three modestly priced bench-top instruments with throughputs and workflows that are well suited to microbial applications have recently hit the market. The 454 GS Junior was released in early 2010 and is a smaller, lower-throughput version of the 454 GS FLX+ machine, exploiting similar emulsion PCR and pyrosequencing approaches but with lower set-up and running costs (Loman et al., 2012). The Ion PGM was launched in early 2011 and saw almost immediate use in the crowd-sourced analysis of the Shiga toxin-producing *Escherichia coli* (STEC) outbreak in Germany (Rohde et al., 2011; Mellmann et al., 2011). This platform has also shown the greatest improvement in performance in recent months: an assembly for the STEC outbreak strain was generated in May 2011 using data from five Ion Torrent 314 chips and consisted of more than 3,000 contigs, whereas comparable data from a single newer 316 chip assembled into fewer than 400 contigs. The MiSeq, which began to ship to customers in late 2011, is based on the existing Solexa chemistry but has dramatically reduced run times compared with the HiSeq (hours rather than days). This is made possible by the use of a smaller flow cell, leading to a reduced imaging time and faster microfluidics.

Each of these bench-top instruments is capable of sequencing a whole bacterial genome in days. The performance of all three instruments was recently compared by sequencing a British isolate from the German STEC outbreak of 2011 (Loman et al., 2012). In this evaluation, all three bench-top sequencing platforms generated useful draft genome sequences with assemblies that mapped to ≥95% of the reference genome, so by these criteria all could be judged fit for purpose. However, no instrument was able to generate accurate one-contig-per-replicon assemblies that might equate to a finished genome.

The MiSeq was found to have the highest throughput per run, lowest error rate and most user-friendly workflow of the three instruments: hands-on time is low because template amplification is carried out directly on the instrument without manual intervention. However, a paired-end 150-base sequencing run took more than 27 hours. The MiSeq is notable for being able to sequence fragments from both ends (paired-end mode) without changes to the library preparation stage or additional intervention during sequencing.

The 454 GS Junior produced the longest reads (mean 522 bases) and generated the least fragmented assemblies but had the lowest throughput and a cost-per-base that was at least one order of magnitude higher than the cost for the other two platforms. The Ion PGM delivered the fastest throughput per hour (80–100 Mb) and had the shortest run time (around 3 hours) but also had the shortest reads (mean 121 bases), although kits producing 200 bases have since been made available for this instrument. The Ion PGM and 454 GS Junior were both prone to making mistakes in homopolymeric tracts, and these mistakes caused assembly

errors that resulted in frame-shifts in coding regions, even when data were assembled at high read coverage.

Coping with the Data

The high-end sequencing platforms make considerable demands on the local information technology infrastructure in terms of data tracking and analysis, short-term storage and long-term archiving. Bench-top instruments have more modest information technology requirements. However, each platform delivers data in a slightly different format, and saying that one has sequenced a bacterial genome means different things on different platforms and can create difficulties when comparing or combining data generated on different platforms (Table A9-2).

There are two main analytical approaches to the exploitation of high-throughput sequencing data: reads can be aligned—that is, mapped—to a known reference sequence or subjected to *de novo* assembly. The choice of strategy depends on the read length obtained (short reads are better mapped to a reference), the availability of a good reference sequence and the intended biological application (for example, genomic epidemiology versus pathogen biology).

To document genetic variation in the genomes of multiple highly related strains, a mapping approach is efficient and often sufficient. In this situation, sequence variants can be called by aligning reads to a reference genome using short-read-mapping tools (see Supplementary information S1 (table)). A mapping approach is problematic when dealing with reads from repetitive regions or from parts of the genome that are absent from the reference genome, or when a closely related reference genome is unavailable.

De novo assembly is more informative when dealing with a new pathogen or a new strain of a well-known pathogen. Sequencing errors can have a significant impact on assembly. When platforms produce random errors, the effect of these errors on assembly can be overcome by increasing the depth of coverage. However, when errors are systematic and occur in predictable contexts (for example, in homopolymers), increasing the depth of coverage is unlikely to help, and it may be necessary to sequence the troublesome regions using an alternative technology. Very high-quality, near complete references may be obtained by a hybrid approach, such as in recent studies combining Pacific Biosciences and Illumina data (Bashir et al., 2012; Koren et al., 2012).

A variety of commonly used assemblers is now available (see Supplementary information S1 (table)), ranging from the platform specific (for example, Newbler from Roche) to the more generally applicable (for example, MIRA (Chevreax, et al., 2004), Velvet (Zerbino and Birney, 2008), and the CLC Genomics Workbench from CLC Bio). *De novo* assemblies can be compared using Mauve (Darling et al., 2004) or Mugsy (Angiuoli and Salzberg, 2011), and the assemblies can be manually examined using the Tablet viewer (Milne et al., 2010). For

annotation of assemblies, Glimmer (Delcher et al., 1999) works well for coding-sequence prediction, while tRNAScan-SE (Schattner et al., 2005) and RNAmmer (Lagesen et al., 2007) work well for stable-RNA prediction. There are numerous pipelines for automatic annotation of *de novo* assemblies, including RAST (Aziz et al., 2008), IMG/ER (Markowitz et al., 2009) and the IGS Annotation Engine (developed by the Institute for Genome Sciences, University of Maryland School of Medicine, USA), although care must be taken when interpreting results from such services, as the public databases used contain annotation errors that are then propagated to newly sequenced genomes (Richardson and Watson, 2012).

For microbial applications, all of the above programs run quickly (in minutes or hours) and are not particularly processor intensive. Some workflows combine a series of programs and provide an accessible interface for microbiologists who are not bioinformatics specialists. For example, xBASE-NG provides a "one-stop shop" for assembly, annotation and comparison of bacterial genome sequences (Chaudhuri et al., 2008). Sophisticated phylogenetic analyses are more demanding and may be beyond the capability of the average research group. One particular issue when constructing bacterial whole-genome phylogenies is the clouding of phylogenetic signal by recombination events and homoplasy (Marttinen et al., 2012). Algorithms such as ClonalFrame (Didelot and Falush, 2007) and ClonalOrigin (Didelot et al., 2010) take multiple whole-genome alignments as input and attempt to identify blocks of recombination. These approaches are computationally very expensive, and there is no "off the shelf" solution to comparing hundreds or thousands of bacterial genomes. There is a growing interest in alignment-free approaches for constructing bacterial phylogenies, as it is thought that these approaches may help address the computational challenges of these analyses (Köser et al., 2012).

A recurring problem with data from high-throughput sequencing is meeting the requirement, as stipulated by journals and funders, that data be lodged in the public domain. Unannotated assembled sequences can be uploaded to conventional sequence databases, such as GenBank, fairly easily. However, submission of annotated sequences can be onerous, slowing down the process of publication even further. Submission of sequence reads to short-read archives may be hampered by slow data transfer rates, and it remains uncertain how sustainable such archives will prove to be in the future. There may come a time when the easiest way to obtain such data will be to re-sequence the sample, rather than upload, archive and retrieve large data sets.

Current Applications and Future Prospects

High-throughput sequencing has already transformed microbiology. Rapid, low-cost genome sequencing has helped make genomic epidemiology a reality, allowing us to track the spread of pathogens through hospitals (Köser et al., 2012; Lewis et al., 2010), communities (Gardy et al., 2011; Mellmann et al., 2011;

Rohde et al., 2011) and across the globe (Beres et al., 2010; Harris et al., 2011; Mutreha et al., 2011). High-throughput sequencing has already had a huge impact on our understanding of microbial evolution, whether within a single patient over years or decades (for example, *Pseudomonas aeruginosa* in a patient with cystic fibrosis [Cramer et al., 2011]) or globally across the centuries (for example, influenza virus in the 1918 influenza pandemic [Dunham et al., 2009] or mediaeval *Yersinia pestis* in the Black Death [Bos et al., 2011]). Genome sequences have even been obtained from single microbial cells (Woyke et al., 2009).

There are many applications beyond mere genome sequencing. High-throughput sequencing has opened up new avenues for sequence-based profiling and metagenomics of complex microbial communities, including those associated with human health and disease (Hess et al., 2011; Qin et al., 2010). Particularly exciting is the promise of culture-independent approaches to pathogen discovery and detection (Lipkin, 2010). In the research laboratory, sequencing is taking over from microarrays as the method of choice for studying gene expression (using RNA sequencing (RNA-seq)) (Passalacqua et al., 2009; Sharma et al., 2010; Sorek and Cossart, 2010), mutant libraries (using Tn-seq and transposon-directed insertion site sequencing (TraDIS)) (Langridge et al., 2009; van Opinjnen et al., 2009) and protein–DNA interactions (using chromatin immunoprecipitation followed by sequencing (ChIP–Seq)) (Grainger et al., 2009).

So, what does the future hold? For current platforms, we can expect to see cheaper, easier library preparation methods and ever-higher sequencing throughputs. However, with the arrival of transformative new technologies (Branton et al., 2008) (Box A9-2), this might be seen as tinkering around the edges. The tipping point has already been reached such that the staff and infrastructure costs of handling and analysing sequence data outweigh the costs of generating that data. If the promise of portable, single-molecule, long-read-length sequencing bears fruit and these technologies show the same steady increase in functionality and cost-effectiveness that we have seen with earlier high-throughput sequencing platforms, we could be just a few years away from user-friendly, "$1-a-pop" bacterial genome sequencing.

As we have argued elsewhere (Pallen and Loman, 2011), high-throughput sequencing may well be poised to make a decisive impact on clinical microbiology, but there are still many difficulties to be overcome—for example, in presenting complex information to clinicians, in agreeing common formats for data sharing, in integrating genomics with clinical informatics and clinical practice, in benchmarking novel technologies and in gaining regulatory approval (from the US FDA and other bodies) for clinical applications of these technologies. One thing is certain: thanks to the expected relentless progress in sequencing technology, microbiology in the next 20 years will look nothing like it does now.

BOX A9-2
Oxford Nanopore: The Game Changer?

In February 2012, at a conference in the United States, the British company Oxford Nanopore Technologies announced a new, near-market "strand sequencing" technology that exploits protein nanopores embedded in an industrially fabricated polymer membrane. As a DNA strand is fed through a nanopore by a processive enzyme, the trinucleotides in contact with the pore are detected through electrochemistry.

The manufacturers have already claimed that they can sequence the 50 kb phage λ genome on both strands, and they claim that there is no theoretical read length limit. They also claim that sequencing can be paused, the sample recovered and replaced, and sequencing then started again. Plus, there is no need for onerous sample preparations: sequences can be read directly from blood (and probably also bacterial lysates).

Oxford Nanopore Technologies has announced two products, both scheduled to ship in late 2012. The MinION is a disposable US$900 sequencer housed in a USB stick, with 512 nanopores, each capable of running 120–1,000 bases per minute per pore for up to 6 hours. The MinION can generate 150 Mb of sequence per hour, all without fluidics or imaging, and bases are streamed live to a laptop through the USB connection. The GridION is a rack-mountable sequencer with 2,000 nanopores and is capable of generating tens of gigabases over 24 hours. Both machines promise astonishing read lengths at low cost and with minimal sample preparation. However, this technology currently suffers from a high error rate (~4%) that is chiefly due to deletion errors but, according to their February 2012 press conference, the manufacturers are confident that they can fix this.

How will access to a disposable sequencer change the way we do microbiology? With no capital costs or cumbersome set-up and installation, this technology certainly has the power to democratize sequencing even further. Will prices fall enough for it to be worth sequencing one bacterial genome per MinION, or will the long read lengths mean that we can mix samples and then disaggregate the genomes with little effort? If read lengths really can be obtained in the ≥100 kb range, then all the existing problems of short-read assembly in genomics and metagenomics will be rendered obsolete.

Furthermore, we can now take the sequencer to the patient's bedside or out into the field. Microbial ecologists need no longer depend on molecular bar codes such as the 16S rRNA gene when they can have whole genomes instead, and latter-day John Snows can use disposable sequencing, not just to detect cholera, but also to track its evolution and spread.

Of course, the reality may not match the hype, and we eagerly await the first independent evaluation of this technology. But if the dream comes true, most of the rest of this article will soon be redundant.

Acknowledgements

The authors thank the anonymous reviewers for their help and suggestions.

Competing Interests Statement

The authors declare competing financial interests. Mark J. Pallen was a winner of an Ion Personal Genome Machine (PGM) (from Ion Torrent, part of Life Technologies) in the European Ion PGM Grant Programme. Nicholas J. Loman has received expenses to speak at an Ion Torrent meeting organized by Life Technologies and has received honoraria and expenses from Illumina to speak at Illumina meetings. Chrystala Constantinidou, Jacqueline Z. M. Chan, Mihail Halachev, Martin Sergeant, Charles W. Penn and Esther R. Robinson declare no competing financial interests.

References

Adey, A. *et al*. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Res.* **11**, R119 (2010).

Angiuoli, S. V. & Salzberg, S. L. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27**, 334–342 (2011).

Aziz, R. K. *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

Bashir, A. *et al*. A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotech.* 1 Jul 2012 (doi: 10.1038/nbt.2288).

Bentley, D. R. *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

Beres, S. B. *et al*. Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. *Proc. Natl Acad. Sci. USA* **107**, 4371–4376 (2010).

Bos, K. I. *et al*. A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478**, 506–510 (2011).

Bowers, J. *et al*. Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods* **6**, 593–595 (2009).

Branton, D. *et al*. The potential and challenges of nanopore sequencing. *Nature Biotech.* **26**, 1146–1153 (2008).

Caruccio, N. in *High-Throughput Next Generation Sequencing: Methods and Applications. Methods in Molecular Biology* Vol. **733** (eds Kwon, Y. M. & Ricke, S. C.) 241–255 (Humana Press, 2011).

Chaudhuri, R. R. *et al*. xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res.* **36**, D543–D546 (2008).

Chevreux, B. *et al*. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, 1147–1159 (2004).

Cramer, N. *et al*. Microevolution of the major common *Pseudomonas aeruginosa* clones C and PA14 in cystic fibrosis lungs. *Environ. Microbiol.* **13**, 1690–1704 (2011).

Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).

Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).

Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266 (2007).

Didelot, X., Lawson, D., Darling, A. & Falush, D. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186**, 1435–1449 (2010).

Domazet-Lošo, M. & Haubold, B. Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics* **27**, 1466–1472 (2011).

Drmanac, R. *et al*. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).

Dunham, E. J. *et al*. Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. *J. Virol.* **83**, 5485–5494 (2009).

Eid, J. *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

Fleischmann, R. D. *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).

Gardy, J. L. *et al*. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).

Glenn, T. C. A field guide to next generation DNA sequencers. *Mol. Ecol. Res.* **11**, 759–769 (2011).

Grainger, D. *et al*. Direct methods for studying transcription regulatory proteins and RNA polymerase in bacteria. *Curr. Opin. Microbiol.* **12**, 531–535 (2009).

Harris, S. R. *et al*. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2011).

Harris, S. R. *et al*. Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. *Nature Genet.* **44**, 413–419 (2012).

Hess, M. *et al*. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–467 (2011).

Koren, S. *et al*. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotech.* 1 Jul 2012 (doi: 10.1038/nbt.2280).

Köser, C. U. *et al*. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.* **366**, 2267–2275 (2012).

Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).

Langridge, G. C. *et al*. Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. *Genome Res.* **19**, 2308–2316 (2009).

Levene, M. J. *et al*. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).

Lewis, T. *et al*. High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. *J. Hosp. Infect.* **75**, 37–41 (2010).

Lipkin, W. I. Microbe hunting. *Microbiol. Mol. Biol. Rev.* **74**, 363–377 (2010).

Loman, N. J. *et al*. Performance comparison of bench-top high-throughput sequencing platforms. *Nature Biotech.* **30**, 434–439 (2012).

Margulies, M. *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

Markowitz, V. M. *et al*. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25**, 2271–2278 (2009).

Marttinen, P. *et al*. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40**, e6 (2012).

Mellmann, A. *et al*. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* **6**, e22751 (2011).

Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res.* **15**, 1767–1776 (2005).

Milne, I. *et al*. Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).

Mutreja, A. *et al*. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465 (2011).

Pallen, M. J. & Loman, N. J. Are diagnostic and public health bacteriology ready to become branches of genomic medicine? *Genome Med.* **3**, 53 (2011).

Passalacqua, K. D. *et al*. Structure and complexity of a bacterial transcriptome. *J. Bacteriol.* **191**, 3203–3211 (2009).

Qin, J. *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

Richardson, E. J. & Watson, M. The automatic annotation of bacterial genomes. *Brief. Bioinform.* 9 Mar 2012 (doi: 10.1093/bib/bbs007).

Rohde, H. *et al*. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.* **365**, 718–724 (2011).

Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **281** 363–365 (1998).

Rothberg, J. M. *et al*. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).

Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33**, W686–W689 (2005).

Sharma, C. M. *et al*. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255 (2010).

Sorek, R. & Cossart, P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Rev. Genet.* **11**, 9–16 (2010).

Valouev, A. *et al*. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).

van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods* **6**, 767–772 (2009).

Venter, J. C. Multiple personal genomes await. *Nature* **464**, 676–677 (2010).

Woyke, T. *et al*. Assembling the marine metagenome, one cell at a time. *PLoS ONE* **4**, e5299 (2009).

Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

A10

EVIDENCE FOR SEVERAL WAVES OF GLOBAL TRANSMISSION IN THE SEVENTH CHOLERA PANDEMIC[34]

Ankur Mutreja,[35,*] *Dong Wook Kim*,[36,37,*] *Nicholas R. Thomson*,[35,*] *Thomas R. Connor*,[35] *Je Hee Lee*,[36,38] *Samuel Kariuki*,[39] *Nicholas J. Croucher*,[35] *Seon Young Choi*,[36,38] *Simon R. Harris*,[35] *Michael Lebens*,[40] *Swapan Kumar Niyogi*,[41] *Eun Jin Kim*,[36] *T. Ramamurthy*,[41] *Jongsik Chun*,[38] *James L. N. Wood*,[42] *John D. Clemens*,[36] *Cecil Czerkinsky*,[36] *G. Balakrish Nair*,[41] *Jan Holmgren*,[40] *Julian* ***Parkhill***,[35] *and Gordon Dougan*[35]

Vibrio cholerae **is a globally important pathogen that is endemic in many areas of the world and causes 3–5 million reported cases of cholera every year. Historically, there have been seven acknowledged cholera pandemics; recent outbreaks in Zimbabwe and Haiti are included in the seventh and ongoing pandemic (Chin et al., 2011). Only isolates in serogroup O1 (consisting of two biotypes known as "classical" and "El Tor") and the derivative O139 (Chun et al., 2009; Hochhut and Waldor, 1999) can cause epidemic cholera (Chun et al., 2009). It is believed that the first six cholera pandemics were caused by the classical biotype, but El Tor has subsequently spread globally and replaced the classical biotype in the current pandemic (Chin et al., 2011). Detailed molecular epidemiological mapping of cholera has been compromised by a reliance on sub-genomic regions such as mobile elements to infer relationships, making El Tor isolates associated with the seventh pandemic seem superficially diverse. To understand the underlying phylogeny of the lineage responsible for the current pandemic, we identified high-resolution**

---

[34] Reprinted with kind permission from Nature Publishing Group.

[35] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

[36] International Vaccine Institute, SNU Research Park, Bongchun 7 dong, Kwanak, Seoul 151-919, Korea.

[37] Department of Pharmacy, College of Pharmacy, Hanyang University, Kyeonggi-do 426-791, Korea.

[38] Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea.

[39] Centre for Microbiology Research, KEMRI at Kenyatta Hosp Compound, Off Ngong Road, PO Box 43640-00100, Kenya.

[40] Department of Microbiology and Immunology and University of Gothenburg Vaccine Research Institute, The Sahlgrenska Academy at the University of Gothenburg, Box 435, 40530 Göteborg, Sweden.

[41] National Institute of Cholera and Enteric Diseases, P-33, CIT Scheme XM, Beliaghata, Kolkata 700 010, India.

[42] University of Cambridge, Department of Veterinary Medicine, Madingley Road, Cambridge CB3 0ES, UK.

* These authors contributed equally to this work.

markers (single nucleotide polymorphisms; SNPs) in 154 whole-genome sequences of globally and temporally representative *V. cholerae* isolates. Using this phylogeny, we show here that the seventh pandemic has spread from the Bay of Bengal in at least three independent but overlapping waves with a common ancestor in the 1950s, and identify several transcontinental transmission events. Additionally, we show how the acquisition of the SXT family of antibiotic resistance elements has shaped pandemic spread, and show that this family was first acquired at least ten years before its discovery in *V. cholerae*.

Main

Whole-genome analysis is perhaps the ultimate approach to building a robust phylogeny in recently emerged pathogens, through the identification of SNPs and other rare genetic variants (Harris et al., 2010). Therefore, we sequenced the genomes of 136 isolates of *V. cholerae*, the causative agent of several million cholera cases each year (http://www.who.int/mediacentre/ factsheets/fs107/en/). These sequences, including 113 isolates from the seventh pandemic, were added to 18 previously published genomes (CDC, 2010; Chin et al., 2011; Chun et al., 2009) to produce a global genomic database from isolates collected in the course of a century. We included representative El Tor isolates collected in the past four decades and compared these to previously reported and novel genome sequences of both classical and non-O1 types (Chin et al., 2011; Chun et al., 2009).

The sequence reads were mapped to the reference sequence of El Tor N16961 (Heidelberg et al., 2000), a seventh-pandemic *V. cholerae* that was isolated in Bangladesh in 1975 (see footnote to Supplementary) and the resulting consensus tree identified eight distinct phyletic lineages (L1–L8, see Supplementary Fig. 1 and Supplementary Table 1 for strain and lineage information), six of which incorporated O1 clinical isolates. The classical isolates formed a distinct, highly clustered group (L1), distant from the El Tor isolates of the seventh pandemic (L2). It is clear from Supplementary Fig. 1 that the classical and El Tor clades did not originate from a recent common ancestor and instead seem to be independent derivatives with distinct phylogenetic histories, consistent with previous proposals (Chun et al., 2009). Isolates of L4 share a common ancestor with previously reported non-conventional O1 isolates (Chun et al., 2009) (Supplementary Fig. 2), and are likely to have acquired the O1 antigen genes by a recombination event onto a genetically distinct genome backbone. Isolates of L7 also have a distinct backbone, whereas L2, L3 (USA Gulf coast strains), L5, L6 and L8 share a more "El-Tor-like" genome backbone, and the L1 backbone is of the "classical" type.

Genome-wide SNP analysis showed that the 123 El Tor isolates in the L2 cluster (Supplementary Fig. 1) differed from the reference by only 50–250 SNPs. With this large sample size we were able to construct a high-resolution phylogeny that shows unequivocally that the current pandemic is monophyletic and

originated from a single source, providing a framework for future epidemiological and phenotypic analysis of *V. cholerae*, including transmission-tracking and typing.

Predicted recombined regions were identified, and along with genomic islands and mobile genetic elements, these were initially excluded from the phylogenetic analysis of seventh-pandemic isolates, to determine the underlying phylogeny. Notably, analysis of the tree (Figure A10-1; see Supplementary Fig. 3 for a tree with strain names) provides clear evidence of a clonal expansion of the

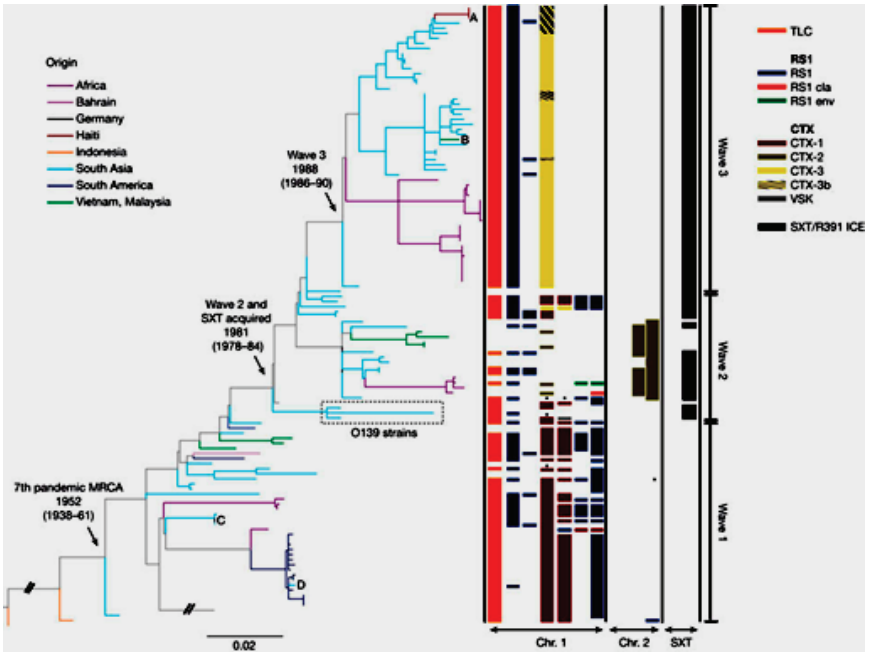


FIGURE A10-1 A maximum-likelihood phylogenetic tree of the seventh pandemic lineage of *V. cholerae* based on SNP differences across the whole core genome, excluding probable recombination events. The pre-seventh-pandemic isolate M66 was used as an outgroup to root the tree. Branches are coloured on the basis of the region of isolation of the strains. The branches representing the three major waves are indicated on the far right. The nodes representing the MRCAs of the seventh pandemic, and subsequent waves 2 and 3, are indicated with arrows and labelled with inferred dates. The presence and type of CTX and SXT elements in each strain are shown to the right of the tree. The presence of toxin-linked cryptic (TLC) and repeated sequence 1 (RS1) elements is shown, but their number and position, respectively, are arbitrarily assigned. Cases of sporadic intercontinental transmission are marked A–D. The dates shown are the median estimates for the indicated nodes, taken from the results of the BEAST analysis. The scale is given as the number of substitutions per variable site; asterisks indicate that no data were available.

lineage, with a strong temporal signature. This is most clearly illustrated by the fact that the most divergent isolates from the N16961 reference are represented by the oldest seventh-pandemic isolate in our collection, A6, collected in 1957, together with the most recent Haitian isolates (CDC, 2010) from late 2010. We performed a linear regression analysis on all the L2 isolates to calculate the rate of SNP accumulation on the basis of the date of isolation and the root-to-tip distance. The shape of the tree and temporal signatures in Fig. A10-1 show a very consistent rate of SNP accumulation, 3.3 SNPs year^{-1} ($R^2 = 0.73$, Supplementary Fig. 4) in the core genome, emphasizing the tree's robustness and utility for transmission studies. The only exception to this is *V. cholerae* A4, a repeatedly passaged laboratory strain that was originally isolated in 1973 (Supplementary Figs 3 and 4). The estimated rate of mutation for our seventh-pandemic *V. cholerae* collection was 8.3×10^{-7} SNPs site^{-1} year^{-1}: between 5 and 2.5 times slower than the rate estimated for recent clonal expansions of some other human-pathogenic bacteria (Croucher et al., 2011; Harris et al., 2010).

The seventh-pandemic tree can be subdivided into three major groups or clades by clustering using Bayesian analysis of population structure (Corander et al., 2003, 2008) (shown as waves 1–3 in Figure A10-1); this clustering is mostly consistent with the cholera toxin (CTX) type of the three clades, which represent independent waves of transmission. Although examples of genetic determinants differentiating these three CTX types have previously been published (Safa et al., 2010), they have not been put into a phylogenetic context, undermining efforts to investigate the evolutionary aspects of their emergence. Perhaps as a result, there has been substantial uncertainty in naming new CTX types as they have been discovered. Our data shows that the first CTX type is canonical CTX El Tor and we propose that it is renamed CTX-1; for the other two we propose a new expandable nomenclature and class them as CTX-2 and CTX-3 (Supplementary Table 2).

Isolates spanning A18 to PRL5 (the lower clade in Figure A10-1) represent wave 1, covering about 16 years (1977–1992). All isolates in this group lack the integrative and conjugative element (ICE) of the SXT/R391 family, encoding resistance to several antibiotics (Garriss et al., 2009; Wozniak et al., 2009). It is within this time period that seventh-pandemic cholera occurred in South America (Heidelberg et al., 2000). Our data show that the South American isolates form a discrete cluster, which also includes a single Angolan isolate collected in 1989. The position of the Angolan isolate at the base of the South American group indicates that transmission to South America may have been via Africa, as previously proposed (Lam et al., 2010). We used BEAST (Drummond et al., 2006) to translate evolutionary distance in SNPs into time (Supplementary Fig. 5) and this indicated that transmission to South America is likely to have occurred between 1981 and 1985. The branch harbouring this West African–South American (WASA) clade is distinguished from all other *V. cholerae* by the acquisition of novel VSP-2 genes (O'Shea et al., 2004) and a novel genomic island that we

have denoted WASA1 (Supplementary Table 3). Notably, the Angolan isolate A5 and all the South American isolates are discriminated by just ten SNPs. Based on the accumulation rate of 3.3 SNPs year^{-1} (Supplementary Fig. 4), the 3-year time period between the isolation of A5 and the oldest South American isolate included in this study, A32, is consistent with previous studies indicating that cholera spread as a single epidemic (Lam et al., 2010).

The first acquisition of an SXT/R391 ICE lies at the point of transition from the wave-1 cluster to the wave-2 cluster. Using our dated phylogeny (Supplementary Fig. 5) (Drummond et al., 2006), we were able to date this transition and the first acquisition of SXT/R391 ICE to 1978–84, ten years before its discovery in O139 strains, which also fits with the otherwise surprising discovery of SXT in a Vietnamese strain isolated before 1992 (Bani et al., 2007). This date would also correspond to the most recent common ancestor (MRCA) of the O1 and O139 serogroup isolates. Analysis of the diversity of the common regions of SXT/R391 ICEs in our seventh-pandemic collection (Supplementary Fig. 6) shows that they are discriminated by 3,161 SNPs, compared to only 1,757 SNPs used to define the core whole-genome phylogeny in Figure A10-1. This indicates either that there have been several recombination events within these ICEs, or that they have been acquired independently several times on the tree (Garriss et al., 2009). Isolates from wave 2 represent a discrete cluster that shows a complex pattern of accessory elements in the CTX locus (Figure A10-1) and a wide phylogeographical distribution. It is also notable that isolates collected in Vietnam in 1995–2004 and strain A109 are the only wave-2 isolates studied from this time period that lack an SXT/R391 ICE. We examined the genomic locus in these clones that marks the point of insertion of SXT/R391 ICE in all other *V. cholerae* isolates and found no remnants of this conjugative element, which may have been lost from this lineage (no "scar" in DNA sequence is expected after the precise excision of SXT/R391 ICE).

Ignoring the CTX-related genomic regions, the seventh-pandemic L2 isolates show relatively little evidence of recombination either within or from outside the tree. On the basis of the SNP distribution, 1,930 out of 2,027 SNPs (Supplementary Table 4) are congruent with the tree, leaving 97 homoplasies that could be due to selection or homologous recombination among the L2 isolates. Only 270 SNPs were predicted to be due to homologous recombination from outside the tree. The only two branches in which the SNP distribution indicated considerable recombination were those leading to the WASA cluster (Supplementary Fig. 7) and the O139 serogroup. Aside from the acquisitions of CTX and the SXT/R391 ICEs, we found evidence of gene flux affecting only 155 other genes (Supplementary Figs 8 and 9 and Supplementary Table 3).

Also represented in our collection are two isolates of serogroup O139, which are known to have arisen from a homologous replacement of their O-antigen determinant into an El Tor genomic backbone (Chun et al., 2009; Hochhut and Waldor, 1999; Lam et al., 2010). CTX types that are different from El Tor,

classical, CTX-2 and CTX-3 have been reported for the O139 serogroup (Basu et al., 2000; Faruque and Mekalanos, 2003; Faruque et al., 2000; Nair et al., 1994); however, the phylogenetic position of the two strains included in this study shows that O139 was derived from O1 El Tor and therefore represents another distinct but spatially restricted wave from the common source.

We were also able to date the ancestor of the El Tor seventh-pandemic lineage, L2, as having existed in 1827–1936 (Supplementary Fig. 5), which is consistent with the predicted date of origin from the linear regression plot (1910, Supplementary Fig. 4). This also corresponds well with the date of isolation of the first El Tor biotype strain in 1905 (Cvjetanovic and Barua, 1972).

It is apparent from Figure A10-1 that *V. cholerae* wave 1, which spread globally, was later replaced by the more geographically restricted wave 2 and wave 3, a phenomenon supported by local clinical observations and phage analysis (Safa et al., 2010). This also reflects the fact that *V. cholerae* epidemics since 2003–2010 have been restricted to Africa and south Asia. Notably, the rates of SNP accumulation calculated independently for wave 1, wave 3 and wave 2 (2.3, 2.6 and 3.5 SNPs year^{-1} respectively) are consistent with the rate calculated over the whole collection period (Supplementary Fig. 4).

The clonal clustering of L2 isolates, the constant rate of SNP accumulation and the temporal and geographical distribution support the concept that the seventh pandemic has spread by periodic radiation from a single source population located in the Bay of Bengal, followed by local evolution and ultimately local extinction in non-endemic areas. This is evidenced by the disappearance of wave-1 isolates, followed by the independent expansion of waves 2 and 3, both derived from the same original population, occurring within seven years of each other. These two waves are clearly distinguished from the first by the acquisition of SXT/R391 ICEs (Figure A10-1). Plotting the intercontinental spread of each wave onto the world map (Figure A10-2) clearly shows that the *V. cholerae* seventh pandemic is sourced from a single, restricted geographical location but has spread in overlapping waves. In these ancestral waves, there are at least four recent long-range transmission events (A–D in Figure A10-1), in which isolates clearly share a common ancestor with recent strains at distant locations, indicating that such events are not uncommon. The most recent example of this is the Haitian outbreak, in which strains share a very recent common ancestor with south-Asian strains at the tip of wave 3. The number of SNP differences, even at whole-genome resolution, between the Haitian and the most closely related Indian and Bangladeshi strains is very low. This demonstrates that the Haitian strains must have come from south Asia, at most within the last six years. However, the limited discrimination means that it may prove challenging to make country-specific inferences as to the origins of the Haitian strains on the basis of DNA sequence alone. For such conclusions to be robust, great care must be taken in the selection of samples for analysis.

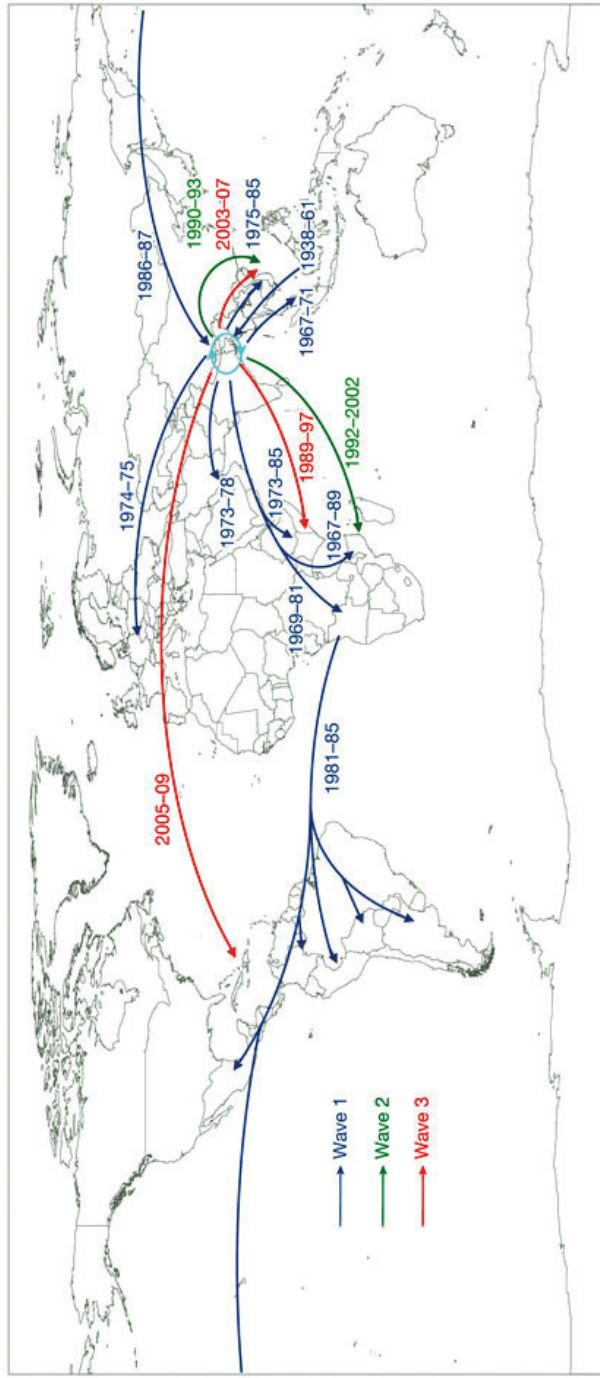


FIGURE A10-2 Transmission events inferred for the seventh-pandemic phylogenetic tree, drawn on a global map. The date ranges shown for transmission events are taken from the BEAST analysis, and represent the median values for the MRCA of the transmitted strains (later bound), and the MRCA of the transmitted strains and their closest relative from the source location (earlier bound).

Despite clear evidence of sporadic long-range transmission events that are likely to be associated with direct human carriage, the overall pattern seen in our data is one of continued local evolution of *V. cholerae* in the Bay of Bengal, with several independent waves of global transmission resulting in short-term epidemics in non-endemic countries. Although our sample set is substantial, there are clearly areas where geographical coverage is limited. However, the structure of the tree, with deep branches between the major waves, means that increasing the number of strains and the resolution further should only identify further independent waves of transmission. Indeed, we cannot rule out the possibility of an El Tor population persisting or evolving as a new wave of the seventh pandemic; for example, in areas such as China that were not sampled in this study.

One notable factor in the ongoing evolution of pandemic cholera was the acquisition of the SXT/R391-family antibiotic resistance element. The clinical use of the antibiotics tetracycline and furazolidone for cholera treatment started in 1963 and 1968 respectively, about 15 years before our prediction of the first acquisition of an SXT/R391 ICE (1978–1984). Our analysis provides a robust framework for elucidating the evolution of the seventh pandemic further, and for studying the local evolution, particularly in the Bay of Bengal, that has such a key role in the evolution of cholera.

Methods

Genomic Library Creation and Multiplex Sequencing

Unique index-tagged libraries for each sample were created, and up to 12 separate libraries were sequenced in each of eight channels in Illumina Genome Analyser GAII cells with 54-base paired-end reads. The index-tag sequence information was used for downstream processing to assign reads to the individual samples (Harris et al., 2010).

Detection of SNPs in the Core Genome

The 54-base paired-end reads were mapped against the N16961 El Tor reference (accession numbers AE003852 and AE003853) and SNPs were identified as described in Croucher et al. (2011). The unmapped reads and the sequences that were not present in all genomes were not considered a part of the core genome, and therefore SNPs from these regions were not included in the analysis. Appropriate SNP cutoffs were chosen to minimize the number of false-positive and false-negative calls; SNPs were filtered to remove those at sites with a SNP quality score lower than 30, and SNPs at sites with heterogeneous mappings were filtered out if the SNP was present in fewer than 75% of reads at that site. From the seventh-pandemic data set, high-density SNP clusters indicating possible recombination were excluded (Croucher et al., 2011). In total, 2,027 SNPs

were detected in the core genome of the El Tor lineage. Of these, 270 SNPs were predicted to be due to recombination. Removing these provided a data set characterized by 1,757 SNPs: these were used to produce the final phylogeny.

Comparative Genomics

Raw Illumina data were split to generate paired-end reads, and assembled using a *de novo* genome-assembly program, Velvet v0.7.03 (Zerbino et al., 2008), to generate a multi-contig draft genome for each of 133 *V. cholerae* strains (Harris et al., 2010). The overlap parameters were optimized to give the highest N50 value. Because seventh-pandemic *V. cholerae* strains are closely related in the core, Abacas (Assefa et al., 2009) was used to order the contigs using the N16961 El Tor strain as a reference, followed by annotation transfer from the reference strain to each draft genome (Harris et al., 2010). Using the N16961 sequence as a database to perform a TBLASTX (Altschul et al., 1990) for each draft genome, a genome comparison file was generated that was subsequently used in the Artemis comparison tool (Carver et al., 2008) to compare the genomes manually and search for novel genomic islands.

Phylogenetic Analysis

A phylogeny was drawn for *V. cholerae* using RAxML v0.7.4 (Stamatakis, 2006) to estimate the trees for all SNPs called from the core genome. The general time-reversible model with gamma correction was used for among-site rate variation for ten initial trees (Harris et al., 2010). USA Gulf coast strains A215 and A325, which have substantially different core genomes from all other strains in our collection, were used as an outgroup to root the global phylogeny (Supplementary Fig. 1), whereas a pre-seventh-pandemic strain, M66 (accession numbers CP001233 and CP001234), and strain A6 (from our collection), were used to root the seventh-pandemic phylogenetic tree (Figure A10-1).

CTX Prophage Analysis

For each strain, the CTX structure and the sequence of *rst*A, *rst*R and *ctx*B was determined as in Lee et al. (2009) and Nguyen et al. (2009).

Linear Regression and Bayesian Analysis

The phylogram for the seventh pandemic was exported to Path-O-Gen v1.3 (http://tree.bio.ed.ac.uk/software/pathogen) and a linear regression plot for isolation date versus root-to-tip distance was generated. The same plot was also constructed individually for the three waves, but A4, being a laboratory strain, was excluded from the latter analysis.

The presence of three waves was checked, and their makeup was determined, using a BAPS analysis performed on the SNP alignment containing the unique SNP patterns from the seventh-pandemic isolates. The program was run using the BAPS individual mixture model, and three independent iterations were performed using an upper limit for the number of populations of 20, 21 and 22 to obtain optimal partitioning of the sample. The dates for the acquisition of SXT and the ancestors of the three waves were inferred using the Bayesian Markov chain Monte Carlo framework BEAST (Drummond and Rambaut, 2007). We used the final SNP alignment with recombinant sites removed and fixed the tree topology to the phylogeny produced by RAxML, as described above. We used BEAST to estimate the rates of evolution on the branches of the tree using a relaxed molecular clock (Drummond et al., 2006), which allows rates of evolution to vary amongst the branches of the tree. BEAST produced estimates for the dates of branching events on the tree by sampling dates of divergence between isolates from their joint posterior distribution, in which the sequences are constrained by their known date of isolation. The data were analysed using a coalescent constant population size and a general time-reversible model with gamma correction. The results were produced from three independent chains of 50 million steps each, sampled every 10,000 steps to ensure good mixing. The first 5 million steps of each chain were discarded as a burn-in. The results were combined using Log Combiner, and the maximum clade credibility tree was generated using Tree Annotator, both parts of the BEAST package (http://tree.bio.ed.ac.uk/software/beast/). Convergence and the effective sample-size values were checked using Tracer 1.5 (available from http://tree.bio.ed.ac.uk/software/tracer). ESS values in excess of 200 were obtained for all parameters.

Nomenclature

The seventh-pandemic cholera strains were clearly distinguished by three waves and we therefore propose their CTX types to be CTX-1, CTX-2 and CTX-3 under the new nomenclature scheme (see Supplementary Table 2). Our nomenclature system is expandable and would be suitable for naming any new seventh-pandemic *V. cholerae* strains. With CTX-1 representing canonical El Tor, we followed the rationale: (1) For CTX-1 to CTX-2, because there was a shift of *rst*R$^{\text{El Tor}}$ to *rst*R$^{\text{Classical}}$, *rst*A$^{\text{El Tor}}$ to *rst*A$^{\text{Classical + El Tor}}$ and *ctx*B$^{\text{El Tor}}$ to *ctx*B$^{\text{Classical}}$, we called it CTX-2; (2) for CTX-1 to CTX-3, because there was a shift of *ctx*B$^{\text{El Tor}}$ to *ctx*B$^{\text{Classical}}$, we called it CTX-3; (3) for CTX-3 to CTX-3b, because there was only one SNP mutation in *ctx*B$^{\text{Classical}}$ from CTX-2 and rest was identical, we called it the next variant of CTX-3, which is CTX-3b.

In summary, if there is a shift of any gene from one biotype to another, the new CTX will be called CTX-n: thus the next strains fitting these criteria will be called CTX-4. However, if there is a mutation(s) that does not lead to a shift

of the gene to another biotype gene, CTX-1b, CTX-1c or CTX-2b; CTX-2c or CTX-3b; CTX-3c and so on should be followed as appropriate.

Methods Summary

Genomic libraries were created for each sample, followed by multiplex sequencing on an Illumina GAIIx analyser. The 54-base paired-end reads obtained were mapped against N16961 El Tor as a reference and SNPs in the core genome were identified as described in Methods. The SNPs were used to draw a whole coregenome phylogeny as described in Harris et al. (2010). The final SNP alignment was used to perform BEAST (Drummond et al., 2006) analysis and to confirm the output of linear regression analysis. The three cholera waves reported in the seventh-pandemic phylogeny were confirmed using BAPS (Corander et al., 2003, 2008). The raw Illumina data were also assembled *de novo* (see Methods) so that pairwise genome comparisons could be made. A new and expandable nomenclature system describing the CTX trends seen in the last 40 years was proposed following the rationale described in Methods.

Full methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Acknowledgements

This work was supported by The Wellcome Trust grant 076964. The IVI is supported by the Governments of Korea, Sweden and Kuwait. D.W.K. was partially supported by grant RTI05-01-01 from the Ministry of Knowledge and Economy (MKE), Korea and by R01-2006-000-10255-0 from the Korea Science and Engineering Foundation; and J.L.N.W. was supported by the Alborada Trust and the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security. Thanks to A. Camilli at Tufts University Medical School for providing the corrected N16961 sequence, to B.M. Nguyen at NIHE, Vietnam, and M. Ansaruzzaman at ICDDR, Bangladesh for providing strains, and to M. Fookes at WTSI for training support.

Author Contributions

A.M., D.W.K. and N.R.T. collected the data, analysed it and performed phylogenetic analyses and comparative genomics. J.H.L., S.Y.C., E.J.K. and J.C. analysed the CTX types. S.K., S.K.N. and T.R. were involved in strain collection and serogroup analysis. T.R.C. performed Bayesian analysis; N.J.C. and S.R.H. did the computational coding. J.L.N.W., J.D.C., C.C., G.B.K., J.H., N.R.T., J.P. and G.D. were involved in the study design. A.M., N.R.T., J.P., G.D., J.H., G.B.K., N.J.C., S.R.H., T.R.C., D.W.K. and M.L. contributed to the manuscript writing.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990)

Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**, 1968–1969 (2009)

Bani, S. *et al.* Molecular characterization of ICEVchVie0 and its disappearance in *Vibrio cholerae* O1 strains isolated in 2003 in Vietnam. *FEMS Microbiol. Lett.* **266**, 42–48 (2007)

Basu, A. *et al. Vibrio cholerae* O139 in Calcutta, 1992–1998: incidence, antibiograms, and genotypes. *Emerg. Infect. Dis.* **6**, 139–147 (2000)

Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**, 2672–2676 (2008)

CDC. 2010. Update: cholera outbreak—Haiti, 2010. *MMWR Morb. Mortal Wkly Rep.* **59**, 1473–1479 (2010)

Chin, C. S. *et al.* The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**,33–42 (2011)

Chun, J. *et al.* Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc. Natl Acad. Sci. USA* **106**,15442–15447 (2009)

Corander, J., Marttinen, P., Siren, J. & Tang, J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* **9**, 539 (2008)

Corander, J., Waldmann, P. & Sillanpaa, M. J. Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374 (2003)

Croucher, N. J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011)

Cvjetanovic, B. & Barua, D. The seventh pandemic of cholera. *Nature* **239**, 137–138 (1972)

Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007)

Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006)

Faruque, S. M. & Mekalanos, J. J. Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends Microbiol.* **11**, 505–510 (2003)

Faruque, S. M. *et al.* The O139 serogroup of *Vibrio cholerae* comprises diverse clones of epidemic and nonepidemic strains derived from multiple *V. cholerae* O1 or non-O1 progenitors. *J. Infect. Dis.* **182**, 1161–1168 (2000)

Garriss, G., Waldor, M. K. & Burrus, V. Mobile antibiotic resistance encoding elements promote their own diversity. *PLoS Genet.* **5**, e1000775 (2009)

Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010)

Heidelberg, J. F. *et al.* DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477–483 (2000)

Hochhut, B. & Waldor, M. K. Site-specific integration of the conjugal *Vibrio cholerae* SXT element into *prf. C. Mol. Microbiol.* **32**, 99–110 (1999)

Lam, C., Octavia, S., Reeves, P., Wang, L. & Lan, R. Evolution of seventh cholera pandemic and origin of 1991 epidemic, Latin America. *Emerg. Infect. Dis.* **16**, 1130–1132 (2010)

Lee, J. H. *et al.* Classification of hybrid and altered *Vibrio cholerae* strains by CTX prophage and RS1 element structure. *J. Microbiol.* **47**, 783–788 (2009)

Nair, G. B., Bhattacharya, S. K. & Deb, B. C. *Vibrio cholerae* O139 Bengal: the eighth pandemic strain of cholera. *Indian J. Public Health* **38**, 33–36 (1994)

Nguyen, B. M. *et al.* Cholera outbreaks caused by an altered *Vibrio cholerae* O1 El Tor biotype strain producing classical cholera toxin B in Vietnam in 2007 to 2008. *J. Clin. Microbiol.* **47**, 1568–1571 (2009)

O'Shea, Y. A. *et al.* The *Vibrio* seventh pandemic island-II is a 26.9 kb genomic island present in *Vibrio cholerae* El Tor and O139 serogroup isolates that shows homology to a 43.4 kb genomic island in *V. vulnificus*. *Microbiology* **150**, 4053–4063 (2004)

Safa, A., Nair, G. B. & Kong, R. Y. Evolution of new variants of *Vibrio cholerae* O1. *Trends Microbiol.* **18**, 46–54 (2010)

Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006)

Wozniak, R. A. *et al.* Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs. *PLoS Genet.* **5**, e1000786 (2009)

Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008)

A11

MULTI-PARTNER INTERACTIONS IN CORALS IN THE FACE OF CLIMATE CHANGE[43]

Koty H. Sharp[44,*] *and Kim B.* ***Ritchie***[45]

Abstract

Recent research has explored the possibility that increased sea-surface temperatures and decreasing pH (ocean acidification) contribute to the ongoing decline of coral reef ecosystems. Within corals, a diverse microbiome exerts significant influence on biogeochemical and ecological processes, including food webs, organismal life cycles, and chemical and nutrient cycling. Microbes on coral reefs play a critical role in regulating larval recruitment, bacterial colonization, and pathogen abundance under ambient conditions, ultimately governing the overall resilience of coral reef systems. As a result, microbial processes may be involved in reef ecosystem-level responses to climate change. Developments of new molecular technologies, in addition to multidisciplinary collaborative research on coral reefs, have led to the rapid advancement in our understanding of bacterially mediated reef responses to environmental change. Here we review new discoveries regarding (1) the onset of coral-bacterial associations; (2) the functional roles that bacteria play in healthy corals; and (3) how bacteria influence coral reef response to environmental change, leading to a model describing how reef microbiota direct ecosystem-level response to a changing global climate.

[43] Sharp, K. H., and K. B. Ritchie. 2012. *Biological Bulletin* 174: 319–329. Used with permission from the Marine Biological Laboratory, Woods Hole, MA.

[44] Eckerd College, 4200 54th Avenue South, St. Petersburg, Florida 33711.

[45] Mote Marine Laboratory, 1600 Ken Thompson Parkway, Sarasota, Florida 34236.

* To whom correspondence should be addressed. E-mail: sharpkh@eckerd.edu.

Introduction

The health of coral reefs is declining on a global scale and continues to be threatened by overfishing and habitat destruction. Anthropogenically induced global climate change has been identified as a significant threat to these sensitive ecosystems. As temperatures rise, bleaching and diseases are increasing, and excess atmospheric carbon dioxide is greatly altering reef ecosystems by changing seawater chemistry through decreases in pH (Anthony et al., 2011).

In a recent review, Bosch and McFall-Ngai (2011) highlight the significance of viewing animals as "metaorganisms"—multicellular organisms consisting of a macroscopic host and multiple microorganisms that interact synergistically to shape the ecology and evolution of the entire association. In this sense, the term *metaorganism* can be applied to a broad range of animal-microbe symbioses, ranging from humans to sponges (Bosch and McFall-Ngai, 2011). Coral research within this perspective has revolutionized the way that researchers study corals. In scleractinian (hard) corals, the term "holobiont" (Knowlton and Rohwer, 2003) was adapted to indicate that corals are dynamic, multi-domain assemblages consisting of an animal host, symbiotic dinoflagellates in the genus *Symbiodinium*, bacteria, archaea, fungi, and viruses (Rohwer et al., 2001, 2002; Stat et al., 2006; Wegley et al., 2007; Thurber et al., 2009). The term metaorganism is especially useful for describing corals and reflecting that corals' response to environmental change is driven by physiological interactions among the various microorganisms associated with the tissue, skeleton, and mucous layer. Corals harbor *Symbiodinium*, which provides fixed carbon to the host via photosynthesis, serving as the trophic foundation for coral reef ecosystems. It has been proposed that corals have additionally evolved to exploit specific bacterial metabolic capabilities that, in turn, directly modulate the survival of the coral holobiont in the marine environment (Zilber-Rosenberg and Rosenberg, 2008). An extensive characterization of the diverse microorganisms in corals will guide our understanding of the ecology of corals and coral reef ecosystems in response to a changing global climate.

Coral microbiology is a rapidly growing area of study. Early culture-based studies of coral-associated bacteria provided a foundation from which genomics, metagenomics, and transcriptomics approaches were established in corals, leading to exciting new advances in our current understanding of the diversity and dynamics of coral-associated bacterial communities. Evidence is accumulating that bacteria have an enormous influence on coral health and resilience, particularly with respect to changing reef environments (Azam and Worden, 2004; Rosenberg et al., 2007; Bourne et al., 2009; Ainsworth et al., 2010; Garren and Azam, 2012). The field of marine microbial ecology underwent a revolution in the 1990s, when culture-independent molecular techniques revealed that bacterial diversity from culture-based assessments was largely underestimated (Azam, 1998). Studies of persistent associations between corals and bacteria, both beneficial and pathogenic, were enhanced by new methods and approaches from this revolution. Those techniques were adopted by coral microbiologists, resulting in

the discovery that particular components of bacterial communities are specific to some coral host species (Rohwer et al., 2002).

The cost and time associated with characterizing these complex bacterial assemblages initially posed a challenge to scientists attempting to identify patterns of diversity across a large scale. However, the gradually decreasing cost and increasing efficiency of high-throughput methods, including 454 pyrosequencing technology, allowed researchers to perform community 16S rRNA gene profiling and metagenome sequencing in a broad range of coral specimens. Recent applications of 16S pyrosequencing in corals have produced hundreds of thousands of 16S sequences—in contrast to hundreds of sequences from cloning methods. Results from pyrosequencing-based studies provide evidence of the presence of "coral-specific" groups of bacterial ribotypes (Reis et al., 2009; Kvennefors et al., 2010; Sunagawa et al., 2010; Ceh et al., 2011). Experiments investigating the bacterial component of coral surface mucous layers suggest that the composition of bacterial communities in coral mucus is distinct from other surface-associated biofilms and is influenced by the physical and biochemical properties of the mucus (Barott et al., 2011; Sweet et al., 2011b). Although corals maintain specific groups of bacteria, variation among individuals of a coral species may occur according to location (Guppy and Bythell, 2006; Littman et al., 2009; Kvennefors et al., 2010; Ceh et al., 2011).

Bacterial communities are maintained in microhabitats within an individual coral host, spatially structured within chemical micro-niches, or compartments, in the skeleton, tissues, and surface mucous layer of corals (Rohwer et al., 2001, 2002; Daniels et al., 2011; Sweet et al., 2011a). This spatial microheterogeneity is similar to previously described trends in the speciation of the dinoflagellate *Symbiodinium* in branching acroporid corals (Rowan and Knowlton, 1995). With that in mind, new collection techniques and apparatuses have recently been developed to enable collection from specific compartments of the coral, with minimized contamination by bacteria from other compartments (Sweet et al., 2011a).

Recent research surveying bacterial communities in a large number of marine sponges suggests that bacteria detected in sponges can be classified in three categories (Schmitt et al., 2011): *core* (groups of bacteria that are shared across many sponges), *species-specific* (groups of bacteria that are specific to certain sponge hosts), and *variable* (groups of bacteria that are transiently associated with the host, probably due to passive attachment from seawater). The recent composition analyses of bacterial assemblages in corals indicate that a similar classification scheme can be applied to coral-associated bacteria. An interesting difference between corals and sponges is that while many sponges have been documented to transmit diverse, specific bacterial communities in their gametes or larvae (Schmitt et al., 2007; Sharp et al., 2007), most corals appear to acquire specific bacteria from the seawater each generation (Apprill et al., 2009; Sharp et al., 2010). The mechanisms by which corals selectively and specifically recruit their core and specific bacterial components are largely undescribed, but they

likely involve the physical properties and the chemical structure of the mucous layer, which is thought to be unique in specific coral species (Bythell and Wild, 2011). Bacteria that successfully colonize the mucus are, in turn, involved in cycling nutrients and organic compounds in corals and on the reefs, and the resident microbes have the potential to modulate the bacterial community structure in coral mucus and tissue.

Here we review recent advances in the study of the coral metaorganism and specifically address (1) the onset of coral-bacterial associations; (2) the functional roles that bacteria play in healthy corals; and (3) how bacteria influence coral reef response to environmental change. These new discoveries are the basis for a model of how coral-associated and reef-inhabiting microbiota influence ecosystem-level responses to global climate change.

Onset of Coral-Bacterial Associations

The Caribbean coral *Porites astreoides* has been shown to transmit a bacterial component to its offspring (Sharp et al., 2012). However, this seems to be an exception to the rule in scleractinian corals. In eight other coral species that have been examined (Apprill et al., 2009; Sharp et al., 2010), corals do not appear to inherit bacteria from parents; rather, bacterial colonization occurs in planula larvae or post-settlement stages. Many bacterial phylotypes detected in planulae and post-settlement stages of *P. astreoides* have also been documented in the adult (Wegley et al., 2007), suggesting that corals acquire specific bacterial phylotypes.

Exploration of bacterial communities in early life stages of corals has not only provided new information about bacterial infection in corals, but it has also simplified analysis of diversity and dynamics of bacterial communities in corals across spatiotemporal scales. In contrast to their adult counterparts, swimming planula larvae of most corals have not yet accumulated a high bacterial load from the surrounding environment or by feeding (Apprill et al., 2009; Sharp et al., 2010); as a result, it is more tractable to characterize and quantify the associated bacterial component in these larvae. Similar phylogenetic clades of bacteria were detected in 16S rRNA gene sequence clone libraries from multiple larval specimens of the Caribbean coral *Porites astreoides* (Sharp et al., 2012) and in the Pacific coral *Pocillopora meandrina* (Apprill et al., 2009), suggesting that some groups of bacteria are common across different coral species. A number of bacterial types have been commonly detected in multiple species of corals, but of particular interest are those belonging to the phylum α-proteobacteria (Apprill et al., 2009; Raina et al., 2009; Sharp et al., 2012). The α-proteobacteria (particularly the Roseobacteriales) are abundant in the oceans, often constituting a third of the bacterioplankton (Wagner-Dobler and Biebl, 2006). This same group of bacteria is also closely associated with phytoplankton, including the dinoflagellate coral endosymbiont *Symbiodinium* (Webster et al., 2004). Many of these bacteria, now classified as *Ruegeria* spp., were originally designated *Silicibacter*

spp. (Yi et al., 2007). It is unknown whether these bacteria play a functional role in corals, but their consistent detection in early life stages of corals and in seawater during coral spawning may be an indication that they are significant to the health of larvae, or even to adult colonies (Apprill et al., 2009; Apprill and Rappe, 2011; Sharp et al., 2012).

New research focusing on the molecular basis of bacterial colonization of the coral tissues or surface mucous layer indicates that coral mucous biofilm communities are a result of selection processes driven by the coral holobiont rather than by incidental attachment by bacteria in the seawater (Sweet et al., 2011b). This is consistent with recent findings from studies in the cnidarians *Hydra*, in which researchers found that the composition of the surface-associated bacterial community is driven directly by host metabolism and production of compounds in the surface layer of *Hydra* (Augustin et al., 2010). It is likely that there are specific molecules that influence colonization in the coral mucous layer. Lectin-mediated uptake of *Symbiodinium* has been demonstrated in corals (Wood-Charlson et al., 2006), but very little is known about bacterial uptake or invasion in corals.

Functional immunological molecules with bacterial binding capacity have been found in corals, describing a means by which the host may control associated microbial composition (Kvennefors et al., 2008; Kvennefors and Roff, 2009). Molecules that control the activities of other coral-associated microbes are thought to be derived from the coral host and in some cases from the associated bacteria (Ritchie, 2006; Teplitski and Ritchie, 2009; Vidal-Dupiol et al., 2011a,b). As previously described in a broad range of other animal-microbe systems (McFall-Ngai et al., 2012), molecules that direct bacterial infection of animal tissue-associated bacteria may be conserved, regardless of whether the bacteria are beneficial, commensal, or pathogenic.

Role of Bacteria in Health of Coral and Coral Reefs

Recent coral microbiology research has described how bacterial communities contribute to the overall physiology and ecology of apparently healthy corals. These discoveries were made possible both by new molecular technologies and by novel fieldwork-based approaches. Bacteria within corals govern the biogeochemical cycling within coral tissues. In addition, bacteria on surfaces in the reef environment influence and facilitate settlement of coral larval, and resident microbes in corals play a role in defining the composition of the bacterial community in corals.

Studies over the past several years indicate that coral-associated bacteria influence biogeochemical cycling within corals and on reefs. Metagenomic data from the bacterial fraction of DNA from the coral *Porites astreoides* indicate the presence of numerous genes capable of degrading diverse aromatic compounds (Wegley et al., 2007). Coral-associated bacteria have been shown to be involved in cycling mucous-derived particulate and dissolved organic compounds in the

reef environment (Wild et al., 2004, 2009; Huettel et al., 2006). In addition, the bacterial metagenome of *P. astreoides* consists of genes encoding enzymes involved in cycling nitrogen via nitrogen fixation, ammonification, nitrification, and denitrification (Wegley et al., 2007). The detection of bacterial nitrogen fixation genes is consistent with previous biochemical research in which cyanobacterial nitrogen fixation was detected (Lesser et al., 2007). Further research focusing on *nifH* gene diversity in two species of *Montipora* (Olson et al., 2009) suggests that nitrogen-fixing bacteria in corals are not limited to cyanobacteria but also belong to taxa representing the α-, β-, γ-, and δ-proteobacterial classes (Olson et al., 2009). Bacteria have been shown to be significant players in transforming nitrogen (Fiore et al., 2010) as well as sulfur and carbon compounds (Ferrier-Pages et al., 2001; Raina et al., 2009; Kimes et al., 2010) in corals and on coral reefs.

Bacteria outside of the coral animal also exert influence on the behavior of corals during their early life stages. Particular species of crustose coralline algae (CCAs) have been shown to facilitate larval settlement of the threatened coral species *Acropora cervicornis* and *A. palmata* in the Florida Keys and the Caribbean (Ritson-Williams et al., 2010). The integration of microbiological and chemical ecology approaches suggests that the facilitation of larval settlement by CCAs may be regulated by bacteria growing in biofilms on the surface of CCAs (Negri et al., 2001; Webster et al., 2004; Tebben et al., 2011). To date, all of the CCA-associated bacteria implicated in inducing coral metamorphosis and settlement belong to the γ-proteobacteria. A strain of the γ-proteobacterium *Pseudoalteromonas* sp. isolated from the surface of the CCA species *Hydrolithon onkodes* induces significant levels of larval metamorphosis in the corals *Acropora willisae* and *A. millepora* in laboratory experiments (Negri et al., 2001). Researchers have recently shown that exposure to *Pseudoalteromonas* isolates cultured from *Negoniolithon fosliei* and *Hydrolithon onkodes* significantly increases rates of metamorphosis on the Pacific coral *Acropora millepora* (Tebben et al., 2011). Bioassay-guided isolation identified the inductive molecule as tetrabromopyrrole (Tebben et al., 2011). Other strains of *Pseudoalteromonas* and *Thalassomonas* have also been shown to induce larval settlement and metamorphosis in the coral *Pocillopora damicornis* (Tran and Hadfield, 2011). Not all tested isolates of *Pseudoalteromonas* and *Thalassomonas* were inductive in that study, indicating that the ability to induce settlement is taxon-specific. In addition, the isolation source of the bacteria (algal surface vs. coral surface) was not linked to the strains' inductive properties (Tran and Hadfield, 2011). Together, these studies indicate that coral recruitment and successful larval attachment and metamorphosis (which is crucial for continued repopulation of coral reef ecosystems) is strongly governed by the activity of specific bacteria in reef environments.

Recent research has focused on the role of bacteria native to the coral surface mucous layer that control bacterial colonization within the mucus, ultimately regulating resistance to disease. Corals have been shown to protect themselves

against pathogen infection *via* the presence of allelopathic properties in the mucus (Geffen and Rosenberg, 2005; Ritchie, 2006) or the coral tissue (Koh, 1997; Kelman et al., 2006; Gochfeld and Aeby, 2008). However, antimicrobial assays with numerous Red Sea corals reveal that the capabilities of coral species for antibiotic production are highly variable (Kelman et al., 2006). Bacteria isolated from corals are able to inhibit the colonization and growth of many other types of bacteria, including potentially invasive coral pathogens (Reshef et al., 2006; Ritchie, 2006; Wegley et al., 2007; Gochfeld and Aeby, 2008; Nissimov et al., 2009; Shnit-Orland and Kushmaro, 2009; Sharon and Rosenberg, 2010; Kvennefors et al., 2012). In addition, the presence of a high number of genes involved in antibacterial compound biosynthesis have been detected in metagenomes from multiple corals (Wegley et al., 2007; Thurber et al., 2009). It is not clear to what extent these bacteria and the metabolites they produce play a role in community structure. In situ antibiotic production by bacteria is known to be a means of securing a niche by controlling microbial populations competing for the same resources (Nielsen et al., 2000; Rao et al., 2005). It is therefore likely that bacteria in and on the coral host govern the dynamics of coral microbiota.

Although the mechanisms by which mucous-associated bacteria prevent pathogenic infection are still unknown, the data indicate that a sophisticated system of bacterial cell-cell chemical signaling known as quorum sensing (QS) may be involved in microbial pathogenesis in corals. QS is modulated by small diffusible compounds called autoinducers, which are molecules that, when accumulated to a threshold concentration within a diffusion-limited environment, result in synchronized group behaviors. This density-dependent regulation allows bacterial populations to act in unison, effectively magnifying their ecological impact. Though the cell-cell communication systems differ among bacterial species, QS has been demonstrated to regulate many bacterial behaviors, including biofilm formation, antibiotic production, bioluminescence, and pathogenesis (Ng and Bassler, 2009), and it commonly drives important interactions between bacterial communities and their hosts (Rasmussen and Givskov, 2006; Dobretsov et al., 2009).

Quorum sensing in bacterial pathogens is the mechanism by which virulence genes are expressed relative to pathogen density in the host, thereby initiating a coordinated attack once bacterial cell numbers reach a critical mass (Dobretsovet al., 2009). Both eukaryotes and prokaryotes have evolved to recognize and counter QS in pathogens, and there is evidence that eukaryotic signal-mimics can stimulate QS responses in bacteria (Teplitski et al., 2011). Other bacteria can counter-attack by producing quorum-quenching acylases or lactonases that break down signaling molecules (Teplitski et al., 2011). In addition to the signal-degrading enzymes, eukaryotes can inhibit or activate bacterial QS by producing compounds that mimic QS signals. For example, Rajamani et al. (2008) demonstrated that lumichrome, a derivative of the vitamin riboflavin that

is produced by the unicellular alga *Chlamydomonas reinhardtii* (as well as other prokaryotes and eukaryotes) can interact with the bacterial receptor for QS signals and elicit QS responses.

Quorum sensing may inhibit or activate pathogenesis, antibiotic production, exoenzyme production, and attachment by beneficial bacteria within coral tissues and on surfaces. Coral extracts contain compounds capable of interfering with QS activities (Skindersoe et al., 2008; Alagely et al., 2011) that may be involved in regulating the colonization of coral mucus by pathogens, commensal bacteria, or beneficial bacteria. The source of this activity is difficult to pinpoint and could originate from the coral, the dominant endosymbiont, or any associated bacteria. Alagely et al. (2011) recently showed that both coral- and *Symbiodinium*-associated bacteria alter swarming and biofilm formation in the coral pathogen *Serratia marcescens*. These phenotypes are typically controlled by QS, although inhibition of QS by these isolates remains to be demonstrated. There are few studies on the *in situ* roles of QS in corals, but this process is likely to be used in both pathogenesis and mutualistic interactions (Krediet et al., 2009a,b; Teplitski and Ritchie, 2009; Tait et al., 2010). While it is clear that at least some coral-associated commensals and pathogens produce QS signals under laboratory conditions (Tait et al., 2010; Alagely et al., 2011), it is not clear whether these signals accumulate to threshold concentrations in natural environments.

It is feasible that *Symbiodinium* spp. also produce signaling molecules that control bacterial cell-cell communication, which would influence the specific complement of bacteria that associate with corals. Perhaps bacterial species-specificity in corals is, in part, driven by *Symbiodinium* within the coral, but this has yet to be tested. The potential for *Symbiodinium* to be a source of antibacterial compounds in corals represents an aspect of bioactive compound production that is not yet described. It is likely that the source of antibacterial activity in corals is a combination of allelopathic chemicals produced by the coral, by associated bacteria, or by endosymbiotic dinoflagellates. In a study conducted by Marquis et al. (2005), eggs from 11 coral species were tested for antibacterial activity, and the only species exhibiting antibiotic activity was the one coral species in the study that incorporates *Symbiodinium* into the egg before the egg is released, suggesting a potential allelopathic contribution of *Symbiodinium*. It is also possible that coral-associated bioactive compounds are derived from bacteria whose presence or activity is influenced by *Symbiodinium*, but this has yet to be tested.

Role of Bacteria in Reef Ecosystem Responses to Environmental Change

The latest research on how coral-associated bacterial communities mediate responses of corals and coral reef ecosystems to environmental change addresses shifts in both the phylogenetic structure and metabolic capabilities of bacterial assemblages in corals. Multiple approaches and tools from microbiology, molecular biology, microscopy, and chemical ecology have been used to identify the role

of bacterial communities in response to threats such as increased sea-surface temperature, increased organic carbon and nutrient levels in seawater, increased macroalgal and cyanobacterial cover on reefs, and decreased seawater pH.

Rising sea-surface temperatures are linked to increases in coral diseases worldwide. However, the study of microbial coral diseases has been challenging due to many factors including microbial dynamics in the marine environment, the complications of proving unequivocal disease causation, and insufficient diagnostic tools (Pollock et al., 2011; Weil and Rogers, 2011). Some bacteria identified as coral pathogens include *Serratia marcescens* (Sutherland et al., 2011), *Aurantimonas coralicida* (Denner et al., 2003), and a consortium of bacterial and cyanobacteria phylotypes that make up what is known as Black Band Disease (Sekar et al., 2008). The most common bacteria present and problematic for corals are members of the Vibrionaceae that have been implicated in coral bleaching (Kushmaro et al., 1997; Ben-Haim and Rosenberg, 2002) and a myriad of coral diseases (Patterson et al., 2002; Frias-Lopez et al., 2003, 2004; Kline et al., 2006; Cervino et al., 2008). The Vibrionaceae are a common but diverse group of heterotrophic marine bacteria, collectively referred to as vibrios. Vibrios have been shown to be present in higher abundance on coral surfaces before obvious signs of distress (Ritchie, 2006; Mao-Jones et al., 2010). This group includes human pathogens and benign planktonic and animal-associated marine bacteria. Bleaching of the scleractinian coral *Oculina patagonica* in the eastern Mediterranean Sea was shown to be caused by *Vibrio shiloi* (Kushmaro et al., 1997). *Vibrio coralliilyticus* was isolated from bleached corals of the genus *Pocillopora damicornis* and shown to cause coral bleaching and tissue sloughing (Ben-Haim and Rosenberg, 2002). In these pathogens, toxin production and the ability to infect coral tissue have a strong temperature dependence (Kushmaro et al., 1997; Ben-Haim and Rosenberg, 2002). *Vibrio* dynamics are affected by water temperature and salinity, yet little else is known about environmental drivers of their abundance and distribution in the marine environment (Johnson et al., 2010). These organisms are often cultured rapidly and are able to utilize a wide range of carbon sources, suggesting that the biogeochemical significance of vibrios may vary with the nutrient state of the environment (Thompson et al., 2004). Some reef organisms are thought to be vectors for coral disease agents, specifically vibrios. These include organisms that come into contact with, or feed on, corals such as fireworms, snails, and corallivorous fishes (Weil and Rogers, 2011). Several recent reviews offer a comprehensive summary of the occurrence and possible environmental determinants of coral diseases (Rosenberg et al., 2009; Pollock et al., 2011; Weil and Rogers, 2011). Research on processes governing pathogen dynamics, abundance, and pathogenesis has informed us on coral defense mechanisms.

The coral surface mucous layer and its resident microbes appear to be significant in defending corals from microbial diseases. Mucus harvested from the coral *Acropora palmata* during a period of increased seawater temperatures does

not exhibit significant antibiotic activity compared to mucus sampled at lower temperatures (Ritchie, 2006). This suggests that the protective capacity of some corals may be lost when temperatures increase, providing a mechanism to explain how increased temperatures lower coral resistance and increase susceptibility to diseases. In addition, when temperatures increase, the dominant bacterial flora in coral mucus shifts from antibiotic-producing bacteria to pathogens (Ritchie, 2006). This finding indicates that a balance of potentially beneficial microbes may be important for the overall physiological health of reef corals. Rising sea-surface temperatures can cause a breakdown of coral-*Symbiodinium* symbiosis. In addition, shifting seawater temperatures can simultaneously affect interactions among other microbes, particularly bacteria present in or on the coral, rendering the host susceptible to opportunistic or secondary infection by certain bacteria (Ritchie, 2006; Lesser et al., 2007). Research on the Pacific coral *Acropora millepora* indicates that after bleaching (the loss of *Symbiodinium*) there is a dramatic shift to a *Vibrio*-dominated community (Bourne et al., 2007), but it is unclear whether the bacterial communities are responding to the absence of the *Symbiodinium*, to physiological changes in the coral host, or to the increased light and sea-surface temperature. Following bleaching-induced coral mortality, nitrogen-fixing bacteria increase in abundance on coral skeletons (Holmes and Johnstone, 2010). The resulting increase in available nitrogen in the seawater has the potential to affect the growth of macroalgae and other nitrogen-limited primary producers, including benthic cyanobacteria (Holmes and Johnstone, 2010). Taken together, these results demonstrate that temperature stress and coral bleaching have the potential to alter the composition and metabolism of coral-associated bacterial assemblages, with significant impacts on the health of corals and coral reef communities.

As a result of heightened fishing pressure, decline in herbivore populations, and increased nutrient levels, reefs are undergoing a "phase shift" from coral-dominated ecosystems to algal-dominated ecosystems (Pandolfi et al., 2003). Overgrowth by turf macroalgae and benthic cyanobacteria has been documented on adult coral colonies on reefs (Ritson-Williams et al., 2005). Concern is growing for how this shift in ecosystems affects bacterial communities within coral reefs (Dinsdale et al., 2008). Recent research demonstrates that allelochemicals from macroalgae and benthic cyanobacteria have the potential to mediate shifts in abundance and community composition of microbiota associated with adult corals (Morrow et al., 2011). When tested against a library of strains isolated from algal surfaces, from mucus of the Caribbean corals *Montastraea faveolata* and *Porites astreoides* in direct contact with algal surfaces, and from the mucus without direct contact of algae, chemical extracts from six species of macroalgae and two species of benthic cyanobacteria stimulated the growth of some strains but inhibited the growth of other strains (Morrow et al., 2011). While some of the algal extracts had broad-spectrum activity against the collection of test isolates from phylogenetically diverse environmental bacteria, other extracts specifically

increased the growth rates of the bacterial genus *Vibrio* (Morrow et al., 2011). Many of the active compounds in the study were hydrophilic, indicating that the bioactive compounds from algae or cyanobacteria may be readily solubilized and transported throughout seawater, providing a potential mechanism for algae to regulate microbial activity without direct contact, especially in low-flow benthic systems (Morrow et al., 2011). Allelopathic interactions among algae and corals have been shown to have detrimental effects on coral larval behavior, recruitment, and survival (Kuffner and Paul, 2004; Kuffner et al., 2006; Ritson-Williams et al., 2009). It is unknown how the bioactive compounds influence health of the early life stages, but it is feasible that the observed effects are linked to shifting bacterial communities associated with the coral planulae and recruits.

Smith et al. (2006) explored the effects of macroalgae on bacterial growth in the coral surface mucopolysaccharide layer. The results of that research, together with prior work on controlled exposure of coral fragments to seawater with increased dissolved organic carbon (DOC) levels (Kline et al., 2006), suggest that an excess of DOC, exuded from macroalgae, leads to coral mortality (Smith et al., 2006). In addition, Barott et al. (2011) found that the community composition of bacteria on surfaces of multiple reef macroalgal species is distinct from those found on coral surface mucous layers.

On the basis of these studies, it is clear that macroalgae have the potential to act as reservoirs of specific bacteria (beneficial, commensal, or pathogenic) not usually native to the coral mucous layer. Macroalgae also release compounds into the surrounding seawater that can have direct inhibitory or stimulatory effects on the coral-associated microbiota and, hence, on the health of the coral host.

Ocean acidification is a major concern for marine ecosystems in general—particularly those dependent on calcifying organisms, as secretion of calcium carbonate skeletons depends directly on carbonate saturation state in seawater (Caldeira et al., 2007). Recent research suggests that a decrease in seawater pH can alter marine bacterial communities, but very little is known about the large-scale impacts of those changes (Joint et al., 2011). Laboratory manipulations of seawater pH have shown that acidification can result in loss of *Symbiodinium* endosymbionts, decrease in calcification, depression of overall net productivity in corals (Anthony et al., 2008), and dissolution or slowed deposition of coral skeletons (Fine and Tchernov, 2007). In addition, decreased seawater pH levels have been attributed to a decline in overall abundance of crustose coralline algae (Kuffner et al., 2008), some of which have been shown to facilitate coral recruitment in reefs (Ritson-Williams et al., 2010). Experiments demonstrate that lower PCO_2 levels in seawater result in significant detrimental effects on early life stages of the coral *Porites astreoides*, including fertilization success, larval settlement rates, post-settlement growth, and post-settlement skeleton deposition (Albright et al., 2008, 2010).

Several laboratory-based studies have focused specifically on the impacts of ocean acidification on coral microbiota. Meron et al. (2011) explored shifts in

microbial assemblages associated with the coral *Acropora eurystoma* exposed to ambient seawater and seawater with pH 7.3 over a period of 2 mon using denaturing gradient gel electrophoresis profiles and 16S rRNA gene clone libraries. According to the resulting cluster analysis, a decrease in pH results in an increase in detection of Rhodobacteraceae and a decrease in detection of Bacteroidetes and Deltaproteobacteria (Meron et al., 2011). Relative to libraries from corals exposed to ambient seawater, clone libraries from *A. eurystoma* exposed to pH 7.3 conditions exhibited a higher percentage of clones representing bacteria closely related to those detected in stressed, injured, or diseased invertebrates (Meron et al., 2011). In another study with the Pacific coral *Porites compressa*, individuals exposed to an extremely low pH (6.7) exhibited shifts in bacterial community diversity (Thurber et al., 2009). Though the mechanism by which this occurs is not yet clear, it has been suggested that the altered seawater pH indirectly causes a shift in the bacterial diversity by impacting host metabolism, which results in a shift of nutrients and carbon available to the associated microbiota (Meron et al., 2011).

Metagenomic analysis of *P. compressa* mucus revealed potential functional shifts in the associated microbiota as a result of decreased pH and increased temperature (Thurber et al., 2009), most notably an increase in the number of detected genes for antibiotic and toxin production. Mucus from corals exposed to a decreased pH exhibits low antimicrobial activity (Meron et al., 2011), and mucus of *Acropora palmata* exhibits lower antibacterial activity after prolonged warm periods (Ritchie, 2006). Together, these results warn that even slight changes in seawater pH and temperature can have ecologically significant effects on coral-associated microbiota and, hence, on coral's susceptibility to bacterial pathogens. The shift in the coral microbiome phylogenetic profile has been proposed as a potential indicator for declining coral health before the corals exhibit more obvious signs of stress or disease (Thurber et al., 2009; Ainsworth et al., 2010; Garren and Azam, 2012).

A Model for Climate-Change-Induced Shifts in the Coral Metaorganism

The research reviewed here suggests that alterations in sea surface temperature, algal and cyanobacterial abundance on reefs, and seawater pH can have detrimental effects on corals by decreasing protective qualities of the coral mucous layer, *via* inhibition of growth or compound production in beneficial bacteria or by alteration of host-associated compound biosynthesis. Another aspect of coral-bacterial interactions that has garnered much attention is the ability of bacteria on reef substrates to influence successful larval recruitment. These surfaces include crustose coralline algae (CCAs), which are coated with microbial biofilms and are thought to be involved in mediating coral larval settlement (Webster et al., 2001, 2011; Ritson-Williams et al., 2009, 2010; Tebben et al., 2011).

Figure A11-1 represents the current model of corals and their interdependence on associated microbes. Both coral tissue and coral mucus contain abundant and diverse microbial communities (Figure A11-1a). When sea-surface temperatures increase, antibacterial compounds in the coral mucus disappear. Simultaneously, antibacterial-producing bacteria normally associated with healthy corals decrease while bacteria with pathogenic capabilities increase (Figure A11-1b).

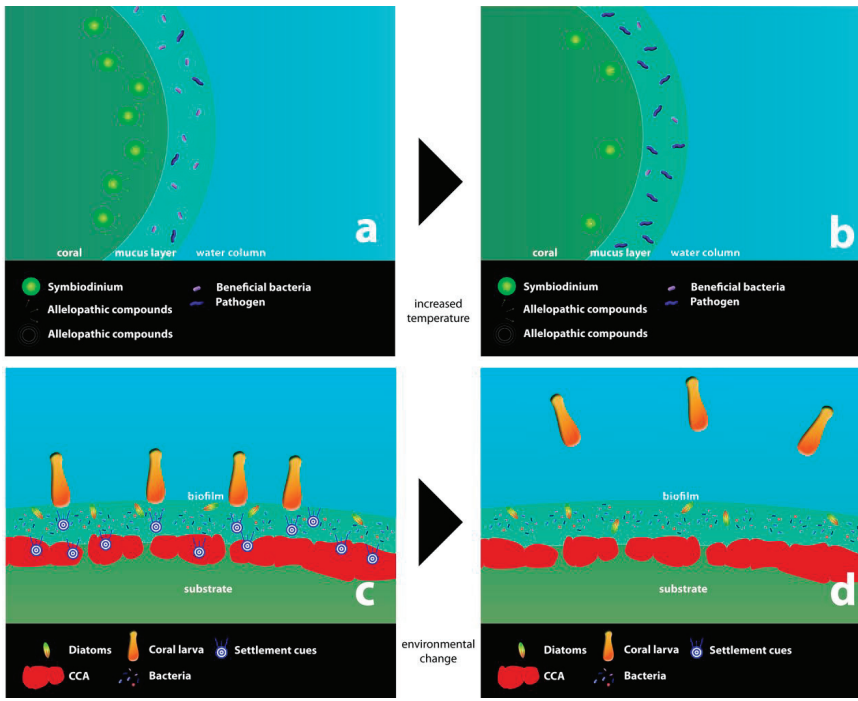


FIGURE A11-1 Schematic of coral surfaces and associated microbes. (a) Under normal conditions, the coral animal, associated endosymbiotic algae, or native bacteria may produce allelopathic compounds that regulate the abundance and activities of other microbes that come into contact with the coral. (b) Under conditions of coral stress (such as increased temperature or decreased pH), production of allelopathic compounds may be altered within the coral holobiont, either by affecting production by the coral host or by the associated microbes. Simultaneously, native beneficial bacteria are replaced by pathogenic bacteria on the coral surfaces. (c) Crustose coralline algae (CCA) and biofilm microbial communities facilitate attachment and settlement of coral larvae via inductive compounds (settlement cues) produced by the CCA or by recruiting specific bacteria that release these cues. (d) Certain types of environmental change (decreased pH, lower PCO_2, increased temperature) may alter the abundance of the inductive bacteria or their production of settlement cue compounds, resulting in reduction of attachment, metamorphosis, and successful settlement of coral larvae.

Mathematical modeling of this system suggests that once this shift to pathogen dominance is established, this state persists long after conditions return to those favorable for the reestablishment of beneficial microbes (Mao-Jones et al., 2010). Recent data from coral mucus bacterial metagenomes exposed to decreased pH (Thurber et al., 2009; Meron et al., 2011) indicate that ocean acidification may also result in a similar shift in the protective properties of coral mucus.

On the basis of this model and the data reviewed in this paper, we present a second model of coral-bacterial interactions in which environmental changes lead to shifts in bacterial communities on reef surfaces (Figure A11-1c and d). It has been shown that increased temperatures change the phylogenetic composition of CCA-associated bacterial communities and the success of larval recruitment (Webster et al., 2011). In addition, it was recently shown that decreased pH inhibits settlement of the coral *Porites astreoides* (Albright et al., 2008, 2010). Temperature may affect the growth, abundance, or bioactive metabolite biosynthesis of beneficial bacteria, particularly *Pseudoalteromonas* spp., on reef surfaces that are important for successful recruitment, which can ultimately result in a decline of new recruitment on reefs. Though the effects of decreased pH on surface biofilms have not been well described, this condition may alter the bacterial biofilm community and influence larval settlement success. Figure A11-1c and d shows a schematic model of reef surface-associated microbes before (c) and after (d) increased sea-surface temperature or ocean acidification. In ambient conditions on the reef, CCAs, or bacteria growing on CCA surfaces, produce compounds that facilitate larval settlement (Figure A11-1c). When sea-surface temperatures increase, bacterial communities on CCAs change, resulting in lower larval recruitment rates (Figure A11-1d). Similarly, as pH decreases, larval settlement decreases (Albright et al., 2008, 2010). It is hypothesized that the inductive properties of CCAs, whether they are due to compounds released by bacterial biofilms on CCAs or by the CCAs themselves, decrease (Figure A11-1d). As in the coral mucus (Figure A11-1a and b), there is a shift in the bacterial community of the reef surfaces. In this case, under increased sea-surface temperatures, the bacterial community dominated by inductive bacteria, such as *Pseudoalteromonas* and *Thalassomonas*, moves to a community dominated by bacteria that may not have inductive properties.

Next Questions: Microbe-Microbe Interactions in Corals

One of the next steps in increasing our understanding of coral fitness is a comprehensive characterization of coral-associated microbial interactions. For example, it is unclear if *Symbiodinium* plays a role in selectively recruiting bacteria to corals, if *Symbiodinium* affects bacterial physiology or secondary metabolite biosynthesis, or if bacterial metabolism influences *Symbiodinium* activity.

Little is known about the nature of free-living *Symbiodinium*, including what bacterial mutualisms may be present before coral acquisition of *Symbiodinium*,

in the case that the algal symbiont is not transmitted vertically. Members of the Roseobacteriales group are specifically present in association with *Symbiodinium* cultures and are able to increase *Symbiodinium* growth rates *in vivo* (Ritchie, 2011). This observed association between α-proteobacteria and dinoflagellates may be a true mutualism with benefits for both the bacteria and the algal host. The bacteria may benefit by having a readily available source of organic compounds such as dimethylsulfoniopropionate (DMSP), a preferred source of reduced sulfur (Miller and Belas, 2004; Raina et al., 2010). The algae may derive benefits from the bacterial production of antimicrobials such as tropodithietic acid (Geng and Belas, 2010) and bioactive compounds such as vitamin B-12 (Geng and Belas, 2010). A genomic comparison of the *Roseobacter* clade of α-proteobacteria indicates that some type of surface-associated lifestyle is central to the ecology of all members of the group (Slightom and Buchan, 2009).

Very little is known about how *Symbiodinium* affects bacterial communities in corals (or vice versa) or how these interactions impact the fitness of the coral host. Recent studies suggest that bacterial communities in juvenile corals differ significantly if they were initially colonized by different strains of *Symbiodinium* (Littman et al., 2009) with different photosynthetic efficiencies (Littman et al., 2010). It has been hypothesized that DMSP production by *Symbiodinium* plays a role in structuring bacterial communities in corals by attracting certain bacteria to the surface mucous layer of corals (Raina et al., 2009, 2010).

An important adaptive property of many α-proteobacteria is the presence of a bacterial system for diversity generation facilitated by gene transfer agents (GTAs) (Paul, 2008). GTAs are defective bacteriophages that are able to randomly package bacterial host DNA and transfer DNA to other α-proteobacteria (Paul, 2008). It has recently been shown that *Symbiodinium*-associated α-proteobacteria produce GTAs and are able to transfer genes to a range of bacteria in the marine environment (McDaniel et al., 2010). Furthermore, gene transfer via this mechanism is much higher in the coral reef environment than in other marine environments, suggesting an alternate mode of adaptation via swapping of potentially beneficial genes among marine bacteria (McDaniel et al., 2010) and possibly the coral holobiont.

A fundamental requirement of model systems is that they address interspecies interactions in a metaorganism. Research on host-microbe interactions can greatly benefit from a well-documented host-microbe study that spans the spectrum from pathogenicity to mutualism. Much work has been done on the basal metazoan *Hydra* to illustrate the value of a model systems approach (Weis et al., 2008; Bosch et al., 2009). Because *Hydra* is associated with a limited number of bacteria, it has provided valuable insight into the molecular basis of immunity and symbiosis in simple animals. Cnidarian and dinoflagellate models can also be used to elucidate roles of bacteria in both coral and *Symbiodinium* biology. Ideally, these models require cultured symbionts (bacterial and dinoflagellate) and an easily maintained cnidarian host (Weis et al., 2008). Our ability to culture many

of these bacterial symbionts will aid in exploring functions that are otherwise impossible to study due to the complex nature of the coral holobiont. Generation of genome sequence data from animal hosts and their associated microorganisms will exponentially enhance our basic understanding of symbiotic associations at the molecular level. This includes reconstruction of host-symbiont phylogenies, analysis of genes important in specific interactions, comparative genomics, and advanced technologies. The sea anemone *Aiptasia pallida* has recently been proposed as a model for coral biology for a number of reasons (Weis et al., 2008). While corals are difficult to grow in captivity, this species is hardy to laboratory manipulation and grows quickly in aquaria. Many protocols have been developed to manipulate *Symbiodinium* density in *A. pallida* without lethal effects on the host, and as a result, this organism has successfully been used to describe mechanisms of coral bleaching (Dunn et al., 2007) and disease (Alagely et al., 2011). *Aiptasia pallida* represents an opportunity to integrate a model systems approach with novel technologies from the "omics age" to learn more about multipartner interactions in corals in a moment of great environmental change.

Acknowledgments

This work was funded in part by the Mote Marine Laboratory Protect Our Reefs Grants Program and by the Dart Foundation. We thank Cathleen Sullivan (MML) for assistance with EndNote formatting, and two anonymous reviewers for improvements in the manuscript.

References

Ainsworth, T. D., R. V. Thurber, and R. D. Gates. 2010. The future of coral reefs: a microbial perspective. *Trends Ecol. Evol.* **25**: 233-240.

Alagely, A., C. J. Krediet, K. B. Ritchie, and M. Teplitski. 2011. Signaling-mediated cross-talk modulates swarming and biofilm formation in a coral pathogen *Serratia marcescens*. *ISME J.* **5**: 1609-1620.

Albright, R., B. Mason, and C. Langdon. 2008. Effect of aragonite saturation state on settlement and post-settlement growth of *Porites astreoides* larvae. *Coral Reefs* **27**: 485-490.

Albright, R., B. Mason, M. Miller, and C. Langdon. 2010. Ocean acidification compromises recruitment success of the threatened Caribbean coral *Acropora palmata*. *Proc. Natl. Acad. Sci. USA* **107**: 20400-20404.

Anthony, K. R. N., D. I. Kline, G. Diaz-Pulido, S. Dove, and O. Hoegh-Guldberg. 2008. Ocean acidification causes bleaching and productivity loss in coral reef builders. *Proc. Natl. Acad. Sci. USA* **105**: 17442-17446.

Anthony, K. R. N., J. A. Kleypas, and J.-P. Gattuso. 2011. Coral reefs modify their seawater carbon chemistry—implications for impacts of ocean acidification. *Glob. Change Biol.* **17**: 3655-3666.

Apprill, A., and M. S. Rappe. 2011. Response of the microbial community to coral spawning in lagoon and reef flat environments of Hawaii, USA. *Aquat. Microb. Ecol.* **62**: 251-266.

Apprill, A., H. Q. Marlow, M. Q. Martindale, and M. S. Rappe. 2009. The onset of microbial associations in the coral *Pocillopora meandrina*. *ISME J.* **3**: 685-699.

Augustin, R., S. Fraune, and T. C. Bosch. 2010. How *Hydra* senses and destroys microbes. *Semin. Immunol.* **22**: 54-58.

Azam, F. 1998. Microbial control of oceanic carbon flux: the plot thickens. *Science* **280**: 694-696.

Azam, F., and A. Z. Worden. 2004. Microbes, molecules, and marine ecosystems. *Science* **303**: 1622-1624.

Barott, K. L., B. Rodriguez-Brito, J. Janouskovec, K. L. Marhaver, J. E. Smith, P. Keeling, and F. L. Rohwer. 2011. Microbial diversity associated with four functional groups of benthic reef algae and the reef-building coral *Montastraea annularis*. *Environ. Microbiol.* **13**: 1192-1204.

Ben-Haim, Y., and E. Rosenberg. 2002. A novel *Vibrio* sp. pathogen of the coral *Pocillopora damicornis*. *Mar. Biol.* **141**: 47-55.

Bosch, T. C. G., and M. J. McFall-Ngai. 2011. Metaorganisms as the new frontier. *Zoology* **114**: 185-190.

Bosch, T. C., R. Augustin, F. Anton-Erxleben, S. Fraune, G. Hemmrich, H. Zill, P. Rosenstiel, G. Jacobs, S. Schreiber, M. Leippe *et al.* 2009. Uncovering the evolutionary history of innate immunity: the simple metazoan *Hydra* uses epithelial cells for host defence. *Dev. Comp. Immunol.* **33**: 559-569.

Bourne, D., Y. Iida, S. Uthicke, and C. Smith-Keune. 2007. Changes in coral-associated microbial communities during a bleaching event. *ISME J.* **2**: 350-363.

Bourne, D. G., M. Garren, T. M. Work, E. Rosenberg, G. W. Smith, and C. D. Harvell. 2009. Microbial disease and the coral holobiont. *Trends Microbiol.* **17**: 554-562.

Bythell, J. C., and C. Wild. 2011. Biology and ecology of coral mucus release. *J. Exp. Mar. Biol. Ecol.* **408**: 88-93.

Caldeira, K., D. Archer, J. P. Barry, R. G. J. Bellerby, P. G. Brewer, L. Cao, A. G. Dickson, S. C. Doney, H. Elderfield, V. J. Fabry *et al.* 2007. Comment on "Modern-age buildup of CO2 and its effects on seawater acidity and salinity" by Hugo A. Loaiciga. *Geophys. Res. Lett.* **34**: 10.1029/2006gl027288. L18608.

Ceh, J., M. van Keulen, and D. G. Bourne. 2011. Coral-associated bacterial communities on Ningaloo Reef, Western Australia. *FEMS Microbiol. Ecol.* **75**: 134-144.

Cervino, J. M., F. L. Thompson, B. Gomez-Gil, E. A. Lorence, T. J. Goreau, R. L. Hayes, K. B. Winiarski-Cervino, G. W. Smith, K. Hughen, and E. Bartels. 2008. The *Vibrio* core group induces yellow band disease in Caribbean and Indo-Pacific reef-building corals. *J. Appl. Microbiol.* **105**: 1658-1671.

Daniels, C. A., A. Zeifman, K. Heym, K. B. Ritchie, C. A. Watson, I. Berzins, and M. Breitbart. 2011. Spatial heterogeneity of bacterial communities in the mucus of *Montastraea annularis*. *Mar. Ecol. Prog. Ser.* **426**: 29-40.

Denner, E. B. M., G. W. Smith, H. J. Busse, P. Schumann, T. Narzt, S. W. Polson, W. Lubitz, and L. L. Richardson. 2003. *Aurantimonas coralicida* gen. nov., sp nov., the causative agent of white plague type II on Caribbean scleractinian corals. *Int. J. Syst. Evol. Microbiol.* **53**: 1115-1122.

Dinsdale, E. A., O. Pantos, S. Smriga, R. A. Edwards, F. Angly, L. Wegley, M. Hatay, D. Hall, E. Brown, M. Haynes *et al.* 2008. Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS One* **3**: e1584.

Dobretsov, S., M. Teplitski, and V. Paul. 2009. Mini-review: quorum sensing in the marine environment and its relationship to biofouling. *Biofouling* **25**: 413-427.

Dunn, S. R., C. E. Schnitzler, and V. M. Weis. 2007. Apoptosis and autophagy as mechanisms of dinoflagellate symbiont release during cnidarian bleaching: every which way you lose. *Proc. R. Soc. B Biol. Sci.* **274**: 3079-3085.

Ferrier-Pages, C., V. Schoelzke, J. Jaubert, L. Muscatine, and O. Hoegh-Guldberg. 2001. Response of a scleractinian coral, *Stylophora pistillata*, to iron and nitrate enrichment. *J. Exp. Mar. Biol. Ecol.* **259**: 249-261.

Fine, M., and D. Tchernov. 2007. Ocean acidification and scleractinian corals—Response. *Science* **317**: 1032-1033.

Fiore, C. L., J. K. Jarett, N. D. Olson, and M. P. Lesser. 2010. Nitrogen fixation and nitrogen transformations in marine symbioses. *Trends Microbiol.* **18**: 455-463.

Frias-Lopez, J., G. T. Bonheyo, Q. S. Jin, and B. W. Fouke. 2003. Cyanobacteria associated with coral black band disease in Caribbean and Indo-Pacific reefs. *Appl. Environ. Microbiol.* **69**: 2409-2413.

Frias-Lopez, J., J. S. Klaus, G. T. Bonheyo, and B. W. Fouke. 2004. Bacterial community associated with black band disease in corals. *Appl. Environ. Microbiol.* **70**: 5955-5962.

Garren, M., and F. Azam. 2012. New directions in coral reef microbial ecology. *Environ. Microbiol.* **14:** 833–844.

Geffen, Y., and E. Rosenberg. 2005. Stress-induced rapid release of antibacterials by scleractinian corals. *Mar. Biol.* **146**: 931-935.

Geng, H. F., and R. Belas. 2010. Molecular mechanisms underlying roseobacter-phytoplankton symbioses. *Curr. Opin. Biotechnol.* **21**: 332-338.

Gochfeld, D., and G. Aeby. 2008. Antibacterial chemical defenses in Hawaiian corals provide possible protection from disease. *Mar. Ecol. Prog. Ser.* **362**: 119-128.

Guppy, R., and J. C. Bythell. 2006. Environmental effects on bacterial diversity in the surface mucus layer of the reef coral *Montastraea faveolata*. *Mar. Ecol. Prog. Ser.* **328**: 133-142.

Holmes, G., and R. W. Johnstone. 2010. The role of coral mortality in nitrogen dynamics on coral reefs. *J. Exp. Mar. Biol. Ecol.* **387**: 1-8.

Huettel, M., C. Wild, and S. Gonelli. 2006. Mucus trap in coral reefs: formation and temporal evolution of particle aggregates caused by coral mucus. *Mar. Ecol. Prog. Ser.* **307**: 69-84.

Johnson, C. N., A. R. Flowers, N. F. Noriea III, A. M. Zimmerman, J. C. Bowers, A. DePaola, and D. J. Grimes. 2010. Relationships between environmental factors and pathogenic Vibrios in the northern Gulf of Mexico. *Appl. Environ. Microbiol.* **76**: 7076-7084.

Joint, I., S. C. Doney, and D. M. Karl. 2011. Will ocean acidification affect marine microbes? *ISME J.* **5**: 1-7.

Kelman, D., Y. Kashman, E. Rosenberg, A. Kushmaro, and Y. Loya. 2006. Antimicrobial activity of Red Sea corals. *Mar. Biol.* **149**: 357-363.

Kimes, N. E., J. D. Van Nostrand, E. Weil, J. Z. Zhou, and P. J. Morris. 2010. Microbial functional structure of *Montastraea faveolata*, an important Caribbean reef-building coral, differs between healthy and yellow-band diseased colonies. *Environ. Microbiol.* **12**: 541-556.

Kline, D. I., N. M. Kuntz, M. Breitbart, N. Knowlton, and F. Rohwer. 2006. Role of elevated organic carbon levels and microbial activity in coral mortality. *Mar. Ecol. Prog. Ser.* **314**: 119-125.

Knowlton, N., and F. Rohwer. 2003. Multispecies microbial mutualisms on coral reefs: the host as a habitat. *Am. Nat.* **162**: S51-S62.

Koh, E. G. L. 1997. Do scleractinian corals engage in chemical warfare against microbes? *J. Chem. Ecol.* **23**: 379-398.

Krediet, C. J., K. B. Ritchie, M. Cohen, E. K. Lipp, K. P. Sutherland, and M. Teplitski. 2009a. Utilization of mucus from the coral *Acropora palmata* by the pathogen *Serratia marcescens* and by environmental and coral commensal bacteria. *Appl. Environ. Microbiol* **75**: 3851-3858.

Krediet, C. J., K. B. Ritchie, and M. Teplitski. 2009b. Catabolite regulation of enzymatic activities in a white pox pathogen and commensal bacteria during growth on mucus polymers from the coral *Acropora palmata*. *Dis. Aquat. Org.* **87**: 57-66.

Kuffner, I. B., and V. J. Paul. 2004. Effects of the benthic cyanobacterium *Lyngbya majuscula* on larval recruitment of the reef corals *Acropora surculosa* and *Pocillopora damicornis*. *Coral Reefs* **23**: 455-458.

Kuffner, I. B., L. J. Walters, M. A. Becerro, V. J. Paul, R. Ritson-Williams, and K. S. Beach. 2006. Inhibition of coral recruitment by macroalgae and cyanobacteria. *Mar. Ecol. Prog. Ser.* **323**: 107-117.

Kuffner, I. B., A. J. Andersson, P. L. Jokiel, K. u. S. Rodgers, and F. T. Mackenzie. 2008. Decreased abundance of crustose coralline algae due to ocean acidification. *Nat. Geosci.* **1**: 114-117.

Kushmaro, A., E. Rosenberg, M. Fine, and Y. Loya. 1997. Bleaching of the coral *Oculina patagonica* by *Vibrio* AK-1. *Mar. Ecol. Prog. Ser.* **147**: 159-165.

Kvennefors, E. C. E., and G. Roff. 2009. Evidence of cyanobacterialike endosymbionts in Acroporid corals from the Great Barrier Reef. *Coral Reefs* **28**: 547-547.

Kvennefors, E. C. E., W. Leggat, O. Hoegh-Guldberg, B. M. Degnan, and A. C. Barnes. 2008. An ancient and variable mannose-binding lectin from the coral *Acropora millepora* binds both pathogens and symbionts. *Dev. Comp. Immunol.* **32**: 1582-1592.

Kvennefors, E. C. E., E. M. Sampayo, T. Ridgway, A. C. Barnes, and O. Hoegh-Guldberg. 2010. Bacterial communities of two ubiquitous Great Barrier Reef corals reveals both site- and species-specificity of common bacterial associates. *PLoS One* **5**: e10401.

Kvennefors, E., E. Sampayo, C. Kerr, G. Vieira, G. Roff, and A. Barnes. 2012. Regulation of bacterial communities through antimicrobial activity by the coral holobiont. *Microb. Ecol.* **63**: 605-618.

Lesser, M. P., L. I. Falcon, A. Rodriguez-Roman, S. Enriquez, O. Hoegh-Guldberg, and R. Iglesias-Prieto. 2007. Nitrogen fixation by symbiotic cyanobacteria provides a source of nitrogen for the scleractinian coral *Montastraea cavernosa*. *Mar. Ecol. Prog. Ser.* **346**: 143-152.

Littman, R. A., B. L. Willis, and D. G. Bourne. 2009. Bacterial communities of juvenile corals infected with different *Symbiodinium* (dinoflagellate) clades. *Mar. Ecol. Prog. Ser.* **389**: 45-59.

Littman, R. A., D. G. Bourne, and B. L. Willis. 2010. Responses of coral-associated bacterial communities to heat stress differ with *Symbiodinium* type on the same coral host. *Mol. Ecol.* **19**: 1978-1990.

Mao-Jones, J., K. B. Ritchie, L. E. Jones, and S. P. Ellner. 2010. How microbial community composition regulates coral disease development. *PLoS Biol.* **8**: e1000345.

Marquis, C. P., A. H. Baird, R. de Nys, C. Holmstrom, and N. Koziumi. 2005. An evaluation of the antimicrobial properties of the eggs of 11 species of scleractinian corals. *Coral Reefs* **24**: 248-253.

McDaniel, L. D., E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, and J. H. Paul. 2010. High frequency of horizontal gene transfer in the oceans. *Science* **330**: 50-50.

McFall-Ngai, M., E. A. Heath-Heckman, A. A. Gillette, S. M. Peyer, and E. A. Harvie. 2012. The secret languages of coevolved symbioses: insights from the *Euprymna scolopes-Vibrio fischeri* symbiosis. *Semin. Immunol.* **24:** 3-8.

Meron, D., E. Atias, L. Iasur Kruh, H. Elifantz, D. Minz, M. Fine, and E. Banin. 2011. The impact of reduced pH on the microbial community of the coral *Acropora eurystoma*. *ISME J.* **5**: 51-60.

Miller, T. R., and R. Belas. 2004. Dimethylsulfoniopropionate metabolism by *Pfiesteria*-associated *Roseobacter spp. Appl. Environ. Microbiol.* **70**: 3383-3391.

Morrow, K. M., V. J. Paul, M. R. Liles, and N. E. Chadwick. 2011. Allelochemicals produced by Caribbean macroalgae and cyanobacteria have species-specific effects on reef coral microorganisms. *Coral Reefs* **30**: 309-320.

Negri, A. P., N. Webster, R. T. Hill, and A. J. Heyward. 2001. Metamorphosis of broadcast spawning corals in response to bacteria isolated from crustose algae. *Mar. Ecol. Prog. Ser.* **223**: 121-131.

Ng, W.-L., and B. L. Bassler. 2009. Bacterial quorum-sensing network architectures. *Annu. Rev. Genet.* **43**: 197-222.

Nielsen, A. T., T. Tolker-Nielsen, K. B. Barken, and S. Molin. 2000. Role of commensal relationships on the spatial structure of a surface-attached microbial consortium. *Environ. Microbiol.* **2**: 59-68.

Nissimov, J., E. Rosenberg, and C. B. Munn. 2009. Antimicrobial properties of resident coral mucus bacteria of *Oculina patagonica*. *FEMS Microbiol. Lett.* **292**: 210-215.

Olson, N. D., T. D. Ainsworth, R. D. Gates, and M. Takabayashi. 2009. Diazotrophic bacteria associated with Hawaiian *Montipora* corals: diversity and abundance in correlation with symbiotic dinoflagellates. *J. Exp. Mar. Biol. Ecol.* **371**: 140-146.

Pandolfi, J. M., R. H. Bradbury, E. Sala, T. P. Hughes, K. A. Bjorndal, R. G. Cooke, D. McArdle, L. McCLenachan, M. J. H. Newman, G. Paredes, R. R. Warner, and J. B. C. Jackson. 2003. Global trajectories of the long-term decline of coral reef ecosystems. *Science* **301**: 955-958.

Patterson, K. L., J. W. Porter, K. E. Ritchie, S. W. Polson, E. Mueller, E. C. Peters, D. L. Santavy, and G. W. Smiths. 2002. The etiology of white pox, a lethal disease of the Caribbean elkhorn coral, *Acropora palmata*. *Proc. Natl. Acad. Sci. USA* **99**: 8725-8730.

Paul, J. H. 2008. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J.* **2**: 579-589.

Pollock, F. J., P. J. Morris, B. L. Willis, and D. G. Bourne. 2011. The urgent need for robust coral disease diagnostics. *PLoS Pathog.* **7**: e1002183.

Raina, J.-B., D. Tapiolas, B. L. Willis, and D. G. Bourne. 2009. Coral-associated bacteria and their role in the biogeochemical cycling of sulfur. *Appl. Environ. Microbiol* **75**: 3492-3501.

Raina, J. B., E. A. Dinsdale, B. L. Willis, and D. G. Bourne. 2010. Do the organic sulfur compounds DMSP and DMS drive coral microbial associations? *Trends Microbiol.* **18**: 101-108.

Rajamani, S., W. D. Bauer, J. B. Robinson, J. M. Farrow III, E. C. Pesci, M. Teplitski, M. Gao, R. T. Sayre, and D. A. Phillips. 2008. The vitamin riboflavin and its derivative lumichrome activate the LasR bacterial quorum-sensing receptor. *Mol. Plant Microbe Interact.* **21**: 1184-1192.

Rao, D., J. S. Webb, and S. Kjelleberg. 2005. Competitive interactions in mixed-species biofilms containing the marine bacterium *Pseudoalteromonas tunicata*. *Appl. Environ. Microbiol* **71**: 1729-1736.

Rasmussen, T. B., and M. Givskov. 2006. Quorum-sensing inhibitors as anti-pathogenic drugs. *Int. J. Med. Microbiol.* **296**: 149-161.

Reis, A. M., S. D. Araujo, Jr., R. L. Moura, R. B. Francini-Filho, G. Pappas, Jr., A. M. Coelho, R. H. Kruger, and F. L. Thompson. 2009. Bacterial diversity associated with the Brazilian endemic reef coral *Mussismilia braziliensis*. *J. Appl. Microbiol.* **106**: 1378-1387.

Reshef, L., O. Koren, Y. Loya, I. Zilber-Rosenberg, and E. Rosenberg. 2006. The coral probiotic hypothesis. *Environ. Microbiol.* **8**: 2068-2073.

Ritchie, K. B. 2006. Regulation of microbial populations by coral surface mucus and mucus-associated bacteria. *Mar. Ecol. Prog. Ser.* **322**: 1-14.

Ritchie, K. B. 2011. Bacterial symbionts of corals and *Symbiodinium*. Pp. 139-150 in *Beneficial Microorganisms in Multicellular Life Forms*, E. Rosenberg and U. Gophna, eds. Springer, Heidelberg.

Ritson-Williams, R., V. J. Paul, and V. Bonito. 2005. Marine benthic cyanobacteria overgrow coral reef organisms. *Coral Reefs* **24**: 629-629.

Ritson-Williams, R., S. N. Arnold, N. D. Fogarty, R. S. Steneck, M. J. A. Vermeij, and V. J. Paul. 2009. New perspectives on ecological mechanisms affecting coral recruitment on reefs. Pp. 437-457 in *Proceedings of the Smithsonian Marine Science Symposium: Smithsonian Contributions to the Marine Sciences*, M. A. Lang, I. G. Macintyre, and K. Ru¨tzler, eds. Smithsonian Institution Scholarly Press, Washington, D.C.

Ritson-Williams, R., V. J. Paul, S. N. Arnold, and R. S. Steneck. 2010. Larval settlement preferences and post-settlement survival of the threatened Caribbean corals *Acropora palmata* and *A. cervicornis*. *Coral Reefs* **29**: 71-81.

Rohwer, F. R., M. B. Breitbart, J. J. Jara, F. A. Azam, and N. K. Knowlton. 2001. Diversity of bacteria associated with the Caribbean coral *Montastraea franksi*. *Coral Reefs* **20**: 85-91.

Rohwer, F., V. Seguritan, F. Azam, and N. Knowlton. 2002. Diversity and distribution of coral-associated bacteria. *Mar. Ecol. Prog. Ser.* **243**: 1-10.

Rosenberg, E., O. Koren, L. Reshef, R. Efrony, and I. Zilber-Rosenberg. 2007. The role of microorganisms in coral health, disease and evolution. *Nat. Rev. Microbiol.* **5**: 355-362.

Rosenberg, E., A. Kushmaro, E. Kramarsky-Winter, E. Banin, and L. Yossi. 2009. The role of microorganisms in coral bleaching. *ISME J.* **3**: 139-146.

Rowan, R., and N. Knowlton. 1995. Intraspecific diversity and ecological zonation in coral-algal symbiosis. *Proc. Natl. Acad. Sci. USA* **92**: 2850-2853.

Schmitt, S., J. B. Weisz, N. Lindquist, and U. Hentschel. 2007. Vertical transmission of a phylogenetically complex microbial consortium in the viviparous sponge *Ircinia felix*. *Appl. Environ. Microbiol.* **73**: 2067-2078.

Schmitt, S., P. Tsai, J. Bell, J. Fromont, M. Ilan, N. Lindquist, T. Perez, A. Rodrigo, P. J. Schupp, J. Vacelet, N. Webster, U. Hentschel, and M. W. Taylor. 2011. Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME J.* **6**: 564-576.

Sekar, R., L. T. Kaczmarsky, and L. L. Richardson. 2008. Microbial community composition of black band disease on the coral host *Siderastrea siderea* from three regions of the wider Caribbean. *Mar. Ecol. Prog. Ser.* **362**: 85-98.

Sharon, G., and E. Rosenberg. 2010. Healthy corals maintain *Vibrio* in the VBNC state. *Environ. Microbiol. Rep.* **2**: 116-119.

Sharp, K. H., B. Eam, D. J. Faulkner, and M. G. Haygood. 2007. Vertical transmission of diverse microbes in the tropical sponge *Corticium* sp. *Appl. Environ. Microbiol.* **73**: 622-629.

Sharp, K. H., K. B. Ritchie, P. J. Schupp, R. Ritson-Williams, and V. J. Paul. 2010. Bacterial acquisition in juveniles of several broadcast spawning coral species. *PLoS One* **5**: e10898.

Sharp, K. H., D. Distel, and V. J. Paul. 2012. Diversity and dynamics of bacterial communities in early life stages of the Caribbean coral *Porites astreoides*. *ISME J.* **6**: 790–801.

Shnit-Orland, M., and A. Kushmaro. 2009. Coral mucus-associated bacteria: a possible first line of defense. *FEMS Microbiol. Ecol.* **67**: 371-380.

Skindersoe, M. E., P. Ettinger-Epstein, T. B. Rasmussen, T. Bjarnsholt, R. de Nys, and M. Givskov. 2008. Quorum sensing antagonism from marine organisms. *Mar. Biotechnol.* **10**: 56-63.

Slightom, R. N., and A. Buchan. 2009. Surface colonization by marine roseobacters: integrating genotype and phenotype. *Appl. Environ. Microbiol.* **75**: 6027-6037.

Smith, J. E., M. Shaw, R. A. Edwards, D. Obura, O. Pantos, E. Sala, S. A. Sandin, S. Smriga, M. Hatay, and F. L. Rohwer. 2006. Indirect effects of algae on coral: algae-mediated, microbe-induced coral mortality. *Ecol. Lett.* **9**: 835-845.

Stat, M., D. Carter, and O. Hoegh-Guldberg. 2006. The evolutionary history of *Symbiodinium* and scleractinian hosts—symbiosis, diversity, and the effect of climate change. *Perspect. Plant Ecol. Evol. Syst.* **8**: 23-43.

Sunagawa, S., C. M. Woodley, and M. Medina. 2010. Threatened corals provide underexplored microbial habitats. *PLoS One* **5**: e9554.

Sutherland, K. P., S. Shaban, J. L. Joyner, J. W. Porter, and E. K. Lipp. 2011. Human pathogen shown to cause disease in the threatened elkhorn coral *Acropora palmata*. *PLoS One* **6**: e23468.

Sweet, M. J., A. Croquer, and J. C. Bythell. 2011a. Bacterial assemblages differ between compartments within the coral holobiont. *Coral Reefs* **30**: 39-52.

Sweet, M. J., A. Croquer, and J. C. Bythell. 2011b. Development of bacterial biofilms on artificial corals in comparison to surface-associated microbes of hard corals. *PLoS One* **6**: e21195.

Tait, K., Z. Hutchison, F. L. Thompson, and C. B. Munn. 2010. Quorum sensing signal production and inhibition by coral-associated vibrios. *Environ. Microbiol. Rep.* **2**: 145-150.

Tebben, J., D. M. Tapiolas, C. A. Motti, D. Abrego, A. P. Negri, L. L. Blackall, P. D. Steinberg, and T. Harder. 2011. Induction of larval metamorphosis of the coral *Acropora millepora* by tetrabromopyrrole isolated from a *Pseudoalteromonas* bacterium. *PLoS One* **6**: e19082.

Teplitski, M., and K. Ritchie. 2009. How feasible is the biological control of coral diseases? *Trends Ecol. Evol.* **24**: 378-385.

Teplitski, M., K. Warriner, J. Bartz, and K. R. Schneider. 2011. Untangling metabolic and communication networks: interactions of enterics with phytobacteria and their implications in produce safety. *Trends Microbiol.* **19**: 121-127.

Thompson, J. R., M. A. Randa, L. A. Marcelino, A. Tomita-Mitchell, E. Lim, and M. F. Polz. 2004. Diversity and dynamics of a north Atlantic coastal *Vibrio* community. *Appl. Environ. Microbiol* **70**: 4103-4110.

Thurber, R. V., D. Willner-Hall, B. Rodriguez-Mueller, C. Desnues, R. A. Edwards, F. Angly, E. Dinsdale, L. Kelly, and F. Rohwer. 2009. Metagenomic analysis of stressed coral holobionts. *Environ. Microbiol.* **11**: 2148-2163.

Tran, C., and M. G. Hadfield. 2011. Larvae of *Pocillopora damicornis* (Anthozoa) settle and metamorphose in response to surface-biofilm bacteria. *Mar. Ecol. Prog. Ser.* **433**: 85-96.

Vidal-Dupiol, J., O. Ladriere, D. Destoumieux-Garzon, P. E. Sautiere, A. L. Meistertzheim, E. Tambutte, S. Tambutte, D. Duval, L. Foure, M. Adjeroud, and G. Mitta. 2011a. Innate immune responses of a scleractinian coral to vibriosis. *J. Biol. Chem.* **286**: 22688-22698.

Vidal-Dupiol, J., O. Ladriere, A. L. Meistertzheim, L. Foure, M. Adjeroud, and G. Mitta. 2011b. Physiological responses of the scleractinian coral *Pocillopora damicornis* to bacterial stress from *Vibrio coralliilyticus*. *J. Exp. Biol.* **214**: 1533-1545.

Wagner-Dobler, I., and H. Biebl. 2006. Environmental biology of the marine *Roseobacter* lineage. *Annu. Rev. Microbiol.* **60**: 255-280.

Webster, N. S., R. I. Webb, M. J. Ridd, R. T. Hill, and A. P. Negri. 2001. The effects of copper on the microbial community of a coral reef sponge. *Environ. Microbiol.* **3**: 19-31.

Webster, N. S., L. D. Smith, A. J. Heyward, J. E. M. Watts, R. I. Webb, L. L. Blackall, and A. P. Negri. 2004. Metamorphosis of a scleractinian coral in response to microbial biofilms. *Appl. Environ. Microbiol.* **70**: 1213-1221.

Webster, N. S., R. Soo, R. Cobb, and A. P. Negri. 2011. Elevated seawater temperature causes a microbial shift on crustose coralline algae with implications for the recruitment of coral larvae. *ISME J.* **5**: 759-770.

Wegley, L., R. Edwards, B. Rodriguez-Brito, H. Liu, and F. Rohwer. 2007. Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. *Environ. Microbiol.* **9**: 2707-2719.

Weil, E., and C. S. Rogers. 2011. Coral reef diseases in the Atlantic-Caribbean. Pp. 465-491 in *Coral Reefs: An Ecosystem in Transition*, Z. Dubinsky and N. Stambler, eds. Springer, Science Business Media, Dordrecht, The Netherlands.

Weis, V. M., S. K. Davy, O. Hoegh-Guldberg, M. Rodriguez-Lanetty, and J. R. Pringe. 2008. Cell biology in model systems as the key to understanding corals. *Trends Ecol. Evol.* **23**: 369-376.

Wild, C., M. Huettel, A. Klueter, S. G. Kremb, M. Y. M. Rasheed, and B. B. Jorgensen. 2004. Coral mucus functions as an energy carrier and particle trap in the reef ecosystem. *Nature* **428**: 66-70.

Wild, C., M. S. Naumann, A. Haas, U. Struck, F. W. Mayer, M. Y. Rasheed, and M. Huettel. 2009. Coral sand O2 uptake and pelagicbenthic coupling in a subtropical fringing reef, Aqaba, Red Sea. *Aquat. Biol.* **6**: 133-142.

Wood-Charlson, E. M., L. L. Hollingsworth, D. A. Krupp, and V. M. Weis. 2006. Lectin/glycan interactions play a role in recognition in a coral/dinoflagellate symbiosis. *Cell. Microbiol.* **8**: 1985-1993.

Yi, H., Y. W. Lim, and J. Chun. 2007. Taxonomic evaluation of the genera *Ruegeria* and *Silicibacter*: a proposal to transfer the genus *Silicibacter* Petursdottir and Kristjansson 1999 to the genus *Ruegeria* Uchino *et al.* 1999. *Int. J. Syst. Evol. Microbiol.* **57**: 815-819.

Zilber-Rosenberg, I., and E. Rosenberg. 2008. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS Microbiol. Rev.* **32**: 723-735.

A12

GENOMIC TRANSITION TO PATHOGENICITY IN CHYTRID FUNGI[46]

Suzanne Joneson,[47] *Jason E. Stahich*,[48] *Shin-Han Shiu*,[49] *and Erica Bree* ***Rosenblum***[47,*]

Abstract

Understanding the molecular mechanisms of pathogen emergence is central to mitigating the impacts of novel infectious disease agents. The chytrid fungus *Batrachochytrium dendrobatidis* (*Bd*) is an emerging pathogen of amphibians that has been implicated in amphibian declines worldwide. *Bd* is the only member of its clade known to attack vertebrates. However, little is known about the molecular determinants of—or evolutionary transition to—pathogenicity in *Bd*. Here we sequence the genome of *Bd*'s closest known relative—a non-pathogenic chytrid *Homolaphlyctis polyrhiza* (*Hp*). We first describe the genome of *Hp*, which is comparable to other chytrid genomes in size and number of predicted proteins. We then compare the genomes of *Hp*, *Bd*, and 19 additional fungal genomes to identify unique or recent evolutionary elements in the *Bd* genome. We identified 1,974 *Bd*-specific genes, a gene set that is enriched for protease, lipase, and microbial effector gene ontology terms. We describe significant lineage-specific expansions in three *Bd* protease families (metallo-, serine-type, and aspartyl proteases). We show that these protease gene family expansions occurred after the divergence

[46] Reprinted from *PLoS Pathogens*. Originally published as Joneson S, Stajich JE, Shiu S-H, Rosenblum EB (2011) Genomic Transition to Pathogenicity in Chytrid Fungi. *PLoS Pathogens* 7(11): e1002338. doi:10.1371/

journal.ppat.1002338.

Funding: We acknowledge NIH funding from the COBRE Program administered by the Initiative for Bioinformatics and Evolutionary Studies (P20RR0116454) and the INBRE Program of the National Center for Research Resources (P20RR016448) to E.B.R., Initial Complement funding provided by the University of California to J.E.S., and NSF funding (DBI-0939454) to the BEACON Center for the Study of Evolution in Action. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rosenblum@uidaho.edu

[47] Department of Biological Sciences, University of Idaho, Moscow, Idaho, USA.

[48] Department of Plant Pathology and Microbiology, University of California, Riverside, California, USA.

[49] Department of Plant Biology, Michigan State University, East Lansing, Michigan, USA.

of *Bd* and *Hp* from their common ancestor and thus are localized to the *Bd* branch. Finally, we demonstrate that the timing of the protease gene family expansions predates the emergence of *Bd* as a globally important amphibian pathogen.

Author Summary

The chytrid fungus *Batrachochytrium dendrobatidis* (*Bd*) is an emerging pathogen that has been implicated in decimating amphibian populations around the world. *Bd* is the only member of an ancient group of fungi (called the Chytridiomycota) that is known to attack vertebrates. The question of how an amphibian-killing fungus evolved from non-pathogenic ancestors is vital to protecting the world's remaining amphibians from *Bd*. We sequenced the genome of *Bd*'s closest known relative—a non-pathogenic chytrid named *Homolaphlyctis polyrhiza* (*Hp*). We compared the genomes of *Bd*, *Hp* and 18 additional fungi to identify what makes *Bd* unique. We identified a large number of *Bd*-specific genes, a gene set that contains a number of possible pathogenicity factors. In particular, we describe a large number of protease genes in the *Bd* genome and show that these genes were duplicated after the divergence of *Bd* and *Hp* from their common ancestor. Studying *Bd*'s pathogenesis in an evolutionary context provides new evidence for the role of protease genes in *Bd*'s ability to kill amphibians.

Introduction

Understanding the emergence of novel pathogens is a central challenge in epidemiology, disease ecology, and evolutionary biology. Emerging pathogens of humans, wildlife, and agriculturally important crops generally have a dynamic recent evolutionary past. For example, many emerging pathogens have become adapted to new environmental conditions, shifted their host range, and/or evolved more virulent forms (Hoskisson and Trevors, 2010; Smith and Guegan, 2010; Woolhouse and Gaunt, 2007). Identifying the genetic basis of these evolutionary shifts can lend insight into the mechanisms of pathogen emergence.

Studies of the amphibian-killing fungus *Batrachochytrium dendrobatidis* (*Bd*) provide an opportunity to better understand evolutionary transitions to pathogenicity. *Bd* is considered the leading cause of amphibian declines worldwide and is found on every continent where amphibians occur (Berger et al., 1998; Lips et al., 2006). *Bd* infects amphibian skin and the resulting disease, chytridiomycosis, is responsible for population declines and extirpations in hundreds of amphibian species (Lötters et al., 2004; Skerratt et al., 2007). *Bd* is the only documented vertebrate pathogen in a diverse, early-branching lineage of fungi called the Chytridiomycota. Some chytrids are pathogens of plants, but most chytrids are primarily known to survive on decaying organic material as saprobes (James et al., 2006). The question of how an amphibian-killing fungus evolved

from an ancestor that was not a vertebrate pathogen is vital to understanding and mitigating the chytridiomycosis epidemic and will also shed light on the evolution of novel pathogens more broadly.

Investigating the transition to pathogenicity in chytrid fungi requires an explicitly evolutionary perspective. Specifically, identifying elements of the genome that have undergone recent evolution in the branch leading to *Bd* may help us determine how *Bd* attacks its amphibian hosts. Previously we identified several families of proteases that may be involved in *Bd*'s ability to infect amphibian skin. Specifically, we found expanded gene families of metallo- and serine proteases in the *Bd* genome that exhibit life-stage specific gene expression patterns (Rosenblum et al., 2008). These proteases have been hypothesized to play a role in the ability of other fungal pathogens to invade and degrade host tissue (Burmester et al., 2011; da Silva et al., 2006; Monod, 2008; Monod et al., 2002). However, previous studies could not resolve if these gene family expansions occurred along the branch leading to *Bd* because the fungal genomes available for comparison were only distantly related to *Bd*.

To determine what unique features of the *Bd* genome might relate to its ability to colonize amphibian skin, we compared genomes of *Bd* and its closest known relative, *Homolaphlyctis polyrhiza* (*Hp*) (this isolate has been described by Joyce Longcore [pers. comm.] and has been referred to as "JEL142" in previous publications [James et al., 2006]). *Bd* and *Hp* are in the same Rhizophydiales order (Letcher et al., 2006), and *Bd* is the only member of this clade known to be a vertebrate pathogen (James et al., 2006). We first confirmed that *Hp* cannot survive on amphibian skin alone. We then sequenced and characterized the genome of *Hp* using Roche-454 pyrosequencing. Finally, we used a comparative genomics approach to identify differences between *Bd* and *Hp* using additional fungal species as outgroups. Based on identified unique elements of the *Bd* genome, we develop hypotheses for the mechanisms and evolution of *Bd* pathogenicity.

Materials and Methods

Taxon Sampling

Our focal isolates were the JAM81 strain of *Bd* and the JEL142 strain of *Hp*. JAM81 was isolated from *Rana muscosa* in the Sierra Nevada Mountains in California, where *Bd* has caused catastrophic declines in *R. muscosa* populations (Rachowicz et al., 2006). *Hp* was collected from leaf litter in Maine and is a presumed saprobe. We also used the information from publically available genomes of an additional *Bd* isolate—JEL423 (http://www.broadinstitute.org/annotation/genome/batrachochytrium_dendrobatidis/MultiHome.html), and an additional chytrid, *Spizellomyces punctatus*, a terrestrial saprobe (Origins of Multicellularity Sequencing Project, Broad Institute of Harvard and MIT [http://www.broadinstitute.org/]). Finally we used the genome information from 17 additional

publicly available fungal genomes (Table S1). We chose these outgroups to represent a broad phylogenetic survey of fungi that span four additional fungal phyla: Blastocladiomycota (*Allomyces macrogynus*), Zygomycota (*Phycomyces blakesleeanus*), Basidiomycota (*Coprinopsis cinerea*, *Cryptococcus neoformans*, *Puccinia graminis* f. sp. *tritici*, *Ustilago maydis*), and Ascomycota (*Arthroderma benhamiae*, *Aspergillus nidulans*, *Blastomyces dermatitidis*, *Botrytis cinerea*, *Coccidioides immitis*, *Fusarium graminearum*, *Microsporum canis*, *Neurospora crassa*, *Pyrenophora tritici-repentis, Trichophyton rubrum*, and *Uncinocarpus reesii*). *Arthroderma benhamiae*, *M. canis*, and *T. rubrum* were chosen in particular because they are dermatophytes (i.e., fungal pathogens that infect skin).

We reconstructed the phylogenetic relationships among the 19 taxa used in this study using Bayesian phylogenetic analyses of 51 single-copy genes. The alignment was comprised of 21,182 total trimmed amino acid residues. The orthologous sequences were aligned with T-Coffee (Notredame et al., 2000), concatenated, and trimmed with trimAl (Capella-Gutiérrez et al., 2009). The Basidiomycota phylum was constrained by members *Ustilago maydis* and *Puccinia graminis*, and the tree rooted with the Chytridiomycota clade based on James et al. (2006). Bayesian posterior probabilities are shown below internal nodes and ML bootstrap values from 100 replicates above the nodes.

Growth of Bd *and* Hp *on Amphibian Skin*

We grew *Bd* (JAM81) and *Hp* on the standard growth medium PmTG (made from peptonized milk, tryptone and glucose) (Barr, 1986). After one week of growth, we transferred 3.8×10^6 zoospores from each isolate to 3 mL of two liquid growth conditions: standard growth media and amphibian skin. For standard growth media we used 1% liquid PmTG, and for amphibian skin we used 10% w/v pulverized and autoclaved cane-toad skin in water. We established six technical replicates of each isolate in each condition. Liquid cultures were gently shaken in 6-well tissue culture plates. To test how long *Bd* and *Hp* survived in each growth condition, we tested an aliquot from each culture every day for 14 days. Each day we removed 15 µL from each of the technical replicates, pooled aliquots for each isolate in each treatment group, and inoculated PmTG-agar growth plates. We inspected growth plates every day using 200× magnification to visualize whether active zoospores were produced.

Hp *Genome Sequence, Assembly, and Annotation*

We grew *Hp* at room temperature (23–25C) in liquid PmTG medium with gentle agitation for approximately 2 weeks. We extracted *Hp* DNA using a Zolan and Pukkila (Zolan and Pukkila, 1986) protocol modified by the use of 2% sodium dodecyl sulphate as extraction buffer in place of CTAB. We sequenced the

Hp genome using a Roche 454 Genome Sequencer FLX with Titanium chemistry and standard Roche protocol. We screened and trimmed 1,100,797 reads of vector sequences and assembled them with Roche's GS De Novo Assembler. We improved the assembly by synteny-based alignment to the JAM81 genome sequence with Mercator (Dewey, 2007).

We annotated the *Hp* genome with predicted proteins using the MAKER annotation pipeline (Cantarel et al., 2008). MAKER predicts proteins based on homology with protein-coding sequences of other species, and with the consensus of the *ab initio* gene prediction algorithms GeneMark, AUGUSTUS, and SNAP. GeneMark is self-training so we simply applied it to determine *ab initio* parameters. We trained AUGUSTUS using parameters provided in the MAKER package and previously determined *Bd* training parameters. We trained SNAP by iteratively running MAKER with SNAP *Bd* models and then retraining on the most confident gene model parameters from the initial run. All parameters files are available in http://fungalgenomes.org/public/Hp_JEL142/. Because MAKER's final set of predicted proteins (referred to hereafter as "Hp_Maker") is a conservative estimate that relies upon the consensus of different prediction algorithms, we also used the set of *ab initio* predicted proteins in MAKER by GeneMark-ES (Ter-Hovhannisyan et al., 2008) as an upper limit (referred to hereafter as "Hp_GeneMark"). Hp_Maker is not a perfect subset of Hp_GeneMark, so we considered both datasets when characterizing the proteome of *Hp*. We annotated *Hp* protein models by comparison to the Pfam database of protein domains (Finn et al., 2010) using HMMER 3.0 (http://hmmer.org/).

We used two methods that rely on different algorithms to confirm that we successfully identified the majority of *Hp* proteins. First, we used the eukaryotic genome annotation pipeline CEGMA to predict the number of core eukaryotic genes in the *Hp* alignment (Parra et al., 2007). Second we determined the number of "chytrid-specific" orthologous groups that were present in the *Hp* genome. We defined chytrid-specific orthologous groups as those groups shared between all available Chytridiomycota genomes: two *Bd* isolates (JAM81 and JEL423) and one *Spizellomyces punctatus* isolate (DAOM BR117) (Table S1). We identified chytrid-specific orthologous groups using BLASTP (Altschul et al., 1990) and OrthoMCL (Li et al., 2003), and determined how many of these were also found within either set of *Hp* predicted proteins (i.e., Hp_Maker and Hp_GeneMark).

Bd *Unique Genomic Features*

We also used BLASTP and OrthoMCL to determine orthologous groups for all sampled taxa. These orthologous groups were used to determine "*Bd*-specific" genes which we defined as those groups or genes that were present in both sequenced *Bd* genomes (JAM81 and JEL423) but absent from all other sampled fungi. [Note that the *Bd*-specific gene set is distinct from the more broadly defined

chytrid-specific gene set discussed above]. We used GO::TermFinder (Boyle et al., 2004) to determine if the Pfam annotations for the set of *Bd*-specific genes showed enrichment for particular GO terms.

Bd *Gene Family Expansions*

We identified several gene family expansions in *Bd* through inspection of the top ten largest *Bd*-specific orthologous groups and inspection of enriched GO categories. We found gene family expansions in families with genes containing M36, S41, and Asp (both Asp and Asp_protease) protease Pfam signature domains (see Table S2 for sequences and their Pfam domain delimitation). We conducted an exhaustive search in the focal genomes for M36, S41, and Asp domains using HMMER3 (http://hmmer.org/). For *Hp* we conducted the HMMER3 search in both the MAKER and GENEMARK datasets. For S41 and Asp, the predicted proteins from Maker were subsets of those from GeneMark, so we only report GeneMark names. For M36 there were several Maker predicted proteins that were not included in the GeneMark set, so we report both Maker and GeneMark names. We then aligned the sequences of the protein domains for all members in each expanded family for the three Chytridiomycota genomes (*Bd*, *Hp*, and *Spizellomyces punctatus*) and one Blastocladiomycota outgroup (*Allomyces macrogynus*). We generated these alignments using the iterative alignment program MUSCLE (Edgar, 2004). After inspecting the alignments, we found that 8 M36 and 13 Asp protein sequences were missing >50% of their domain sequences. These partial sequences were likely mis-annotation or pseudogenes so we excluded them from further analysis (see Table S2B for identities of excluded partial sequences). After aligning the protein domain sequences of the remaining proteins (see Figure S1 for alignments), we reconstructed gene trees for each family using the Maximum Likelihood method implemented in RAxML (Stamatakis et al., 2005). We used the rapid bootstrap algorithm (400 replicates) with the Jones-Taylor-Thornton substitution matrix assuming a gamma model of rate heterogeneity. We report the Maximum Likelihood trees with the highest log likelihood score and bootstrap support values.

We calculated synonymous and non-synonymous substitution rates (*Ks* and *Ka*, respectively) with the yn00 program implemented in the PAML package (Yang, 2007) using full length annotated coding sequences. For each expanded protease gene family (containing M36, S41, and Asp domains) we calculated *Ks* and *Ka* of putative orthologs between all focal taxa pairs [i.e., chytrids (*Bd*, *Hp*, and *Spizellomyces punctatus*) and between all focal taxa and the outgroup (*Allomyces macrogynus*)]. We identified putative orthologs based on a cross-species reciprocal best match between any species pairs (Hanada et al., 2008). In addition, we used a second, more stringent approach that required sequence distances between reciprocal best matches to follow the relationships between the

four focal species. Because the rate distributions from these two approaches were similar, we only report results from the first approach. Because yn00 does not robustly correct for multiple substitutions (Yang and Bielawski, 2000), and because *Ks* values are large between our focal taxa, we use *Ks* values to make a general comparison (within versus between species) for rates of molecular evolution.

We made rough divergence time estimates for the duplication events in the three expanded protease gene families using "node-*Ks*" as a proxy of time. The node *Ks* is defined as follows: for each node N in the mid-point rooted phylogeny, its *Ks* is the averaged *Ks* values between all operational taxonomic unit pairs across the two lineages that originated from N. There are no empirical estimates of chytrid substitution rates, so we do not propose specific dates for the duplication events. However, we do use a rough approximation for a reasonable substitution rate (following previous molecular evolution studies in fungi [Lynch and Conery, 2000]) to test whether the timing of gene duplications was likely coincident with the emergence of *Bd* as a deadly amphibian pathogen.

Results

Taxon Sampling

The phylogenetic relationship among all 19 taxa in this study can be seen in Figure A12-1. As described above, we sampled genomes from across the diversity of five fungal phyla (i.e., Chytridiomycota, Blastocladiomycota, Zygomycota, Basidiomycota, Ascomycota). Our sampling scheme allowed us to determine, in a phylogenetic context, which elements of *Bd*'s genome are shared with *Hp* and other fungal taxa.

Growth of Bd *and* Hp *on Amphibian Skin*

Both *Bd* and *Hp* grew well in standard PmTG growth media and produced viable zoospores throughout the entire 14 day observation period. However, only *Bd* survived on frog skin alone. *Bd* produced viable zoospores in the cane-toad skin treatment throughout the entire observation period, and after 14 days of incubation the *Bd*—frog skin solution was cloudy with chytrid growth and degraded skin (Figure A12-2). Conversely, *Hp* did not survive and reproduce on cane-toad skin alone. We observed viable zoospores for *Hp* in the cane-toad skin treatment only for the first three days (these zoospores most likely persisted from the initial inoculation), and after 14 days of incubation the *Hp*—cane-toad skin solution remained clear of chytrid growth and the cane-toad skin remained intact and not further degraded (Figure A12-2). We did not observe the growth of any bacterial or fungal contaminants in any of the treatments.

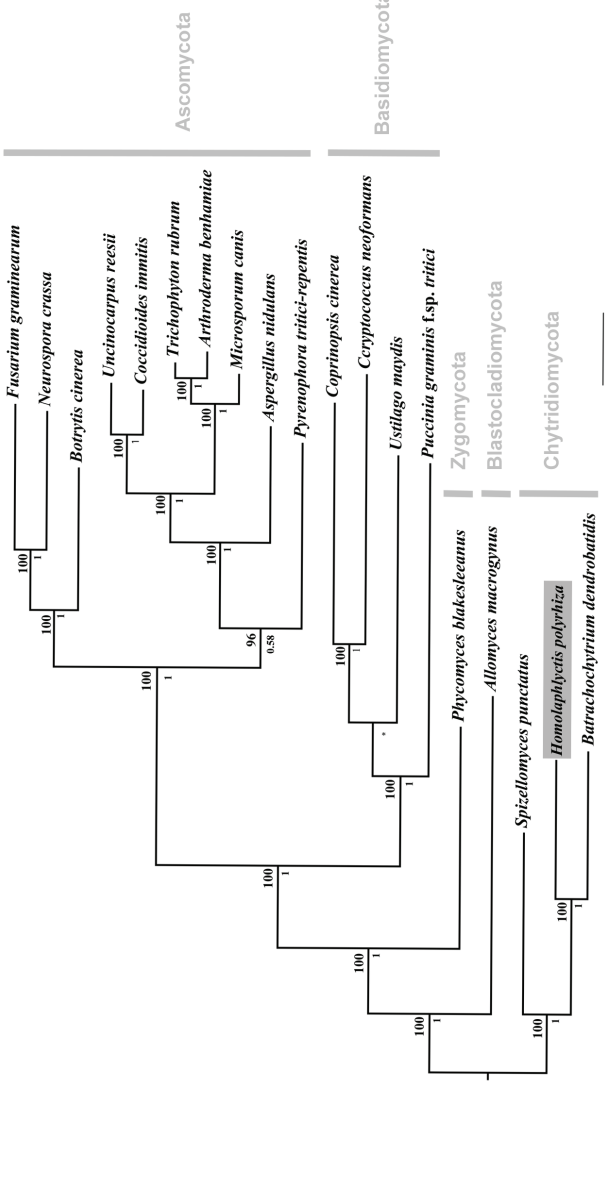


FIGURE A12-1 Phylogenetic relationships among the 19 taxa used in comparative genomics analyses. Focal taxon, *Hp*, boxed in grey. We compared the *Hp* genome to the genome of the amphibian pathogen *Bd* and to a diverse group of other fungal genomes including representatives from all major fungal lineages. Phylogenetic relationship constrained at node between *U. maydis* and *P. graminis* and marked with an asterisk. Bayesian posterior probabilities are shown below internal nodes and ML bootstrap values from 100 replicates are shown above the nodes.

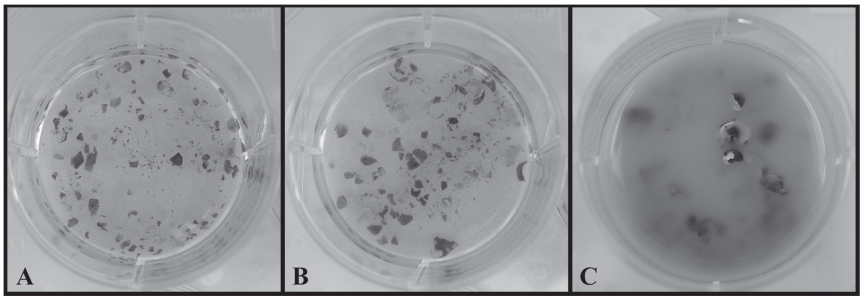


FIGURE A12-2 Chytrid growth on cane-toad-skin. **A.** Negative control (no chytrid): intact skin after 14 days. **B.** *Hp* treatment: intact skin and no *Hp* growth after 14 days. **C.** *Bd* treatment: degraded skin and *Bd* growth after 14 days.

Hp *Genome Sequence, Assembly, and Annotation*

We achieved a roughly 11.2× coverage of the *Hp* genome (total number of aligned bases divided by final genome length, assuming that most of the genome is represented in the aligned reads). We assembled 922,085 screened and trimmed sequencing reads into 16,311 contigs (N50 = 36,162). We inferred a haploid genome size for *Hp* of 26.7 Mb, comparable to other Chytridiomycota genomes [*Bd* (JAM81) = 24.3 Mb, and *Spizellomyces punctatus* = 24.1 Mb]. We have deposited the *Hp* 454 reads in GenBank through the NCBI Sequence Read Archives under the accession SRA037431.1, and we have deposited the Whole Genome Shotgun project at DDBJ/EMBL/GenBank under the accession AFSM00000000 (the version described here is the first version, AFSM01000000).

We generated 5,355 high confidence MAKER predictions and 11,857 GeneMark *ab initio* predictions for *Hp*'s protein coding genes. The number of predicted *Hp* proteins falls within the range of other annotated chytrid genomes (8,732 predicted proteins in *Bd* (JAM81) and 8,804 in *Spizellomyces punctatus*). The difference in number of *Hp* predicted protein numbers between MAKER and GeneMark is due to MAKER's conservative approach, which relies upon homology with protein-coding sequences of other species, and with the consensus of multiple *ab initio* gene prediction algorithms. We did not directly validate the number of expressed genes in our predicted protein sets with EST or RNA sequencing. However, we did compare the *Hp* predicted protein set to gene content in other species, which provides confidence in the *Hp* annotation and assembly. We recovered 92% (228/249) of the core eukaryotic genes using CEGMA in the Hp_Maker dataset. Similarly, we identified 3,216 orthologous groups of "chytrid-specific" proteins shared among both *Bd* isolates and *S. punctatus* (Table S3). Of the predicted chytrid-protein set we recovered 90% (2,885/3,216) in one or both *Hp* predicted protein sets (2,271 in Hp_Maker and 2,817 in Hp_GeneMark).

Together, these results indicate that our sequencing efforts recovered a large proportion of genes that are predicted to occur in the *Hp* genome.

Bd *Unique Genomic Features*

We identified *Bd*-specific genes using the genomes of *Hp* and 17 additional fungi. We considered genes to be *Bd*-specific if they were present in orthologous groups in both sequenced *Bd* genomes (JAM81 and JEL423) and absent from all other fungi including *Hp*. Using OrthoMCL clustered proteins we defined 6,556 orthologous groups in *Bd* (Table S4). Of the 6,556 orthologous groups in *Bd*, 1700 were *Bd*-specific by the above definition. The *Bd*-specific orthologous groups were comprised of 1,974 protein encoding genes, 417 (21%) of which could be functionally categorized by a Pfam domain (with an e-value <0.01) (Table S4). We did not find any orthologous groups uniquely shared between *Bd* and the dermatophytes to the exclusion of all other fungal outgroups (Table S4). Although we defined orthologous groups using the sequenced genomes of both *Bd* isolates (JAM81 and JEL423), below we report gene IDs from JAM81 for simplicity.

We conducted enrichment analyses using gene ontology (GO) terms from the set of 417 *Bd*-specific genes associated with a Pfam domain and found enrichment in all 3 GO structured vocabularies: Cellular Component, Biological Process, and Molecular Function. We present all significantly enriched GO terms (with a corrected P-value of ≤ = 0.05) for the *Bd*-specific gene set in Table A12-1. Briefly, in the Biological Process ontology we found enrichment for genes involved in metabolic processes and regulation of carbohydrates, proteins, and transcription. In the Cellular Component ontology we found enrichment of genes located extracellularly, in the nucleus, and in membranes. In the Molecular Function ontology we found enrichment for genes involved in zinc-ion binding, protein dimerization, DNA-binding, hydrolase activity, and protease and triglyceride lipase activity.

Within the set of *Bd*-specific and GO-enriched genes were several functional groups of particular interest for their possible role in *Bd* pathogenesis. First, many *Bd*-specific genes were proteases and were found in expanded gene families (see below). Second, the *Bd*-specific gene set was enriched for genes containing the Lipase_3 Pfam domain found in triacylglyceride lipases (6 of 417 in the *Bd*-specific gene list, vs 20 of 8732 in the genome, p<0.03) (BATDEDRAFT 93190, BATDEDRAFT 26490, BATDEDRAFT_86691, BATDEDRAFT 93191, BATDEDRAFT_89307, BATDEDRAFT_26489). Third, we identified 62 genes from the *Bd*-specific gene set that encode Crinkler or CRN-like microbial effectors (CRN), a class of genes previously reported only in oomycetes and not found in any of the other fungi considered here (Figure A12-3 and Table S5).

TABLE A12-1 The Enrichment of Cellular Component, Biological Process and Molecular Function GO Terms of 417 *Bd* Specific Genes Associated with a Pfam Domain

GOID	Term	Corrected p-value	# in *Bd* specific gene set	# in *Bd* genome
Biological Process				
GO:0006508	Proteolysis	1.3E-57	99	292
GO:0019538	Protein metabolic process	1.6E-31	126	845
GO:0044238	Primary metabolic process	5.5E-29	179	1644
GO:0043170	Macromolecule metabolic process	1.0E-26	147	1229
GO:0008152	Metabolic process	4.5E-24	196	2075
GO:0019219	Regulation of nucleobase, nucleoside, Nucleotide and nucleic acid metabolic process	6.6E-17	46	205
GO:0051171	Regulation of nitrogen compound metabolic process	6.6E-17	46	205
GO:0009889	Regulation of biosynthetic process	8.2E-17	43	180
GO:0010556	Regulation of macromolecule biosynthetic process	8.2E-17	43	180
GO:0031326	Regulation of cellular biosynthetic process	8.2E-17	43	180
GO:2000112	Regulation of cellular macromolecule biosynthetic process	8.2E-17	43	180
GO:0031323	Regulation of cellular metabolic process	1.8E-16	46	210
GO:0080090	Regulation of primary metabolic process	1.8E-16	46	210
GO:0045449	Regulation of transcription	2.3E-16	42	176
GO:0060255	Regulation of macromolecule metabolic process	6.0E-16	43	189
GO:0010468	Regulation of gene expression	2.1E-15	42	186
GO:0019222	Regulation of metabolic process	4.9E-15	46	227
GO:0065007	Biological regulation	6.8E-15	67	454
GO:0050789	Regulation of biological process	1.5E-14	66	449
GO:0050794	Regulation of cellular process	3.3E-14	62	409
GO:0006355	Regulation of transcription, DNA-dependent	3.4E-14	35	139
GO:0051252	Regulation of RNA metabolic process	3.4E-14	35	139
GO:0051704	Multi-organism process	3.6E-2	5	12
GO:0005975	Carbohydrate metabolic process	4.5E-2	26	253
Cellular Component				
GO:0005623	Cell	1.6E-17	150	1571
GO:0044464	Cell part	1.6E-17	150	1571
GO:0005622	Intracellular	5.5E-10	105	1149
GO:0043231	Intracellular membrane-bounded organelle	4.4E-08	51	425
GO:0043227	Membrane-bounded organelle	5.2E-08	51	427
GO:0005634	Nucleus	1.2E-07	38	272
GO:0043229	Intracellular organelle	1.9E-05	61	659
GO:0043226	Organelle	2.1E-05	61	661
GO:0016021	Integral to membrane	3.9E-05	33	271
GO:0016020	Membrane	1.0E-04	58	645
GO:0044425	Membrane part	1.2E-04	40	381
GO:0031224	Intrinsic to membrane	1.6E-04	33	288

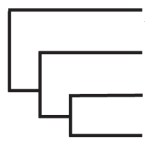
continued

TABLE A12-1 Continued

GOID	Term	Corrected p-value	# in *Bd* specific gene set	# in *Bd* genome
GO:0005576	Extracellular region	2.3E-04	14	69
GO:0044421	Extracellular region part	1.4E-03	12	60
GO:0044424	Intracellular part	7.6E-03	66	880
Molecular Function				
GO:0004190	Aspartic-type endopeptidase activity	9.0E-93	82	98
GO:0070001	Aspartic-type peptidase activity	9.0E-93	82	98
GO:0070011	Peptidase activity, acting on L-amino acid peptides	5.0E-63	102	282
GO:0004175	Endopeptidase activity	1.1E-61	87	198
GO:0008233	Peptidase activity	1.3E-60	102	296
GO:0016787	Hydrolase activity	4.1E-32	143	1056
GO:0003824	Catalytic activity	2.2E-21	200	2254
GO:0001071	Nucleic acid binding transcription factor activity	3.6E-20	32	79
GO:0003700	Sequence-specific DNA binding transcription factor activity	3.6E-20	32	79
GO:0046914	Transition metal ion binding	1.8E-14	56	342
GO:0008270	Zinc ion binding	1.5E-13	47	261
GO:0005488	Binding	2.2E-13	164	1964
GO:0043565	Sequence-specific DNA binding	2.5E-12	20	49
GO:0046872	Metal ion binding	3.3E-11	57	417
GO:0043167	Ion binding	4.1E-11	57	419
GO:0043169	Cation binding	4.1E-11	57	419
GO:0003676	Nucleic acid binding	1.8E-09	70	633
GO:0003677	DNA binding	6.3E-09	43	298
GO:0005515	Protein binding	1.9E-06	44	370
GO:0008236	Serine-type peptidase activity	2.7E-04	16	85
GO:0017171	Serine hydrolase activity	2.7E-04	16	85
GO:0016810	Hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	9.8E-04	11	46
GO:0046983	Protein dimerization activity	1.6E-03	9	32
GO:0004871	Signal transducer activity	3.4E-03	11	52
GO:0060089	Molecular transducer activity	3.4E-03	11	52
GO:0004806	Triglyceride lipase activity	3.3E-02	6	20

Bd *Gene Family Expansions*

We conducted more detailed analyses for three protease gene families that were identified in the *Bd*-specific gene set and showed GO term enrichment: metallo-, serine-type, and aspartyl proteases (M36, S41, and Asp Pfam domains, respectively). The *Bd* genome contained 38 metalloproteases, 32 serine-type proteases, and 99 aspartyl proteases, in all cases at least 4 times as many family members as *Hp* (Figure A12-3). We found that expansions of metalloproteases,



	M36	S41	Asp	CRN
Allomyces macrogynus	31	0	6	0
Spizellomyces punctatus	3	3	10	0
Homolaphlyctis polyrhiza	5	3	22	0
Batrachochytrium dendrobatidis	38	32	99	62

FIGURE A12-3 Gene family copy numbers for metalloproteases (M36), serine-type proteases (S41), aspartyl proteases (ASP) and CRN-like proteins (CRN) in the Chytridiomycota (*Bd*, *Hp* and *S. punctatus*), and a Blastocladiomycota outgroup (*A. macrogynus*). Phylogenetic relationship of taxa adapted from James et al. (2006). Focal taxa highlighted in grey.

serine-type proteases and aspartyl proteases were largely *Bd* specific, having occurred after the split between the *Bd* and *Hp* lineages from their most recent common ancestor (Summary in Figure A12-4, gene-names available in tree, Figure S2). In all three families, *Bd* had a greater number of gene copies than any of the other focal taxa, and the *Bd* gene copies were generally clustered together to the exclusion of homologues from other taxa. This clustering is consistent with lineage-specific gene family expansions in *Bd* (Figure A12-4). We observed a large number of metalloprotease genes not only in *Bd* but also in *Allomyces macrogynus* (38 and 31 gene family members, respectively) (Figure A12-3). However, the gene tree indicates that the expansion of metalloprotease genes in *Bd* and *A. macrogynus* were independent with most duplication events occurring after the divergence of *Bd* and *A. macrogynus* from their common ancestor (Figure A12-4A).

In addition to identifying many lineage-specific duplicates of proteases in *Bd*, we demonstrate that these *Bd* duplication events likely occurred significantly more recently than the divergence time between the species analyzed (Figure A12-5). To assess the timing of expansion in each protease gene family, we calculated synonymous substitution rate, *Ks*, between homologs and based on the phylogeny, we calculated a node *Ks* value for each lineage-specific duplication node (Figure A12-5, left panel). The median *Ks* values for the metallo-, serine-type, and aspartyl—proteases derived from *Bd*-specific duplications were 0.37, 0.14 and 0.24, respectively (Figure A12-5, left panel). We note that in the metalloprotease family there were similar numbers of lineage-specific duplications in *Bd* (24) and *A. macrogynus* (28). However the *Ks* values of *Bd*-specific duplicates were significantly lower *A. macrogynus* duplicates (median *Ks*: 0.37 and 1.56, respectively; Kolmogorov-Smirnov tests, $p<4.6e-5$), indicating that *Bd*-specific M36 duplications took place much more recently than the *A. macrogynus* duplications.

We also examined *Ks* values of putative orthologs between *Bd* and *Hp*, *Bd* and *S. punctatus*, and *Bd* and *A. macrogynus* (Figure A12-5B, right panel). As

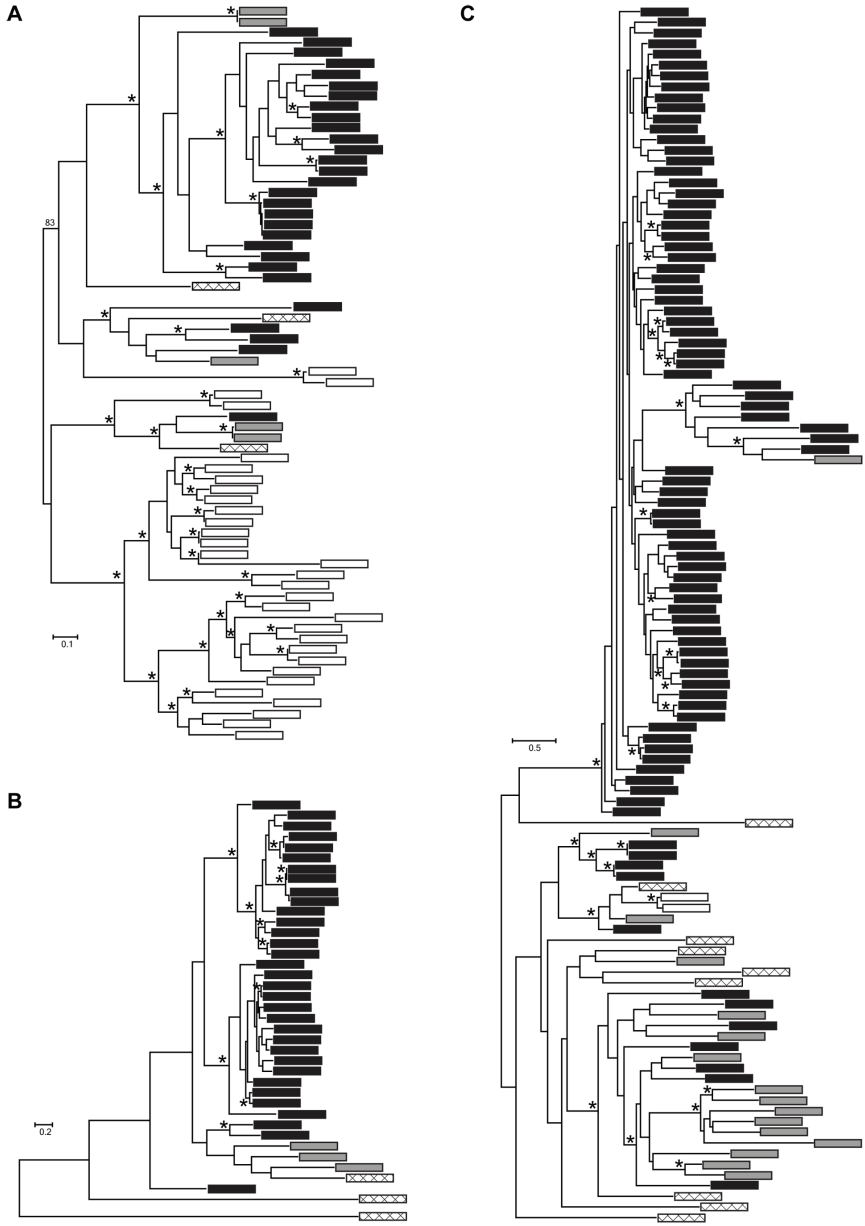


FIGURE A12-4 Maximum likelihood phylogenies of gene families containing (A) M36, (B) S41, and (C) Asp Pfam domains. Each tip represents a single gene copy and each source species is denoted by shaded or hatched boxes (*Bd*: black, *Hp*: grey, *S. punctatus*: hatched, *A. macrogynus*: white). Bootstrap values over 80% indicated with asterisks at internal nodes.

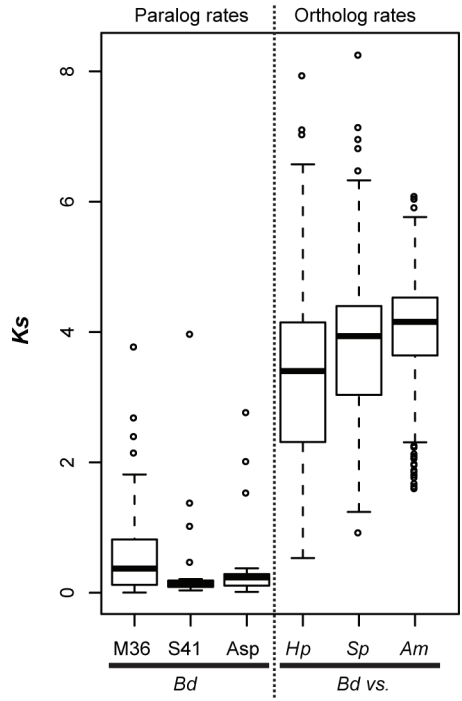


FIGURE A12-5 Left panel (paralog rates) shows box plots of synonymous substitution rates (*Ks*) for *Bd* lineage-specific duplicates in three protease families. Right panel (ortholog rates) shows box plots of *Ks* values for putative orthologs between *Bd* and *Hp*, *Bd* and *S. punctatus*, and *Bd* and *A. macrogynus*. Box and whisker plots show median (line), inter-quartile range (box), 1.5 inter-quartile range (whiskers), and outliers (open circles).

expected from the phylogenetic relationships of these four species (James et al., 2006), the median *Ks* for all species was high, but the median *Ks* for *Bd-Hp* (3.40) was significantly lower than that of *Bd-S. punctatus* (3.94) and *Bd-A. macrogynus* (4.03) (Kolmogorov-Smirnov tests, $p<2.2e-16$). Importantly, the median *Bd-Hp* orthologous *Ks* values were ~9–24 fold higher than the median *Ks* of *Bd* lineage-specific duplicates. Therefore, the *Bd*-specific duplications occurred substantially more recently than the divergence of *Bd* and *Hp*. Previous molecular evolution studies in fungi have used a synonymous nucleotide substitution rate of 8.1e-9 substitutions per site per year to estimate the timing of molecular events (Lynch and Conery, 2000). If this substitution rate is reasonable for chytrid fungi, the duplication events leading to the metallo-, serine-type, and aspartyl protease gene family expansions in *Bd* would be millions of years old (Table S6). Even if the true substitution rate differs by several orders of magnitude, it is important to

recognize that these protease duplication events occurred long before *Bd* emerged as a global pathogen of amphibians.

Discussion

To investigate the genomic changes that accompanied the evolution of pathogenicity, we compared *Bd*, the deadly chytrid pathogen of amphibians, with *Hp*, a closely related chytrid that is not a known pathogen of vertebrates. We confirm that *Bd* and *Hp* have different nutritional modes (Figure A12-2); unlike *Hp*, *Bd* is capable of growing on amphibian skin alone. Given the most chytrids are saprobes like *Hp*, *Bd*'s ability to infect vertebrate skin likely arose after the divergence of *Bd* and *Hp* from their common ancestor. Fungal growth on vertebrate skin requires the expression of enzymes that break-down host epidermal tissue (Burmester et al., 2011; da Silva et al., 2006; Monod, 2008; Monod et al., 2002). Because *Bd* causes chytridiomycosis by infecting frog skin (Longcore et al., 1999; Voyles et al., 2009) we were particularly interested in elements of the *Bd* genome whose evolution might have allowed *Bd* to colonize and degrade amphibian skin.

We compared the genomes of *Bd* and *Hp* in a broad taxonomic context of 18 diverse fungal genomes to identify genomic factors that make *Bd* unique. The *Bd* and *Hp* genomes are similar in size and number of predicted genes but show important differences in gene content. Therefore we could identify *Bd*-specific genes (i.e., genes that were found in *Bd* but not in *Hp* or other fungal outgroups). *Bd*-specific genes are enriched for GO terms related to extracellular and enzymatic activity. Many *Bd*-specific genes are members of recently expanded gene families (i.e., gene families with significantly more members than other fungal species). Below we discuss *Bd*-specific genes with particular emphasis on understanding how *Bd* may interact with its amphibian hosts.

Proteases are the most dramatically enriched class of *Bd*-specific genes. The *Bd* genome contains expanded gene families of metalloproteases, serine-type proteases, and aspartyl proteases. Each of these *Bd* gene families contains more than 30 family members and contains 4–10 times as many family members as found in *Hp* (Figure A12-3). Extracellular fungal proteases have been implicated in the adherence to, invasion of, and degradation of host cells by other fungal pathogens (Burmester et al., 2011; da Silva et al., 2006; Monod, 2008; Monod et al., 2002). In particular, protease gene family expansions have been suggested as a link to pathogenesis in other fungal pathogens. Several fungal pathogens of vertebrates (e.g., *Arthroderma benhamiae*, *Coccidioides* spp, and *Trichophyton* spp.) exhibit gene family expansions specifically for metalloproteases and serine-type proteases (Burmester et al., 2011; Jousson et al., 2004; Sharpton et al., 2009).

Here we strengthen the evidence implicating proteases in *Bd* pathogenesis in several ways. First and most importantly we demonstrate that protease gene families are not expanded in *Hp* and thus polarize the expansion events to a

much shorter phylogenetic branch leading to *Bd*. Second, we present an additional protease gene family expansion. We previously reported the *Bd* gene family expansions for metallo- and serine proteases (Rosenblum et al., 2008), and we now describe a dramatic expansion of aspartyl proteases in the *Bd* genome (Figure A12-3). Aspartyl proteases are of particular interest because they have been implicated in the adherence to and invasion of human host tissue by fungal pathogens (*Candida* spp.) (Kaur et al., 2007; Monod and Borg-von Zepelin, 2002). Many genes in the expanded metallo-, serine- and aspartyl-protease gene families are highly expressed, and in some cases differentially expressed between *Bd* life stages (Rosenblum et al., 2008). Finally, we more rigorously document the dynamics of protease gene family expansions. Calculations using a range of reasonable substitution rates show that the three protease gene family expansions occurred substantially more recently than the divergence of *Bd* and *Hp* from their common ancestor.

It is important to caution that our comparative genomics results do not conclusively demonstrate a role for proteases as pathogenicity factors. First, we lack a specific mechanism by which proteases mediate *Bd* host invasion. Understanding the functional consequences of protease gene family expansions will require molecular assays to determine how specific enzymes contribute to host substrate metabolism. Second, protease gene family expansions are not always obviously correlated with fungal pathogenicity. For example we observed a large number of metalloprotease genes in *Allomyces macrogynus* and a variable number of aspartyl proteases in several of our outgroup taxa (Figure A12-4). These are independent expansion events relative to the *Bd* gene duplications and are not associated with a specific shift in substrate metabolism. Third, the estimated timing of the *Bd* protease gene duplications does not unambiguously link particular genes to the recent emergence of *Bd* as a global frog pathogen. Although the gene duplication events are relatively recent, most still likely occurred millions of years ago. More ancient duplication of protease genes may have set the stage for *Bd*'s ability to infect frogs, but finer scale intraspecific data will be required to determine whether particular paralogs exhibit molecular signatures of recent selection.

While proteases may play the most obvious role in pathogen invasion and metabolism of host tissue, we also observed an enrichment of genes with triglyceride lipase activity in the *Bd*-specific gene set (Table A12-1). These enzymes are known to play a role in fungal-plant interactions (Gaillardin, 2000), and have been hypothesized to play a role in at least one fungal-vertebrate interaction—between *Malassezia furfur* and the skin of its human host (Brunke and Hube, 2006). *M. furfur* incorporates host lipids into its own cell wall; this is thought to assist *M. furfur* in adhering to the host and evading the host's immune system. The extent to which *Bd* can utilize the products of triglyceride lipase activity for nutrition or adhesion remains to be tested. However, the enrichment of triglyceride lipase genes in the *Bd*-specific gene set suggests considering whether lipases could play a role in *Bd*'s invasion of host tissue.

In addition to the genes that may be involved in host tissue metabolism we observed a large number of *Bd*-specific genes with similarity to microbial proteins known as Crinklers and Crinkler-like effectors (CRN). Microbial effectors in general act within the host cytoplasm to suppress host defenses and alter normal host cell metabolism (Haas et al., 2009; Kamoun, 2006). It is unusual that a fungus contains CRN effectors as these proteins have so far only been reported from oomycetes, a group of important plant and fish pathogens within the kingdom Chromista (Cavalier-Smith and Chao, 2006). CRN effectors are modular proteins consisting of a signal-peptide, a downstream translocator domain that allows CRN proteins to gain entry into host cells, and a C-terminus domain that interacts with host proteins (Haas et al., 2009). While 62 unique *Bd* proteins show similarity to CRN effectors at the protein level, only one is predicted to be secreted (BATDEDRAFT_23205). Therefore while the function of putative CRN effectors in *Bd* remains to be determined, the possibility that they function as microbial effectors and interact with host elements merits further investigation.

We have sequenced the genome of *Bd*'s closest known relative to develop hypotheses for genomic determinants of *Bd*'s ability to infect and kill amphibians. However, the divergence between *Bd* and *Hp* is still substantial (James et al., 2006). Recent research indicates that chytrids may be more ubiquitous than previously appreciated in both aquatic and terrestrial environments (Freeman et al., 2009), and much chytrid diversity remains to be characterized. The discovery of additional taxa more closely related to *Bd* than *Hp* would help further localize genomic changes to the *Bd* lineage. Interspecific comparisons such as the one presented here can be complemented by intraspecific comparisons among *Bd* isolates to understand the evolutionary dynamics of genes hypothesized to play a role in *Bd* pathogenicity. However, robust hypothesis testing will require functional characterization of genes that may be important to *Bd*'s ability to infect frogs. *Bd* currently lacks a transformation system in which to study gene function, but heterologous expression systems could potentially be used to determine specific gene functions. Additionally, understanding expression patterns of candidate genes under different nutrient conditions and during different stages of host invasion are likely to yield important insights. Ultimately, identifying the molecular mechanisms of host-pathogen interactions will provide new avenues for mitigating the devastating effects of chytridiomycosis.

Acknowledgments

We thank Joyce Longcore (University of Maine) for providing *Hp* strain JEL142. We thank Matt Settles (University of Idaho) for bioinformatics support. We thank Joyce Longcore, Tim James (University of Michigan), and Jamie Voyles (University of Idaho) for comments on the manuscript and input throughout the project. We acknowledge Igor Grigoriev and the Joint Genome Institute for access to the *B. dendrobatidis* (JAM81) genome.

Author Contributions

Conceived and designed the experiments: SJ JES EBR. Performed the experiments: SJ. Analyzed the data: SJ JES SHS. Contributed reagents/materials/analysis tools: JES SHS SJ EBR. Wrote the paper: SJ EBR.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

Barr DJS (1986) *Allochytridium expandens* rediscovered—morphology, physiology and zoospore ultrastructure. Mycologia 78: 439–448.

Berger L, Speare R, Daszak P, Green DE, Cunningham AA, et al. (1998) Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and Central America. Proc Natl Acad Sci U S A 95: 9031–9036.

Boyle EI, Weng SA, Gollub J, Jin H, Botstein D, et al. (2004) GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710–3715.

Brunke S, Hube B (2006) MfLIP1, a gene encoding an extracellular lipase of the lipid-dependent fungus *Malassezia furfur*. Microbiology 152: 547–554.

Burmester A, Shelest E, Glöckner G, Heddergott C, Schindler S, et al. (2011) Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. Genome Biol 12: R7.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, et al. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18: 188–196.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

Cavalier-Smith T, Chao EEY (2006) Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). J Mol Evol 62: 388–420.

da Silva BA, dos Santos ALS, Barreto-Bergter E, Pinto MR (2006) Extracellular peptidase in the fungal pathogen *Pseudallescheria boydii*. Curr Microbiol 53: 18–22.

Dewey C (2007) Aligning multiple whole genomes with Mercator and MAVID. Methods Mol Biol 395: 221–236.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–D222.

Freeman KR, Martin AP, Karki D, Lynch RC, Mitter MS, et al. (2009) Evidence that chytrids dominate fungal communities in high-elevation soils. Proc Natl Acad Sci U S A 106: 18315–18320.

Gaillardin C (2010) Lipases as Pathogenicity Factors of Fungi. In: Timmis KN, editor. Handbook of Hydrocarbon and Lipid Microbiology. Springer. pp. 3259–3268.

Haas BJ, Kamoun S, Zody MC, Jiang RHY, Handsaker RE, et al. (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature 461: 393–398.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148: 993–1003.

Hoskisson PA, Trevors JT (2010) Shifting trends in pathogen dynamics on a changing planet. Antonie Van Leeuwenhoek 98: 423–427.

James TY, Letcher PM, Longcore JE, Mozley-Standridge SE, Porter D, et al. (2006) A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota). Mycologia 98: 860–871.

Jousson O, Léchenne B, Bontems O, Capoccia S, Mignon B, et al. (2004) Multiplication of an ancestral gene encoding secreted fungalysin preceded species differentiation in the dermatophytes *Trichophyton* and *Microsporum*. Microbiology 150: 301–310.

Kamoun S (2006) A catalogue of the effector secretome of plant pathogenic oomycetes. Annu Rev Phytopathol 44: 41–60.

Kaur R, Ma B, Cormack BP (2007) A family of glycosylphosphatidylinositol-linked aspartyl proteases is required for virulence of *Candida glabrata*. Proc Natl Acad Sci U S A 104: 7628–7633.

Letcher PM, Powell MJ, Churchill PF, Chambers JG (2006) Ultrastructural and molecular phylogenetic delineation of a new order, the Rhizophydiales (Chytridiomycota). Mycol Res 110: 898–915.

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178–2189.

Lips KR, Brem F, Brenes R, Reeve JD, Alford RA, et al. (2006) Emerging infectious disease and the loss of biodiversity in a Neotropical amphibian community. Proc Natl Acad Sci U S A 103: 3165–3170.

Longcore JE, Pessier AP, Nichols DK (1999) *Batrachochytrium dendrobatidis* gen et sp nov, a chytrid pathogenic to amphibians. Mycologia 91: 219–227.

Lötters S, La Marca E, Stuart S, Gagliardo R, Veith M (2004) A new dimension of current biodiversity loss. Herpetotropicos 1: 29–31.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

Monod M (2008) Secreted proteases from dermatophytes. Mycopathologia 166: 285–294.

Monod M, Borg-von Zepelin M (2002) Secreted aspartic proteases as virulence factors of *Candida* species. Biol Chem 383: 1087–1093.

Monod M, Capoccia S, Léchenne B, Zaugg C, Holdom M, et al. (2002) Secreted proteases from pathogenic fungi. Int J Med Microbiol 292: 405–419.

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–217.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. Bioinformatics 23: 1061–1067.

Rachowicz LJ, Knapp RA, Morgan JAT, Stice MJ, Vredenburg VT, et al. (2006) Emerging infectious disease as a proximate cause of amphibian mass mortality. Ecology 87: 1671–1683.

Rosenblum EB, Stajich JE, Maddox N, Eisen MB (2008) Global gene expression profiles for life stages of the deadly amphibian pathogen *Batrachochytrium dendrobatidis*. Proc Natl Acad Sci U S A 105: 17034–17039.

Sharpton TJ, Stajich JE, Rounsley SD, Gardner MJ, Wortman JR, et al. (2009) Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. Genome Res 19: 1722–1731.

Skerratt LF, Berger L, Speare R, Cashins S, McDonald KR, et al. (2007) Spread of chytridiomycosis has caused the rapid global decline and extinction of frogs. EcoHealth 4: 125–134.

Smith KF, Guegan JF (2010) Changing Geographic Distributions of Human Pathogens. Annual Review of Ecology, Evolution, and Systematics, Vol 41. Palo Alto: Annual Reviews. pp. 231–250.

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–463.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Res 18: 1979–1990.

Voyles J, Young S, Berger L, Campbell C, Voyles WF, et al. (2009) Pathogenesis of Chytridiomycosis, a Cause of Catastrophic Amphibian Declines. Science 326: 582–585.

Woolhouse M, Gaunt E (2007) Ecological origins of novel human pathogens. Crit Rev Microbiol 33: 231–242.

Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.

Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.

Zolan ME, Pukkila PJ (1986) Inheritance of DNA methylation in *Coprinus cinereus*. Mol Cell Biol 6: 195–200.

A13

NATURAL AND EXPERIMENTAL INFECTION OF *CAENORHABDITIS* NEMATODES BY NOVEL VIRUSES RELATED TO NODAVIRUSES[50]

Marie-Anne Félix,[51,†,‡] *Alyson Ashe*,[52,‡] *Joséphine Piffaretti*,[51,‡] *Guang Wu*,[53] *Isabelle Nuez*,[51] *Tony Békucard*,[51] *Yanfang Jiang*,[53] *Guoyan Zhao*,[53] *Carl J. Franz*,[53] *Leonard D. Goldstein*,[52] *Mabel Sanroman*,[51] *Eric A. Miska*,[52,†] *and David* ***Wang***[53,†]

Abstract

An ideal model system to study antiviral immunity and host-pathogen coevolution would combine a genetically tractable small animal with a virus capable of naturally infecting the host organism. The use of *C. elegans* as a model to define host-viral interactions has been limited by the lack of viruses known to

---

[50] Reprinted from PLoS Biology. Originally published as Félix M-A, Ashe A, Piffaretti J, Wu G, Nuez I, et al. (2011) Natural and Experimental Infection of Caenorhabditis Nematodes by Novel Viruses Related to Nodaviruses. PLoS Biol 9(1): e1000586. doi:10.1371/journal.pbio.1000586

Funding: This work was supported in part by the CNRS (Félix lab), National Institutes of Health grant U54 AI057160 to the Midwest Regional Center of Excellence for Biodefense and Emerging Infectious Disease Research (Wang lab), and a Cancer Research UK Programme Grant (Miska lab). DW holds an Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund. AA was supported by a fellowship from the Herchel-Smith Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Eric Miska's spouse is a member of the PloS Biology editorial staff; in accordance with the PLoS policy on competing interests she has been excluded from all stages of the review process for this article.

†E-mail: felix@ijm.univ-paris-diderot (MAF); eam29@cam.ac.uk (EAM); davewang@wustl.edu (DW)

‡These authors contributed equally to this work.

[51] Institut Jacques Monod, CNRS-University of Paris-Diderot, Paris, France.

[52] Gurdon Institute, University of Cambridge, Cambridge, United Kingdom.

[53] Departments of Molecular Microbiology and Pathology & Immunology, Washington University in St. Louis School of Medicine, St. Louis, Missouri, United States of America.

infect nematodes. From wild isolates of *C. elegans* and *C. briggsae* with unusual morphological phenotypes in intestinal cells, we identified two novel RNA viruses distantly related to known nodaviruses, one infecting specifically *C. elegans* (Orsay virus), the other *C. briggsae* (Santeuil virus). Bleaching of embryos cured infected cultures demonstrating that the viruses are neither stably integrated in the host genome nor transmitted vertically. 0.2 μm filtrates of the infected cultures could infect cured animals. Infected animals continuously maintained viral infection for 6 mo (~50 generations), demonstrating that natural cycles of horizontal virus transmission were faithfully recapitulated in laboratory culture. In addition to infecting the natural *C. elegans* isolate, Orsay virus readily infected laboratory *C. elegans* mutants defective in RNAi and yielded higher levels of viral RNA and infection symptoms as compared to infection of the corresponding wild-type N2 strain. These results demonstrated a clear role for RNAi in the defense against this virus. Furthermore, different wild *C. elegans* isolates displayed differential susceptibility to infection by Orsay virus, thereby affording genetic approaches to defining antiviral loci. This discovery establishes a bona fide viral infection system to explore the natural ecology of nematodes, host-pathogen co-evolution, the evolution of small RNA responses, and innate antiviral mechanisms.

Author Summary

The nematode *C. elegans* is a robust model organism that is broadly used in biology. It also has great potential for the study of host-microbe interactions, as it is possible to systematically knockout almost every gene in high-throughput fashion to examine the potential role of each gene in infection. While *C. elegans* has been successfully applied to the study of bacterial infections, only limited studies of antiviral responses have been possible since no virus capable of infecting any *Caenorhabditis* nematode in laboratory culture has previously been described. Here we report the discovery of natural viruses infecting wild isolates of *C. elegans* and its relative *C. briggsae*. These novel viruses are most closely related to the ssRNA nodaviruses, but have larger genomes than other described nodaviruses and clearly represent a new taxon of virus. We were able to use these viruses to infect a variety of laboratory nematode strains. We show that mutant worms defective in the RNA interference pathway, an antiviral system known to operate in a number of organisms, accumulate more viral RNA than wild type strains. The discovery of these viruses will enable further studies of host-virus interactions in *C. elegans* and the identification of other host mechanisms that counter viral infection.

Introduction

Model organisms such as *D. melanogaster* (Hao et al., 2008; Sabin et al., 2009) and *C. elegans* (Kim et al., 2002; Powell et al., 2009) have been increasingly

used in recent years to examine features of the host immune system and host-pathogen co-evolution mechanisms, due to the genetic tractability and ease of manipulation of these organisms. A prerequisite to fully exploit such models is the identification of an appropriate microbe capable of naturally infecting the host organism. Analysis in *C. elegans* of bacterial pathogens such as *Pseudomonas*, *Salmonella*, or *Serratia* has been highly fruitful, in some instances revealing the existence of innate immune pathways in *C. elegans* that are also conserved in vertebrates (Kim et al., 2002). The recent report of natural infections of *C. elegans* intestinal cells by microsporidia makes it a promising model for microsporidia biology (Troemel et al., 2008). Efforts to use *C. elegans* to understand anti-viral innate immunity, however, have been hampered by the lack of a natural virus competent to infect and replicate in *C. elegans*.

In the absence of a natural virus infection system, some efforts to define virus-host responses in *C. elegans* have been pursued using artificial methods of introducing viruses or partial virus genomes into animals (Liu et al., 2006; Lu et al., 2005). For example, the use of a transgenic Flock House virus RNA1 genome segment has clearly established a role for RNAi in counteracting replication of Flock House virus RNA (Lu et al., 2005) and has defined genes essential for the RNAi response (Lu et al., 2009). However, this experimental system can only examine replication of the viral RNA and is fundamentally unable to address the host response to other critical aspects of the virus life cycle such as virus entry, virion assembly, or egress. The ability of a host to target steps other than genome replication to control viral infections is highlighted by recent discoveries such as the identification of tetherin, which plays a critical role at the stage of viral egress by blocking the release of fully assembled HIV virions from infected human cells (Neil et al., 2008). Furthermore, the artificial systems used to date for analysis of virus-nematode interactions cannot be used to examine transmission dynamics of virus infection. These limitations underscore the need to establish an authentic viral infection and replication system in nematodes.

Natural populations of *C. elegans* have proven hard to find until recent years. The identification of *C. elegans* habitats and the development of simple isolation methods (MAF, unpublished) (Barrière and Félix, 2006) has now enabled extensive collection of natural isolates of *C. elegans*. Here we report the discovery of natural populations of *C. elegans* and of its close relative *C. briggsae* that display abnormal morphologies of intestinal cells. These abnormal phenotypes can be maintained in permanent culture for several months, without detectable microsporidial or bacterial infection. We show that these populations are infected by two distinct viruses, one specific for *C. elegans* (Orsay virus), one for *C. briggsae* (Santeuil virus). These viruses resemble viruses in the *Nodaviridae* family, with a small, bipartite, RNA (+sense) genome. Infection by each virus is transmitted horizontally. In both nematode species, we find intraspecific variation in sensitivity to the species-specific virus. We further show that infected worms mount a small RNA response and that RNAi mechanisms act as antiviral immunity in

nematodes. Finally, we demonstrate that the *C. elegans* isolate from which Orsay virus was isolated is incapable of mounting an effective RNAi response in somatic cells. We thus find natural variation in host antiviral defenses. Critically, these results establish the first experimental viral infection system in *C. elegans* suitable for probing all facets of the host antiviral response.

Results

Natural Viral Infections of C. briggsae *and* C. elegans

From surveys of wild nematodes from rotting fruit in different regions of France, multiple *Caenorhabditis* strains were isolated that displayed a similar unusual morphology of the intestinal cells and no visible pathogen by optical microscopy. Intestinal cell structures such as storage granules disappeared (Figures A13-1A–J, A13-2A–C) and the cytoplasm lost viscosity and became fluid (Figure A13-1B, I), moving extensively during movement of the animal. The intestinal apical border showed extensive convolutions and intermediate filament disorganization (Figures A13-1A, A13-2H, as described in some intermediate filament mutants [Hüsken et al., 2008]). Multi-membrane structures were sometimes apparent in the cytoplasm (Figure A13-1C). Elongation of nuclei and nucleoli, and nuclear degeneration, were observed using Nomarski optics, live Hoechst 33342 staining, and electron microscopy (Figures A13-1E–H, A13-2D–F). Finally, some intestinal cells fused together (Figure A13-1I). This suite of symptoms was first noticed during sampling of *C. briggsae*. Indeed, more individuals appeared affected in *C. briggsae* than in *C. elegans* cultures, and to a greater extent (Figure A13-1K).

One representative, stably infected, strain of each nematode species, *C. elegans* JU1580 (isolated from a rotting apple in Orsay, France) and *C. briggsae* JU1264 (isolated from a snail on a rotting grape in Santeuil, France), were selected for detailed analysis. Bleaching of adult animals resulted in phenotype-free progeny from both strains, demonstrating that the phenotype was not vertically transmitted (embryos are resistant to the bleaching treatment) (Figure A13-1K). Addition of dead infected animals, or homogenates from infected animals after filtration through 0.2 μm filters, to plates containing previously bleached animals recapitulated the morphological phenotype, raising the possibility that a virus might play a role in inducing the morphological phenotype (Figure A13-1K). We found that the infectious agent could be passed on horizontally through live animals by incubating GFP-labeled animals (strain JU1894, Table A13-1) with 10 non-GFP-infected worms (JU1580), checking that the latter did not die before removing them 24 h later. The GFP-labeled culture displayed the intestinal symptoms after a week. One possibility is that the intestinal infectious agent is shed from the intestine through the rectum and may enter the next animal during feeding.

In support of the hypothesis that these wild *Caenorhabditis* were infected by a virus, small virus-like particles of approximately 20 nm diameter were visible by electron microscopy of the intestinal cells (Figures A13-2H, S1). Such particles were not observed in bleached animals, nor in *C. elegans* animals infected by bacteria, which showed a strong reduction of intestinal cell volume (strain JU1409, unpublished data).

While a clear morphological phenotype was visible by microscopy, infection did not cause a dramatic decrease in adult longevity (unpublished data), nor a change in brood size (Figure S2A,B). However, progeny production was significantly slowed down during adulthood, most clearly in the infected *C. briggsae* JU1264 isolate compared to the uninfected control (Figure S2D).

Molecular Identification of Two Divergent Viruses

An unbiased high-throughput pyrosequencing approach was used to determine whether any known or novel viruses were present in the animals. From JU1264, 28 unique sequence reads were identified initially that shared 30%–48% amino acid sequence identity to known viruses in the family *Nodaviridae*. Nodaviruses are bipartite positive strand RNA viruses. The RNA1 segment of all previously described nodaviruses is ~3.1 kb and encodes ORF A, the viral RNA-dependent RNA polymerase. Some nodaviruses also encode ORFs B1/B2 at the 3′ end of the RNA1 segment. The B1 protein is of unknown function while the B2 protein is able to inhibit RNAi (Li et al., 2002). The RNA2 segment of all previously described nodaviruses is ~1.4 kb and possesses a single ORF encoding the viral capsid protein. Assembly of the initial JU1264 pyrosequencing reads followed by additional pyrosequencing, RT-PCR, 5′ RACE, and 3′ RACE yielded two final contigs, which were confirmed by sequencing of overlapping RT-PCR amplicons. The two contigs corresponded to the RNA1 and RNA2 segments of a novel virus. The first contig (3,628 nt) encoded a predicted open reading frame of 982 amino acids that shared 26%–27% amino acid identity to the RNA-dependent RNA polymerase of multiple known nodaviruses by BLAST alignment. All known nodavirus B2 proteins overlap with the C-terminus of the RNA-dependent RNA polymerase and are encoded in the +1 frame relative to the polymerase. No open reading frame with these properties was predicted in the 3′ end of the RNA1 segment. The second contig of 2,653 nt, which was presumed to be the near-complete RNA2 segment, encoded at its 5′ end a predicted protein with ~30% identity to known nodavirus capsid proteins (Figure A13-3A). This contig was ~1 kb larger than the RNA2 segment of all previously described nodaviruses and appeared to encode a second ORF of 332 amino acids at the 3′ end. This second predicted ORF, named ORF δ, had no significant BLAST similarity to any sequence in Genbank.

Pyrosequencing of JU1580 demonstrated the presence of a second distinct virus that shared the same general genomic organization as the virus detected in

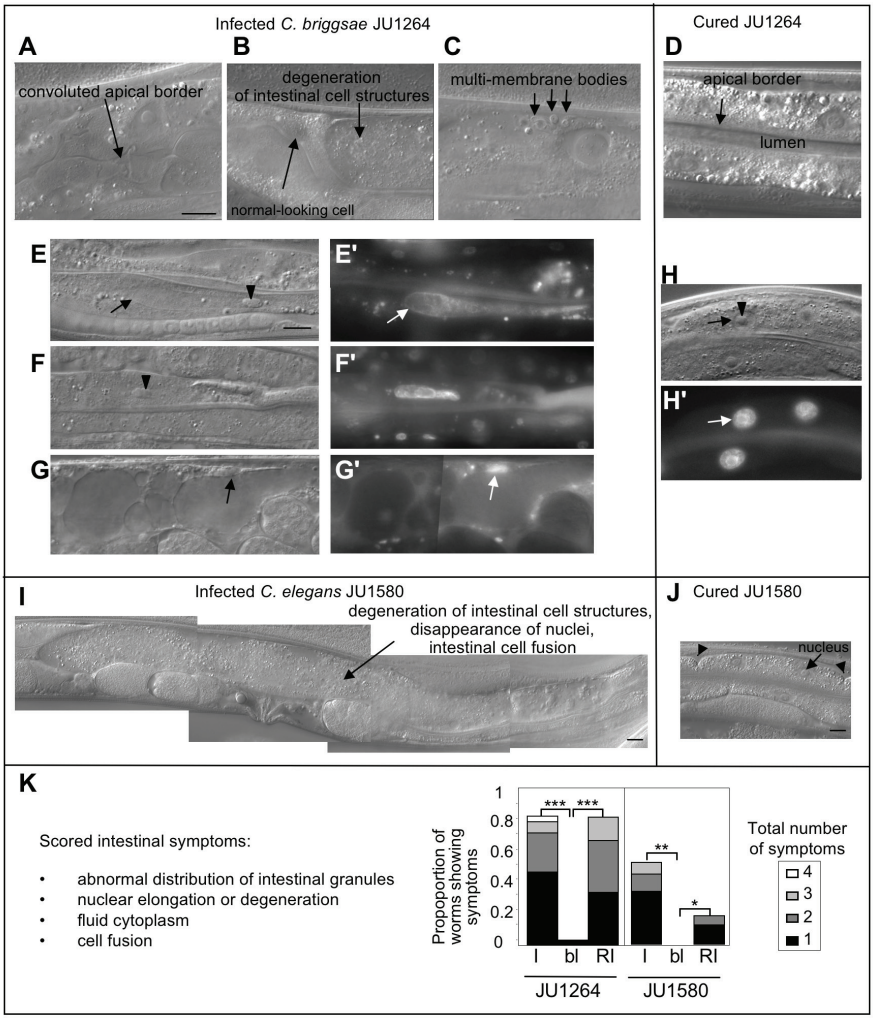


FIGURE A13-1 Intestinal cell infection phenotypes in wild *Caenorhabditis* isolates. (A–H) *C. briggsae* JU1264 and (I, J) *C. elegans* JU1580 observed by Nomarski microscopy. (A–C, E–G, I) Infected adult hermaphrodites from the original cultures, with the diverse infection symptoms: convoluted apical intestinal border (A), degeneration of intestinal cell structures and liquefaction of the cytoplasm (B, G, I), presence of multi-membrane bodies (C). The animals in (E–H) were also observed in the fluorescence microscope after live Hoechst 33342 staining of the nuclei, showing the elongation and degeneration of nuclei (E′–H′). In (E), the nucleus and nucleolus are abnormally elongated. In (F), the nuclear membrane is no longer visible by Nomarski optics. In (G), the cell cytoplasmic structures are highly abnormal (apparent vacuolisation) and the nucleus is very reduced in size. In (E–H′), arrows denote nuclei and arrowheads nucleoli. The infected animal in (I)

TABLE A13-1 Strain List

Strain	Genotype
JU1264	*C. briggsae* wild isolate, Santeuil, France
JU1580	*C. elegans* wild isolate, Orsay, France
AF16	*C. briggsae* wild reference isolate, India
N2	*C. elegans* wild reference isolate, England
CB4856	*C. elegans* wild isolate, Hawaii
AB1	*C. elegans* wild isolate, Australia
PB303	*C. elegans* wild isolate, USA
PB306	*C. elegans* wild isolate, USA
JU258	*C. elegans* wild isolate, Madeira
PS2025	*C. elegans* wild isolate, California, USA
JU1894	*mfEx50*[*let858*::GFP, *myo-2*::DsRed] in JU1580 background
JU1895	*mfEx51*[*let858*::GFP, *myo-2*::DsRed] in N2 background
WM27	*rde-1(ne219)* V
WM29	*rde-2(ne221)* I
WM49	*rde-4(ne301)* III
NL936	*unc-32(e189) mut-7(pk204)* III

JU1264. Partial genome sequences of 2,680 nucleotides of the RNA1 segment and 2,362 nucleotides of the RNA2 segment were obtained and confirmed by RT-PCR. The putative RNA-dependent RNA polymerases of the two viruses shared 44% amino acid identity by BLAST analysis. Like the virus in JU1264, the virus in JU1580 was predicted to encode a capsid protein at the 5′ end of the RNA2 segment as well as a second ORF in the 3′ half of the RNA2 segment. The ORF δ encoded proteins from the two viruses shared 37% amino acid identity when compared using BLAST. Thus, the genomic organization of these two viruses, while sharing substantial commonality with known nodaviruses, also displayed novel genomic features. Phylogenetic analysis of the predicted RNA polymerase and capsid proteins demonstrated that the virus sequences in JU1580 and JU1264

---

displays an abnormally large intestinal cell that is probably the result of cell fusions, with degeneration of cellular structures including nuclei. (D, H, J) Uninfected (bleached) adults. Arrowheads in (J) indicate antero-posterior boundaries between intestinal cells, each of which generally contains two nuclei. Bars: 10 µm. (K) Proportion of worms showing the indicated cumulative number of morphological infection symptoms in at least one intestinal cell, in the original wild isolate (I), after bleaching (bl) and after re-infection by a 0.2 µM filtrate (RI). Note that not all symptoms shown in (A–I) were scored, because some are difficult to score or may also occur in healthy animals. The animals were scored 4 d after re-infection for *C. briggsae* JU1264, and 7 d after re-infection for *C. elegans* JU1580, at 23°C. The symptoms are similar in both species, and generally more frequent in JU1264. *** p value on number of worms showing infection symptoms $<7.10^{-11}$, Fisher's exact test; ** p value$<3.10^{-6}$; * p value$<3.10^{-2}$.

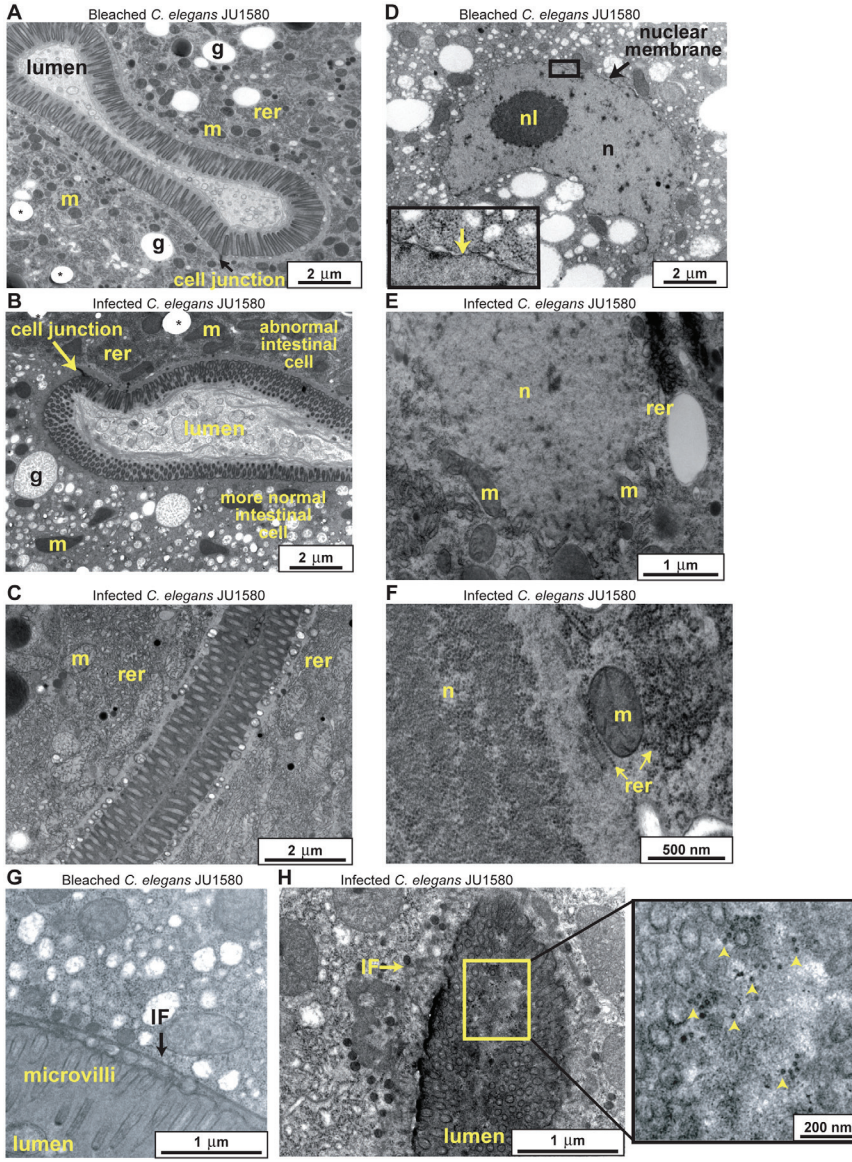


FIGURE A13-2 Transmission electron micrographs of intestinal cells of *C. elegans* JU1580 adult hermaphrodites. (A,D,G) Bleached animals. (B–C, E–F, H) Naturally infected animals. (A–C) The infection provokes a reorganization of cytoplasmic structures, most visibly the loss of intestinal lipid storage granules (g). The cytoplasm of infected intestinal cells mostly contains rough endoplasmic reticulum (rer) and mitochondria (m).

were highly divergent from all previously described nodaviruses and most closely related to each other (Figure A13-3B,C). We propose that these sequences represent two novel virus species and have tentatively named them Santeuil virus (from JU1264) and Orsay virus (from JU1580).

Viral Detection and Confirmation of Viral Infection

RT-PCR assays were used to analyze RNA extracted from JU1580, JU1264, their corresponding bleached control strains JU1580bl and JU1264bl, and the same strains following reinfection with viral filtrates. Orsay virus RNA could be detected by RT-PCR in the original JU1580 culture, disappeared in the bleached strains, and stably reappeared following re-infection with the corresponding viral filtrate (Figure A13-4A). The same pattern applied for the Santeuil virus and JU1264 animals (Figure A13-4B). JU1580 and JU1264 cultures continuously propagated for 6 mo by transferring a piece of agar (approx. 0.1 cm^3) to the next plate twice a week continued to yield positive RT-PCR results (unpublished data).

Northern blotting confirmed the presence of Orsay and Santeuil virus RNA sequences in the infected animals. Hybridization with a DNA probe targeting the RNA1 segment of Santeuil virus yielded multiple bands in JU1264 animals but not in the corresponding bleached control strain. The strongest band detected migrated between 3.5 and 4 kb consistent with the 3,628 nt sequence we generated for the putative complete RNA1 segment (Figure A13-4C). Multiple higher molecular weight bands were also detected that may represent multimeric forms of the viral genomic RNAs, which have previously been described for some nodaviruses (Ball, 1994; Johnson et al., 2000). Northern blotting with a probe targeting the RNA2 segment (Figure A13-4C) yielded a major band that migrated at ~2.5 kb as well as fainter, higher molecular weight bands. Similar patterns were seen for both segments of Orsay virus (unpublished data).

---

* hole in the resin used for inclusion in electron microscopy. (D–F) A nucleus in a non-infected animal is surrounded by a nuclear membrane (see inset in D), whereas the nuclear membrane disappears upon infection (E–F). Absence or incomplete nuclear membrane was observed repeatedly in infected animals, while the nuclear membrane could be observed on bleached animals (using both fixation methods). The nuclear material (n) in (F) may represent nucleolar material and at lower magnification (not shown) matches the shape of elongated nucleoli as observed by Nomarski optics (Figure A13-1E–G). The rough endoplasmic reticulum (rer) on the left of the mitochondrion (m) in (F) may be a remnant of the nuclear envelope. (G–H) The infection may result in disorganization of the intermediate filament (IF) network normally located below the apical plasma membrane. On the right of (H) is shown a higher magnification of the intestinal lumen, showing putative viral particles (arrowheads). The animals were fixed using high-pressure freezing (A–C, E–F) or conventional fixation (D, G, H).

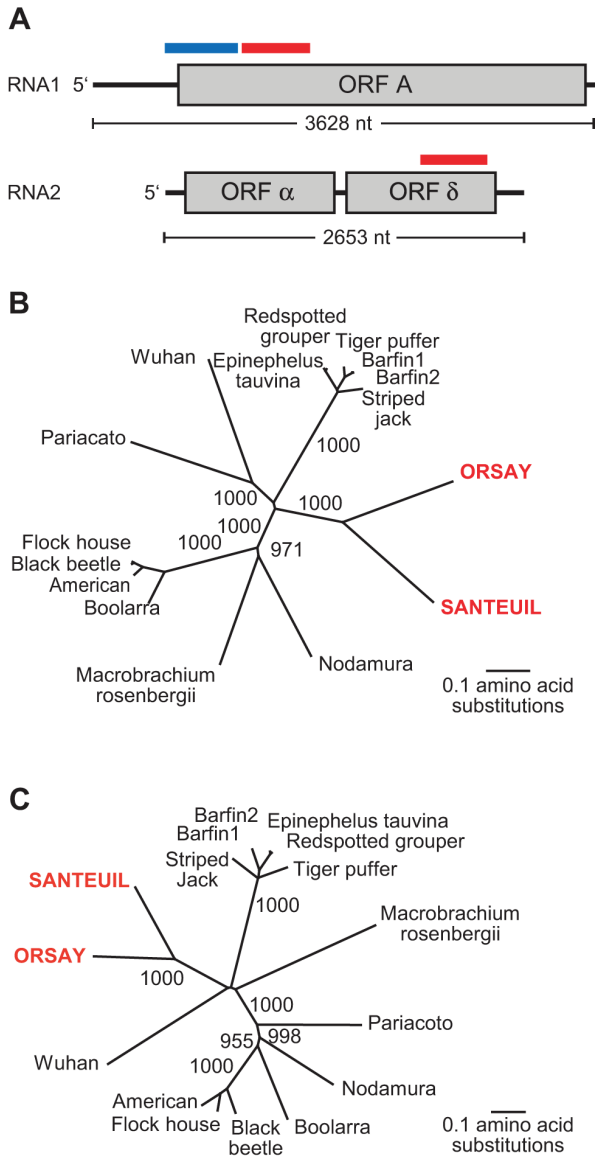


FIGURE A13-3 Genomic organization and phylogenetic analysis of novel viruses. (A) Schematic of genomic organization of Santeuil virus. Predicted open reading frames are displayed in gray boxes. Red bar indicates sequence used to generate double-stranded DNA probes for Northern blotting. Blue bar indicates sequence used to generate single-stranded riboprobes. (B) Neighbor-joining phylogenetic analysis of the predicted RNA-dependent RNA polymerases encoded by the RNA1 segments. (C) Neighbor-joining phylogenetic analysis of the predicted capsid proteins encoded by the RNA2 segments.

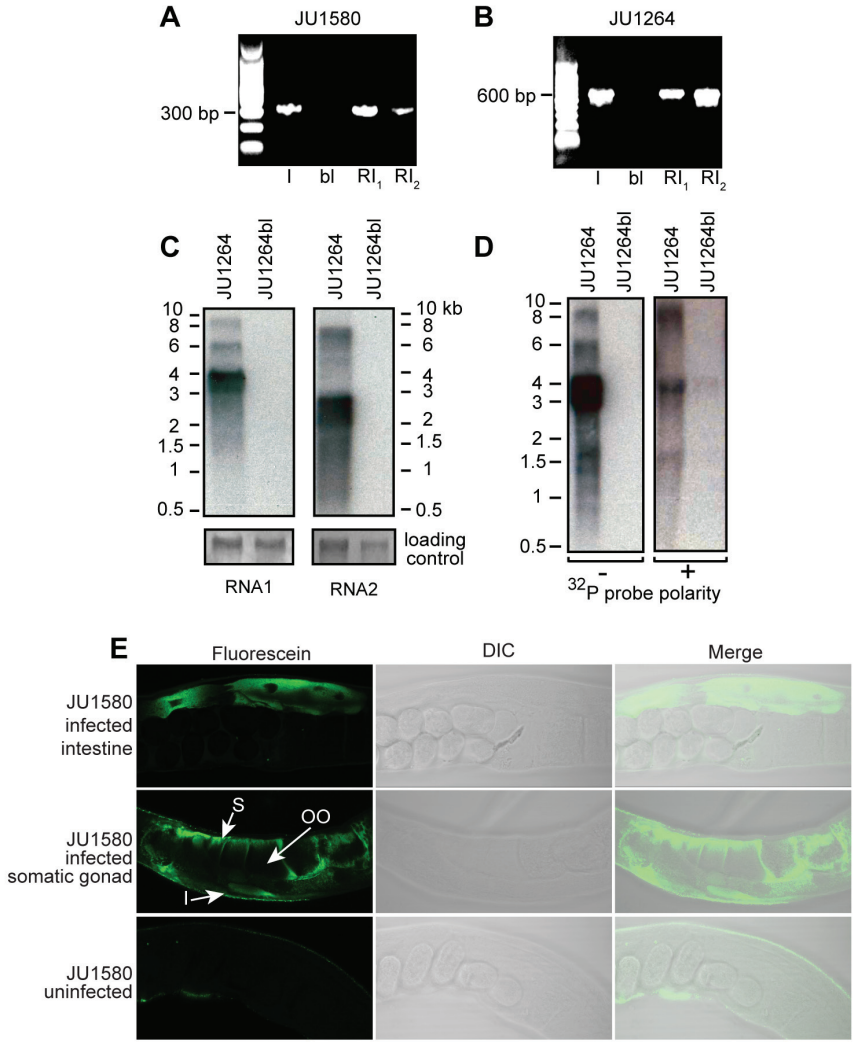


FIGURE A13-4 Molecular evidence of viral infection. (A) RT-PCR detection of the Orsay virus in the original JU1580 wild isolate (I), after bleaching (bl) and after re-infection by a 0.2 µM filtrate after 7 d (RI_1) and 3 wk (RI_2) of culture at 23°C. (B) RT-PCR detection of the Santeuil virus in the original wild isolate (I), after bleaching (bl) and after re-infection by a 0.2 µM filtrate after 4 d (RI_1) and 4 wk (RI_2) at 23°C. (C) Northern blots of Santeuil virus RNA1 and RNA2 segments hybridized using a double-stranded DNA probe. (D) Northern blots of Santeuil virus RNA1 segment using + and − sense riboprobes. (E) RNA FISH with a probe targeting Orsay virus RNA1 segment. Representative JU1580bl animals following infection by Orsay virus (top and middle rows) or uninfected (bottom row). S corresponds to ovary sheath cells, OO is an oocyte, and I is an intestinal cell.

To demonstrate virus replication in the infected animals, we performed Northern blotting using strand-specific riboprobes. For positive sense RNA viruses like nodaviruses, the negative sense RNA is only synthesized during active viral replication. It is not packaged in virions and typically exists in much lower quantities than the positive strand. Robust levels of the positive strand of the Santeuil virus RNA1 segment were detected (Figure A13-4D). Northern blotting with a riboprobe designed to hybridize to the negative sense strand detected a band of ~3.5 kb as well as higher molecular weight bands and a lower band of ~1.5 kb (~30-fold longer exposure than the positive sense blot; Figure A13-4D). While the precise nature of the high and low molecular weight species remains to be defined, the presence of multiple RNA species of negative sense polarity in JU1264 animals demonstrates bona fide replication of Santeuil virus in JU1264.

In order to determine the localization of Orsay viral RNA in infected animals, we performed RNA fluorescent in situ hybridization (FISH) using a probe complementary to the positive sense RNA1 segment of Orsay virus. Viral RNA was robustly detected in intestinal cells of JU1580bl animals infected 4 d previously with Orsay viral filtrate (Figure A13-4E, top panels). Interestingly, some animals also showed localization of viral RNA in the somatic gonad (Figure A13-4E, middle panels). JU1580bl animals not treated with the viral filtrate displayed no fluorescent signal (Figure A13-4E, bottom panels).

High Specificity of Infection by Orsay and Santeuil Nodaviruses

We tested whether the Orsay and Santeuil viruses could cross-infect the cured wild isolate of the other *Caenorhabditis* species, as well as the reference laboratory strains of *C. elegans* and *C. briggsae*. The Orsay and Santeuil viruses could only infect strains of *C. elegans* and *C. briggsae*, respectively (Figure A13-5A,B and Figure S3). Furthermore, each virus showed intraspecific specificity of infection. Indeed, we could not detect any replication of the Santeuil virus in *C. briggsae* AF16. The N2 laboratory *C. elegans* strain, while infectable by Orsay virus, appeared to be more resistant to viral infection than JU1580bl. Quantitative RT-PCR demonstrated that viral RNA accumulated in the N2 strain at levels above background but 50–100-fold lower than in JU1580bl (Figure A13-5C).

Small RNA Response upon Infection

One key defense mechanism of plants and animals against RNA viruses is the small RNA response (Aliyari and Ding, 2009). We therefore determined by deep sequencing of small RNAs whether the infected animals produced small RNAs in response to viral infection. We generated small RNA libraries from mixed-stage JU1580 animals infected with the Orsay virus and from the bleached control strain and analyzed them using Illumina/Solexa high-throughput sequencing. These libraries represent small RNAs of 18–30 nucleotides in length

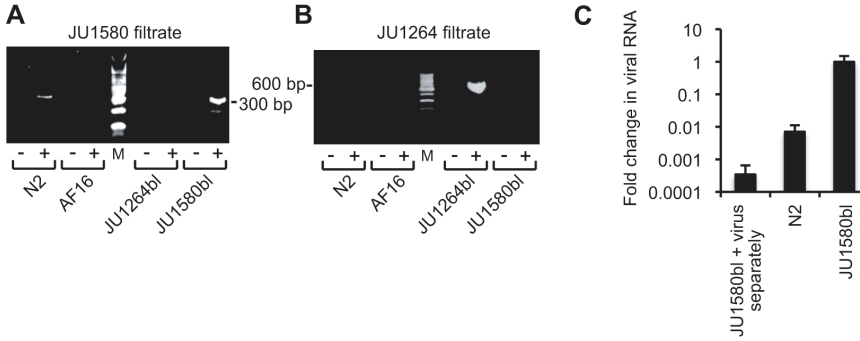


FIGURE A13-5 Specificity of infection by the Orsay and Santeuil viruses. (A) Specificity of infection by the Orsay virus. Each *Caenorhabditis* strain (name indicated below the gel) was mock-infected (–) or infected with a virus filtrate (+). RT-PCR on cultures after 7 d at 23°C. See Figure S3 for corresponding morphological symptom scoring. (B) Specificity of infection by the Santeuil virus. RT-PCR results after 4 d at 23°C. (C) Quantitative variation in viral replication N2 versus JU1580. N2 and JU1580 were tested by qRT-PCR for infection with Orsay virus extract (n = 10 independent replicates for each strain). By conventional RT-PCR assay, Orsay virus infection of N2 yielded positive bands in 3 out of 10 replicate infections whereas 7 out of 10 replicate infections of JU1580bl were positive in these conditions. Control RNA (n = 6) was extracted from JU1580bl animals grown in parallel without virus filtrate, and to which filtrate was added at the time of sample collection. RNA levels were normalized to *ama-1* and shown as average fold-change relative to JU1580bl. Error bars represent SEM.

independent of their 5′ termini. Small RNAs from infected JU1580 animals that mapped to viral RNA1 or RNA2 and had no match to the *C. elegans* genome are shown in Figure A13-6A and A13-6B, respectively. Of a total of 1,149,633 unambiguously mapped unique sequences, almost 2% (21,392) mapped to the two RNA segments of Orsay virus. Such RNAs were virtually absent from a library generated from bleached JU1580 animals (<0.001%) (unpublished data). Small RNAs that corresponded to the sense strand of the viral RNAs had a broad length distribution and no 5′ nucleotide preference. These sense small RNAs might represent Dicer cleavage products or other viral RNA degradation intermediates. In contrast, most antisense small RNAs were 22 nt long and showed a bias for guanidine as the first base (Figure A13-6A, B). This signature is reminiscent of a class of secondary RNAs named 22G RNAs that are thought to be downstream effectors of exogenous and endogenous small RNA pathways (Gu et al., 2009; Pak and Fire, 2007; Sijen et al., 2007). Such RNAs are not associated with transgenes expressed in the soma of *C. elegans* from extrachromosomal arrays (Pak and Fire, 2007) nor generally a feature associated with active transcription of endogenous genes (Gu et al., 2009; Pak and Fire, 2007; Sijen et al., 2007). These

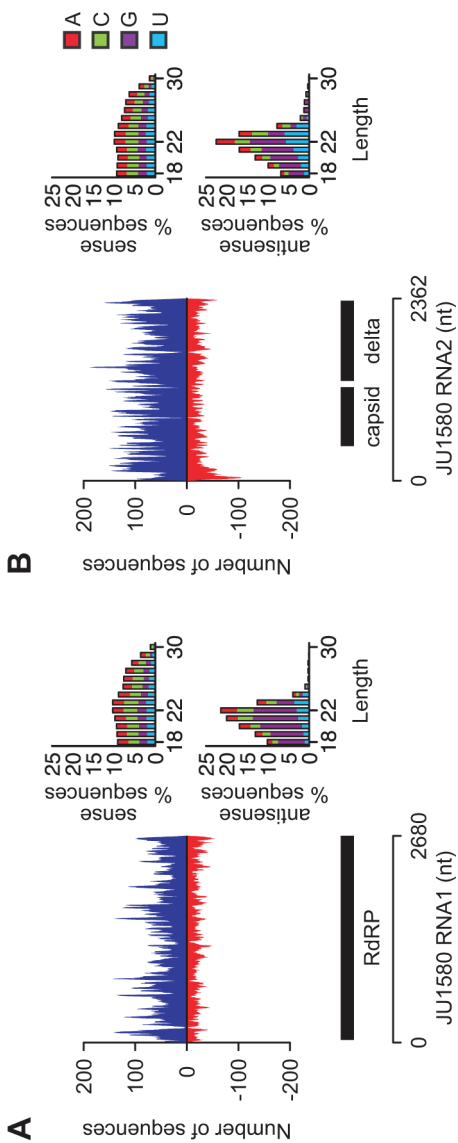


FIGURE A13-6 Small RNAs produced upon viral infection. Number of unique sequences obtained by Illumina/Solexa high-throughput sequencing of a 5′-independent small RNA library from JU1580 matching a given position in the Orsay virus segment RNA1 (A) or RNA2 (B). The number of sequences in sense and antisense orientation are shown on the positive (blue) and negative (red) *y*-axis, respectively. Only sequences with a perfect and unambiguous match to the virus genome were considered. The location of virus protein-coding genes is indicated below each graph as black bars and the RNA genome as a line. Features of sense and antisense sequences (length and identity of first nucleotide) are shown to the right of each graph.

data suggest that JU1580 animals raise a small RNA response to viral infection. We also detected small RNAs of both sense and antisense polarity that mapped to the Santeuil virus genome in the JU1264 wild *C. briggsae* isolate but not in bleached animals (unpublished data).

RNAi Competency of the Host Is an Antiviral Defense

As viral infection appears to invoke a small RNA response in JU1580 animals, we next tested if mutations in small RNA pathways could affect replication of the Orsay virus. Orsay virus infection of the N2 reference strain was reduced compared to JU1580, as assayed by viral RNA qRT-PCR (Figure A13-5C) and infection symptoms (Figures S3A and A13-7B). Mutation of the *rde-1* gene—which encodes an Argonaute protein required for the initiation of exogenous RNAi (Tabara et al., 1999)—in the N2 background increased viral RNA abundance and morphological symptoms to levels comparable to JU1580 using both assays (Figure A13-7A,B). The infected *rde-1* strain produced infectious viral particles, as reinfection of the cured JU1580 strain by filtrates of infected *rde-1* animals yielded positive RT-PCR results (unpublished data). In addition, mutation of other exogenous RNAi pathway genes including *rde-2*, *rde-4*, and *mut-7* (Table A13-1) also led to increased viral RNA accumulation as determined by quantitative RT-PCR (Figure A13-7A). We thus conclude that RNAi mechanisms

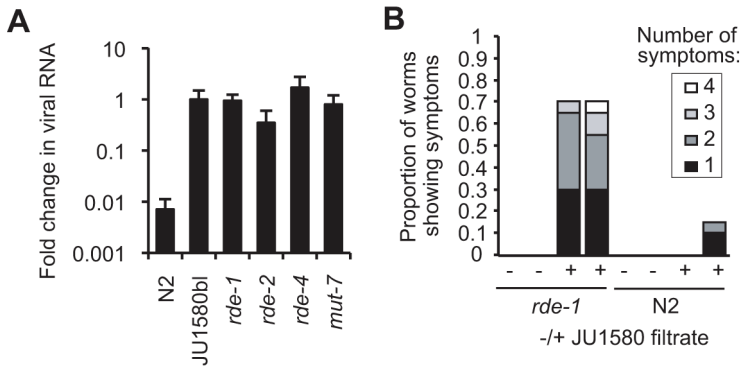


FIGURE A13-7 RNAi-deficient mutants of *C. elegans* can be infected by the Orsay virus. (A) JU1580bl, N2, *rde-1(ne219)* (n = 10 independent replicates each), *rde-2(ne221)*, *rde-4(ne301)*, and *mut-7(pk204)* (n = 5 independent replicates each) were tested by qRT-PCR for infection with Orsay virus extract. RNA levels were normalized to *ama-1* and shown as average fold-change relative to JU1580bl. Error bars represent SEM. Same results as displayed in Figure A13-5C for N2 and JU1580. (B) Scoring of symptoms in two independent replicates of infection of *rde-1* mutant and wild-type N2 animals by the Orsay virus filtrate, after 4 d.

provide antiviral immunity to *C. elegans* and that Orsay virus infection of mutant animals can be used to define genes important for antiviral defense.

Natural Variation in Somatic RNAi Efficiency in C. elegans

Since a functional RNAi pathway limits the accumulation of viral RNA in the N2 reference strain, we assessed the exogenous RNAi competency of the bleached culture of JU1580 (JU1580bl) relative to the reference N2 strain. Using external application of dsRNAs by feeding, JU1580bl was found to be highly resistant to RNAi of a somatically expressed gene (*unc-22*) but competent for RNA inactivation of a germline-expressed gene (*pos-1*) (Figure A13-8A, B). *C. elegans* wild isolates, such as CB4856, were previously known to be variably sensitive to germline RNAi (Tijsterman et al., 2002). Here we thus observed for the first time a large variation in sensitivity to somatic RNAi, which does not correlate with germline RNAi sensitivity and thus cannot be due to inability to intake dsRNA from the intestinal lumen. We confirmed insensitivity to somatic RNAi of the JU1580bl isolate using a ubiquitously expressed GFP transgene (*let-858::GFP*), which was inactivated by GFP RNAi in the *C. elegans* N2 reference background, but only modestly repressed in the JU1580bl isolate (Figure A13-8C, Figure S4). We confirmed that the insensitivity to somatic RNAi also applied when *unc-22* dsRNAs were directly injected into the syncytial germline (Figure A13-8D). Therefore, the robust accumulation of Orsay virus RNA observed in infected JU1580 may be rendered possible in part by the partial defect in the somatic RNAi pathway of this wild isolate. The accumulation of small RNAs in response to the virus in infected JU1580 indicates, however, that its RNAi response is at least partially active in some tissues, perhaps including the germline.

The germline RNAi competence of JU1580 together with the presence of Orsay virus RNA1 in the somatic gonad raises the possibility that vertical transmission of viral infection could occur in a strain defective for germline RNAi. To examine this possibility, JU1580bl, N2, and *rde-1* were exposed to Orsay virus filtrate. A subset of adult animals from each plate was bleached and their adult offspring collected 4 d later. No evidence for vertical transmission was observed by qRT-PCR for Orsay virus RNA in any strain (Figure S5).

We further tested the efficiency of the RNAi response in six other wild C. *elegans* isolates representative of its worldwide diversity (Figure A13-8A, B, D). Our results suggest that the somatic RNAi response varies quantitatively in *C. elegans* and is not correlated with germline RNAi sensitivity. Under experimental conditions that yield efficient infection of JU1580bl by Orsay virus, none of the other strains yielded significant levels of morphological symptoms (Figure A13-8E). Only JU1580bl and JU258 were positive by RT-PCR (unpublished data). Thus, factors other than RNAi competency also contribute to the sensitivity of *C. elegans* to the Orsay virus.

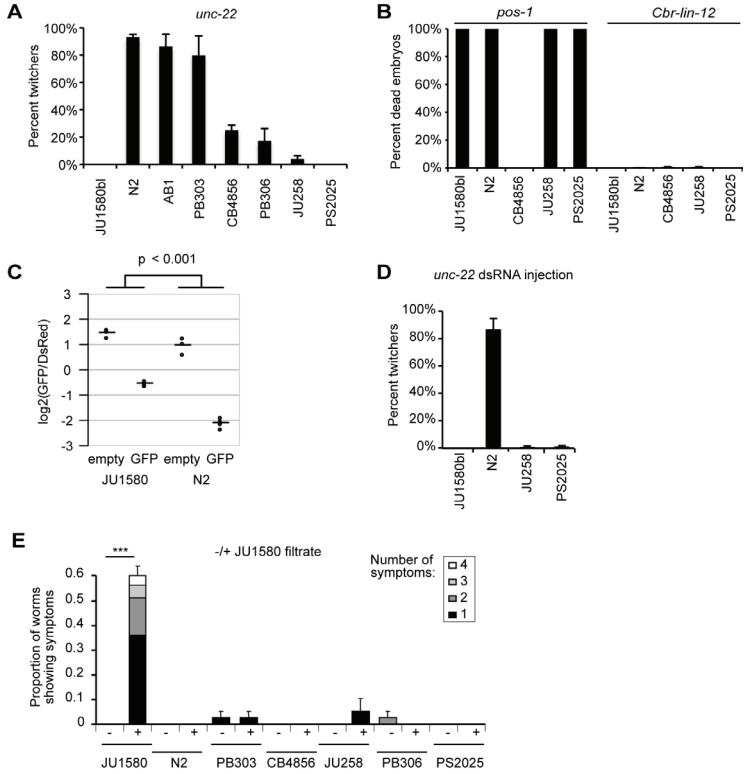


FIGURE A13-8 Natural variation in somatic RNAi efficacy in *C. elegans*. (A) Somatic RNAi was tested using bacteria expressing dsRNA specific for the *unc-22* gene (acting in muscle [Fire et al., 1998]). The percentage of animals with the corresponding twitcher phenotype is shown for different *C. elegans* wild isolates (representative of the species' diversity [Milloz et al., 2008]). Bar: standard error over four replicate plates. (B) Germline RNAi was tested by feeding the animals with bacteria expressing dsRNA specific for the *pos-1* gene. The percentage of animals with the corresponding embryonic-lethal phenotype is shown for five wild genetic backgrounds of *C. elegans*. *Cbr-lin-12* RNAi is a negative control. Bar: standard error over six replicate plates (too small to be seen). n>450 observed individuals for each treatment. (C) Somatic RNAi was tested using bacteria expressing dsRNA specific for GFP. Each point corresponds to the median $\log_2$(GFP/DsRed) intensity ratio from one flow cytometry run of strains carrying the *let-858::GFP* transgene in the JU1580 and N2 backgrounds, after treatment with GFP RNAi or empty vector. Horizontal bars indicate group means. The difference in $\log_2$ intensity ratios between GFP RNAi and empty vector is reduced in JU1580 compared to N2 (p<0.001, see Methods). (D) *unc-22* dsRNA was administered by injection into the syncytial germline of the mother. 10–14 animals of each genotype were injected and 30 progeny were scored for the twitcher phenotype on each plate. (E) Orsay virus sensitivity of seven wild *C. elegans* isolates representative of the species' diversity. Morphological symptoms were scored 5 d after infection of clean cultures by the Orsay virus filtrate at 23°C. The JU1580 control was performed in duplicate. Bar: standard error on total proportion. *** p<0.001.

Discussion

The First Viruses Infecting Caenorhabditis

Here we report the first molecular description, to our knowledge, of viruses that naturally infect nematodes in the wild. The two novel viruses we identified, while clearly related to known nodaviruses, possess unique genomic features absent from all other previously described nodaviruses. These viruses may thus define a novel genus within the family *Nodaviridae* or may even represent prototype species of a new virus family (pending formal classification by the International Committee for the Taxonomy of Viruses). The same range of intestinal symptoms was observed in animals that were infected by the Orsay and Santeuil viruses, further suggesting that these viral infections were causing the cellular symptoms. We observed putative viral particles of the size expected for nodaviruses, and a strong RNA FISH signal in intestinal cells and the somatic gonad of infected animals demonstrating that the virus is present intracellularly. It is likely that further sampling of natural populations of *Caenorhabditis* will yield other viruses of this and other groups. In fact, these symptoms were seen repeatedly in *C. briggsae* animals sampled from different locations in France, and in one instance, a Santeuil virus variant has been identified (unpublished data).

A characteristic feature of these two viruses is the presence of the novel ORF δ. Conservation of sequence length and identity of the ORF δ in these two viruses, and the absence of this ORF in all other described nodaviruses, suggests that this predicted protein is likely to be important for the ability of the virus to infect or replicate in nematodes. Its function is currently unknown, but it is tempting to speculate that this protein may play a role in antagonizing an innate antiviral pathway.

A Laboratory Viral Infection of a Small Model Animal

The infection of *C. elegans* by the Orsay nodavirus provides an exciting prospect for studies in virology, host cell biology, and antiviral innate immunity. Genetic screens to identify anti-viral factors in model organisms have been limited in large part by the lack of natural infection systems. Although *Drosophila* has been used with great success to examine host-virus interactions for various insect viruses (Huszar and Imler, 2008) and influenza (Hao et al., 2008), none of these studies has examined viral infection of the host organism by natural transmission routes. Here we present a novel association between *C. elegans* and a virus that persists in culture through horizontal transmission, causing high damage in intestinal cells yet remarkably little effect on the animal, which continues moving, eating, and producing progeny, although at a lower rate.

De novo infection of naïve animals can be affected by the simple addition of either dead infected animals or homogenized lysates made from infected animals to culture dishes. This is sufficient to seed sustained complete cycles of

viral replication, shedding, and infection. With this system, it is now possible to embark on whole genome genetic screens to identify host factors that block any facet of the viral life cycle. Using the current experimental conditions, infection of JU1580bl and *rde-1* mutants in N2 background was highly reproducible. The fact that the reference wild type N2 strain may only sustain a very low yet detectable viral titer makes it a particularly favorable genetic background in which to screen for genes involved in interaction with the virus.

The intestine is a tissue that is particularly exposed to microbes through ingestion, and is a main entry point for pathogens in *C. elegans* as in other animals. In *C. elegans*, the intestinal cells are large and easily amenable to observations by optical microscopy. The viral parasites affect the organization of the polarized epithelial intestinal cells and will likely provide interesting mechanisms and tools to study their cell biology. Clear reorganization occurs in the intermediate filaments that line the apical brush border, as well as in the lipid storage granules, the nuclear membrane, and other intracellular compartments.

The abnormal state of the intestinal cells may slow down progeny production by decreasing the food intake. Alternatively, the presence of viral RNA in the somatic gonad may explain the delay in progeny production, although no gonadal cellular phenotypes have been observed. The presence of viral RNA in the somatic gonad is particularly interesting given the lack of vertical transmission.

Targeted Mutant Screens with Orsay Virus Confirm a Role for RNAi in Antiviral Defense

Although prior studies have clearly demonstrated a role for *C. elegans* RNAi in counteracting viral infection, these studies utilized either a transgenic system of viral RNA expression (Lu et al., 2005) or primary culture cells (Schoot et al., 2005; Wilkins et al., 2005). The observed susceptibility of Orsay virus RNA to RNAi processing in JU1580 animals provides the first evidence in a completely natural setting, without any artificial manipulations, that RNAi serves an antiviral role in nematodes. Coupled to the increase in accumulation of Orsay virus RNA in RNAi pathway mutant strains as compared to wild type N2, these studies demonstrate that the RNAi pathway is an important antiviral defense against Orsay virus. Moreover, these results demonstrate the feasibility of identifying antiviral genes or pathways in this experimental infection system. The mechanism by which the animals prevent transmission to their offspring is unclear, but our initial results with *rde-1* mutants suggest that perturbing germline RNAi is not sufficient to enable vertical transmission.

Evolution of Viral Sensitivity and Specificity in Natural Populations

The quantitative difference in Orsay nodavirus sensitivity between the N2 and JU1580 wild *C. elegans* genetic backgrounds will allow the identification of

a set of host genes that modulate viral sensitivity during evolution of natural host populations. Based on the defect in exogenous RNAi of the JU1580 strain, we speculate that this set will include, but is unlikely to be limited to, genes involved in exogenous RNAi pathways. Support for the role of other genes outside the RNAi pathway comes from our data on natural isolates. Despite the fact that the magnitude of the somatic RNAi defect of the natural isolate PS2025 was comparable to that of JU1580, no evidence of viral RNA accumulation or morphological symptoms was observed following addition of Orsay virus filtrate. Whether PS2025 lacks one or more crucial receptors for viral infection or has alternative antiviral pathways that suppress viral replication is currently unknown.

In addition, the Orsay and Santeuil viruses appear to specifically infect *C. elegans* and *C. briggsae*, respectively. Moreover, the *C. elegans rde-1* mutation in the N2 background confers susceptibility to the Orsay virus, but not to the Santeuil virus (Figure S3C). The two viruses thus provide a system to study host-parasite specificity and its evolution. With the isolation of additional variants of each virus (our unpublished data), viral evolution studies can also be undertaken. Host-parasite evolutionary and ecological interactions can thus be explored at two evolutionary scales, within and between species of both host and parasite. The rapid life cycle of *C. elegans* also allows experimental evolution in the laboratory (Azevedo et al., 2002; Schulte et al., 2010). This model system, which can include both natural and engineered variants of both virus and host, is thus favorable for combining studies of host-pathogen co-evolution in the laboratory and in natural populations.

Materials and Methods

Nematode Field Isolation

Caenorhabditis nematodes were isolated on *C. elegans* culture plates seeded with *E. coli* strain OP50 using the procedures described in (Barrière and Félix, 2006). JU1264 was isolated from a snail collected on rotting grapes in Santeuil (Val d'Oise, France) on 14 Oct 2007. JU1580 was isolated from a rotting apple sampled in Orsay (Essonne, France) on 6 Oct 2008. When required, cultures were cleared of natural bacterial contamination by frequent passaging of the animals and/or antibiotic treatment (LB plates with 50 µg/ml tetracycline, ampicilline, or kanamycine for 1 h). Infected cultures were kept frozen at −80°C and in liquid N2 as described in Wood (1988). Bleaching was performed as in Wood (1998).

Light Microscopy

When observed with a transillumination dissecting microscope, infected animals displayed a paler intestine than healthy worms. This lack of intestinal coloration occurred all along the entire intestinal tract in *C. briggsae* JU1264

and preferentially in the anterior intestinal tract in *C. elegans* JU1580. Intestinal cells were observed with Nomarski optics with a 63× or 100× objective. The four symptoms used for scoring were 1, the disappearance of gut granules in at least part of a cell; 2, degeneration of the nucleus including a very elongated nuclear or nucleolus (when the rest of the nucleus has degenerated) or the apparent disappearance of the nucleus; 3, the loss of cytoplasmic viscosity visible as a very fluid flow of cytosol within the cell; and 4, the fusion of intestinal cells. Some of these traits may sometimes appear in uninfected animals. We systematically tested for a significant increase after infection of the proportion of animals with symptoms (Fisher's exact test). Note that some of these symptoms can also be caused by microsporidial and bacterial infections. Thus, the diagnostic of a viral infection based on the cellular symptoms requires an otherwise clean culture.

Live Hoechst 33342 Staining of Nuclei

Animals were washed off a culture plate in 10 ml of ddH_20, pelleted and incubated in 10 ml of 10 µg/ml Hoechst 33342 in ddH_20 for 45 min with soft agitation, protecting the tube from light with an aluminum foil. The animals were then pelleted and transferred to a new culture plate seeded with *E. coli* OP50. After 2 h, they were mounted and observed with a fluorescence microscope.

Electron Microscopy

A few adults were washed in 0.2 ml of M9 solution, suspended in 2% paraformaldehyde +0.1% glutaraldehyde, and cut in two on ice under a dissecting microscope for better reagent penetration (Hall, 1995). Worm pieces were then resuspended overnight in 2% OsO_4 at 4°C, washed, embedded in 2% low melting point agar, dehydrated in solutions of increasing ethanol concentrations, and embedded in resin (Epon-Araldite). High-pressure freezing was performed using a Leica PACT2 high-pressure freezer (Weimer, 2006).

Progeny Counts

The time course was started by isolating single L4 larvae for *C. elegans* JU1580 and single L3 larvae for *C. briggsae* JU1264. The parent animal then transferred every day to a new plate until the end of progeny production. The plates were incubated at 20°C for 2 d and kept at 4°C until scoring. The few cases where the parent died before the end of its laying period were not included. Some progeny died as embryos in both infected and non-infected cultures (non-significant effect of treatment; unpublished data). The timing of progeny production was analyzed in R using a Generalized Linear Model using infection status, day, individual (nested in infection status), and Infection Status×Day as explanatory variables, assuming a Poisson response variable and a log link

function. Individual, day and Infection Status×Day were the significant explanatory variables for both JU1264 and JU1580 ($p<0.001$).

Infectious Filtrate Preparation and Animal Infections

Nematodes were grown on 10 plates (90 mm diameter) until just starved, resuspended in 15 ml of 20 mM Tris-Cl pH 7.8, and pelleted by low-speed centrifugation (5,000 g). The supernatant was centrifuged twice at 21,000 g for 5 min (4°C) and pellets discarded. The supernatant was passed on a 0.2 µm filter. 55 mm culture plates were prepared with 2–5 young adults of N2, *rde-1(ne219)*, or JU1580bl. At the same time (Figures A13-1K, A13-4A,B, A13-5, A13-7B, and S3), or the following day (Figures A13-5C, A13-7A), 30 µl of infectious filtrate was pipetted onto the bacterial lawn. The cultures were incubated at 20°C except otherwise indicated. When both *C. elegans* and *C. briggsae* were grown in parallel, an incubation temperature of 23°C (indicated in the figure legends) was used so that both species developed at similar speeds. Maintenance over more than 4 d after re-infection was performed by transferring a piece of agar (approx. 0.1 cm^3) every 2–3 d to a new plate with food.

High-Throughput Sequencing

Phenol-chloroform purified DNA and RNA from infected JU1580 and JU1264 animals were subject to random PCR amplification as described (Wang et al., 2003). The amplicons were then pyrosequenced following standard library construction on a Roche Titanium Genome Sequencer. Raw sequence reads were filtered for quality and repetitive sequences. BLASTn and BLASTx were used to identify sequences with limited similarity to known viruses in Genbank. Contigs were assembled using the Newbler assembler. To confirm the assembly, primers for RT-PCR were designed to amplify overlapping fragments of ~1.5 kb. Amplicons were cloned and sequenced.

5′ and 3′ RACE

5′ RACE was performed according to standard protocols (Invitrogen 5′ RACE kit). 3′ RACE was performed by first adding a polyA tail using PolyA polymerase (Ambion) and then using Qiagen 1-step RT-PCR kit with gene specific primers and an oligo-dT-adapter primer. Products were cloned into pCR4 and sequenced using standard Sanger chemistry.

Small RNA Sequencing

4–6 90 mm plates with 15–20 adults (JU1580 or bleached JU1580) were grown for 4 d at 20°C. Mixed stage animals from all plates were collected, pooled, and frozen at −80°C. Total RNA was extracted using the mirVana miRNA

isolation kit (Ambion). Small RNAs were size selected to 18–30 bases by denaturing polyacrylamide gel fractionation. A cDNA library that did not depend on 5′-monophosphates was constructed by tobacco acid pyrophosphatase treatment using adapters recommended for Solexa sequencing as described previously (Das et al., 2008). Each sample was labeled with a unique four base pair barcode. cDNA was purified using the NucleoSpin Extract II kit (Macherey & Nagel). Small RNA libraries were sequenced using the Illumina/Solexa GA2 platform (Illumina, Inc., San Diego, CA). Fastq data files were processed using custom Perl scripts. Reads with missing bases or whose first four bases did not match any of the expected barcodes were excluded. Reads were trimmed by removing the first four nucleotides and any 3′ As. The obtained inserts were collapsed to unique sequences, retaining the number of reads for each sequence. Sequences in the expected size range (18–30 nucleotides) were aligned to the *C. elegans* genome (WS190) downloaded from the UCSC Genome Browser website (http://genome.ucsc.edu/) (Kent et al., 2002) and the JU1580 partial virus genome using the ELAND module within the Illumina Genome Analyzer Pipeline Software, v0.3.0. Figure A13-6 is based on unique sequences (multiple reads of the same sequence were collapsed) with perfect and unambiguous alignment to the Orsay virus genome. Small RNA sequence data were submitted to the Gene Expression Omnibus under accession number GSE21736.

Neighbor-Joining Phylogenetic Analysis

The predicted amino acid sequences from Orsay and Santeuil nodaviruses were aligned using ClustalW to the protein sequences of the following nodaviruses. Capsid Protein: Barfin1 flounder nervous necrosis virus NC_013459, Barfin2 flounder virus BF93Hok RNA2 NC_011064, Black beetle virus NC_002037, Boolarra virus NC_004145, *Epinephelus tauvina* nervous necrosis virus NC_004136, Flock house virus NC_004144, *Macrobrachium rosenbergii* nodavirus RNA-2 NC_005095, Nodamura virus RNA2 NC_002691, Pariacoto virus RNA2 NC_003692, Redspotted grouper nervous necrosis virus NC_008041, Striped Jack nervous necrosis virus RNA2 NC_003449, Tiger puffer nervous necrosis virus NC_013461, Wuhan nodavirus ABB71128.1, and American nodavirus ACU32796.1. 1,000 bootstrap replicates were performed.

RNA Polymerase: Barfin flounder nervous necrosis virus YP_003288756.1, Barfin flounder virus BF93Hok YP_002019751.1, Black beetle virus YP_053043, Boolarra virus NP_689439, *Epinephelus tauvina* nervous necrosis virus NP_689433.1, Flock house virus NP_689444.1, Nodamura virus NP_077730, Pariacoto virus NP_620109.1, Redspotted grouper nervous necrosis virus YP_611155.1, Striped Jack nervous necrosis virus NP_599247.1, Tiger puffer nervous necrosis virus YP_003288759.1, *Macrobrachium rosenbergii* nodavirus NP_919036.1, Wuhan_Nodavirus AAY27743, and American nodavirus SW-2009a ACU32794.1. 1,000 bootstrap replicates were performed.

RT-PCR

Nematodes from two culture plates were resuspended in M9 and then washed three times in 10 ml M9. RNA was extracted using Trizol (Invitrogen) (5–10 vol:vol of pelleted worms) and resuspended in 20 µl in RNAse-free ddH_2O. 5 µg of RNA were reverse transcribed using SuperscriptIII (Invitrogen) in a 20 µl volume. 5 µl were used for PCR in a 20 µl volume (annealing temperature 60°C, 35 cycles). For the Orsay nodavirus, the reverse transcription used the GW195 primer (5′ GACGCTTCCAAGATTGGTATTGGT) and the PCR oTB3 (5′ CGGATTCTCGACATAGTCG) and oTB4 (5′GTAGGCGAGGAAGGAGATG). For the Santeuil nodavirus, reverse transcription used oTB6RT (5′ GGTTCTGGTGGTGATGGTG) and PCR oTB5 (5′ GCGGATGTTCTTCACGGAC) and oTB6 (5′ GTCAGTAGCGGACCAGATG).

One-Step RT-PCR

Animals from one 55 mm culture plate plus viral filtrate (see infection procedure) were washed twice in M9. RNA was extracted using 1 ml Trizol (Invitrogen) and resuspended in 10 µl DEPC-treated H_2O. 0.1 µl was used for RT-PCR using the OneStep RT-PCR Kit (Qiagen). Primers annealed to viral RNA1 (GW194 and GW195).

qRT-PCR

cDNA was generated from 1 µg total RNA with random primers using Superscript III (Invitrogen). cDNA was diluted to 1:100 for qRT-PCR analysis. qRT-PCR was performed using either QuantiTect SYBR Green PCR (Qiagen) or ABsolute Blue SYBR Green ROX (Thermo Scientific). The amplification was performed on a 7300 Real Time PCR System (Applied Biosystems). Each sample was normalized to *ama-1*, and then viral RNA1 (primers GW194: 5′ ACC TCA CAA CTG CCA TCT ACA and GW195: 5′ GAC GCT TCC AAG ATT GGT ATT GGT) levels were compared to those present in re-infected bleached JU1580 animals.

Northern Blotting

For Northern blots, 0.5 µg of total RNA extracted from JU1264 and JU1264bl animals were electrophoresed through 1.0% denaturing formaldehyde-MOPS agarose gels. RNA was transferred to Hybond nylon membranes and then subject to UV cross-linking followed by baking at 75°C for 20 min. Double stranded DNA probes targeting the RNA1 segment of Santeuil nodavirus (nt 1141–1634) and the RNA2 segment of Santeuil nodavirus (nt 1833–2308) were generated by random priming in the presence of α-^{32}P dATP using the Decaprime kit (Ambion). Blots were hybridized for 4 h at 65°C in Rapid hyb buffer (GE

Healthcare) and washed in 2XSSC/0.1%SDS 5 min×2 at 25°C, 1XSSC/0.1%SDS 10 min×2 at 25°C, 0.1XSSC/0.1%SDS 5 min×4 at 25°C, and 0.1XSSC/0.1%SDS 15 min×2 at 42°C and 0.1XSSC/0.1%SDS 15 min×1 at 68°C. For strand specific riboprobes, ^{32}P labeled RNA was generated by in vitro transcription with either T7 or T3 RNA polymerase (Ambion) in the presence of α-^{32}P UTP. The target plasmid contained a cloned region of the Santeuil nodavirus RNA1 segment (nt 523–1022) and was linearized with either PmeI or NotI, respectively. For the riboprobes, blots were hybridized at 70°C and then sequentially washed as follows: 2XSSC/0.1%SDS 5 min×2 at 68°C, 1XSSC/0.1%SDS 10 min×2 at 68°C, 0.1XSSC/0.1%SDS 10 min×2 at 68°C, and 0.1XSSC/0.1%SDS 20 min×1 at 73°C. The Santeuil RNA1 segment migrates at approximately the same position as the 28S ribosomal RNA. Under the extended exposure time (72 h) needed to visualize the negative sense genome, low levels of non-specific binding to the 28S RNA become apparent (Figure A13-4D).

RNA Interference

For *pos-1* and *unc-22* RNAi using bacteria as the dsRNA source, bacterial clones from the Ahringer library expressing dsRNAs (Kamath et al., 2003) (available through MRC Geneservice) were used to feed *C. elegans* on agar plates. For the *pos-1* experiment, bacteria were concentrated 10-fold by centrifugation prior to seeding the plates. A *C. briggsae Cbr-lin-12* fragment (Félix, 2007) was used as a negative control as it does not match any sequence in *C. elegans*. Three or four L4s were deposited on an RNAi plate, singly transferred the next day to a second RNAi plate, and their progeny scored after 2 d (*pos-1*) or 3 d (*unc-22*) at 23°C.

For *unc-22* dsRNA synthesis and injection, the *unc-22* fragment in the Ahringer library clone was amplified by PCR using the T7 primer and in vitro transcribed with the T7 polymerase using the Ambion MEGAscript kit, according to the manufacturer's protocol (Ahringer, 2006). *Cel-unc-22* dsRNAs were injected at 50 ng/µl into both gonadal arms of young hermaphrodite adults of the relevant strain. The animals were incubated at 20°C. The adults were transferred to a new plate individually on the next day, and the proportion of twitching progeny scored 3 d later, touching each animal with a platinum-wire pick to induce movement.

For GFP RNAi, transgenic N2 and JU1580 strains were generated expressing the ubiquitously expressed *let-858::GFP* and the pharyngeal marker *myo-2::DsRed* as an extrachromosomal array. Bacteria expressing dsRNA against GFP cDNA were used to feed animals on agar plates. An empty vector was used as a negative control. Two or three L4s were deposited on a 55 mm RNAi plate, grown at 20°C for 3 d, and the GFP/DsRed expression levels in their offspring measured using flow cytometry (Union Biometrica) as described previously (Lehrbach et al., 2009). Offspring from two RNAi plates were combined for sorting. Each

combination of RNAi vector and strain was repeated in at least triplicate. GFP and DsRed intensities were obtained from 14 wormsorter runs including 3–4 replicate runs for N2 and JU1580 after treatment with GFP RNAi or empty vector. A larger proportion of N2 animals showed reporter expression compared to JU1580 animals (Figure S4, top). To control for this difference between strains, animals with no reporter expression were excluded by requiring DsRed intensities to exceed a cutoff set to the median 99th percentile from three control runs of animals with no array present (Figure S4). A linear regression model was fitted to the median log2(GFP/DsRed) intensity ratios including strain, treatment, and an interaction term as explanatory variables. The interaction term was significantly different from zero at $p<0.001$.

RNA Fluorescent In Situ Hybridization (FISH)

A segment of Orsay virus RNA1 was generated with primers GW194 and GW195 and cloned into pGEM-T Easy (Promega). Fluorescein labeled probe was generated from linearized plasmid using the Fluorescein RNA Labeling Mix (Roche) and MEGAscript SP6 transcription (Ambion). JU1580bl animals were infected with Orsay virus filtrate and grown for 4 d at 20°C on 90 mm plates. Control animals were grown under the same conditions in the absence of virus. In situ hybridization was performed essentially as previously described (Motohashi et al., 2006). The fluorescent RNA probe was visualized directly on an Olympus FV1000 Upright microscope.

Genbank Sequences: Accession numbers for Orsay and Santeuil virus contigs: HM030970–HM030973. Small RNA sequencing data at GEO: GSE21736.

Acknowledgments

M.A.F. thanks the caretakers of the Orsay orchard for access to it and J.-L. Bessereau, F. Duveau, R. Legouis, N. Naffakh, and B. Samuel for helpful discussions.

Author Contributions

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: MAF EAM DW. Performed the experiments: MAF AA JP GW IN TB YJ CJF MS. Analyzed the data: MAF AA JP GW IN TB GZ CJF LDG MS EAM DW. Wrote the paper: MAF AA EAM DW.

References

Ahringer J, editor. (2006) Reverse genetics. Wormbook, ed The *C elegans* research community. Available http://www.wormbook.org. Accessed 29 December 2010.

Aliyari R, Ding S. W (2009) RNA-based viral immunity initiated by the Dicer family of host immune receptors. Immunol Rev 227: 176–188.

Azevedo R. B, Keightley P. D, Lauren-Maatta C, Vassilieva L. L, Lynch M, et al. (2002) Spontaneous mutational variation for body size in *Caenorhabditis elegans*. Genetics 162: 755–765.

Ball L. A (1994) Replication of the genomic RNA of a positive-strand RNA animal virus from negative-sense transcripts. Proc Natl Acad Sci U S A 91: 12443–12447.

Barrière A, Félix M-A, Community The *C. elegans* Research Community, editor (2006) Isolation of *C. elegans* and related nematodes. Wormbook.

Das P. P, Bagijn M. P, Goldstein L. D, Woolford J. R, Lehrbach N. J, et al. (2008) Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. Mol Cell 31: 79–90.

Félix M-A (2007) Cryptic quantitative evolution of the vulva intercellular signaling network in *Caenorhabditis*. Curr Biol 17: 103–114.

Fire A, Xu S, Montgomery M. K, Kostas S. A, Driver S. E, et al. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391: 806–811.

Gu W, Shirayama M, Conte D Jr, Vasale J, Batista P. J, et al. (2009) Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans germline*. Mol Cell 36: 231–244.

Hall D. H (1995) Electron microscopy and three-dimensional image reconstruction. In: Epstein H. F, Shakes D. C, editors. Methods in cell biol, *Caenorhabditis elegans* modern biological analysis of an organism. San Diego: Academic Press. pp. 395–436.

Hao L, Sakurai A, Watanabe T, Sorensen E, Nidom C. A, et al. (2008) Drosophila RNAi screen identifies host genes important for influenza virus replication. Nature 454: 890–893.

Hüsken K, Wiesenfahrt T, Abraham C, Windoffer R, Bossinger O, Leube R. E (2008) Maintenance of the intestinal tube in *Caenorhabditis elegans*: the role of the intermediate filament protein IFC-2. Differentiation 76: 881–896.

Huszar T, Imler J. L (2008) Drosophila viruses and the study of antiviral host-defense. Adv Virus Res 72: 227–265.

Johnson K. N, Zeddam J. L, Ball L. A (2000) Characterization and construction of functional cDNA clones of Pariacoto virus, the first Alphanodavirus isolated outside Australasia. J Virol 74: 5123–5132.

Kamath R. S, Fraser A. G, Dong Y, Poulin G, Durbin R, et al. (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421: 231–237.

Kent W. J, Sugnet C. W, Furey T. S, Roskin K. M, Pringle T. H, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.

Kim D. H, Feinbaum R, Alloing G, Emerson F. E, Garsin D. A, et al. (2002) A conserved p38 MAP kinase pathway in *Caenorhabditis elegans* innate immunity. Science 297: 623–626.

Lehrbach N. J, Armisen J, Lightfoot H. L, Murfitt K. J, Bugaut A, et al. (2009) LIN-28 and the poly(U) polymerase PUP-2 regulate *let-7* microRNA processing in *Caenorhabditis elegans*. Nat Struct Mol Biol 16: 1016–1020.

Li H, Li W. X, Ding S. W (2002) Induction and suppression of RNA silencing by an animal virus. Science 296: 1319–1321.

Liu W. H, Lin Y. L, Wang J. P, Liou W, Hou R. F, et al. (2006) Restriction of vaccinia virus replication by a *ced-3* and *ced-4*-dependent pathway in *Caenorhabditis elegans*. Proc Natl Acad Sci U S A 103: 4174–4179.

Lu R, Maduro M, Li F, Li H. W, Broitman-Maduro G, et al. (2005) Animal virus replication and RNAi-mediated antiviral silencing in *Caenorhabditis elegans*. Nature 436: 1040–1043.

Lu R, Yigit E, Li W. X, Ding S. W (2009) An RIG-I-Like RNA helicase mediates antiviral RNAi downstream of viral siRNA biogenesis in *Caenorhabditis elegans*. PLoS Pathog 5: e1000286. doi:10.1371/journal.ppat.1000286.

Milloz J, Duveau F, Nuez I, Félix M-A (2008) Intraspecific evolution of the intercellular signaling network underlying a robust developmental system. Genes Dev 22: 3064–3075.

Motohashi T, Tabara H, Kohara Y (2006) Protocols for large scale in situ hybridization on *C. elegans* larvae. WormBook. pp. 1–8. Available: http://wormbook.org. Accessed 29 December 2010.

Neil S. J, Zang T, Bieniasz P. D (2008) Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature 451: 425–430.

Pak J, Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. Science 315: 241–244.

Powell J. R, Kim D. H, Ausubel F. M (2009) The G protein-coupled receptor FSHR-1 is required for the *Caenorhabditis elegans* innate immune response. Proc Natl Acad Sci U S A 106: 2782–2787.

Sabin L. R, Zhou R, Gruber J. J, Lukinova N, Bambina S, et al. (2009) Ars2 regulates both miRNA- and siRNA-dependent silencing and suppresses RNA virus infection in *Drosophila*. Cell 138: 340–351.

Schott D. H, Cureton D. K, Whelan S. P, Hunter C. P (2005) An antiviral role for the RNA interference machinery in *Caenorhabditis elegans*. Proc Natl Acad Sci U S A 102: 18420–18424.

Schulte R. D, Makus C, Hasert B, Michiels N. K, Schulenburg H (2010) Multiple reciprocal adaptations and rapid genetic change upon experimental coevolution of an animal host and its microbial parasite. Proc Natl Acad Sci U S A 107: 7359–7364.

Sijen T, Steiner F. A, Thijssen K. L, Plasterk R. H (2007) Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. Science 315: 244–247.

Tabara H, Sarkissian M, Kelly W. G, Fleenor J, Grishok A, et al. (1999) The *rde-1* gene, RNA interference, and transposon silencing in *C. elegans*. Cell 99: 123–132.

Tijsterman M, Okihara K. L, Thijssen K, Plasterl R. H. A (2002) PPW-1, a PAZ/PIWI protein required for efficient germline RNAi, is defective in a natural isolate of *C. elegans*. Curr Biol 12: 1535–1540.

Troemel E. R, Felix M. A, Whiteman N. K, Barriere A, Ausubel F. M (2008) Microsporidia are natural intracellular parasites of the nematode *Caenorhabditis elegans*. PLoS Biol 6: 2736–2752. doi:10.1371/journal.pbio.0060309.

Wang D, Urisman A, Liu Y. T, Springer M, Ksiazek T. G, et al. (2003) Viral discovery and sequence recovery using DNA microarrays. PLoS Biol 1: E2. doi:10.1371/journal.pbio.0000002.

Weimer R. M (2006) Preservation of *C. elegans* tissue via high-pressure freezing and freeze-substitution for ultrastructural analysis and immunocytochemistry. Methods Mol Biol 351: 203–221.

Wilkins C, Dishongh R, Moore S. C, Whitt M. A, Chow M, et al. (2005) RNA interference is an antiviral defence mechanism in *Caenorhabditis elegans*. Nature 436: 1044–1047.

Wood W. B (1988) The nematode *Caenorhabditis elegans*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory. 667 p.

A14

GENOMIC APPROACHES TO STUDYING THE HUMAN MICROBIOTA[54]

George M. ***Weinstock***[55]

Abstract

The human body is colonized by a vast array of microbes, which form communities of bacteria, viruses and microbial eukaryotes that are specific to each anatomical environment. Every community must be studied as a whole because many organisms have never been cultured independently, and this poses formidable challenges. The advent of next-generation DNA sequencing has allowed more sophisticated analysis and sampling of these complex systems by culture-independent methods. These methods are revealing differences in community structure between anatomical sites, between individuals, and between healthy and diseased states, and are transforming our view of human biology.

The microbes that exist in the human body are collectively known as the human microbiota. This amazingly complex and poorly understood group of communities has an enormous impact on humans. An increasing number of conditions are being examined for correlative and causative associations with the microbiome—which, in this Review, is used to refer to the microbiota and the habitat it colonizes (Box A14-1). Each one of the many microbial communities has its own structure and ecosystem, depending on the body environment it exists in. The fundamental goal of human microbiome research is to measure the structure and dynamics of microbial communities, the relationships between their members, what substances are produced and consumed, the interaction with the host, and differences between healthy hosts and those with disease. Despite an explosion in human-microbiome research, these communities are still the dark matter of the body. The microbiome has been called another organ (Backhed et al., 2005; Foxman et al., 2008; Possemiers et al., 2011; Shanahan, 2002) because of its products, its responsiveness to the environment and its integration with other systems. Sometimes referred to as our second genome (Bruls and Weissenbach, 2011), the genes of microbes that make up the microbiome outnumber human

[54] Reprinted with kind permission from Nature Publishing Group.

[55] The Genome Institute, Washington University, 4444 Forest Park Avenue, Campus Box 8501, St. Louis, Missouri 63108, USA.

Competing financial interests: The author declares no competing financial interests.

BOX A14-1
Terminology

- **Biodiversity** is a measure of the complexity of a community. It is affected by the number of taxa (richness) and their range of abundance (evenness). High biodiversity occurs when many taxa (high richness) are present at similar abundances (an even distribution).
- **Commensals** are organisms that benefit from another organism but that have no harm or benefit themselves. Microbes of the microbiome were thought to be commensals that benefited from the human host but did no harm. Many of these organisms provide benefits to the human host and so have a mutualistic relationship.
- **Contig** is a stretch of contiguous sequence in a genome assembly.
- **Coverage** is the number of times a genome or gene is sequenced. In a genome sequenced to coverage, each nucleotide in the sequence appears, on average, in 100 reads.
- **Genome assembly** is the process of constructing a genome sequence from short subsequences by sequencing many random fragments from a sheared genome. The random short sequences are compared, and overlapping common sequences are used to determine their orientation and order with respect to each other. A consensus sequence is constructed from this layout. Usually there are gaps, but when contigs can be arranged in the correct order and orientation, these longer stretches are called scaffolds.
- **Metagenomics** was defined (Hooper et al., 2012) as a process for identifying genes specifically by their function by cloning them directly from the environment and expressing genes in a surrogate host (Riesenfeld et al., 2004). Therefore, gene function was known even if the sequence was not sufficient for functional inference, such as when it encoded a protein of previously unknown function. This definition, also known as functional metagenomics, is widely

genes by more than 100-fold, with over 3 million bacterial genes in the gut alone (Human Microbiome Project Consortium, 2012a; Qin et al., 2010). These extensive microbial ecosystems are not limited to the human body. Microbes and their communities dominate the environment and occupy a vast range of niches. Environmental metagenomics was developed extensively before being applied to the human body (Stein et al., 196; Vergin et al., 1998), and methods from other disciplines have had a significant effect on human-microbiome research. Defining complicated microbial ecosystems and developing tools to probe their workings is an important research enterprise of twenty-first century microbiology.

The complexity of microbial communities makes studying them challenging. There may be hundreds of different species, and enumerating what organisms are present with standard microbiological techniques is not possible because many organisms have never been grown in culture and may require special, as yet

used. More recently, metagenomics refers to general analyses of microbial communities by culture-independent methods, which do not necessarily focus on function. The combined genomes of the microbes in a community are thought of as the community metagenome. Another type of metagenomic analysis focuses on the structure of these aggregate genomes in a community.

- **Microbiome** in this Review refers to the microbiota and the habitat it colonizes and is analogous to the term biome in ecology. Microbiome is also used to refer to the collective genomes of the microbes—what is now the metagenome, and may have originally been coined by Joshua Lederberg (cited by Hooper and Gordon, 2001). However, it is also used for the more ecologically consistent meaning. A microbiome can be a specific body site, such as the gut microbiome, but the human microbiome is often used to refer to the collection of microbiomes of the human body.
- **Mutualism** is a type of symbiosis in which both organisms benefit. This is one type of relationship seen in the human microbiome.
- **Operational taxonomic unit** in microbiome research is a group of organisms with 16S ribosomal RNA gene sequences that show a certain level of identity. This group is often used as a surrogate for a species when the 16S rRNA sequences are at least 97% identical.
- **Pathogenic microbe** is one with the potential to cause disease.
- **Read** is the primary output of DNA sequencing, consisting of a short stretch of DNA sequence that is produced from sequencing a region of a single DNA fragment.
- **Shotgun sequencing** is the process of randomly breaking (often by shearing) a long DNA molecule (for example, a complete chromosome) and then sequencing the resultant DNA fragments, which each come from a different location in the original long DNA molecule.
- **Virome** is the collection of viruses in the microbiota.

unknown, growth conditions. In addition, the abundance of some microbes can range over orders of magnitude, so deep sampling is required to detect the less-abundant members. Culture-independent methods of taking a microbial census began about 25 years ago and were based on targeted sequencing of 5S and 16S ribosomal RNA genes (Olsen et al., 1986), which differ for each species and are a convenient identifier. As this became a tractable research area, next-generation sequencing (NGS) technologies (Table A14-1) were developed and allowed more extensive analyses, both targeted 16S rRNA gene sequencing and whole-genome shotgun sequencing of microbes in communities en masse. The number of culture-independent metagenomic investigations of the human microbiome has mushroomed, and it is one of the most studied areas of microbiology with significant potential to benefit clinical practice. This culture-independent methodology is broadly applied outside human-microbiome research and is expanding our

TABLE A14-1 DNA Sequencing Platforms Used for Microbiome Analysis

Platform	Method	Characteristics	16S rRna	Shotgun	Comments
Established					
Sanger-based or capillary-based instrument	Fluorescent, dideoxy terminator	750-base reads High accuracy	Full length sequenced with 2–3 reads	Long reads help with database comparisons	Most costly method Relatively low throughput, so low coverage of 16S or shotgun
Roche–454	Pyrosequencing light emission	400-base reads	Up to 3 variable regions per read	Long reads help with database comparisons	Cost limits shotgun coverage but 16S coverage is good
Illumina	Fluorescent, stepwise sequencing	100–150-base reads	Only 1 variable region per read	Short reads do not seem to limit analysis	Very high coverage owing to high instrument output and very low cost
Not yet widely used					
Ion Torrent	Proton detection	More than 200-base reads	Like other NGS	Like Illumina	Expect high coverage, but longer reads than Illumina
PacBio	Fluorescent, single-molecule sequencing	Up to 10-kilobase reads Low accuracy	Accuracy an issue for correct taxon identification	Long reads could help assembly	Attractive for long reads, but lower accuracy limits applications
Oxford Nanopore*	Electronic signal as DNA passes through pore Single-molecule sequencing	Long reads	Unknown	Long reads could help assembly	Not yet available

*At the time of publication, the Oxford Nanopore system was not available, and information provided is based on company presentations. Ion Torrent and PacBio are both available but have not been widely used for microbiome analysis. The Illumina MiSeq instrument is expected to provide 250-base reads in the near future.

knowledge of the environment. This Review describes how NGS approaches are transforming human-microbiome studies, and posing questions and challenges for the future.

Single Organisms and Microbial Communities

In the past, research on microbial interactions with humans has focused on single pathogenic organisms. Studies of communities of non-pathogenic microbes in the body were limited because the organisms were thought to be benign, with minor effects on human health compared with pathogens. Microbiome research has led to new interest in the communities of non-pathogenic microbes that inhabit the human body, and the need to describe the genomes of these organisms to understand the human microbiome has been recognized.

Every community of the microbiome has its own characteristics (Table A14-2). For the gut community, for example, high biodiversity is associated with a healthy state and reduced biodiversity occurs in patients with conditions such as Crohn's disease (Manichanh et al., 2006), whereas for tissues of the vagina, a lower biodiversity exists in healthy individuals and a bloom of organisms occurs in patients with vaginosis (Fredricks et al., 2005). To understand why different sites have different properties, the mechanisms that lead to the disruption of

TABLE A14-2 Characteristics of Bacteria, Microbial Eukaryotes, and Viruses in the Human Microbiome

Characteristic	Bacteria	Viruses	Eukaryotic Microbes
Genome size	0.5–10 megabases	1–1,000 kilobases	10–50 megabases
Number of taxa in the human microbiome	At least thousands	Unknown, but could be as many as bacteria	Unknown, but may be fewer than bacteria
Relative abundances	Highly variable	Highly variable	Unknown
Targeted detection methods	Sequencing of genes such as 5S and 16S rRNA	No universal method for genes, but virus-specific polymerase chain reaction assays for some	Sequencing of 18S rRNA gene Spacer region in rRNA
Shotgun approach to analyses	Alignment to reference genomes or database comparison	Database comparison	Alignment to reference genomes or database comparison
Subspecies or strain diversity	Modest sequence variation Horizontal gene transfer also contributes	High sequence variation	Unknown

ecosystems and to disease, and exceptions to generalities about a tissue, researchers require knowledge of the structure and behaviour of microbial communities.

Microbial communities benefit the host by providing functions such as digestion of nutrients (Flint et al., 2008) or protection against infection (Srikanth and McCormick, 2008). Antibiotic treatment perturbs the microbiome (Dethlefsen et al., 2008; Jakobsson et al., 2010) by reducing its size and altering its composition. This disturbance can lead to infection (Croswell et al., 2009; Miller et al., 1956; Sekirov et al., 2008), and antibiotic-resistant organisms such as *Clostridium difficile*—normally controlled by the microbiome—can overgrow and create problems (Mulligan, 1984). More complex community contributions also exist, such as interactions with host immune and inflammatory systems (Jarchum and Pamer, 2011; Marsland, 2012) or production of metabolites involving hybrid pathways from multiple organisms, including host–microbe pathways (Wang et al., 2011). Understanding these phenomena will ultimately allow the microbiome to be manipulated so that, for example, transplants of microbial communities could treat *C. difficile* infections (Brandt and Reddy, 2011; Gough et al., 2011).

Whether the microbial ecology of the human body can be simplified to the properties of single organisms is unknown. Many organisms have never been cultured and may be adapted to life in a community environment rather than a pure culture. For organisms for which growth requirements are understood, there is a dependence on secreted products from other community members. For example, secreted siderophores (D'Onofrio et al., 2010) are small molecules that help microbes to scavenge iron, which is a limiting factor for growth in the body. So even the study of individual organisms can be dependent on studying the community.

Dissecting a Microbiome

Analysis of community structure (Figure A14-1) focuses on either targeted regions (such as the 16S rRNA gene) or shotgun sequencing to catalogue the genes that are present. Additional analysis involves sequencing genomes of individual organisms to produce a catalogue of reference genomes (Human Microbiome Jumpstart Reference Strains Consortium, 2010), and analysing RNA to describe the transcriptome and identify RNA viruses. Non-genomic analyses include proteomic and metabolomic studies, but these are not discussed here. Every sample should be well-annotated with clinical metadata, so that, ultimately, the microbiome's genetic and community structures can be correlated with the individual's phenotype.

Census of Organisms

Modern metagenomic analyses of microbial communities were developed from culture-independent methods for taking a census of organisms present in a community and their abundances. Although DNA reassociation kinetics provides

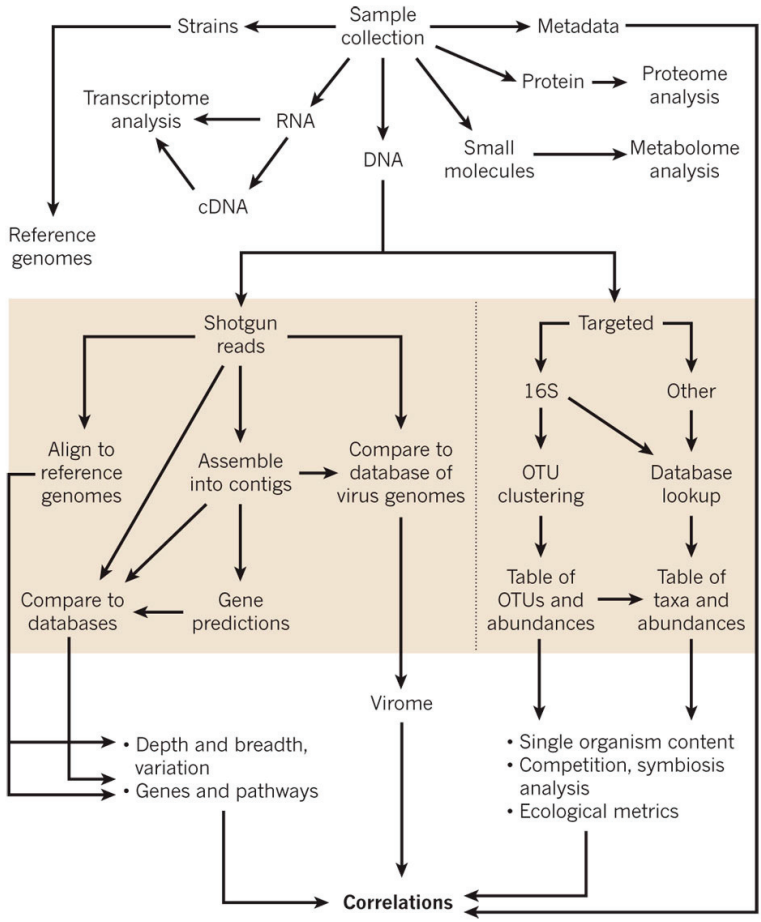


FIGURE A14-1 Data and analysis workflow for microbiome analysis. From a microbiota sample, DNA, RNA and protein can be extracted, and metadata and strains of bacteria obtained. Data from DNA can be supplemented with proteome and transcriptome analysis. During primary analysis, shotgun techniques can produce reads from DNA, which are then aligned to reference genomes to identify variants and community population genetics, assembled into contigs to make gene predictions or compared with databases. Alternatively, targeted sequencing such as 16S rRNA gene sequencing can be used to take a community census, and these data are then compared with databases to create tables of taxa and abundance, or analysed with software programs to cluster the reads into OTUs to create tables of abundance. The derivative data is used in secondary analysis for ecological metrics or competition and symbiosis analysis. In addition, shotgun reads and comparisons with reference genomes and databases can be used to build pathways and reconstruct the capabilities of a community. The combination of these analyses will contribute to understanding the differences within and between individuals.

information on community diversity and structure (Gans et al., 2005), there is no accounting for organisms that may be tracked between samples. Methods more useful for providing information on the entire structure often focus on signature sequences that distinguish taxa (detected by hybridization to arrays of diagnostic oligonucleotides [Nelson et al., 2011]), various methods for fingerprinting polymerase chain reaction (PCR) products (such as single-strand conformation polymorphisms or terminal restriction fragment length polymorphisms) or DNA sequencing of targeted PCR products. Sequencing of 16S rRNA genes is the main method of taking a community census because fingerprinting methods do not adequately measure low-abundance organisms (Bent et al., 2007).

16S rRNA differs for each bacterial species. A bacterial species is hard to define, but is often taken as organisms with 16S rRNA gene sequences having at least 97% identity—an operational taxonomic unit (OTU). A 16S rRNA gene sequence of about 1.5 kilobases has nine short hypervariable regions that distinguish bacterial taxa; the sequences of one or more of these regions are targeted in a community census.

Before the introduction of NGS methods, the prevailing approach was to clone full-length 16S rRNA genes after PCR with primers that would amplify genes from a wide range of organisms. Cloned 16S rRNA genes were sequenced by the Sanger method, which required two or three reads to cover the entire gene. Accuracy was crucial because sequencing errors led to misclassification. The cost and effort required for the Sanger method limited the depth of sampling, and studies often produced about 100 sequences per specimen. This method identified the dominant organisms in a community, but analysis of less abundant organisms was limited.

Introducing NGS to 16S rRNA gene analysis led to marked improvements in cost and depth of sampling. The Roche–454 platform has dominated microbial community analysis (Sogin et al., 2006). As the read length for 454 pyrosequencing is about 400 bases, only a portion of the 16S rRNA gene can be sampled, and many different studies have targeted between one and three of the hypervariable regions, with different hypervariable regions targeted in different studies. Using a portion of the 16S rRNA gene led to a loss of sensitivity (some taxa cannot be reliably defined at the species level, although high confidence identification of higher taxonomic ranks is possible), nevertheless gains in depth of sampling and cost savings outweigh this caveat. The US Human Microbiome Project (HMP) (The NIH HMP Working Group et al., 2009) has sequenced more than 10,000 specimens from healthy adults on the 454 platform by targeting V3 to V5 regions in the 16S rRNA gene and producing, on average, 7,000 sequences per specimen (Human Microbiome Project Consortium, 2012b), which is a vast expansion on the Sanger method of sequencing analysis. The results of the HMP, which sampled 18 body sites, provide an in-depth definition of the human microbiome. Another study16 that focused on the effects of the antibiotic ciprofloxacin reported the “rare biosphere” in the gut. This study documented perturbation of

taxa and recovery from antibiotic treatment, as well as minor constituents that did not recover after antibiotic treatment. Such analyses will be important in identifying individuals who are at risk of side effects from antibiotic treatment, for example overgrowth of pathogens such as *C. difficile* or life-threatening antibiotic-associated diarrhoea.

When using 16S rRNA gene sequencing to compare individuals it is not necessary to know which organisms are present, only whether the spectra of 16S rRNA gene sequences are similar and the degree of difference between samples. Projects that compare healthy cohorts and those with disease to determine whether there is a difference in the microbiome, or examine the effects of diet, antibiotic treatment or environmental factors on the microbiome, all focus on detecting differences in communities, rather than identifying actual taxa. A loss of sensitivity for organism identification can be tolerated, and NGS allows cost-effective deep sampling of large cohorts, which is needed to reach statistically significant conclusions. The Illumina sequencing platform has been applied to metagenomics projects (Claesson et al., 2010; Gloor et al., 2010; Lazarevic et al., 2009), but because this sequencing platform currently produces reads of 100 bases (HiSeq system) to 150 bases (MiSeq system), only a single hypervariable region can be sequenced. However, this further loss of sensitivity does not preclude the use of the Illumina platform for the comparative projects already described in this Review. An early application of this platform was its use in a study of vaginal microbiomes in patients with HIV, for which comparisons of patients with conditions such as vaginosis before and after antibiotic therapy were examined (Hummelen, 2010). As a result of the exceptional increases in numbers of reads and the lower cost associated with the Illumina platform, it is becoming more widely used for 16S rRNA gene-sequence profiling and continues the microbiome-analysis trend of deeper sampling at lower costs.

Shotgun Sequencing for Cataloguing Organisms

Targeted sequencing is a powerful tool for assessing the organisms that are present in microbial communities, but it is limited in terms of the functional and genetic information produced. Organisms for which the genome sequences are known (currently there are several thousand sequenced bacterial genomes) can be used to infer the genes and functional capabilities of the community (Figure A14-1). However, many organisms have no reference sequence. Furthermore, a reference sequence does not completely describe the genes that are contributed by an organism. There is considerable variation in the genomes between strains of the same species. Two strains of *Escherichia coli*, O157:H7 and K-12, both have 16S rRNA gene sequences of E. coli, but differ in hundreds of genes. There are limits to what can be learned about the genetic content of communities from 16S rRNA gene sequences alone.

Moving beyond this level of functional inference requires a gene-based census. This catalogue of genes can be provided by shotgun sequencing of DNA that has been extracted from the community as a whole and samples the mixture of genomes that make up the metagenome (Figure A14-1). In a community in excess of hundreds of species with varying abundance, deep sequencing is needed to sample minor constituents that are not necessarily unimportant. The bacterial concentration in the gut can be 10^{11} cells ml^{-1} (refs. Luckey, 1972; Zubrzycki and Spaulding, 1962), so for an organism that is present at a concentration of 1 per 10^6 there are 10^5 cells ml^{-1}, which is sufficient for the organism's products, such as metabolites and toxins, to have an effect on the community and the host.

Illumina sequencing of faecal samples produced 4 gigabases per sample and 10 Gb per sample in the Metagenomics of the Human Intestinal Tract (MetaHIT) (Qin et al., 2010) and HMP (Human Microbiome Project Consortium, 2012) projects, respectively, which corresponded to tens of millions of reads per sample. At this depth of sequencing, the genomes of minor constituents such as *E. coli* (with an abundance of about 1% or lower) are sampled almost completely, and organisms with an even lower abundance have some of their genome represented. This extraordinary sampling of complex microbial communities is made possible by producing large amounts of data and by the low cost of NGS methods.

Shotgun sequence data, in addition to 16S rRNA gene analysis, provide information on the organisms that make up communities. Extracting 16S rRNA gene sequences from shotgun reads to determine the organisms present is possible; however, targeted 16S rRNA gene sequencing tends to introduce biases (owing to the broad-range PCR used to amplify 16S rRNA gene sequences or the choice of region within the 16S rRNA gene), which shotgun sequencing does not. Shotgun sequencing is less sensitive than targeted rRNA sequencing because a small fraction of the sequences are from 16S rRNA genes. Another approach is to align shotgun sequences to bacterial reference genomes (Arumugam et al., 2011; Human Microbiome Project Consortium, 2012; Martin et al., 2012), allowing the relative abundance of species to be determined on the basis of the number of reads that align to each reference genome (also useful for the comparative studies already described). The MetaHIT project has used this approach to classify individuals into different groups, called enterotypes, on the basis of the community structure in their faecal samples (Arumugam et al., 2011). The same enterotypes have been found in 16S rRNA gene-based analysis (Wu et al., 2011). The vaginal microbiome has also been classified into five groups (Ravel et al., 2011). These observations suggest the human microbiome may exist in distinct states in different people, although correlation with environmental, genetic or health status is not yet clear. Stratifying future studies depending on which community class an individual belongs to may be important for identifying correlations with phenotypic data.

The need for reference genome sequences is clear both to infer genetic content of organisms identified by 16S rRNA genes and to identify sources of

shotgun reads by aligning to reference genomes, and so determining organismal content of communities from shotgun data. NGS techniques have reduced the cost of bacterial sequences to less than US$1,000 per genome and led to an increase in the production of 'complete' genome sequences. Current methodology relies mainly on Illumina shotgun sequencing and a variety of methods to assemble the reads into a genome. The product is not a true complete genome, but a high-quality draft that covers almost all of the genome and results in a high-quality base sequence (Human Microbiome Jumpstart Reference Strains Consortium, 2010). Programmes such as the HMP (The NIH HMP Working Group et al., 2009; Proctor, 2011) and the Genomic Encyclopedia of Bacteria and Archaea (GEBA) (DOE Joint Genome Institute, 2012) are producing reference genomes by the thousands.

Although bacteria are the main components of the human microbiome, eukaryotic microbes and viruses (both human viruses and bacteriophages) are also present (Table A14-2). The study of eukaryotic microbes is not as advanced as that of bacteria (Parfrey et al., 2011), but the organisms are identified by signature sequences (such as fingerprinting and 18S rRNA) and shotgun sequencing analogous to bacteria. The number of reference genomes for eukaryotic microbes is smaller than that for bacteria, and progress will depend on addressing this shortfall.

By contrast, considerable effort is being given to characterizing the genomes of human viruses (Wylie et al., 2012a) and bacteriophages (Breitbart et al., 2003), known as the virome (Box A14-1). This work is based on shotgun sequencing (Figure A14-1), although oligonucleotides microarrays for virus detection are also used (Palacios et al., 2007; Wang et al., 2003). Viral sequences can be detected in shotgun data from different body sites, and viruses can also be enriched by processing samples before DNA extraction (Casas and Rohwer, 2007). Virome analysis by shotgun sequencing of microbial communities (discussed later) has led to the identification of human viruses (Allander et al., 2005; Breitbart and Rohwer, 2005; Finkbeiner et al., 2008), as well as the detection of known viruses in healthy subjects and diseases of unknown aetiology (Wiley et al., 2012b). Likewise, bacteriophages are found to be highly diverse at different body sites (Breitbart et al., 2008; Minot et al., 2012; Pride et al., 2011), with differences between individuals as a result of diet (Minot et al., 2011) or disease states (Lepage et al., 2008; Willner and Furlan, 2010).

Sequencing for Gene Catalogues and Functional Inference

Metagenomic shotgun data also sample community gene content, which is useful to define community capabilities and identify particular members. Deep sequencing, such as that used in the MetaHIT and the HMP, broadly samples the genomes of even minor constituents, facilitating the identification of genes present within a given community (Figure A14-1). By using the sequence reads

themselves, or by first assembling them into contigs (Box A14-1), sequence data can be compared with databases such as the National Institutes of Health's GenBank to identify which genes are present. De novo prediction of genes from metagenomic data is also possible (Human Microbiome Project Consortium, 2012), which provides motifs for functional inference even if the sequence does not find a match in a database. Finally, alignment of reads or contigs to reference genomes identifies which organisms are present, along with their known gene content. These methods convert metagenomic sequence data into catalogues of genes that can be further analysed.

Gene catalogues can be compared with databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2010), which sorts gene products into pathways and processes. Such analyses provides lists of pathways, identify which pathway genes are in the community and quantify the abundances of genes and pathways (Abubucker et al., 2012). Comparing gene catalogues to specialized metabolic databases, such as the Carbohydrate-Active Enzymes database (Cantarel et al., 2009), is also useful. Carbohydrate-degrading capabilities of communities differ between body sites, suggesting the carbohydrate spectrum of each body site has determined which organisms and pathways are present (Cantarel et al., 2012).

In addition to pathway analysis, determining the presence and abundance of genes, such as antibiotic-resistance genes or virulence factors, in a community is possible using similar methods to those already described, and can shed light on pathogen burden in an individual and consequences of antibiotic treatment. The importance of functional analyses cannot be overemphasized, and functional properties of communities are thought to be more important than their taxonomic composition (Turnbaugh and Gordon, 2009).

Computational Tools and Strategies

Metagenomic data are a rich source of information for the sequencing and analysis methods already discussed (Raes et al., 2007; Wooley et al., 2010). The data analysis workflow has three phases. In the first phase, primary data are processed and filtered depending on the application. For 16S rRNA gene sequencing, the quality of analysis is important so that organisms are not misclassified. Initial processing addresses read quality, chimaerism (a read formed from different 16S rRNA genes), read length after removing low-quality bases and related issues (Edgar et al., 2011; Haas et al., 2011; Jumpstart Consortium Human Microbiome Project Data Generation Working Group, 2012; Schloss et al., 2011; Wright et al., 2012). For shotgun sequence data (Human Microbiome Project Consortium, 2012; Qin et al., 2010)—in addition to sequence quality—artefacts such as duplicate reads must also be addressed, as well as computationally removing contamination from human sequences. Removal of human and

bacterial sequences is important in read processing for virome analysis (Wylie et al., 2012a,b) (Figure A14-1).

Following production of processed reads, the second phase involves generating various derivative data sets. For 16S rRNA gene analysis, tables of taxa and abundance are produced by comparisons with 16S rRNA sequence databases or by using software packages to cluster the reads into OTUs (Lozupone et al., 2011; Schloss et al., 2009). Comparing shotgun reads to gene databases, such as GenBank or KEGG, by using the Basic Local Alignment Search Tool (BLAST), for example, produces lists of genes and the number of matched reads (Abubucker et al., 2012; Human Microbiome Project Consortium 2012a,b). Alignment of reads to reference genomes produces tables of breadth and depth of coverage, by reads of each genome (Martin et al., 2012). In each of these data sets, there is more biological information to be gleaned and added through further analysis. Not all reads match sequences in databases because not all organisms have a reference genome sequenced. In addition, reads may match genes whose function has not been elucidated. These sequences of unknown origin or function can be a sizeable fraction and the effect of this uninformative portion of data on analyses and conclusions is not clear.

The third phase of analysis uses these derivative data to produce trees or other representations of the similarity of communities, abundance curves, biodiversity plots, and other ecological and statistical descriptors of community structure (Lozupone et al., 2011; Schloss et al., 2009) (Figure A14-1). A list of hits from BLAST is used to build metabolic pathways for reconstruction of community capabilities (Abubucker et al., 2012). Alignments to reference genomes are further analysed for variants and population genetics of communities. Computational analysis can also be used to determine which organisms co-occur or rarely co-occur as evidence for symbiosis or competition, respectively, or to follow the dynamics of community structure in longitudinal time series (Caporaso et al., 2011).

Some analyses pose significant computational challenges. Comparisons to gene databases at the protein level are particularly demanding because shotgun sequences must be translated into polypeptides in all six reading frames, and each must be compared with a gene database represented at the protein level. Using conventional BLASTx programs for this comparison in large data sets, such as the HMP, could take decades, so supercomputers, accelerated BLAST programs or both must be used (Human Microbiome Project Consortium, 2012). A lack of efficient software and large enough computer clusters are often bottlenecks for metagenomic analysis, because sequencing and data production are not limiting factors. Management of large data sets and computing resources are receiving more attention, with cloud-computing services seeming to be a viable alternative (Angiuoli et al., 2011).

Future Directions and Challenges

The rapid rise in metagenomic studies has solved many problems but, as the field has grown, other questions have been raised. Existing methodology is becoming more sophisticated, and sequencing technology is making exponential advances (Table A14-1). The Illumina platform introduced instruments that were more appropriate for sequencing smaller genomes, with faster run times and longer read lengths, offering more flexibility for metagenomic applications. The long read length of the PacBio platform has the potential to help distinguish the reads from different organisms, which is a challenge for metagenomic shotgun sequencing. The technology produced by Oxford Nanopore promises long reads and short run times in a scalable system, and is therefore a good match for microbial applications. Reducing the amount of DNA needed for shotgun sequencing will allow communities in smaller anatomical regions, such as within the gastrointestinal tract, to be studied separately rather than together with other regions as is the case with the current methodology. Short run-time instruments and reductions in sample size will also hasten the introduction of microbiome analysis to the clinic, where analyses of patient samples must be quick and able to deal with limited amounts of material. Ultimately, the aim of human-microbiome research is its application as a diagnostic, therapeutic and preventive tool in the clinic.

The main limitation of using shotgun data is the large number of organisms that have not been cultured, let alone sequenced. These organisms are therefore under-represented in databases, and their shotgun reads are anonymous. When community shotgun data are assembled into genomes to obtain genome sequences for new organisms, contig sizes are typically small as a result of lower organism abundance and the challenges associated with assembly of a complex mixture. The long read lengths of PacBio and Oxford Nanopore instruments should help with these challenges, as will the development of assembly algorithms for metagenomic data. Expanding the catalogue of reference genomes by producing reference sequences for individual uncultured organisms is an active area. Methods that use cell sorting to isolate organisms, coupled with sequencing and assembly techniques for single-cell DNA preparations, are producing new genome sequences (Chitsaz et al., 2011; Dichosa et al., 2012) and, in high-throughput mode, could complement shotgun metagenomics for analysing communities.

One problem associated with genomic data is that it does not address whether an organism is alive or has succumbed to host defences or antibiotic treatment. However, the data can be complemented with transcriptome analysis, or proteomic and metabolomic data sets, which analyse gene expression and metabolic data that are more likely to be derived specifically from living cells.

The simultaneous advances in human genetics and genomics offer opportunities for combining studies of host genotype with microbiome phenotype. Methods for viewing the microbiome as a quantitative trait and relating this to host genotype are being developed (Benson et al., 2012). Advances in host–microbiome studies are also coming from combining immunology and human-microbiome

research (Elinav et al., 2011; Hooper et al., 2012). Moreover, continued development of statistical methods in microbiome research, such as advances in power analysis, will aid experimental design and future analysis.

Acknowledgments

The author gratefully acknowledges generous support from the National Institutes of Health.

References

Abubucker, S. *et al*. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).

Allander, T. *et al*. Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc. Natl Acad. Sci. USA* **102**, 12891–12896 (2005).

Angiuoli, S. V., White, J. R., Matalka, M., White, O. & Fricke, W. F. Resources and costs for microbial sequence analysis evaluated using virtual machines and cloud computing. *PLoS ONE* **6**, e26624 (2011).

Arumugam, M. *et al*. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).

Backhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host–bacterial mutualism in the human intestine. *Science* **307**, 1915–1920 (2005).

Benson, A. K. *et al*. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl Acad. Sci. USA* **107**,18933–18938 (2010).

Bent, S. J. *et al*. Measuring species richness based on microbial community fingerprints: the emperor has no clothes. *Appl. Environ. Microbiol.* **73**, 2399–2401 (2007).

Brandt, L. J. & Reddy, S. S. Fecal microbiota transplantation for recurrent *clostridium difficile* infection. *J. Clin. Gastroenterol.* **45**, S159–S167 (2011).

Breitbart, M. & Rohwer, F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* **39**, 729–736 (2005).

Breitbart, M. *et al*. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220–6223 (2003).

Breitbart, M. *et al*. Viral diversity and dynamics in an infant gut. *Res. Microbiol.* **159**, 367–373 (2008).

Bruls, T. & Weissenbach, J. The human metagenome: our other genome? *Hum. Mol. Genet.* **20**, R142–R148 (2011).

Cantarel, B. L. *et al*. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233–D238 (2009).

Cantarel, B. L., Lombard, V. & Henrissat, B. Complex carbohydrate utilization by the healthy human microbiome. *PLoS ONE* **7**, e28742 (2012).

Caporaso, J. G. *et al*. Moving pictures of the human microbiome. *Genome Biol.* **12**, R50 (2011).

Casas, V. & Rohwer, F. Phage metagenomics. *Methods Enzymol.* **421**, 259–268 (2007).

Chitsaz, H. *et al*. Efficient *de novo* assembly of single-cell bacterial genomes from short-read data sets. *Nature Biotechnol.* **29**, 915–921 (2011).

Claesson, M. J. *et al*. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res.* **38**, e200 (2010).

Croswell, A., Amir, E., Teggatz, P., Barman, M. & Salzman, N. H. Prolonged impact of antibiotics on intestinal microbial ecology and susceptibility to enteric *Salmonella* infection. *Infect. Immun.* **77**, 2741–2753 (2009).

Dethlefsen, L., Huse, S., Sogin, M. L. & Relman, D. A. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* **6**,e280 (2008).

Dichosa, A. E. *et al*. Artificial polyploidy improves bacterial single cell genome recovery. *PLoS ONE* **7**, e37387 (2012).

DOE Joint Genome Institute. *A Genomic Encyclopedia of Bacteria and Archaea.* http://www.jgi.doe.gov/programs/GEBA/ (US Department of Energy, 2012).

D'Onofrio, A. *et al*. Siderophores from neighboring organisms promote the growth of uncultured bacteria. *Chem. Biol.* **17**, 254–264 (2010).

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).

Elinav, E. *et al*. NLRP6 inflammasome regulates colonic microbial ecology and risk for colitis. *Cell* **145**, 745–757 (2011).

Finkbeiner, S. R. *et al*. Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathogens.* **4**, e1000011 (2008).

Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R. & White, B. A. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Rev. Microbiol.* **6**,121–131 (2008).

Foxman, B., Goldberg, D., Murdock, C., Xi, C. & Gilsdorf, J. R. Conceptualizing human microbiota: from multicelled organ to ecological community. *Interdiscip. Perspect. Infect. Dis.* **2008**, 613979 (2008).

Fredricks, D. N., Fiedler, T. L. & Marrazzo, J. M. Molecular identification of bacteria associated with bacterial vaginosis. *N. Engl. J. Med.* **353**, 1899–1911 (2005).

Gans, J., Wolinsky, M. & Dunbar, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**, 1387–1390 (2005).

Gloor, G. B. *et al*. Microbiome profiling by Illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS ONE* **5**, e15406 (2010).

Gough, E., Shaikh, H. & Manges, A. R. Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent *Clostridium difficile* infection. *Clin. Infect. Dis.* **53**, 994–1002 (2011).

Haas, B. J. *et al*. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* **21**, 494–504 (2011).

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245–R249 (1998).

Hooper, L. V. & Gordon, J. I. Commensal host–bacterial relationships in the gut. *Science* **292**, 1115–1118 (2001).

Hooper, L. V., Littman, D. R. & Macpherson, A. J. Interactions between the microbiota and the immune system. *Science* **336**, 1268–1273 (2012).

Human Microbiome Jumpstart Reference Strains Consortium. A catalog of reference genomes from the human microbiome. *Science* **328**, 994–999 (2010). **This paper presents methods and analysis for large-scale production of reference genome sequences from human-microbiome organisms.**

Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, 215–221 (2012b). **This paper describes the data sets and resources of the HMP.**

Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012a). **This paper presents analysis of data from the HMP.**

Hummelen, R. *et al*. Deep sequencing of the vaginal microbiota of women with HIV. *PLoS ONE* **5**, e12078 (2010).

Jakobsson, H. E. *et al*. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS ONE* **5**, e9836 (2010).

Jarchum, I. & Pamer, E. G. Regulation of innate and adaptive immunity by the commensal microbiota. *Curr. Opin. Immunol.* **23**, 353–360 (2011).

Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS ONE* **7**, e39315 (2012).

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38**, D355–D360 (2010).

Lazarevic, V. *et al*. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J. Microbiol. Methods* **79**, 266–271 (2009).

Lepage, P. *et al*. Dysbiosis in inflammatory bowel disease: a role for bacteriophages? *Gut* **57**, 424–425 (2008).

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: an effective distance metric for microbial community comparison. *ISME J.* **5**, 169–172 (2011).

Luckey, T. D. Introduction to intestinal microecology. *Am. J. Clin. Nutr.* **25**, 1292–1294 (1972).

Manichanh, C. *et al*. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**, 205–211 (2006).

Marsland, B. J. Regulation of inflammatory responses by the commensal microbiota. *Thorax* **67**, 93–94 (2012).

Martin, J. *et al*. Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PloS ONE* **7**, e36427 (2012).

Miller, C. P., Bohnhoff, M. & Rifkind, D. The effect of an antibiotic on the susceptibility of the mouse's intestinal tract to *Salmonella* infection. *Trans. Am. Clin. Climatol. Assoc.* **68**, 51–55 (1956).

Minot, S. *et al*. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Hypervariable loci in the human gut virome. *Proc. Natl Acad. Sci. USA* **109**, 3962–3966 (2012).

Mulligan, M. E. Epidemiology of *Clostridium difficile*-induced intestinal disease. *Clin. Infect. Dis.* **6**, S222–S228 (1984).

Nelson, T. A. *et al*. PhyloChip microarray analysis reveals altered gastrointestinal microbial communities in a rat model of colonic hypersensitivity. *Neurogastroenterol. Motil.* **23**,169–177 (2011).

The NIH HMP Working Group *et al*. The NIH Human Microbiome Project. *Genome Res.* **19**, 2317–2323 (2009).

Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.* **40**, 337–365 (1986).

Palacios, G. *et al*. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg. Infect. Dis.* **13**, 73–81 (2007).

Parfrey, L. W., Walters, W. A. & Knight, R. Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Front. Microbiol.* **2**, 153 (2011).

Possemiers, S., Bolca, S., Verstraete, W. & Heyerick, A. The intestinal microbiome: a separate organ inside the body with the metabolic potential to influence the bioactivity of botanicals. *Fitoterapia* **82**, 53–66 (2011).

Pride, D. T. *et al*. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J.* **6**, 915–926 (2011).

Proctor, L. M. The Human Microbiome Project in 2011 and beyond. *Cell Host Microbe* **10**, 287–291 (2011).

Qin, J. *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010). **This paper presents initial findings on the gut microbiome from the MetaHIT project.**

Raes, J., Foerstner, K. U. & Bork, P. Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr. Opin. Microbiol.* **10**, 490–498 (2007).

Ravel, J. *et al*. Vaginal microbiome of reproductive-age women. *Proc. Natl Acad. Sci. USA* **108**, S4680–S4687 (2011).

Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**, 525–552 (2004).

Schloss, P. D. *et al*. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).

Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* **6**, e27310 (2011).

Sekirov, I. *et al*. Antibiotic-induced perturbations of the intestinal microbiota alter host susceptibility to enteric infection. *Infect. Immun.* **76**, 4726–4736 (2008).

Shanahan, F. The host–microbe interface within the gut. *Best Pract. Res. Clin. Gastroenterol.* **16**, 915–931 (2002).

Sogin, M. L. *et al*. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci. USA* **103**, 12115–12120 (2006).

Srikanth, C. V. & McCormick, B. A. Interactions of the intestinal epithelium with the pathogen and the indigenous microbiota: a three-way crosstalk. *Interdiscip. Perspect. Infect. Dis.* **2008**, 626827 (2008).

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & DeLong, E. F. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.* **178**, 591–599 (1996).

Turnbaugh, P. J. & Gordon, J. I. The core gut microbiome, energy balance and obesity. *J. Physiol. (Lond.)* **587**, 4153–4158 (2009).

Vergin, K. L. *et al*. Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales. *Appl. Environ. Microbiol.* **64**, 3075–3078 (1998).

Wang, D. *et al*. Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.* **1**, E2 (2003).

Wang, Z. *et al*. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57–63 (2011).

Willner, D. & Furlan, M. Deciphering the role of phage in the cystic fibrosis airway. *Virulence* **1**, 309–313 (2010).

Wooley, J. C., Godzik, A. & Friedberg, I. A primer on metagenomics. *PLoS Comput. Biol.* **6**, e1000667 (2010).

Wright, E. S., Yilmaz, L. S. & Noguera, D. R. DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. *Appl. Environ. Microbiol.* **78**, 717–725 (2012).

Wu, G. D. *et al*. Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–108 (2011).

Wylie, K. M., Mihindukulasuriya, K. A., Sodergren, E., Weinstock, G. M. & Storch, G. A. Sequence analysis of the human virome in febrile and afebrile children. *PLoS ONE* **7**, e27735 (2012b).

Wylie, K. M., Weinstock, G. M. & Storch, G. A. Emerging view of the human virome. *Transl. Res.* http://dx.doi.org/10.1016/j.trsl.2012.03.006 (24 April 2012a).

Zubrzycki, L. & Spaulding, E. H. Studies on the stability of the normal human fecal flora. *J. Bacteriol.* **83**, 968–974 (1962).

A15

SEQUENCE ANALYSIS OF THE HUMAN VIROME IN FEBRILE AND AFEBRILE CHILDREN[56]

Kristine M. Wylie,[57,*] *Kathie A. Mihindukulasuriya*,[57] *Erica Sodergren*,[57] *George M. **Weinstock***,[57] *and Gregory A. Storch*[58]

Abstract

Unexplained fever (UF) is a common problem in children under 3 years old. Although virus infection is suspected to be the cause of most of these fevers, a comprehensive analysis of viruses in samples from children with fever and healthy controls is important for establishing a relationship between viruses and UF. We used unbiased, deep sequencing to analyze 176 nasopharyngeal swabs (NP) and plasma samples from children with UF and afebrile controls, generating an average of 4.6 million sequences per sample. An analysis pipeline was developed to detect viral sequences, which resulted in the identification of sequences from 25 viral genera. These genera included expected pathogens, such as adenoviruses, enteroviruses, and roseoloviruses, plus viruses with unknown pathogenicity. Viruses that were unexpected in NP and plasma samples, such as the astrovirus MLB-2, were also detected. Sequencing allowed identification of virus subtype for some viruses, including roseoloviruses. Highly sensitive PCR assays detected low levels of viruses that were not detected in approximately 5 million sequences, but greater sequencing depth improved sensitivity. On average NP and plasma samples

[56] Reprinted from *PLoS ONE*. Originally published as Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA (2012) Sequence Analysis of the Human Virome in Febrile and Afebrile Children. *PLoS ONE* 7(6): e27735. doi:10.1371/journal.pone.0027735

Editor: Chiyu Zhang, Institut Pasteur of Shanghai, Chinese Academy of Sciences, China

Received: August 12, 2011; **Accepted:** October 23, 2011; **Published:** June 13, 2012

Funding: This study was supported by grant number 1UAH2AI083266-01 from the National Institute of Allergy and Infectious Diseases, a Demonstration Project of the Human Microbiome Project to GS, grant number UL1RR024992 from the National Institutes of Health (NIH)-National Center for Research Resources to GS, grant number U54HG003079 from the National Institutes of Health–National Human Genome Research Institute (NIH-NHGRI) to The Genome Institute, and grant number U54HG004968 from the NIH-NHGRI to The Genome Institute. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

* E-mail: kwylie@genome.wustl.edu

[57] The Genome Institute, Washington University School of Medicine, Saint Louis, Missouri, USA.

[58] The Department of Pediatrics, Washington University School of Medicine, Saint Louis, Missouri, USA.

from febrile children contained 1.5- to 5-fold more viral sequences, respectively, than samples from afebrile children. Samples from febrile children contained a broader range of viral genera and contained multiple viral genera more frequently than samples from children without fever. Differences between febrile and afebrile groups were most striking in the plasma samples, where detection of viral sequence may be associated with a disseminated infection. These data indicate that virus infection is associated with UF. Further studies are important in order to establish the range of viral pathogens associated with fever and to understand of the role of viral infection in fever. Ultimately these studies may improve the medical treatment of children with UF by helping avoid antibiotic therapy for children with viral infections.

Introduction

Viruses are thought to be the primary cause of unexplained fever in children under 3 years old, a common problem that results in medical visits and in some cases hospitalization (Baraff, 2000; Krauss et al., 1991). With the implementation of several vaccines against bacterial infections, the frequency of bacterial infections is low (Rudinsky et al., 2009; Waddle and Jhaveri, 2009; Watt et al., 2010; Wilkinson et al., 2009). While viruses are suspected to be the cause of fevers in children with no documented bacterial infection, in clinical practice, tests for viruses are often not performed, and consequently no cause for the fever is determined. In the absence of a clear diagnosis, antibiotics are often prescribed (Colvin et al., manuscript submitted). A comprehensive analysis of viruses in children with fever could improve our understanding of the causes of unexplained fevers (UF) and ultimately lead to modifications of the treatment of children with UF, including restricted use of antibiotics.

Next generation sequencing technologies have been used successfully for viral metagenomic analyses (Angly et al., 2006; Nakamura et al., 2009; Reyes et al., 2010; Willner et al., 2009, 2011) and discovery of novel viruses (Beritmart and Rohwer, 2005; Felix et al., 2011; Loh et al., 2009). The Roche 454 platform has been favored over the Illumina GAIIX platform because its longer read lengths are argued to be advantageous for detecting more remote sequence homologies with known viruses. However, the sequence depth per unit cost is much greater using the Illumina platform, and greater sequencing depth would presumably favor the detection of rare virus sequences in metagenomic samples. In this study we sought to develop a sensitive, cost-effective method for characterizing the human virome, the viral component of the human microbiome, to be applied to the analysis of the virome in samples from febrile and afebrile children.

We analyzed 176 plasma and nasophyaryngeal swab (NP) samples from children with UF (febrile) and afebrile children by high-throughput sequencing using the Illumina platform. Using sequencing for the analysis of viruses

associated with UF has several advantages. First, unlike targeted PCR assays, sequencing can detect unexpected and novel viruses. Second, sequencing can provide additional information such as virus subtype or sequence variation from reference genomes, adding detail to the understanding of the viruses present. Using the protocol we developed, we detected viruses in 86% of plasma samples and 63% of NP samples from febrile and afebrile children. Furthermore, distinctions between the viromes of febrile and afebrile groups were observed. The assessment of known viruses and initial identification of potentially novel viruses using short-read Illumina sequencing moves us toward a more complete understanding of the human virome and its role in health and disease.

Results

Detection of Known Viruses

The goal of our study was to use high-throughput, deep sequencing analysis methods to characterize the human virome in large sample sets. Previous studies demonstrated improved virus detection using deeper sequencing on the 454 pyrosequencing platform compared to Sanger sequencing (Victoria et al., 2009) and we sought to determine whether the greater sampling depth possible with the Illumina GAIIX platform would provide a further improvement. This required development of methods for processing the small clinical sample amounts for Illumina sequencing, computational pipelines for management of the large number of Illumina reads from multiple samples, and accurate detection of viruses from the short Illumina reads.

In preliminary experiments, clinical samples with known viruses were used to develop these methods. Plasma samples were identified that were PCR positive for either an enterovirus (a group of relatively small, single-stranded RNA viruses) or human herpesvirus 6 (HHV-6, a large, double-stranded DNA virus). The total nucleic acid (see Methods) from each sample was amplified in two independent experiments, and the replicates were sequenced on the 454 and Illumina platforms. In the initial Illumina experiment, 75-base, paired-end reads were generated (Table A15-1). The enterovirus sample, in which virus was detected by real-time PCR with an average Ct value of 30.6 from multiple experiments, showed viral reads with both 454 and Illumina platforms (Table A15-1). A single HHV-6 read was found in one of the 454 replicates, but reads were found in both Illumina replicates (see Methods for sequencing analysis pipeline). In both the enterovirus and HHV-6 samples, the number of virus reads was higher on the Illumina platform compared with the 454 platform, indicating that increased depth of sequencing strengthened the virus signal. These results encourgaged us to improve the Illumina library construction and sequencing protocol. Specifically, we increased the length of fragments from 300–400 base pairs (gel purified) to 300–800 base pairs by minimizing shearing. Second, the Illumina read-length

TABLE A15-1 Detection of Viruses with Next Generation Sequencing

			454 reads		Illumina (75 base) reads			Illumina (100 base) reads		
Sample	PCR	Replicate	Total	Viral	Total	Viral	% genome length Covered	Total	Viral	% genome length covered
9008	Enterovirus	A	58,924	11	34,612,722	1473	7.50%	4,792,380	342	42.3%
9008	Enterovirus	B	45,625	2	35,760,754	419	-	-	-	-
9022	HHV-6	A	62,361	1	36,629,654	4	-	3,931,804	8	-
9022	HHV-6	B	50,866	0	34,933,530	7	-	-	-	-

was increased to 100 bases to facilitate identification of viral sequences, especially those that were divergent from reference genomes. 3- to 13-fold more HHV-6 and enterovirus sequences were detected with this modified protocol (Table A15-1). Based on this result, we reduced the read depth for subsequent analyses to 3–5 million per sample, which facilitated processing large numbers of clinical samples. The improved protocol sampled more broadly across the enterovirus genome with 65% of the reads assembling into small contigs covering 1907 bases of the enterovirus genome. In contrast, more than 97 percent of the 1473 reads from the first experiment were concentrated in two contigs that covered only 550 bases. We next sought to apply the 100-base read-length protocol to large-scale analysis of samples from afebrile children and those with UF.

Sequencing and Analysis of Samples from Febrile and Afebrile Children

Nasopharyngeal (NP) swabs and plasma samples from children 2–36 months of age with fever without an apparent source and afebrile controls from the same age group (Table A15-2) were sequenced on the Illumina GAIIX. The samples that were sequenced were a subset of those included in a PCR-based analysis that tested for 15 genera of known viral pathogens, 12 in NP samples and 7 in plasma samples (Table A15-3) (Colvin et al., manuscript submitted). The median number of sequences produced per sample was approximately 4.4 million (Figure S1). The median and mean numbers of reads for the NP febrile, NP afebrile, and plasma febrile groups were each greater than 4 million. However, the number of reads from the plasma afebrile group was lower, with a median of 2.4 million and mean of 3.2 million reads, which may indicate less nucleic available for amplification and sequencing in plasma from healthy children

Sequences from 11 of the 15 viral genera in the PCR panel were detected in samples that had been tested by both sequencing and PCR, allowing for the comparison of methods (Figure A15-1). Only Betacoronavirus, Influenza B virus, and Rubulovirus were not found by either method. Parechovirus was found in NP samples by sequencing, but NP samples were not screened by PCR for this virus. Likewise, blood samples were not tested by PCR for respiratory viruses. These 11 viral genera were detected in 90 samples. In 39 instances the same virus was detected by both methods, 25 were found only by sequencing, and 42 were found only by PCR (Figure A15-1). Beyond the 11 genera targeted by PCR, sequences from a variety of both DNA and RNA viruses were detected by sequencing

TABLE A15-2 Samples

	Afebrile	Unexplained fever (UF)
Nasalpharyngeal swab	81	50
Plasma	22	23

TABLE A15-3 Virus Genera Screened by PCR

Genus	PCR assay targets	Samples assayed by PCR
Alphacoronavirus	229E, NL63	NP
Betacoronavirus	OC43, HKU1	NP
Bocavirus	Human Bocaviruses	NP and Plasma
Enterovirus	Enteroviruses, Rhinovirues	NP—Enteroviruses and Rhinoviruses Plasma—Enteroviruses
Erythrovirus	Parvovirus B19	
Influenza A Virus	Influenza A, H1, H3	Plasma
Influenza B Virus	Influenza B	NP
Mastadenovirus	Human adenoviruses	NP
Metapneumovirus	Human metapneumovirus	NP and Plasma
Parechovirus	Human parechoviruses	NP
Pneumovirus	RSV A, RSV B	Plasma
Polyomavirus	JC, BK, WU, KI	NP—WU and KI Plasma—All 4
Respirovirus	Human parainfluenza virus (HPIV)-1, HPIV-3	NP
Roseolovirus	Human Herpesvirus (HHV)-6, HHV-7	Plasma
Rubulovirus	HPIV-2, HPIV-4	NP

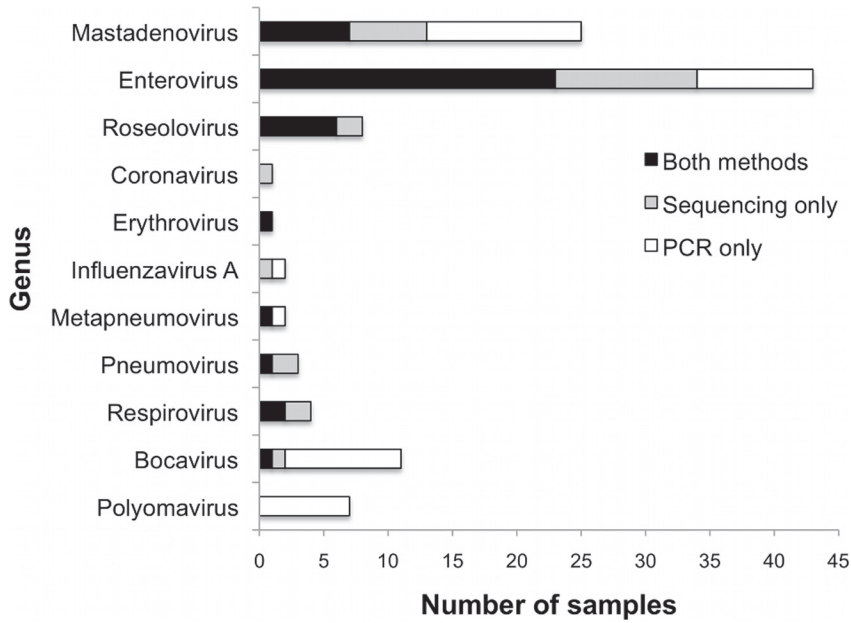


FIGURE A15-1 Comparison of sequencing and PCR results. The number of samples in which each virus was detected by PCR (white bars), sequencing (gray bars), or both (black bars) is shown.

(Figure A15-2A and B, respectively). Nearly 50,000 sequences with similarity to 25 known viral genera were identified.

Comparison of Sequencing and PCR Results for the Most Commonly Detected Viruses

Mastadenoviruses, enteroviruses, and roseoloviruses were frequently detected by both sequencing and PCR, which allows a thorough comparison of the methods. Mastadenoviruses (referred to as adenoviruses) were detected in 19 samples by PCR, and 5 of those were confirmed by sequencing (Figure A15-1). For those samples in which an adenovirus was detected by sequencing, the Ct values in the real-time PCR assay tended to be lower (average Ct = 29.1) compared with those samples in which an adenovirus was not detected (average Ct = 37.8) (P = 0.0023), suggesting that deeper sequencing would be required to detect the low levels of virus present in these samples (Figure A15-3A). When we sequenced one NP sample (sample 9021-581), which had a Ct of 42.6 and was originally negative for adenovirus by sequencing, to a read depth of greater than 21 million reads, we were able to detect 2 adenovirus reads, consistent with the PCR result. However, for the plasma sample from the same subject (sample 9021-895), which had a Ct of 37.5, adenovirus was still not detected after generating more than 50 million reads. Interestingly, 192 and 9,003 adeno-associated virus sequences were detected in these NP and plasma samples respectively, indirectly supporting the presence of adenovirus detected by PCR. The detection of the adeno-associated virus demonstrates that the unbiased nature of shotgun sequencing can reveal the presence of additional viruses not evaluated by PCR.

Enteroviruses (enterovirus and rhinovirus species) were detected in 32 samples by PCR, of which 21 were also detected by sequencing (Figure A15-1). Different commercial PCR assays were used for plasma and NP samples. For the plasma samples, a real-time PCR assay from Cepheid run on the SmartCycler thermal cycler (Cepheid, Sunnyvale, CA) was used (Colvin et al., manuscript submitted). Enterovirus sequences were detected in each of the 5 plasma samples that were positive for enterovirus by PCR. The NP samples were assessed using the xTAG Respiratory Virus Panel, produced by Luminex or the Multicode PLx Respiratory Virus Panel produced by Eragen Inc (Madison, WI) (Colvin et al., manuscript submitted). Of the 27 PCR-positive NP samples, rhinovirus or enterovirus sequences were detected in 16. Although the MFI read out of the Luminex assay is not strictly quantitative, the average MFI for samples with rhinovirus or enterovirus detected by sequencing was significantly higher than the average MFI of those samples missed by sequencing (P = 0.0193), suggesting that enteroviruses are present at low levels in some samples and would require deeper sequencing for detection. In support of this, sequencing one sample to a read depth of greater than 20 million reads produced 2 previously undetected rhinovirus reads (sample 9031-591, MFI 1500).

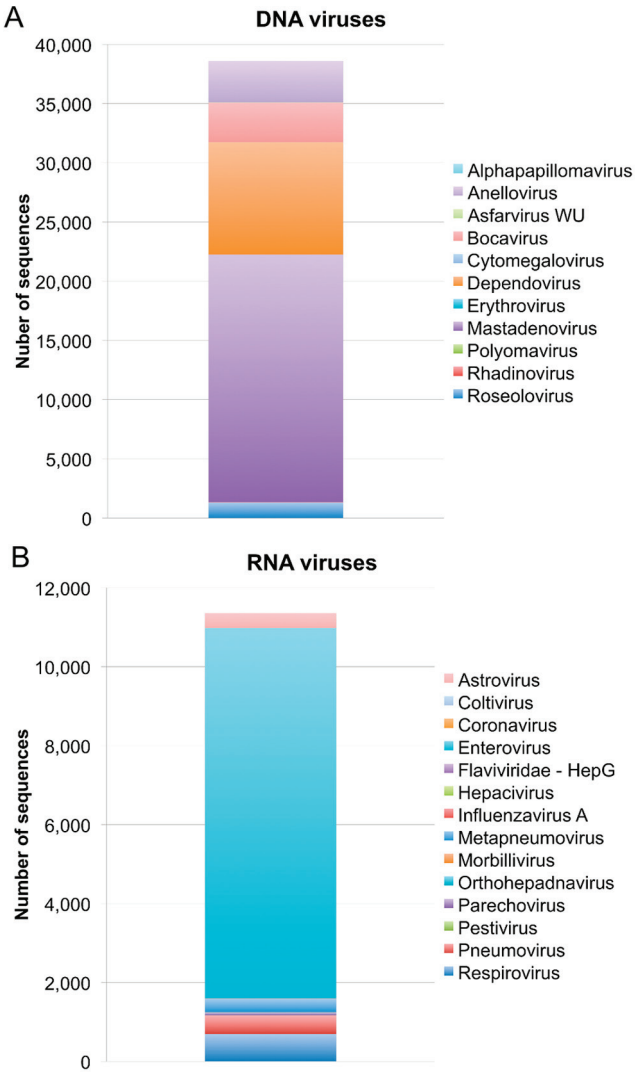


FIGURE A15-2 Sequence analysis identifies a variety of viruses in samples from febrile and afebrile children. Analysis of 176 plasma and NP samples on the Illumina GAIIX and HiSeq 2000 platforms identified approximately 50,000 sequences with similarity to 25 known (A) DNA and (B) RNA virus genera.

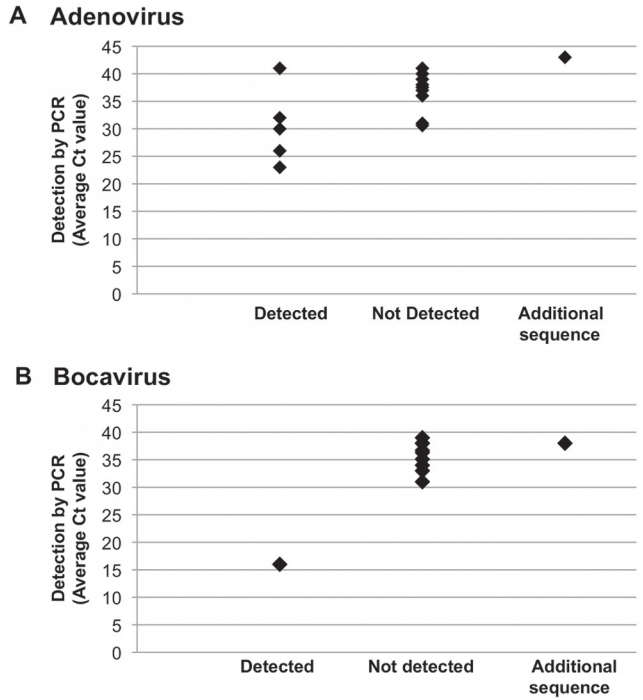


FIGURE A15-3 Comparison of sequencing results with Ct values from real-time PCR assays. The average Ct values from real-time PCR assays for (A) adenovirus and (B) bocavirus are graphed. Each PCR experiment was repeated at least twice per sample. Samples are grouped from left to right according to whether they were detected by sequencing, not detected by sequencing, or only detected when the number of sequence reads was increased to 21 or 51 million sequences for A and B, respectively.

For the roseoloviruses, HHV-6 and HHV-7, sequences were detected in every sample that was positive by PCR and also in 2 additional samples that were negative by PCR (Figure A15-1). The sensitivity of sequencing appeared to be high, possibly because the large genome size of the roseoloviruses allows the virus to be sampled efficiently during sequencing. The roseolovirus assay is a conventional (not real-time) PCR assay that is not quantitative, so the level of virus present in these samples is unknown. Futhermore, roseoloviruses were found in 5 NP samples by sequencing, while NP samples were not screened by PCR for these viruses.

As mentioned above, there were 25 examples of adenovirus, enterovirus, roseolovirus and a number of less frequently detected viruses that were found by sequencing but not by PCR (Figure A15-1). In some cases, significant sequence similarity to the reference genome was found at the amino acid but not nucleotide

level, suggesting that these might be novel or divergent virus sequences not detected by the PCR assay. Future analyses will be aimed at characterizing viruses with remote sequence similarities to reference genomes and validating viruses detected by sequencing but not by PCR.

Polyomaviruses (WU and KI) and bocaviruses were detected by PCR, but rarely or not at all by sequencing (Figure A15-1). We did not detect any polyomavirus sequences in any of the samples that were positive by PCR. This includes one sample in which KI was detected with a Ct of 39.2 that was sequenced to a depth of greater than 21 million reads. WU and KI polyomavirus Ct values were relatively high, with all but one above 30. The sensitivity of the WU PCR assay is 7 plasmid copies per reaction and that of the KI assay is 50 plasmid copies (Hormozdi et al., 2010). Thus, the level of the viruses present in the samples may be relatively low, and this fact, coupled with the small, 8 kb genome size suggests that little nucleic acid is available for detection. Bocavirus was detected 10 samples by a sensitive PCR assay that can detect as few as 5 plasmid copies of the target (Sumino et al., 2010) (Figure A15-1). During initial sequencing of a sample from a febrile child with respiratory symptoms, not included in the UF group, the 5 kb bocavirus genome was detected by sequencing. This sample had higher levels of bocavirus by PCR than other samples (sample 9105-663, Ct 16.17). A second sample containing less bocavirus (sample 9013-573, Ct 37.59) was sequenced to a read depth of greater than 50 million reads, yielding 4 bocavirus reads (Figure A15-3B). These results show sequencing enables detection of viral genomes present at low levels, but requires additional optimization to target rare smaller genomes.

Many samples contained sequence reads that aligned to reference genomes from viruses that were not included in the PCR panels (Figure A15-4). NP samples positive for parechoviruses and roseoloviruses are included in this figure because those viruses were only assessed by PCR in plasma samples. The polyomavirus-positive samples contained SV40-like sequences, which were not included in the polyomavirus PCR assays. Each of the viral sequences represented in Figure A15-4 were detected in samples from febrile children, with the exception of the alphapapillomavirus and some of the roseolovirus sequences. Some of the viruses detected are known to be pathogenic (parechovirus), while the pathogenic significance of others remains undetermined (hepatitis G virus). Further validation and characterization of some of these unexpected viruses may reveal the presence of novel viruses with remote homologies to the known references or a previously unknown role for a virus in febrile illness.

In plasma and NP samples from one febrile subject with UF, we detected astrovirus sequences. The sequences from the plasma sample assembled into contigs that spanned 55% of the recently discovered astrovirus MLB2 genome. Astroviruses have previously been detected in stool samples and are associated with diarrhea. However, this is the first time an astrovirus has been detected in either NP or plasma samples. No other pathogen was detected in these samples, suggesting the astrovirus may have been the cause of the subject's fever. We

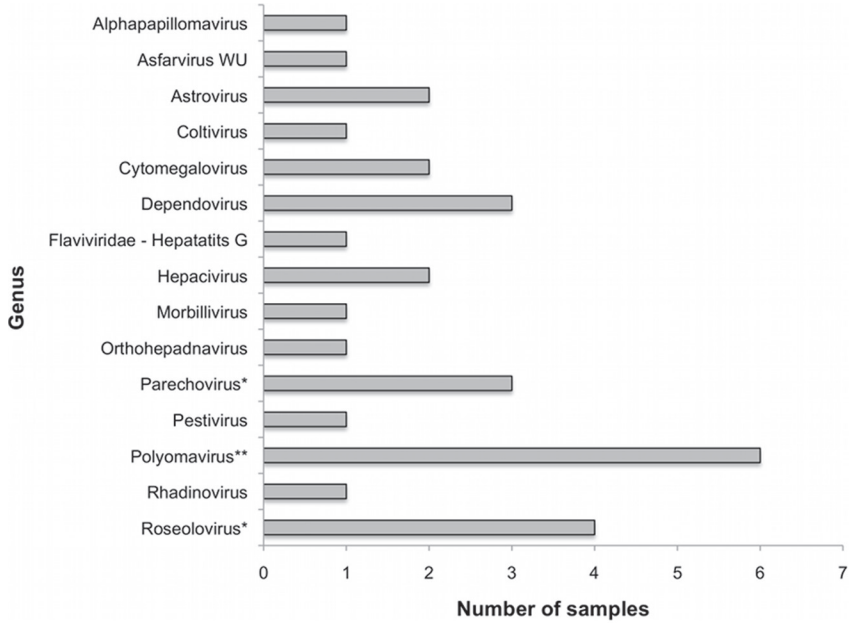


FIGURE A15-4 Viruses detected by sequencing that were not screened by PCR. The bars represent the number of samples in which each virus was detected by sequencing. *Indicates viruses that were not assayed by PCR in NP samples. **Indicates a virus that was not assayed by PCR but belongs to a genus with members that were.

have subsequently extended the sequence of the capsid gene of this MLB2 virus, confirming its presence by both sequencing and PCR of RNA from the plasma sample and allowing us to compare this virus to another MLB2 isolate (Holtz et al., 2011).

One concern about using 100-base reads for virome analysis is that shorter read lengths may not permit the discovery of novel viruses. Therefore, we asked whether we would have detected the astrovirus sequences had MLB2 or closely related MLB1 not been previously discovered. In fact, 162 reads from this sample have amino acid sequence similarity to other human, avian, and mammalian astroviruses (Table A15-4) and, therefore, this method would have been successful at detecting a novel virus.

In contrast to the samples in which viruses were present at low levels with only a few sequence reads detected, many samples yielded sufficient virus sequences to provide more detailed information about the virus in the sample. In some cases the genome coverage was sufficiently deep and broad for contigs to be assembled that covered significant portions of the genome (Table A15-5). The genome coverage allowed us to show that the human bocavirus was type 1, with 99% identity to known reference genomes. Likewise, we were able to determine

TABLE A15-4 Illumina Sequences with Remote Homologies to Astroviruses

Virus	Number of sequences
Astrovirus from dog feces	14
Bat astrovirus	47
Bottlenose dolphin astrovirus	1
California sea lion astrovirus	4
Human-mink astrovirus	3
Human astrovirus	59
Rat astrovirus	17
Sheep astrovirus	2
Swine astrovirus	14
Turkey astrovirus	1

The top alignment from tblastx, excluding MLB1 and MLB2, is reported if the alignment was to an astrovirus.

that one of the rhinoviruses we detected was most similar to the recently discovered group C rhinovirus QPM (McErlean et al., 2007).

Do Viral Sequences Correlate with Fever Without a Source?

To compare afebrile children and children with UF, the number of sequence reads was normalized to 3 million per sample. After the adjustment, samples from children with UF had 1.5- to 5-fold more viral sequences than samples from afebrile children in NP and plasma samples, respectively (Figure A15-5A). Although sequencing is not strictly quantitative, the number of sequences generated was inversely correlated with Ct values from the real-time PCR assays

TABLE A15-5 Genome Coverage

Genome	Contigs	Input sequences	Sequences incorporated into contigs	Smallest contig	Largest contig	Genome size	Coverage
Human bocavirus	5	2733		100 nt	1886 nt	5299 nt	92.6%
Respiratory syncitial virus	24	2588		102 nt	1882 nt	15,191 nt	58.4%
Human rhinovirus QPM	2	7159		798 nt	5962 nt	6948 nt*	94.5%
Human parainfluenza virus	10	189		105 nt	803 nt	15, 462 nt	14.2%

*Full genome sequence not available. Largest Genbank sequence used.

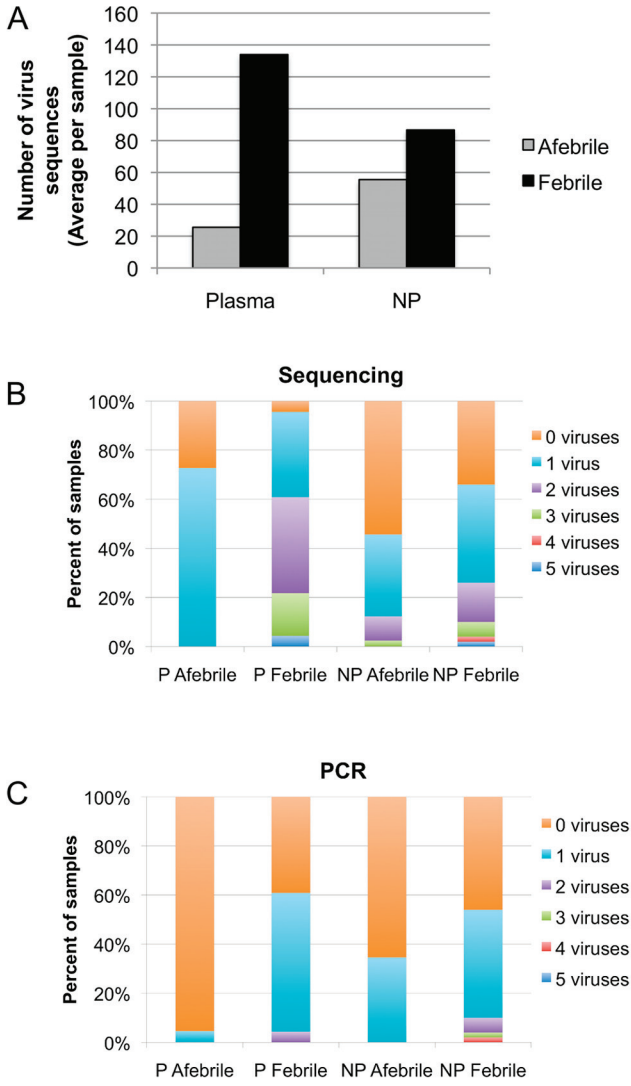


FIGURE A15-5 Febrile children have more viral sequences from a greater range of viruses than do afebrile children. The number of sequences was scaled to 3 million per sample before comparisons were made between groups. (A) The average numbers of viral sequences found in plasma and NP samples from the subjects are represented by gray bars for samples from afebrile children and black bars for samples from febrile children. The percentage of samples in each group for which 0, 1, 2, 3, 4, or 5 viruses was detected is plotted for (B) sequencing data and (C) PCR data.

(Figure S2), which suggests the number of virus reads correlates with the amount of viral genomic material is present. More than one viral genus was found in some samples, and samples from children with fever had a greater number of viruses compared to those from afebrile children (Figure A15-5B). No plasma sample from an afebrile child had more than 1 viral genus detected, compared to 2 to 5 genera detected in 61% of the plasma samples from febrile children (Figure A15-5B). The difference in the percentages of samples with multiple viruses was not as striking in NP samples from febrile and afebrile children, although only 12% of samples from afebrile children had 2 or more viruses compared with 26% of samples from febrile children (Figure A15-5B). In all groups, sequencing detected multiple viral genera in a larger proportion of the samples than did directed PCR assays (Figure A15-5C).

More plasma samples from febrile children were positive for viral sequences than were samples from afebrile children. Anellovirus sequences were the only group found in the plasma from afebrile children. They were found in 80% of all plasma samples, and there was no significant difference between the presence of anellovirus sequences in febrile and afebrile children ($P = 0.2837$, Fisher's Exact Test). The presence of anelloviruses is not surprising, as they infect the majority of children by 1 year of age and establish chronic infections that are detectable in the blood of healthy individuals (Breitbart and Rohwer, 2005; Ninomiya et al., 2008; Vasilyev et al., 2009). By removing the ubiquitous anellovirus sequences from the analysis the difference between the febrile and afebrile groups became even more striking (Figure A15-6A). Most viruses were detected in only a few samples, so differences between the febrile and afebrile groups were not statistically significant for individual viruses in this limited sample set. However, the enterovirus and roseolovirus sequences were more likely to be found in the febrile subjects than the afebrile subjects (Figure A15-6A), consistent with their roles as pathogens that can cause fever.

Viral sequences were also detected more commonly in NP samples from febrile children compared with those from afebrile children (Figure A15-6B). Again, anellovirus sequences were ubiquitous. Enterovirus sequences were found in similar proportions in samples from febrile and afebrile children. Excluding the ubiquitous anelloviruses and enteroviruses, the less common viral sequences were detected more frequently in the febrile subjects compared to the afebrile subjects (Figure A15-6B). Specifically, adenovirus and parechovirus were more commonly associated with NP samples from febrile children (Figure A15-6B). These data indicate that viruses are more commonly associated with samples from febrile children and suggest that viruses are the cause of many fevers in young children for which a source is not determined.

Sequences from febrile children revealed a greater range of viral genera compared to sequences from afebrile children. The difference was most striking in plasma, with sequences from 9 genera found as a result of screening the 23 samples from febrile children and 1 genus found as a result of screening the 22

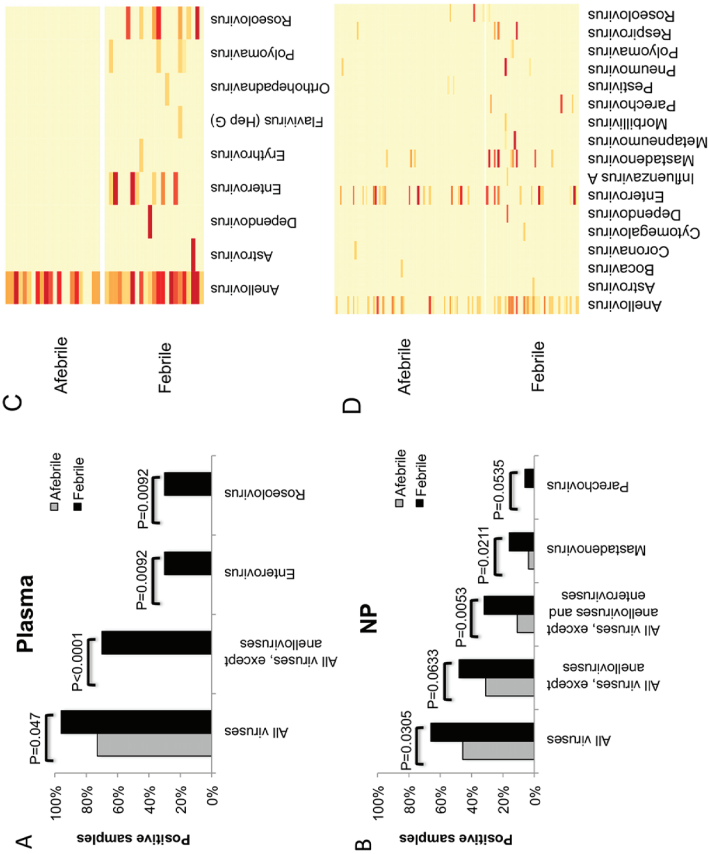


FIGURE A15-6 Prevalence of viruses in samples from febrile compared with afebrile children. The total number of reads per sample was scaled to 3 million to make samples more comparable, and all counts of ≥1 virus sequence were reported. The percent virus-positive samples are graphed for plasma and NP samples, respectively (A and B). P-values were determined using Fisher's exact test. Heatmaps representing the number of virus reads for each virus detected (x-axis) and each sample evaluated (y-axis) are presented for plasma and NP samples, respectively (C and D). The light yellow area is 0 reads, with more intense red representing larger numbers of reads.

samples from afebrile children (Figure A15-6C). In NP samples, sequences from 14 genera were detected by screening the 50 samples from subjects with UF compared with 10 genera detected by screening the 81 samples from afebrile subjects (Figure A15-6D). These data indicate that fever is likely associated with a broad range of viruses, and further studies with larger sample sizes may be important for elucidating the roles of particular viruses in febrile illness.

Discussion

Although it has long been suspected that virus infection is the cause of many unexplained fevers in children under 3 years old, this is the first comprehensive analysis of viruses in samples from children with UF and controls using deep sequencing. We show that more viral sequences from a greater diversity of viruses are found in plasma and NP samples from children with UF than in corresponding samples from afebrile children, which supports the idea that viruses are the cause of many of these unexplained fevers. Children with UF are frequently hospitalized or treated with antibiotics without a positive test for a bacterial infection. The evidence we provide indicates that viruses are commonly associated with UF, and further studies should be done to confirm and elaborate on their role in this clinical syndrome. Ultimately, it would be helpful to identify specific clinical features or tests that could aid diagnosis of virus infection to improve the treatment of children with UF and minimize the unnecessary use of antibiotics.

As expected, the virome of the nasopharynx, which is directly exposed to the environment, is much more complex than the virome detected in plasma. Some viruses found in NP swabs were detected in both febrile and afebrile children. Of particular interest are the Enterovirus sequences, which include rhinoviruses that are known to cause colds. The presence of an enterovirus or rhinovirus in an NP sample from a child with fever would likely lead a physician to conclude that the enterovirus or rhinovirus was the cause of the fever, but we show that Enteroviruses are equally prevalent in the NP samples of afebrile children. These data suggest that in a microbial habitat that is exposed to the environment, the presence of a known pathogen should be interpreted with caution. These data also suggest that we are exposed to a number of known pathogens without showing symptoms of infection, either because the presence of the virus is transient or the particular virus species or strain does not cause symptoms. These observations indicate the importance of future experiments to evaluate the microbiome of the airways over time to look for indicators that a viral infection will become symptomatic, such as correlation of symptoms with specific viral subtypes, correlation with specific biomarkers, or shifts in the larger microbial community structure.

The detection of viruses in the plasma has different implications than in NP samples. Plasma is not generally exposed to the environment, so the presence of a known viral pathogen in the plasma is most likely the result of a disseminated infection. While this study was not designed to determine causation of fever, the

complete absence of known viral pathogens in the plasma of afebrile subjects suggests the viral pathogens detected in the plasma of febrile subjects were the sources of their fevers. While it is more invasive to collect blood than other samples, these data suggest blood samples may provide clearer assessment of viruses that are directly associated with disease in contrast to NP samples where viral pathogens are detected in asymptomatic individuals. Additional studies will need to be done to confirm these ideas. Other viruses, such as anelloviruses, are present chronically in the plasma of healthy people. It remains to be determined what kind of effects long-term exposure to these viruses has on the immune response and human health.

This study could be expanded in several ways in order to better characterize the role of viruses in UF, including detecting viruses in children in whom no viruses have been detected thus far. The first would be to include additional sample types, such as stool. The second would be deeper sequencing of samples, particularly plasma, in which the presence of virus sequences are most likely to be clinically significant. We confirmed that additional sequencing improved virus detection of low abundance virus sequences, and as sequencing costs decrease and analysis tools improve it may be practical to generate and analyze 10 times the number of sequences for each sample to enhance virus detection. It is notable that the use of the Illumina platform in this study enabled the detection of many rare virus sequences, which would likely have been missed using sequencing platforms that generate fewer sequencing reads per unit cost. The third way to improve the study would be further examination of existing sequence data for novel viruses, focusing especially on samples from febrile children with no pathogen detected.

Virus discovery using high-throughput sequencing methods has been very productive in recent years (Briese et al., 2009; Felix et al., 2011; Finkbeiner et al., 2008, 2009; Holtz et al., 2008; Loh et al., 2009). While short-read Illumina sequencing has not been widely adapted for virus discovery in metagenomic samples to date, our findings suggest that this 100-base platform can be applied to virus discovery. For example, the sequences we obtained from the recently discovered astrovirus MLB2 and rhinovirus QPM would have allowed discovery of those viruses based on alignment to other more remotely related reference genomes. In addition, the depth of sequencing gained using the Illumina platform gives the advantage of detecting more virus sequences compared to the 454 platform, which could be advantageous by allowing alignment over different parts of a reference genome, some of which may be more conserved, and by generating enough sequences to enable longer, contiguous sequences to be assembled for further analysis.

An important outcome of this study is to show that deep, Illumina-based sequencing has at least two advantages over targeted, PCR-based assays for the assessment of viruses in clinical samples. First, sequencing does not require prior knowledge of which viruses might be in the sample, thus allowing the detection

of unexpected and novel viruses. Second, sequencing can often provide information such as virus subtype or sequence variation from reference genomes, which adds detail to our understanding of the viruses present. Our study illustrates both of these advantages. First, we identified viruses that would not have been routinely queried by PCR assays for known pathogens. For example, we detected the astrovirus MLB2 in plasma and NP samples from a febrile child, which were subsequently confirmed by PCR in both samples ([Holtz et al., 2011] and data not shown). Because no other cause of the fever has been detected, these data suggest MLB2 is the cause of this subject's fever and further examination of the role of this virus in pediatric fever is warranted.

The second advantage of sequencing, the ability to determine virus subtype or sequence variation from reference genomes, is also evident in our study. For example, we were able to identify specific types or subtypes or strains of rhinovirus and bocavirus. Notably, this can often be accomplished without sequencing most of the viral genome. In the case of HHV-6, all of the positive plasma samples were determined to be serotype 6B, even though 4 of the 8 samples had fewer than 15 HHV-6 sequences. We were also able to make distinctions between anellovirus species TTV, TTMDV, and TTMV with as little as one read. In future studies we will examine how different virus species and subtypes correlate with clinical symptoms.

One challenge in analyzing the virome in metagenomic samples is the speed of alignment tools available. Aligners designed for large data sets with short sequences generally gain processing speed by sacrificing the ability to identify sequences that differ more than slightly from the reference genome. Thus, many of these very fast aligners cannot be used effectively for analysis of virus sequences, which frequently differ considerably from their most closely related reference sequences. We are implementing new tools to be used for virome analysis that improve the speed of nucleotide and amino acid sequence alignments while retaining most of the sensitivity, which will allow the efficient analysis of a greater number of sequences. A second challenge for virome analysis is the use of a more inclusive reference database (such as NCBI's NT) because this would allow identification of more virus sequences based on sequence similarity; however, alignment results from a large database can be problematic for several reasons: (a) taxonomy can be irregular causing computational problems and (b) some of the viral entries contain sequences from the human genome or bacterial cloning vectors, which cause false positive alignments. We have addressed these problems in the present study by manually reviewing the data, but our goal is to develop an easily updated, semi-curated database that would minimize these problems. Future versions of this analysis protocol will be improved with faster alignment tools and improved databases.

This study of deep sequencing of samples from febrile and afebrile children indicates that viruses are frequently detected in both groups, but with greater frequency and diversity in the samples from children with fever of unknown

cause. A causal role for these viruses would have important implications for the medical treatment of these children, since the children would not require antibiotic therapy. In evaluating viral causes of fever, sequencing appears to be advantageous in that it frequently reveals the presence of multiple viruses in a given sample, including unexpected viruses. Highly sensitive and specific PCR assays for a subset of viruses complement the sequencing analysis. As sequencing continues to become less expensive and the speed of computational tools improves, it is possible that its sensitivity could match that of PCR. This could lead to a powerful diagnostic approach: rapid, unbiased sequence analysis of the microbiome in patient samples, which could identify potentially pathogenic viruses and other microbes, followed by confirmation of the results using highly targeted and extremely specific PCR assays.

Methods

Ethics Statement

Samples were collected from human subjects using a protocol that was approved by the Washington University Human Research Protection Office. Written informed consent was obtained from the parents or legal guardians of all subjects.

Sample Collection

The subjects included were febrile and afebrile children 2 to 36 months of age seen at St. Louis Children's Hospital. The group of febrile children was comprised of patients seen in the emergency room who had fever without an obvious source. In order to be included in the study, the physicians must have elected to obtain blood for testing. The afebrile group was comprised of children undergoing surgery. NP swabs and plasma samples were collected as described (Colvin et al., manuscript submitted). NP swabs were collected by inserting flocked swabs into the nasopharyngeal area, rotating the swab, and holding the swab in place for 10–15 seconds to increase specimen collection. Swabs were submerged in Universal Transport Medium (Copan), and the shafts of the swabs were cut or broken off and discarded. The medium containing the swab was briefly vortexed, and then the swabs were removed without wringing out any absorbed medium. Tubes were centrifuged, and the supernatant was aliquotted and frozen at −70°C. Total nucleic acid was extracted using the Qiagen BioRobot M48 the Roche MagNA Pure automated extractor for NP and plasma samples, respectively.

Sample Preparation and Sequencing

For samples from afebrile and febrile children, sequencing libraries were prepared to look for DNA and RNA viruses. DNA and RNA were prepared as

previously described (Wang et al., 2002, 2003). In brief, using total nucleic acid templates, RNA was primed with Primer A for reverse transcription. Sequenase DNA polymerase was used for second strand synthesis. DNA and RNA fragments were amplified with Primer B for 40 cycles. Samples were sequenced on the Roche 454 GS FLX Titanium or Illumina GAIIX. For samples in which additional sequencing reads were generated, the Illumina GAIIX or Illumina HiSeq 2000 was used (Figure S3). The SRR accession numbers for the sequence data are provided in Figure S4.

Sequence Analysis

A pipeline was developed for the analysis of large numbers of short sequence reads. This was adapted from that used for the analysis of 454 sequences, which used BLASTn and tBLASTx (Altschul et al., 1997) to align sequences to references in the NT database,[59] followed by a manual review of the viral alignments. The details of the protocol for analysis of short reads follow. After removal of primer sequences, completely identical sequences were collapsed into a single representative sequence to minimize the number of sequences to be analyzed. Low complexity sequences were then masked using Dust (Morgulis et al., 2006). Sequences with greater than 20 N nucleotides (either from sequencing error or as a result of Dust) were removed. Human sequences were identified for removal by aligning sequences to the Genome Reference Consortium's human build 36[60] including unplaced, human mitochondrial, and 5.8 s, 18 s, and 28 s rDNA sequences using cross_match (Green, 1994) with the following alignment parameters: minscore 70, bandwidth 3, penalty –1, gap_init –1, gap_ext –1, masklevel 0. Non-human sequences were aligned to a metagenomic database consisting of all virus and phage sequences in NCBI NT plus full genomes from other microbes including bacteria, archaea, and small eukaryotes (Mitreva, et al., unpublished). Cross_match was used with the same parameters used for the human alignments. Any sequences that were unaligned using nucleotide alignment were then aligned to NR (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) using WU-BLAST (BlastX) (Altschul et al., 1990) with the following parameters: filter seg, W 6, WINK 6, nogap. Sequences that aligned to microbial references using either cross_match or WU-BLAST were confirmed by WU-BLAST alignment to the larger NT database. Virus alignments were then manually evaluated, and ambiguous alignments were removed. The same protocol was used for the analysis of the 75-mer data, except a minscore of 50 was used in the cross_match alignments. Detailed sequence statistics are presented in Figure S5. Figure S6 shows the number of virus sequences found with cross_match and BlastX, without scaling.

[59] ftp://ftp.ncbi.nlm.nih.gov/blast/db

[60] http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/data.shtml

Viral sequences were assembled into contigs using Tigra (Chen L and Weinstock G, unpublished).

Acknowledgments

We thank David Wang for helpful discussion, Makedonka Mitreva, John Martin, Sahar Abubucker, and Karthik Kota for providing human reference and non-viral microbial reference databases, and Todd Wylie, John Martin, Eric Becker, and Matt Callaway for assistance with programming and parallelization of alignments.

Author Contributions

Conceived and designed the experiments: KMW KAM ES GMW GAS. Performed the experiments: KMW KAM. Analyzed the data: KMW KAM GAS. Wrote the paper: KMW GMW GAS. Contributed patient samples: GAS.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402. doi: 10.1093/nar/25.17.3389.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. PLoS Biol 4: e368. doi: 10.1371/journal.pbio.0040368.

Baraff LJ (2000) Management of fever without source in infants and children. Ann Emerg Med 36: 602–614.

Breitbart M, Rohwer F (2005) Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. Biotechniques 39: 729–736. doi: 10.2144/000112019.

Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, et al. (2009) Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS Pathog 5: e1000455. doi: 10.1371/journal.ppat.1000455.

Felix MA, Ashe A, Piffaretti J, Wu G, Nuez I, et al. (2011) Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses. PLoS Biol 9: e1000586. doi: 10.1371/journal.pbio.1000586.

Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, et al. (2008) Metagenomic analysis of human diarrhea: viral detection and discovery. PLoS Pathog 4: e1000011. doi: 10.1371/journal.ppat.1000011.

Finkbeiner SR, Holtz LR, Jiang Y, Rajendran P, Franz CJ, et al. (2009) Human stool contains a previously unrecognized diversity of novel astroviruses. Virol J 6: 161. doi: 10.1186/1743-422X-6-161.

Green P (1994) Cross_match. Unpublished manuscript. http://wwwphraporg.

Holtz LR, Finkbeiner SR, Kirkwood CD, Wang D (2008) Identification of a novel picornavirus related to cosaviruses in a child with acute diarrhea. Virol J 5: 159. doi: 10.1186/1743-422X-5-159.

Holtz LR, Wylie KM, Weinstock GM, Sodergren E, Jiang Y, et al. (2011) Astrovirus MLB2 Viremia in a Febrile Child. Emerg Inf Dis. (in press).

Hormozdi DJ, Arens MQ, Le BM, Buller RS, Agapov E, et al. (2010) KI polyomavirus detected in respiratory tract specimens from patients in St. Louis, Missouri. Pediatr Infect Dis J 29: 329–333. doi: 10.1097/INF.0b013e3181c1795c.

Krauss BS, Harakal T, Fleisher GR (1991) The spectrum and frequency of illness presenting to a pediatric emergency department. Pediatr Emerg Care 7: 67–71. doi: 10.1097/00006565-199104000-00001.

Loh J, Zhao G, Presti RM, Holtz LR, Finkbeiner SR, et al. (2009) Detection of novel sequences related to african Swine Fever virus in human serum and sewage. J Virol 83: 13019–13025. doi: 10.1128/JVI.00638-09.

McErlean P, Shackelton LA, Lambert SB, Nissen MD, Sloots TP, et al. (2007) Characterisation of a newly identified human rhinovirus, HRV-QPM, discovered in infants with bronchiolitis. J Clin Virol 39: 67–75. doi: 10.1016/j.jcv.2007.03.012.

Morgulis A, Gertz EM, Schaffer AA, Agarwala R (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol 13: 1028–1040. doi: 10.1089/cmb.2006.13.1028.

Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, et al. (2009) Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. PLoS One 4: e4219. doi: 10.1371/journal.pone.0004219.

Ninomiya M, Takahashi M, Nishizawa T, Shimosegawa T, Okamoto H (2008) Development of PCR assays with nested primers specific for differential detection of three human anelloviruses and early acquisition of dual or triple infection during infancy. J Clin Microbiol 46: 507–514. doi: 10.1128/JCM.01703-07.

Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334–338. doi: 10.1038/nature09199.

Rudinsky SL, Carstairs KL, Reardon JM, Simon LV, Riffenburgh RH, et al. (2009) Serious bacterial infections in febrile infants in the post-pneumococcal conjugate vaccine era. Acad Emerg Med 16: 585–590. doi: 10.1111/j.1553-2712.2009.00444.x.

Sumino KC, Walter MJ, Mikols CL, Thompson SA, Gaudreault-Keener M, et al. (2010) Detection of respiratory viruses and the associated chemokine responses in serious acute respiratory illness. Thorax 65: 639–644. doi: 10.1136/thx.2009.132480.

Vasilyev EV, Trofimov DY, Tonevitsky AG, Ilinsky VV, Korostin DO, et al. (2009) Torque Teno Virus (TTV) distribution in healthy Russian population. Virol J 6: 134. doi: 10.1186/1743-422X-6-134.

Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, et al. (2009) Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J Virol 83: 4642–4651. doi: 10.1128/JVI.02301-08.

Waddle E, Jhaveri R (2009) Outcomes of febrile children without localising signs after pneumococcal conjugate vaccine. Arch Dis Child 94: 144–147. doi: 10.1136/adc.2007.130583.

Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, et al. (2002) Microarray-based detection and genotyping of viral pathogens. Proc Natl Acad Sci U S A 99: 15687–15692. doi: 10.1073/pnas.242579699.

Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, et al. (2003) Viral discovery and sequence recovery using DNA microarrays. PLoS Biol 1: E2. doi: 10.1371/journal.pbio.0000002.

Watt K, Waddle E, Jhaveri R (2010) Changing epidemiology of serious bacterial infections in febrile infants without localizing signs. PLoS One 5: e12448. doi: 10.1371/journal.pone.0012448.

Wilkinson M, Bulloch B, Smith M (2009) Prevalence of occult bacteremia in children aged 3 to 36 months presenting to the emergency department with fever in the postpneumococcal conjugate vaccine era. Acad Emerg Med 16: 220–225. doi: 10.1111/j.1553-2712.2008.00328.x.

Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One 4: e7370. doi: 10.1371/journal.pone.0007370.

Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, et al. (2011) Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. Proc Natl Acad Sci U S A 108: 4547–4553. doi: 10.1073/pnas.1000089107.

Appendix B

Agenda

The Science and Applications of Microbial Genomics

June 12–13, 2012
500 Fifth Street, NW
Washington DC

DAY ONE: TUESDAY, JUNE 12, 2012

8:45–9:15: Registration and Continental Breakfast

9:15–9:30: Welcoming Remarks: David Relman, James Hughes, and Lonnie King

9:30–10:00: KEYNOTE: *Yersinia pestis* Population Genetics Across Time and Space
Paul Keim, Northern Arizona University

10:00–10:20: Discussion

10:20–10:30: BREAK

SESSION I: Microbial Genomics—Diversity, Evolution, and Adaptation

Arturo Casadevall, Moderator

10:30–11:00: The Earth Microbiome Project: Modeling the Earth's Microbiome
Jack A. Gilbert, Argonne National Laboratory/University of Chicago

11:00–11:30: Variation in Microbial Communities and Genomes
George Weinstock, Washington University in St. Louis

11:30–12:00: Population Diversity in Deep-Sea Microbial Communities
Peter Girguis, Harvard University

12:00–12:30: The Application of Computational/Theoretical and Experimental Approaches to Study the Evolution of Microorganisms
Eric Alm, Massachusetts Institute of Technology

12:30–1:00: Discussion

1:00–1:45: LUNCH

SESSION II: Microbial Genomics—Molecular Mechanisms of Disease Emergence and Epidemiology

David Relman, Moderator

1:45–2:15: Characterizing Intra-host Influenza Virus Populations to Predict Emergence
Elodie Ghedin, University of Pittsburgh School of Medicine

2:15–2:45: Identifying Signatures of Recent Selection and Transmission in Pathogenic Bacteria
Julian Parkhill, The Sanger Institute

2:45–3:15: Comparative Genomics of *E. coli* and *Shigella*: Identification and Characterization of Pathogenic Variants Based on Whole Genome Sequence Analysis
David Rasko, University of Maryland Institute for Genome Sciences

3:15–3:45: BREAK

3:45–4:15: Coral Health and Disease in the Face of Climate Change
Kim Ritchie, Mote Marine Laboratory

4:15–4:45: Evolution and Pathogenicity in the Deadly Chytrid Pathogen of Amphibians
Erica Bree Rosenblum, University of California–Berkeley

4:45–5:15: Discussion

5:15–6:00: Concluding Remarks

6:15: ADJOURN DAY ONE

DAY TWO: WEDNESDAY, JUNE 13, 2012

8:30–9:00: Registration and Continental Breakfast

9:00–9:15: Welcoming Remarks and Summary of Day One: David Relman

SESSION III: Application of Genomics and High-Throughput Technologies for Microbial Surveillance and Outbreak Traceback

Claire Fraser, Moderator

9:15–9:45: Virulence as an Emergent Property
Arturo Casadevall, Albert Einstein College of Medicine

9:45–10:15: Understanding the Origins, Evolution, and Transmission Dynamics of Outbreak Agents Through Genomic Epidemiology
Jennifer Gardy, British Columbia Centre for Disease Control/University of British Columbia

10:15–10:45: Use of Genomic Platforms to Detect and Discover Emerging/Evolving Viral Diseases
David Wang, Washington University in St. Louis

10:45–11:00: BREAK

11:00–11:30: Genomic Epidemiology of Gram-Negative Pathogens: From *Acinetobacter* to *E. coli*
Mark Pallen, University of Birmingham

11:30–12:00: The Impact of Sequencing Errors on Estimates of Diversity in the Rare Biosphere (and Potential Solutions)
Susan Huse, Marine Biological Laboratory

12:00–12:30: Discussion

12:30–1:15: LUNCH

SESSION IV: Microbial Forensic Tools, Technologies, and Platforms: Problems of Concordance and Discordance

Paul Keim, Moderator

1:15–1:45: Microbial Forensics
Bruce Budowle, University of North Texas Health Science Center

1:45–2:15: Discussion of the Technical Approaches Used in the Amerithrax Investigation
Claire Fraser, University of Maryland Institute for Genome Sciences

2:15–2:45: Analyzing Metagenomic Data: Inferring Microbial Community Function with MG-RAST
Folker Meyer, Argonne National Laboratory

2:45–3:00: BREAK

3:00–5:00: **Panel Discussion:** The Problem of Concordance and Discordance in Data Generated Using Different Platforms and Technologies
DISCUSSANTS:
- George Weinstock
- Susan Huse
- Mark Pallen
- Jack Gilbert

5:00–5:15: Concluding Remarks

5:15: ADJOURN

Appendix C

Acronyms

AEEC	attaching and effacing *Escherichia coli*
AFMIC	Armed Forces Medical Intelligence Center
ANSI	American National Standards Institute
ATCC	American Type Culture Collection
Bd	*Batrachochytrium dendrobatidis*
BFP	bundle-forming pili
BLAST	basic local alignment search tool
CDC	Centers for Disease Control and Prevention
CLIA	clinical laboratory improvement amendment
COG	cluster of orthologous groups of proteins
DEC	diarrheagenic *Escherichia coli*
DHS	Department of Homeland Security
DNA	deoxyribonucleic acid
DOD	Department of Defense
DOJ	Department of Justice
DRISEE	duplicate read inferred sequencing error estimation
DTRA	Defense Threat Reduction Agency
EAEC	enteroaggregative *Escherichia coli*
EHEC	enterohemorrhagic *Escherichia coli*
EID	emerging infectious disease
EIEC	enteroinvasive *Escherichia coli*

EPEC	enteropathogenic *Escherichia coli*
ESS	Environmental Shotgun Sequencing
ETEC	enterotoxigenic *Escherichia coli*
FBI	Federal Bureau of Investigation
FDA	U.S. Food and Drug Administration
GEBA	Genomic Encyclopedia of Bacteria and Archaea
GBS	Group B streptococcus
HAART	highly active antiretroviral therapy
HAI	hospital-acquired infection
HGT	horizontal gene transfer
HHS	Department of Health and Human Services
HIV	human immunodeficiency virus
IOM	Institute of Medicine
ISO	International Organization for Standardization
KEGG	Kyoto Encyclopedia of Genes and Genomes
LEE	locus of enterocyte effacement
MIRU	mycobacterial interspersed repetitive unit
MLST	multi-locus sequence typing
MRSA	methicillin-resistant *Staphylococcus aureus*
NGS	next-generation sequencing
ORF	open reading frame
OTU	operational taxonomic unit
PAUP	phylogenetic analysis using parsimony
PCoA	principal coordinates analysis
PCR	polymerase chain reaction
RNA	ribonucleic acid
rRNA	ribosomal RNA
SNP	single nucleotide polymorphisms
STEC	shiga toxin-producing *Escherichia coli*

TB	tuberculosis
TIGR	The Institute for Genomic Research
USAMRIID	U.S. Army Medical Research Institute of Infectious Diseases
USDA	U.S. Department of Agriculture
VNTR	variable number tandem repeat

Appendix D

Glossary

Amplicon: Pieces of DNA formed as the products of natural or artificial amplification events. For example, they can be formed via polymerase chain reactions (PCRs) or ligase chain reactions (LCRs), as well as by natural gene duplication.

Animalcules: A microscopic or minute organism, such as an amoeba or paramecium, usually considered to be an animal.

Antibiotic: Class of substances that can kill or inhibit the growth of some groups of microorganisms. Originally antibiotics were derived from natural sources (e.g., penicillin from molds), but many currently used antibiotics are semisynthetic and modified with additions of man-made chemical components. See *Antimicrobial*.

Antibiotic resistance: Property of bacteria that confers the capacity to inactivate or exclude antibiotics or a mechanism that blocks the inhibitory or killing effects of antibiotics.

Antimicrobial: Any agent (including an antibiotic) used to kill or inhibit the growth of microorganisms (bacteria, viruses, fungi, or parasites). This term applies whether the agent is intended for human, veterinary, or agricultural applications.

Archaea: One of the three main branches of evolutionary descent (Archaea, Eukaryota, and Bacteria), archaea are single-celled organisms once classified as extremophiles (being found in harsh environments such as hot springs and salt lakes), yet recent evidence shows that archaea are widely distributed in nature.

Astroviruses: Small, single-stranded, positive sense RNA viruses, 6 to 8 kb in length. They cause diarrhea in humans and other animals.

Attaching and effacing *Escherichia coli* (AEEC): *E. coli* that produces Shiga toxin (verotoxin). Certain serotypes cause enteritis, colitis, and diarrhea in a number of different animal species by expressing a virulence factor protein called intimin that allows intimate attachment of the organism to the microvillus brush border of the enterocyte, forming a characteristic attaching and effacing lesion. Diagnosis is by the detection of the Shiga toxin and characteristic lesions.

Bacteria: Microscopic, single-celled organisms that have some biochemical and structural features different from those of animal and plant cells.

Bacteriophage: A virus that infects bacteria.

Bundle-forming pilus: So named because of its tendency to aggregate into rope-like bundles. The bundle-forming pilus is a member of a family of pili produced by important pathogens of humans and domestic animals known as type IV fimbriae.

Chemoautotrophic: Organisms that use energy derived from the oxidation of inorganic compounds to fix carbon.

Contig: A group of overlapping clones representing regions of the genome; the contiguous sequence of DNA created by assembling these overlapping chromosome fragments.

Diarrheagenic *Escherichia coli* (DEC): Any of a number of *E. coli* bacteria that cause diarrhea.

Disease: In medicine, disease is often viewed as an observable change of the normal network structure of a system resulting in damage to the system.

DNA (deoxyribonucleic acid): Any of various nucleic acids that are usually the molecular basis of heredity, are constructed of a double helix held together by hydrogen bonds between purine and pyrimidine bases that project inward from two chains containing alternate links of deoxyribose and phosphate, and that in eukaryotes are localized chiefly in cell nuclei.

Duplicate read inferred sequencing error estimation (DRISEE): A new tool to estimate the amount of "noise" in metagenomic data sets.

Earth Microbiome Project: An initiative to collect natural samples and to analyze the microbial community around the globe.

Emerging infection: Infections that are rapidly increasing in incidence or geographic range.

Enteroaggregative *Escherichia coli* (EAEC): A subgroup of diarrheagenic *E. coli* (DEC) that during the past decade has received increasing attention as a cause of watery diarrhea, which is often persistent. EAEC have been isolated from children and adults worldwide.

Enterohemorrhagic *Escherichia coli* (EHEC): A strain of *E. coli* that causes hemorrhage in the intestines. The organism produces Shiga toxin, which damages bowel tissue, causing intestinal ischemia and colonic necrosis. Symptoms are stomach cramping and bloody diarrhea. An infectious dose may be as low as 10 organisms. Spread by contaminated beef, unpasteurized milk and juice, sprouts, lettuce, salami, and contaminated water, the infection can be serious although there may be no fever. Treatment consists of antibiotics and maintenance of fluid and electrolyte balance. In advanced cases, surgical removal of portions of the bowel may be required.

Enteroinvasive *Escherichia coli* (EIEC): A strain of Escherichia coli that penetrates gut mucosa and multiplies in colon epithelial cells, resulting in shigellosis-like changes of the mucosa. This strain produces a severe diarrheal illness that can resemble shigellosis except for the absence of vomiting and shorter duration of illness.

Enteropathogenic *Escherichia coli* (EPEC): Strains of *E. coli* that cause enteritis by close association with enteric cells. This group includes attaching and effacing *E. coli*.

Enterotoxigenic *Escherichia coli* (ETEC): A strain of *E. coli* that is a frequent cause of diarrhea in travelers.

***Escherichia coli*:** A straight rod-shaped gram-negative bacterium that is used in public health as an indicator of fecal pollution (as of water or food) and in medicine and genetics as a research organism and that occurs in various strains that may live as harmless inhabitants of the human lower intestine or may produce a toxin causing intestinal illness.

Genetic fingerprinting: A method employed to determine differences in amino acid sequences between related proteins; relies upon the presence of a simple tandem-repetitive sequences that are scattered throughout the human genome.

Genome: The complete set of genetic information in an organism. In bacteria, this includes the chromosome(s) and plasmids (extrachromosomal DNA molecules that can replicate autonomously within a bacterial cell).

Genome metastructure: Organization of the genome with respect to where the various structural and functional components are located.

Genomics: The study of genes and their associated functions.

Gram-negative bacteria: Refers to the inability of a microorganism to accept a certain stain. This inability is related to the cell wall composition of the microorganism and has been useful in classifying bacteria.

Gram-positive bacteria: Refers to the ability of a microorganism to retain a certain stain. This ability is related to the cell wall composition of the microorganism and has been useful in classifying bacteria.

Highly active antiretroviral therapy (HAART): The name given to treatment regimens that aggressively suppress HIV replication and progression of HIV disease. The usual HAART regimen combines three or more anti-HIV drugs from at least two different classes.

Horizontal gene transfer: Any process in which an organism incorporates genetic material from another organism without being the offspring of that organism.

Human Genome Project: An international scientific research project with a primary goal of determining the sequence of chemical base pairs that make up DNA, and of identifying and mapping the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint. A working draft of the genome was announced in 2000 and a complete one in 2003, with further, more detailed analysis still being published.

Human immunodeficiency virus (HIV): Retrovirus that causes AIDS by infecting helper T cells of the immune system. The most common serotype, HIV-1, is distributed worldwide, while HIV-2 is primarily confined to West Africa.

Human Microbiome Project: A U.S. National Institutes of Health initiative with the goal of identifying and characterizing the microorganisms that are found in association with both healthy and diseased humans (i.e., their microbial flora). The ultimate goal of this and similar NIH-sponsored microbiome projects is to test how changes in the human microbiome are associated with human health or disease.

Indel: An insertion or deletion of genetic material. Indel refers to the mutation class that includes both insertions, deletions, and the combination thereof, including insertion and deletion events that may be separated by many years.

Metagenomics: A culture-independent analysis method that involves obtaining DNA from communities of microorganisms, sequencing it in a "shotgun" fashion, and characterizing genes and genomes comparisons with known gene sequences. With this information, researchers can gain insights into how members of the microbial community may interact, evolve, and perform complex functions in their habitats.

Metagenomics Rapid Annotation using Subsystem Technology server (MG-RAST): A community resource providing an automated analysis platform for metagenomes providing quantitative insights into microbial populations based on sequence data.

Methicillin-resistant *Staphylococcus aureus* (MRSA): A type of staph bacteria that is resistant to certain antibiotics called beta-lactams. These antibiotics include methicillin and other more common antibiotics such as oxacillin, penicillin, and amoxicillin. See *Antibiotic resistance*.

Microbe: A microorganism or biologic agent that can replicate in humans (including bacteria, viruses, protozoa, fungi, and prions).

Microbial threat: Microbes that lead to disease in humans.

Multilocus sequence typing (MLST): A technique in molecular biology for the typing of multiple loci. The procedure characterizes isolates of microbial species using the DNA sequences of internal fragments of multiple housekeeping genes. Approximately 450–500 bp internal fragments of each gene are used, as these can be accurately sequenced on both strands using an automated DNA sequencer. For each housekeeping gene, the different sequences present within a bacterial species are assigned as distinct alleles and, for each isolate, the alleles at each of the loci define the allelic profile or sequence type (ST).

Mycobacterial interspersed repetitive unit (MIRU): Short tandem repeat structures found at multiple loci throughout the Mycobacterium tuberculosis genome that have been used for typing these pathogens.

Operational taxonomic unit (OTU): Taxonomic level of sampling selected by the user to be used in a study, such as individuals, populations, species, genera, or bacterial strains.

Parasite: An organism that lives in or on and takes its nourishment from another organism. A parasite cannot live independently. Parasitic diseases include infections by protozoa, helminths, and arthropods.

Pathogen: Organism capable of causing disease.

Pathogenic: Capable of causing disease.

Phage: A virus that infects bacteria. Many phage have proved useful in the study of molecular biology and as vectors for the transfer of genetic information between cells. Lytic phage (e.g., the T series phage that infect *Escherichia coli* [coliphages]), invariably lyse a cell following infection; temperate phage (e.g., lambda bacteriophage) can also undergo a lytic cycle or can enter a lysogenic cycle, in which the phage DNA is incorporated into that of the host, awaiting a signal that initiates events leading to replication of the virus and lysis of the host cell.

Phylogenetics: The study of evolutionary relationships among groups of organisms (e.g., species, populations) that is discovered through molecular sequencing data and morphological data matrices.

Polymerase chain reaction (PCR): A scientific technique in molecular biology to amplify a single or a few copies of a piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence.

Principal coordinates analysis: A method to explore and to visualize similarities or dissimilarities of data. It starts with a similarity matrix or dissimilarity matrix (= distance matrix) and assigns for each item a location in a low-dimensional space, such as a 3-D graphics.

Rare biosphere: Rare microbial species in the soil, ocean, and living creatures that were effectively cloaked from previous sampling methods by more abundant species.

Resistance: See *Antibiotic resistance*.

RNA (ribonucleic acid): Any of various nucleic acids that contain ribose and uracil as structural components and are associated with the control of cellular chemical activities.

rRNA (ribosomal RNA): A nucleic acid found in all living cells. Plays a role in transferring information from DNA to the protein-forming system of the cell.

More specifically, rRNA sits in the ribosome, decoding the mRNA into various amino acids and assisting in translation.

Shiga toxin-producing *Escherichia coli* (STEC): A type of enterohemorrhagic *E. coli* (EHEC) bacteria that can cause illness ranging from mild intestinal disease to severe kidney complications. Other types of enterohemorrhagic *E. coli* include the relatively important serotype *E. coli* O157:H7, and more than 100 other non-O157 strains.

Shotgun sequencing: The sequencing of a large DNA segment by sequencing of randomly derived sub-segments, whose order and orientation within the large segment are unknown until the final assembly of overlapping sequences.

Single nucleotide polymorphism (SNP): A DNA sequence variation occurring when a single nucleotide—A, T, C, or G—in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes in an individual.

***Staphylococcus aureus*:** A gram-positive bacteria that is the most common cause of staph infections. It is frequently part of the skin flora found in the nose and on skin. About 20 percent of the human population are long-term carriers of *S. aureus*.

Sympatric speciation: Speciation in the absence of physical barriers to genetic exchange between incipient species.

Taxon: A particular taxonomic grouping, such as a species, genus, family, order, class, phylum, or kingdom.

Tuberculosis (TB): A potentially fatal contagious disease that can affect almost any part of the body but is mainly an infection of the lungs. It is caused by a bacterial microorganism, the tubercle bacillus or *Mycobacterium tuberculosis*.

Vaccine: A vaccine is a biological preparation that improves immunity to a particular disease. A vaccine typically contains an agent that resembles a disease-causing microorganism, and is often made from weakened or killed forms of the microbe or its toxins.

Variable number tandem repeats (VNTRs): Short nucleotide sequences that are present in multiple copies at a particular locus in the genome. The number of repeats can vary from individual to individual, making analysis of VNTRs useful for subtyping of microorganisms.

Virulence factor: Intrinsic characteristic of an infectious bacteria that facilitates its ability to cause disease.

Virus: A small infectious agent that can only replicate inside the cells of another organism. Viruses are too small to be seen directly with a light microscope. Viruses infect all types of organisms, from animals and plants to bacteria and archaea.

Whole genome sequencing: A laboratory process that determines the complete DNA sequence of an organism's genome at a single time. This entails sequencing all of an organism's chromosomal DNA as well as DNA contained in the mitochondria and, for plants, in the chloroplast. Almost any biological sample containing a full copy of the DNA—even a very small amount of DNA or ancient DNA—can provide the genetic material necessary for whole genome sequencing.

Appendix E

Speaker Biographies

Eric Alm, Ph.D., is the Doherty Assistant Professor of Ocean Utilization at the Massachusetts Institute of Technology (MIT). His research includes both computational/theoretical and experimental approaches to understanding the evolution of microorganisms, emphasizing a "systems-level" perspective. Some areas of special interest include: tools for detecting natural selection in microbes; the evolutionary origin of gene families; mining metagenomic sequence data; experimental evolution of microbes; modeling bacterial ecology; gene regulatory networks in bacteria; and protein structure and design. He enjoys teaching a variety of classes at MIT, spanning his diverse interests in microbiology, computer algorithms, and thermodynamics of biomolecules. He is currently looking forward to teaching a new class on microbial evolution and genetics. Dr. Alm has earned the following degrees: B.S., 1995, University of Illinois at Urbana-Champaign; M.S., 1997, University of California, Riverside; Ph.D., 2001, University of Washington, Seattle; and Postdoc, 2005, University of California, Berkeley/Lawrence Berkeley National Laboratory.

Bruce Budowle, Ph.D., received a Ph.D. in genetics in 1979 from Virginia Polytechnic Institute and State University. From 1979 to 1982, Dr. Budowle was a postdoctoral fellow at the University of Alabama at Birmingham. Working under a National Cancer Institute fellowship, he carried out research predominately on genetic risk factors for diseases such as insulin-dependent diabetes mellitus, melanoma, and acute lymphocytic leukemia. From 1983 to 2009, Dr. Budowle was employed at the FBI Laboratory Division and carried out research, development, and validation of methods for forensic biological analyses. Dr. Budowle has worked on laying some of the foundations for the current statistical analyses

in forensic biology and defining the parameters of relevant population groups. He has published approximately 500 articles, made approximately 600 presentations, and testified in more than 250 criminal cases in the areas of molecular biology, population genetics, statistics, quality assurance, and forensic biology. He has been a chair and member of the Scientific Working Group on DNA Methods, Chair of the DNA Commission of the International Society for Forensic Genetics, and a member of the DNA Advisory Board. He was one of the architects of the CODIS National DNA database, which maintains DNA profiles from convicted felons, from evidence in unsolved cases, and from missing persons. Some of Dr. Budowle's efforts over the past decade are in counter-terrorism, primarily in identification of victims from mass disasters and in efforts involving microbial forensics and bioterrorism. Dr. Budowle was an advisor to New York State in the effort to identify the victims from the World Trade Center attack. In the area of microbial forensics, Dr. Budowle has been the chair of the Scientific Working Group on Microbial Genetics and Forensics, whose mission was to set quality assurance guidelines, develop criteria for biologic and user databases, set criteria for a National Repository, and develop forensic genomic applications. In 2009 Dr. Budowle became Executive Director of the Institute of Applied Genetics and Professor in the Department of Forensic and Investigative Genetics at the University of North Texas Health Science Center at Fort Worth, Texas. His current efforts focus on the areas of human forensic identification, microbial forensics, and emerging infectious disease.

Arturo Casadevall, M.D., Ph.D., is Professor and Chair of the Department of Microbiology & Immunology at the Albert Einstein College of Medicine. Dr. Casadevall received both his M.D. and Ph.D. from New York University and completed his internship and residency in internal medicine at Bellevue Hospital in New York, New York. Afterward, he completed subspecialty training in Infectious Diseases at the Albert Einstein College of Medicine. Dr. Casadevall's major research interests are in fungal pathogenesis and the mechanism of antibody action. He has authored more than 540 scientific papers. Dr. Casadevall was elected to membership in the American Society for Clinical Investigation, the American Academy of Physicians, and the American Academy of Microbiology. He is a fellow of the American Association for the Advancement of Science and has received numerous honors including the Solomon A Berson Medical Alumni Achievement Award from New York University, the Maxwell L. Littman Award, the Rhoda Benham Award from Medical Mycology Society of America, and the Kass Lecturership from Infectious Diseases Society of America. He is the Editor in Chief of *mBio* and serves on numerous editorial boards. Dr. Casadevall served on the National Academy of Sciences committee that reviewed the FBI investigation of the anthrax attacks in 2001. He serves on the National Science Advisory Board for Biosecurity and the National Institute of Allergy and Infectious Diseases Board of Scientific Counselors.

Claire Fraser, Ph.D., is Director of the Institute for Genome Sciences at the University of Maryland School of Medicine in Baltimore, Maryland. She has joint faculty appointments at the University of Maryland School of Medicine in the Department of Medicine and Microbiology/Immunology.

Until 2007, she was President and Director of The Institute for Genomic Research (TIGR) in Rockville, Maryland, and led the teams that sequenced the genomes of several microbial organisms, including important human and animal pathogens. She helped launch the new field of microbial genomics and revolutionized the way microbiology has been studied. In a 1995 landmark publication, a group of TIGR investigators reported on the first complete genome sequence of a free-living organism, *Haemophilus influenzae*. This new approach has, to date, produced DNA sequence data from nearly 1,000 different species across the phylogenetic tree.

Her work on the Amerithrax investigation led to the identification of four genetic mutations in the anthrax spores that allowed the FBI to trace the material back to its original source. She is one of the world's experts in microbial forensics and the growing concern about dual uses—research that can provide knowledge and technologies that could be misapplied.

Dr. Fraser has authored more than 200 publications, edited 3 books, and served on the editorial boards of 9 scientific journals. For the past 10 years, she has been the most highly cited investigator in the field of microbiology. Her list of awards includes the E.O. Lawrence Award, the highest honor bestowed on research scientists by the Department of Energy, the Promega Biotechnology Award from the American Society of Microbiology, and the Charles Thom Award from the Society for Industrial Microbiology. She has been selected as one of Maryland's Top 100 Women Circle of Excellence, and in 2010, was named to the Maryland Women's Hall of Fame.

She has served on many advisory panels for all of the major federal funding agencies, the National Research Council, the Department of Defense, and the intelligence community. In addition, she has contributed her time as a board member for universities, research institutes, and other nonprofit groups because of her commitment to the education of our next generation of scientists.

Jennifer Gardy, Ph.D., leads the Genome Research Lab at the British Columbia Centre for Disease Control, where she and her colleagues use emerging genomics technologies as tools for solving problems in public health. Dr. Gardy's particular interests lie in using next-generation sequencing of pathogen isolates from an outbreak situation to understand how outbreaks start, how pathogens transmit from person to person, and community to community, and how to turn this knowledge of transmission dynamics into sustainable and effective public health interventions. Her lab published the first genome-based reconstruction of a large tuberculosis outbreak, a project that also looked at the role of social networks in the spread of disease, and the group is currently working on similar reconstructions

for other outbreaks of both bacterial and viral disease. The Genome Research Lab is also exploring the role of metagenomics in public health, investigating its utility in identifying novel microbial markers of disease. Dr. Gardy completed her Ph.D. at Simon Fraser University in 2006 and is currently an Adjunct Professor in Microbiology and Immunology at the University of British Columbia.

Elodie Ghedin, Ph.D., is an Associate Professor in the Department of Computational and Systems Biology and a member of the Center for Vaccine Research at the University of Pittsburgh School of Medicine. Her research is aimed at generating critical insights about host-pathogen interaction and pathogen population structures. She uses functional and comparative genomics, computational and evolutionary biology, and molecular parasitology techniques to focus on the agents that cause diseases endemic to tropical climates, such as lymphatic filariasis (elephantiasis), onchocerciasis (River blindness), and leishmaniasis, as well as global diseases such as seasonal and pandemic influenza. Dr. Ghedin came to the University of Pittsburgh in 2006 after spending six years at TIGR where she led the Influenza Genome Project, the first of its kind to characterize large collections of an acute RNA virus, overturning outdated models of influenza evolution that were based on limited genetic data. Using NextGen platforms, she and her team are determining the extent and structure of genetic variation in influenza virus populations sampled within individual hosts, and variant transmission. In 2011, citing the creativity and collaborative nature of her work and her contributions to parasitology and virology, the MacArthur Foundation recognized Dr. Ghedin with its fellowship award.

Jack A. Gilbert, Ph.D., earned his Ph.D. from Nottingham University, United Kingdom, in 2002, and received his postdoctoral training in Canada at Queens University. He subsequently returned to the United Kingdom in 2005 and worked for Plymouth Marine Laboratory at a senior scientist until his move to Argonne National Laboratory and the University of Chicago in 2010. Dr. Gilbert is an Environmental Microbiologist at Argonne National Laboratory, Adjunct Professor in the Department of Ecology and Evolution at the University of Chicago, and Fellow of the Institute of Genomic and Systems Biology. Dr. Gilbert is currently applying next-generation sequencing technologies to microbial metagenomics and metatranscriptomics to test fundamental hypotheses in microbial ecology. He has authored more than 70 publications and book chapters on metagenomics and approaches to ecosystem ecology. He has focused on analyzing microbial function and diversity, with a specific focus on nitrogen and phosphorus cycling, with an aim of predicting the metabolic output from a community. He is currently working on generating observational and mechanistic models of microbial communities associated with aquatic and terrestrial ecosystems. He is on the board of the Genomic Standards Consortium (www.gensc.org), is an editor for *PLoS ONE*

and the *ISME Journal*, and is co-leading the Earth Microbiome Project (www.earthmicrobiome.org).

Peter Girguis, Ph.D., is currently a John L. Loeb Associate Professor of Natural Sciences at Harvard University and an Adjunct Research Engineer at the Monterey Bay Aquarium Research Institute. His research focuses on the ecological physiology of microbes that live in extreme environments, to better understand the role they play in mediating deep ocean carbon and nitrogen cycling. He is particularly interested in the physiological and biochemical adaptations (adaptive traits) to life in anaerobic environments. His work on microbial fuel cells has furthered our understanding of how microbes generate energy using solid-state minerals as electron acceptors and—in collaboration with colleagues around the world—has led to the development of systems that enable energy to be harnessed and used from the environment.

He received his B.Sc. from the University of California, Los Angeles, where he also worked with Drs. David Chapman and William Hamner. He received his Ph.D. from the University of California, Santa Barbara, where he worked with Dr. James Childress on the physiological and biochemical adaptation of deep-sea hydrothermal vent tubeworms and their microbial symbionts to the vent environment. He did his postdoctoral research at the Monterey Bay Aquarium Research Institute with Dr. Edward Delong on the growth and population dynamics of anaerobic methanotrophs.

Susan Huse, Ph.D., is an Assistant Research Scientist at Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts. She studies microbial diversities, population structures, evolution, and bioinformatics. Her research is in the role of the microbiome in human health and disease. She is looking to define what is "normal" in the human mouth and gut within the context of a great level of interpersonal microbiome variation. Evaluating changes from the "normal healthy" state are critical for understanding health, dysbiosis, and host-microbiome interactions. Her collaborations have included projects describing the progression of pouchitis in inflammatory bowel disease patients, the impacts of antibiotics, and the effects of changing dentition. She has also been active in developing and advancing best practices for the use of 16S ribosomal RNA gene for studying microbial communities, starting with her participation in the first paper that used the 454 tag pyrosequencing for studying microbial ecology and the rare biosphere. In addition, she published the first assessment of how to filter low-quality 454 sequence data and developed a new method for assigning taxonomy to these tags (GAST). She also demonstrated that the accepted standard method of clustering 16S rRNA tags for taxonomic-independent methods of evaluating microbial communities was overestimating microbial diversity and underestimating the similarities between

communities, and developed the SLP clustering method to more closely estimate the true diversity and inter-community similarities.

Paul Keim, Ph.D., holds the Cowden Endowed Chair in Microbiology at Northern Arizona University (NAU), where he is also a Regents Professor of Biology. In addition, he is now a Professor and Director of the Pathogen Genomics Division at The Translational Genomics Research Institute (TGen). He is an affiliate researcher at the Los Alamos National Laboratory, where he has been engaged in national security research since 1993. Dr. Keim received his B.S. in biology and chemistry from Northern Arizona University in 1977 and his Ph.D. in botany from the University of Kansas in 1981. His NAU laboratory was heavily involved in analysis of evidentiary material from the 2001 anthrax-letter attacks for the FBI and the Department of Justice. His work has helped lead to the development of a new scientific field known as microbial forensics. He has published extensively on the evolution and population genetics of *Bacillus anthracis*, *Yersinia pestis*, *Francisella tularensis*, *Burkholderia pseudomallei*, *Burkholderia mallei*, *Brucella spp.*, and *Coxiella burnetii*. Recently, these same scientific principles have been applied to other public health and clinically important pathogens such as *Vibrio cholerae* (cholera in Haiti), *S. aureus*, and *E. coli*.

Folker Meyer, Ph.D., is a Computational Biologist at Argonne National Laboratory and a Senior Fellow at the Computation Institute at the University of Chicago. He was trained as a computer scientist and with that came his interest in building software systems. He now is interested in building systems that further our understanding of biological data sets. In the past he has been best known for his leadership role in the development of the GenDB genome annotation system and the design and implementation of Bielefeld University's high-performance computing facility. Currently he is most interested in comparative analysis of large numbers of microbial genomes.

Mark Pallen, M.D., Ph.D., is a medically qualified microbiologist (M.D., Ph.D.), educated at Cambridge and in London. Since 2001, he has been Professor of Microbial Genomics at the University of Birmingham. Dr. Pallen's research interests span bioinformatics, bacterial protein secretion, and bacterial pathogenomics. In recent years, his attention has focused on the application of high-throughput sequencing to medical microbiology, particularly genomic epidemiology, sequence-based approaches to surveying complex microbial communities, and new bench-top sequencing platforms. In the summer of 2011, Dr. Pallen's group kick-started and guided an open-source genomics programme on the German *E. coli* outbreak strain, which included innovative crowdsourcing of analysis and culminated in a *New England Journal of Medicine* paper. More recently, Dr. Pallen has published a performance comparison of bench-top sequencers applied to the German outbreak strain. Dr. Pallen is author of *The Rough Guide to*

Evolution, a wide-ranging introduction to this topic, and is currently writing a book titled *The Last Days of Smallpox*.

Julian Parkhill, Ph.D., is a Professor in pathogen genomics at the Sanger Institute. Since joining the Sanger Institute in 1997, he has been involved in the genomic analysis of a large number of bacteria from a wide diversity of genera, including *Bordetella*, *Burkholderia*, *Campylobacter*, *Chlamydia*, *Clostridium*, *Corynebacterium*, *Escherichia*, *Haemophilus*, *Mycobacterium*, *Neisseria*, *Salmonella*, *Staphylococcus*, *Streptococcus*, *Yersinia*, and many others.

His current research uses very high-throughput sequencing and phenotyping technologies to understand the evolution of bacterial pathogens on short and long time scales, how they transmit between hosts on a local and global scale, how they adapt to different hosts, and how they respond to natural and human-induced selective pressures.

Dr. Parkhill gained his Ph.D. in 1991 from the University of Bristol through work on bacterial transcriptional regulation. He subsequently pursued postdoctoral research at the University of Birmingham, first on bacterial transcriptional regulation, and then on the transforming proteins of adenoviruses.

David Rasko, Ph.D., is an Assistant Professor in the Department of Microbiology and Immunology and a member of the Institute for Genome Sciences. During his career he has developed expertise in comparative microbial genomics, bioinformatics, and functional genomics. Dr. Rasko has led comparative genome sequencing and analysis projects for important human diarrheal pathogens, focusing on *Escherichia coli* and *Shigella* species as well as *Bacillus cereus* group isolates including *Bacillus anthracis*. He has developed comparative bioinformatics tools designed to characterize the genetic diversity in closely related bacterial isolates. Dr. Rasko was the first to publish a comparative genomic study that included a genome reference from a true commensal, each of the six diarrheagenic *E. coli* pathogenic variants (pathovars) as well as representatives of the urinary tract and avian derived *E. coli*. Recent comparative genomic studies have focused on the development of genomic epidemiology tools for study of hundreds of enterotoxigenic *E. coli* and *Shigella* species isolates. These comparative works provide the framework for the continued functional study of the evolution of these pathogens, as well as functional studies of identified unique and conserved gene features as vaccine and therapeutic targets.

Kim Ritchie, Ph.D., is Senior Scientist and Manager of the Marine Microbiology program at Mote Marine Laboratory, Florida. She is a molecular biologist investigating the microbial community structure of Florida coral reefs and its role in disease resistance. She received her Ph.D. in the laboratory of Thomas Petes studying telomere length regulation in 2000 followed by postdoctoral work at the Smithsonian's Tropical Research Laboratory in Panama. Past work

includes identification and characterization of coral disease pathogens. Her current studies include characterizations of symbiotic microfauna in multiple coral species (mountainous star coral, sea fan corals, threatened Elkhorn and staghorn corals) as well as culture-based studies on the production of anti-microbial and anti-fungal compounds produced by bacterial symbionts. Dr. Ritchie is also interested in dinoflagellate-bacterial interactions including the relationship between symbiotic bacteria and the coral dinoflagellate, *Symbiodinium* sp, and vital interactions between bacterial symbionts and the red tide causing dinoflagellate, *Karenia brevis*.

Erica Bree Rosenblum, Ph.D., is currently an Assistant Professor at University of California (UC), Berkeley, in the Department of Environmental Science Policy and Management. Research in the Rosenblum lab focuses on the evolutionary processes of speciation and extinction in changing environments. One current emphasis is on the impact of emerging infectious diseases on wildlife. To address questions across spatial and temporal scales, Dr. Rosenblum's research uses integrative methods, and she is involved in a number of interdisciplinary initiatives including the Berkeley Initiative for Global Change Biology and the NSF Bio/computational Evolution in Action CONsortium Center (BEACON). The Rosenblum lab is also highly invested in scientific outreach and facilitates a number of activities to improve public understanding of science including a "Save the Frogs Day" for pre-school students and a "Lizard Camp" for middle school students. Her research has been featured broadly in the popular press including in the *New York Times*, *Science* magazine, the *Discovery Channel*, National Public Radio, and *Ranger Rick magazine*.

Dr. Rosenblum conducted her Ph.D. research at UC Berkeley in the Department of Integrative Biology and her postdoctoral research at the Lawrence Berkeley National Lab in the Department of Genomics.

Tim Stearns, Ph.D., is a Professor in the Stanford University Department of Biology, the Stanford University Medical School Department of Genetics, and the Cancer Biology Program. Dr. Stearns received his Ph.D. at the Massachusetts Institute of Technology, was a postdoctoral fellow at University of California, San Francisco, and joined the Stanford faculty in 1993. Research in Dr. Stearns' lab is focused on the centrosome and cilium, microtubule-based structures that are at the center of cell signaling and cell division, and on using simple microbial models to assess human genetic variation. Dr. Stearns is the recipient of a Howard Hughes Medical Institute Professor award and the Stanford Dean's Award for Distinguished Teaching. In addition to undergraduate and graduate teaching at Stanford, he has taught laboratory courses at Cold Spring Harbor Laboratory, Woods Hole, and internationally in Chile, South Africa, Ghana, and Tanzania, and is a member of the International Affairs Committee of the American Society for Cell Biology. Dr. Stearns is a member of JASON, a panel that advises the U.S.

government on defense and technology issues, and he is on the scientific advisory board of the Temasek Life Sciences Laboratory in Singapore. He currently holds the Frank Lee and Carol Professorship at Stanford University.

David Wang, Ph.D., is currently an Associate Professor at Washington University in St. Louis. He earned a B.S. degree in chemistry from Stanford University in 1992 and a Ph.D. in biological chemistry from the Massachusetts Institute of Technology in 1998. Following postdoctoral training at University of California, San Francisco, he started his own research laboratory at Washington University in St. Louis in 2004. Research in the Wang laboratory focuses upon the identification and characterization of novel viruses. Dr. Wang has pioneered novel genomic approaches for viral discovery, including pan-viral DNA microarray and high–throughput sequencing–based strategies. Application of these methods to unexplained cases of acute respiratory disease and diarrheal disease in humans has lead to the discoveries of many novel viruses, such as WU polyomavirus, Astroviruses VA1, VA2, VA3, MLB1, and MLB2, human klassevirus 1, and human cosavirus E1. Characterization efforts have focused on defining the epidemiology and seroepidemiology as well as the basic virology of these viruses. In addition, his laboratory has analyzed a diverse array of non-human specimens including marine mammals, insect vectors, and environmental samples (e.g., raw sewage). Most recently, his laboratory discovered the first viruses capable of infecting the nematode *C. elegans* and established an infection system for studying host-virus interactions in this genetically robust model organism.

George Weinstock, Ph.D., is currently Professor of Genetics and of Microbiology and Associate Director of The Genome Institute at Washington University in St. Louis, where he leads a number of projects in human genetics and microbial genomics including projects in the Human Microbiome Project and other metagenomic studies of humans, monkeys, and mice. Previously he was Co-Director of the Human Genome Sequencing Center at Baylor College of Medicine, where he was a leader of the Human Genome Project and a number of other genome projects. He has had a long-term interest in genetics, genomics, and bioinformatics, particularly as bioinformatics couples technological advances to new biology through increases in the size and quality of data sets as well as explaining phenotypes through characterization of genotypes.