THE ROLE OF HORIZONTAL GENE TRANSFER IN BACTERIAL EVOLUTION

A Dissertation
Presented to
The Academic Faculty

by

Alejandro Caro-Quintero

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Biology
School of Biology

Georgia Institute of Technology
AUGUST 2013

COPYRIGHT © ALEJANDRO CARO-QUINTERO 2013

THE ROLE OF HORIZONTAL GENE TRANSFER IN BACTERIAL EVOLUTION

Approved by:

Dr. Konstantinos Konstantinidis, Advisor
Department of Civil and Environmental
Georgia Institute of Technology

Dr. Soojin Yi
School of Biology
Georgia Institute of Technology

Dr. King Jordan
School of Biology
Georgia Institute of Technology

Dr. Frank E. Löffler, Advisor
Department of Microbiology
Department of Civil and Environmental Engineering
University of Tennessee, Knoxville

Dr. Thomas J. DiChristina
School of Biology
Georgia Institute of Technology

Date Approved: April 26th 2013

Science with social consciousness should be our Impact Factor.

Alejo

This is for my beautiful wife and daughter.
Thanks for your for unfailing love, support and guidance
during this journey.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

%	Percentage
°C	Degrees Celsius
μg	Micrograms
ACT	*Artemis* DNA Comparison Tool
AIC	Akaike Information Criterion
ANI	Average Nucleotide Identity
ATP	Adenosine Triphosphate
bp	Base Pairs
BLAST, BLASTn, BLASTp	Basic Local Alignment Search Tool
BSA	Bovine serum albumin
CDS	Coding Sequences
COG	Clusters of Orthologous Groups
CRISPER	*Clustered Regularly Interspaced Short Palindromic Repeats*
Cy3	Cyanine 3
Cy5	Cyanine 5
DDH	DNA-DNA Hybridization
DMSO	Dimethyl sulfoxide
Dn	non-synonymous substitution rate
Ds	synonymous substitution rate
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic *acid*
E.C.	*Enzyme* Classification
g	generations

G+C%	Guanine- Cytosine content
gAAI	Genome-Aggregate Amino Acid Identity
GARD	Genetic Algorithm for Recombination Detection
gi	GenInfo Identifier
h	hour
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HGT	Horizontal Gene Transfer
HR	Homologous Recombination
IDs	Identification
incQ	Incompatibility group Q plasmids
kb, kpb or kbps	kilo base or kilo base pairs
Ka	non-synonymous substitution rate
Ks	synonymous substitution rate
IDs	Identification
m	Mutation rate
Mb	Mega bases
MDR	Multi Drug Resistant bacteria
ME	Mobile Elements
mg	milligrams
min	minute
ml	milliliter
MLST	*Multilocus Sequence Typing*
MLSA	*Multilocus Sequence Analysis*
mM	millimolar
MP	Maximum Parsimony

MPN	Most Probable Number
N	nitrate
n	number of counts
NCBI	National Center for Biotechnology Information
NHR	Non-Homologous Recombination
NJ	Neighbor Joining
NR	Non-Redundant
O	Oxygen
PBPs	Penicillin-Binding Proteins
PCR	Polymerase Chain Reaction
pTi	tumor-inducing plasmid
r	recombination rate
RAPD	Randomly Amplified Polymorphic DNA
RBM	Reciprocal Best Match
RNA	Ribonucleic Acid
t-RNA	Transfer-Ribonucleic Acid
SDS	Sodium dodecyl sulfate
SNPs	Single Nucleotide Polymorphisms
SSC	saline-sodium citrate
ST	
T	Thiosulfate
TCA	Tricarboxylic Acid Cycle
USA	United States of America

SUMMARY

Bacteria are well known for their immense genetic and physiological diversity. This diversity has allowed them to colonize all environments, making bacteria the most ubiquitous and abundant living organisms on the planet. Fast adaptation to the environment is an important component of bacterial success and therefore identification of the mechanisms underlying adaptation is essential to understand the evolution of microbial life on our planet. Horizontal gene transfer (HGT) is probably the most important mechanism for functional novelty and adaption in prokaryotes. However, a robust understanding of the rates of HGT for most bacterial species and the influence of the ecological settings on the rates remain elusive. Although preexisting genetic diversity and environmental selective pressure are important for adaptation, little is known about how ecological interactions affect the frequency of genetic exchange, particularly what kind of relationships might produce effective encounters for genetic exchange to occur. An improved understanding of this issue has important broader impacts such as for reliable diagnosis of infectious disease agents, successful bioremediation strategies, and robust modeling of bacterial evolution and speciation.

In this dissertation, I will describe four studies that aimed at evaluating the interplay between ecology and HGT and quantifying HGT at three important levels: i) the species level, where an overlapping ecological niche can be shown to cause HGT to be so rampant that it can serve as the force of species cohesion; ii) the genus level, where HGT appears to mobilize mostly genes with ecological/selective advantage for the host genome and to prevent species convergence; and iii) the phylum level, where HGT is, in general, less frequent than the genus level, but a case was identified where direct inter-phylum genetic exchange has affected more than half of the genome, resulting in chimeric phyla. Subsequently, a novel bioinformatics pipeline was developed to systematically detect and quantify inter-phylum HGT, normalizing for biased representation of phyla among the available genomes. Using this pipeline, I quantitatively evaluated, the preferential exchange between phylogenetic groups, the functions more likely to be transferred, and the correlation of exchange with the organisms' known ecological constraints. The results of this analysis show that, large genetic exchange across phyla is more common than previously anticipated and that ecologically relevant interactions, such as syntrophy, organic matter degradation and fermentation, seem to promote inter-phylum genetic exchange. In conclusion, this dissertation provides new avenues to link ecological preferences with HGT and suggests that genetic diversity within an environment has the potential to affect adaptation, even among very divergent organisms.

CHAPTER 1

INTRODUCTION

1.1 Horizontal Gene Transfer (HGT), the Major Force in Bacterial Evolution

Prokaryotes are the most ubiquitous living organisms of the planet. They catalyze fundamental steps in the geochemical cycles, and participate in key ecological relationships (i.e., symbiosis, protocoperation, competition) that determine the diversity and distribution of higher organisms in most, if not all, of the environments. A key aspect favoring prokaryotes functional and ecological diversity is their ability to incorporate foreign DNA through horizontal gene transfer (HGT). The occurrence of HGT is so frequent that it is thought to be the main process responsible for the large physiological diversity and remarkable adaptability of prokaryotes [1, 2]. In fact, recent analysis of protein families suggests that HGT and not duplication has driven protein expansion and functional novelty in prokaryotes [3]. Genome sequencing has expanded our view of the role of HGT in prokaryotic evolution. The availability of thousands of genomes has allowed the identification of genetic exchange events at different time scales (i.e., from ancestral to recent events), and between organisms with different phylogenetic divergence (i.e., from close related strains to very distantly related groups) [4-7]

This chapter provides the background information for understanding the factors involved in HGT. The first part describes the mechanisms of transfer and incorporation of DNA, along with some case studies exemplifying their role in prokaryotic adaptation.

The second part provides examples of how genetic diversity and overlapping ecology affect the outcome of HGT. A summarizing picture of the current models that use HGT to explain prokaryotic evolution is also provided. The chapter concludes with a description of the specific questions that this dissertation sought to answer related to the role of ecology and divergence in the outcome of HGT.

1.2 Background

The effects of HGT in adaptation are diverse and some of them not completely understood. Some argue that HGT has been so pervasive that a correct reconstruction of the phylogenetic relations of living organisms is out of reach [8, 9]. Along the same lines, it has been suggested that because of rampant HGT, there is no unifying concept for what a "species" is and, as a consequence, such a concept does not exist for prokaryotes [10-12]. On the other hand, others argue that the rates of HGT between close related organisms (i.e., strain of the same species) are very high and decrease exponentially with higher genetic divergence creating coherent and cohesive populations similar to "species" [13, 14]. The available evidence suggests that the effects of HGT on prokaryotic evolution are diverse and that any rules emerging probably apply to only a few organisms and there will be plenty of exceptions.

The large diversity of evolutionary outcomes related to gene transfer is the result of a complex interplay between molecular and ecological factors. Molecular factors encompass those processes that are directly related to the transfer and incorporation of DNA. They include: the mechanisms of transfer (i.e., transformation, transduction and conjugation), the mechanisms of incorporation (i.e., homologous and non-homologous recombination [NHR]) and the defense mechanisms of the host against foreign DNA (e.g., Clustered Regularly Interspaced Short Palindromic Repeats [CRISPER] and restriction modification systems). Ecological factors are those related to the selection and fixation of the transferred DNA. Examples of ecological factors are: the interactions between organisms (e.g., competition, symbiosis), environmental conditions (e.g., physico-chemical conditions, carbon substrate available), and the population size and intra-genetic diversity. Integration of genomics with measurements of these factors is starting to reveal the prevailing mechanisms and controls underlying prokaryotic HGT and adaptation. Here, an overview of these molecular and ecological factors is presented together with recent studies that linked the factors to HGT.

1.3 Molecular Factors Affecting HGT

1.3.1 Mechanisms of Genetic Exchange

HGT encompasses different mechanisms that mediate the transfer of genetic information from a donor to a recipient cell. These mechanisms are mainly classified as transduction (mediated by phages), conjugation (mediated by plasmids), transformation (mediated by uptake of naked DNA), and the recently described virus-like particle transfer agents [1, 15].

1.3.1.1 Transduction

The term transduction refers to the mechanism in which a bacteriophage (phages) transfers DNA from one bacterium to another. In order to infect, the phage attaches to the extracellular receptors of the host. Once inside the cell, the phage can integrate its genome into the host genome and take over the cell machinery to synthesize new copies of its genome as well as all the proteins required for packing and structure. Upon excision from the host genome, the phage genome can pick up adjacent host genes (typically only a few) and eventually transfer them to a newly infected cell. There are two main types of transduction described, generalized and specialized transduction. In generalized transduction, the phages do not require a specific attachment site in the host genomes (random integration), and therefore, they can potentially transfer many types of genes (i.e., host genes flanking the phage genome). In contrast, specialized phages required specific integration sites in the host genome, and therefore they can potentially transfer just a narrow variety of genes. Packing of host DNA in specialized or generalized transduction therefore occurs via a mistake of the mechanisms of excision (aberrant excision). Another important factor in determining the potential of genetic mobilization by phages is their host range. Phages can either infect specific species (or even strains of a species) or can have a broad host-range (i.e., different species, genera or even families) [16]. An example of a broad-range bacteriophage is the ΦOT8 phage that has been shown to successfully transfer genes related to antibiotic resistance between two different species of the *Enterobacteriaceae* family, *Pantoea agglomerans* and *Serratia* sp. [17].

1.3.1.2 Conjugation

The term conjugation refers to the case of DNA transfer mediated by the type IV secretion system, that requires cell-to-cell contact. The type IV secretion drives the transfer of conjugal plasmids or conjugal transposons. The system transfers single-stranded DNA molecules that are generated by relaxase proteins that nick the DNA in a highly conserved and specific motif know as the origin of transfer (oriT). Interestingly, if the plasmid was previously integrated into the chromosome of the donor genome some of the host DNA can also be transferred via conjugation. Once the single-stranded DNA molecule is transferred, a complementary strand is synthesized to produce a double-stranded circular plasmid. Novel conjugation systems that are clearly distinguished from DNA transfer by a type IV secretion system have also been found, for example the TraB conjugation system in *Streptomyces spp.* [18].

Plasmids can also be categorized based on their host range, similar to phages. Plasmids can either be specific (narrow-host-range) or broad range (broad-host-range). Broad-host-range plasmids can be transferred even across phyla or even kingdoms. The most studied case is the transfer of the tumor-inducing plasmid (pTi) from *Agrobacterium tumefaciens* to a plant cell [19, 20]. Another case of broad-host-range plasmids is the incompatibility group Q plasmids (incQ). These plasmids have been found in a wide variety of environments and have been transferred between gram-positive and gram-negative bacteria [21].

1.3.1.3 Transformation

The term transformation refers to the process of HGT in which DNA uptake from the environment occurs [22]. The ability to uptake exogenous DNA is known as "natural competence". In bacteria, natural competence is a complex process that requires the expression of genes involved in the assembly of type IV pili and type II secretion systems [23]. Expression of these sets of genes (about 40 genes in *Bacillus subtilis*) depends on specific physiological and environmental cues such as high cell density and limited nutrient availability. In *Vibrio cholera*, expression of competence genes also requires the presence of chitin surfaces [24].

The components involved in DNA-uptake are not the same for gram-positive and gram-negative bacteria due to the difference in cell wall structure. In gram-positive bacteria, retraction of a pseudopilus opens a cell wall hole that allows DNA to diffuse from the surface. In gram-negative bacteria, due to the presence of an extra membrane, DNA uptake requires the presence of a more complex channel, mainly formed by secretins (PilQ). In contrast to DNA uptake, DNA translocation across the cell membrane is similar in gram-negative and gram-positive bacteria. In both groups, homologues of the ComEC channel proteins mediate the transport of the DNA to the cytoplasm. During this process, one strand of the incoming DNA is degraded by nucleases, and the remaining single-stranded DNA is bound by proteins that protect it from degradation. Incorporation into the chromosome can be catalyzed by the mechanisms of HR if sufficient sequence identity exists.

1.3.2 Mechanisms of Foreign DNA Incorporation

1.3.2.1 Homologous Recombination (HR)

Homologous recombination is a general DNA repair process that plays an important housekeeping role in maintaining functionality of the genetic material. This process depends on a group of proteins (e.g., RecA protein) that catalyze the exchange of donor and recipient DNA through a strand invasion mechanism and requires a high degree of homology (i.e., DNA molecules are evolutionary related due to a shared ancestry; the higher the degree of homology, the higher the sequence identity) between the recombining DNA sequences.

Interestingly, the same process allows the integration of foreign DNA (from the donor cell) to the chromosome of the recipient cell, resulting in the substitution of whole or parts of genes. There are several constrains that affect the frequency of HR happens. For instance, divergence between recombining sequences has a major (negative) effect on the recombination rates [25-29]. Studies in *Bacillus*, *Escherichia*, and *Streptococcus,* have shown that recombination rates decrease with increasing divergence between the recombining DNA sequences [25-29]. This decreased in recombination efficiency is related to the minimum sequence identity that the protein complexes involved in recombination required for successfully catalyzing the exchange [28, 30]. In addition to sequence identity, methylation-restriction mechanisms can influence the overall length of the recombined DNA segments, as demonstrated for recombinant clades of *Neisseria meningitidis* [31]. In addition, HR rates are also affected by the type of gene and its locations in the genome; recent genomic analysis of recombination in *Acinetobacter* *baylyi* showed that the rates of recombination might vary up to 10,000 fold across the genome, and these differences appear to be related to local gene organization and synteny [32]. Homologous recombination patterns have been detected and quantified through various DNA sequencing approaches, e.g., multilocus sequence typing (MLST), genomics and metagenomics (Table 1.1). These approaches offered different resolution in the role of HR in bacterial evolution, and provided evidence that HR is more important and common than previously thought [33, 34] and that it can facilitate the spreading of adaptive mutations and HGT events. For instance, high rates of recombination in several pathogens are linked to the rapid adaptation of virulent populations [35-37]. Similarly, genes under positive selection are often transferred horizontally (mediated by HR). Some examples are the capsule biosynthesis locus of *Streptococcus spp*. [36], and the surface molecule (InlA) from *Listeria monocytogenes* [38]. Direct comparison of HR rates between different prokaryotes reveals that HR is an ubiquitous process whose magnitude may differ between environments and lifestyles [39]. The outcome of HR is diverse and depends on multiple factors such as the selection pressure of the environment and the genetic divergence between the donor and the recipient cells.

Table 1.1 Case studies that quantified homologous recombination and their methods. The letter “r” represents the rate of recombination between populations, while “m” represents the rate of polymorphisms brought in by mutation during the same time period. The ratio between r and m provides an estimate of how much the examined organisms behave like a sexual (ratio > 1) or clonal organism (ratio < 1). For more information see biological species concept section.

Analysis	Organism	(r/m)	Methods	Description
MLST Intra-species	*Pelagibacter ubique (SAR 11)*	1.26[40] 63.8[39]	LDhat, [41] ClonalFrame[42]	This study revealed significant phylogenetic incongruence in seven of the genes, indicating that frequent recombination obscures phylogenetic signals from the linear inheritance of genes in this population.
	Salmonella enterica	n/d[43] 30.2[39]	Linkage-Desequilibrium ClonalFrame	This study showed that HR is a predominant within the subspecies, where lack of phylogenetic congruence was observed between phylogenetic trees of six housekeeping genes
MLST Inter-species	*Streptococcus pneumoniae*	23.1[39] n/d[44]	ClonalFrame BAPS[45, 46]	Mosaic genotypes were identified, emerging as a result of historic hyper-recombination period, where strains acquired divergent versions of alleles and antibiotic resistance determinants.
	Helicobacter pylori	13.6[39] 3.35[47]	ClonalFrame ABC	This microevolutionary analysis revealed higher rates of mutation and HR than quantified by long-term mutation rates, 5-17 times higher.
	Sulfolobulus islandicus (Archaea)	6.6[48] 1.2[39]	LDhat /DnaSP ClonalFrame	Significant incongruence among gene genealogies and lack of association between alleles was consistent with recombination rates greater than the rate of mutation, accounting for genetic cohesion in archaea.
	Halorubrum sp. (Archaea)	n/d[49] 2.1[39]	BURTS -Linkage Equilibrium ClonalFrame	This study showed that *Haloarchaea* exchanged genetic information promiscuously, exhibiting a degree of linkage disequilibrium approaching that of a sexual population.
	N.lactamica- N.meningiditis-N.gonorrheae	n/d[50]	Phylogenybased	Species clusters are not ideal entities with sharp and unambiguous boundaries; instead they may be fuzzy and indistinct in recombinogenic bacteria.
	C.jejuni-C.coli	4[9]	STRUCTURE [51]	Inter-species recombination based on sequence type reflects convergence between the *C.jejuni* and *C.coli*.
Genome Inter-species	*Streptococcus pneumoniae*	7.2*[36]*	Own algorithm	PMEN1 lineage undergoes HR with unknown outsider lineage, based on a reference genome and quantifying changes on isolates from different time points. More polymorphisms were brought by recombination compared to mutation.

1.3.2.2 Non-Homologous Recombination (NHR)

Non-homologous recombination mechanisms incorporate DNA material without the requirement of sequence homology, and therefore, are more frequently responsible for conferring novel metabolic capabilities than HR [52-55]. This incorporation is primarily mediated by the integration of sequences through mobile elements (ME) such as phages, transposases, and integrons. Mobile elements often encode modular sets of genes (e.g., genetic islands or gene cassettes) that can confer immediate adaptations. In pathogenic bacteria, ME have been extensively investigated due to their role in spreading of antibiotic resistance and disease outbreaks [56, 57]. There are many studies showing HGT mediating the acquisition of pathogenicity determinants. Recently, the dynamics of such acquisitions have been confirmed using population genomic approaches [58-60]. One clear example of the role of ME in disease outbreaks is the fast acquisition of resistance to multiple antibiotics in the so-called "superbug", methicillin-resistance *Staphylococcus aureus*. ME can spread between a broad phylogenetic range of organisms within environments. For example, worldwide screenings have documented the spreading of certain elements (i.e., class 2 integrases) found in clinical isolates to non-clinical environments affected by human activities [55]. Furthermore, analysis of 10 million protein-encoding genes and gene tags from sequenced bacteria, archaea, eukaryotes, viruses, and from various metagenomes revealed that genes encoding transposases are the most prevalent in nature, suggesting a quantitatively important role in spreading genes among prokaryotes [61]. Finally, because ME can mediate the acquisition of modular sets of genes, they can play an important role in ecological specialization and phylogenetic divergence of bacteria [62]

1.3.3 Mechanisms of Immunity to HGT

1.3.3.1 Restriction Modification Systems

Restriction endonucleases recognize specific DNA sequences; these sequences are mostly palindromes of four, six or eight base pairs. The endonucleases are accompanied by a modification enzyme that methylates the recognition sequence (palindrome) in the host DNA. The methylation protects the host and allows the identification and degradation of foreign DNA. Therefore, it is expected that bacteria sharing the same restriction-modification system can more effectively exchange and incorporate DNA. Recent studies in *Neisseria meningitidis* have shown that clade-associated restriction modification systems generate a differential barrier to DNA exchange, and that this barrier is consistent with the observed population structure and frequency of HR [63].

1.3.3.2. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) System

The CRISPRs-Cas system is a nucleotide based immune system mechanism that provide defense against foreign phages or plasmids. The CRISPRs are composed of short repeated sequences (21-48 bp length), separated by a sequence spacer (26-72 bp length). Most of the times, the sequence spacer is derived from phages or plasmids that have previously infected the cell lineage. Examples of acquisition of immunity to M102-like phages have been identified, for instance, in strains of *Streptococcus mutans* [64]; however, the process by which a new spacer is integrated into the host genome remains poorly understood. These short sequences are transcribed and the transcript is cleaved to form smaller RNA sequences. These short RNA sequences can then bind to homologous DNA or RNA of plasmid or phages based on base-pairing, the heteroduplex is recognized by a multifunctional complex (Cas proteins) and is degraded [65]. A clear case of how this mechanism can limit HGT has been described for *Staphylococcus epidermidis*. This study shows that a CRISPR present in *S. epidermitis* prevents conjugation and plasmid transformation of known staphylococcal conjugative plasmids by the binding of the spacer RNA to a nickase gene present in almost all staphylococcal conjugative plasmids [66].

1.4 Ecological Factors Affecting HGT

1.4.1 The Role of Intra-Population Genetic Diversity

Little is known about how preexisting diversity influences the genetic adaptation of populations. However, it has been shown that populations capable of HGT can adapt faster than clonal ones, suggesting that genetic diversity of co-occurring organisms in the environment can provide new or advantageous alleles for adaptation through HGT. The multidrug-resistant (MDR) *Acinetobacter baylyi* is one interesting example of how preexisting genetic diversity fosters faster adaptation to new antibiotics in clinical settings. When populations with different chromosomally-encoded drug resistance mechanisms were mixed in culture and selected for resistance to all antibiotics, MDR evolved rapidly in strains with an active HR mechanism through shuffling of the preexisting resistance alleles [67]. Another recent evolution study of the human pathogen *Helicobacter pylori* has shown that strains capable of natural transformation adapt more quickly to new conditions than do mutants lacking genes required for transformation. The authors concluded that the measurable advantage of the transformable strain is best explained by the ability of gene exchange, which facilitates acquisition of novel beneficial mutations [68]. However, a quantitative understanding of the interplay between HGT and preexisting genetic diversity during population adaptation under natural habitats remains still unexplored. For instance, it is not known to what degree intra-population genetic diversity (e.g., a higher variety of secondary metabolic capabilities within a population) increases the ability of the populations to survive environmental stresses. A better understanding of the role of preexisting population genetic diversity in adaptation could have important practical applications in medical, biotechnological and agricultural fields. For instance, antibiotic resistance is frequently acquired by horizontal gene transfer from other bacteria, therefore, new studies to predict the evolution of multi resistant strains should evaluate the community/population diversity of antibiotic resistance genes within the possible habitats of pathogens and opportunistic pathogens [69]. Therefore, the incorporation of population genetic diversity and evolutionary models will allowed a better modeling and prediction of the conditions (i.e., environmental and/or genetic) that favor new pathogen outbreaks or the evolution of new catabolic capabilities.

1.4.2 The Role of Ecology in the Outcome of HGT

Ecological niche overlap and its role in prokaryotic genetic exchange have been evaluated recently for a variety of organisms and environments. Genetic exchange between co-occurring organisms has been observed at different levels of genetic divergence, ranging from same species to different phyla or kingdoms [7]. Recently, a comparison of ~2200 bacterial genomes revealed that those isolated from similar local sites of the human body have higher rates of genetic exchange compared to genomes from different sites [70]. Higher frequency of HGT between niche-overlapping organisms can be the result of more frequent encounters between co-occurring organisms, which favors more conjugation, transduction or transformation events. However, it is possible that the detection of more HGT event within a niche is the result of higher fixation rates due to common selection pressure such as for optimum G+C% content or compatibility between the t-RNA pools as opposed to higher rate of HGT per se [71]. Along the same lines, higher fixation rates may be related to the availability of more abundant ecologically important genes in organisms from the same than different communities. However, not all organisms that co-occur in a habitat engage into HGT, and even if they do, the HGT frequency can vary markedly due to the molecular factors. It is also possible that the frequency of exchange is affected by the strength of ecological interactions among the partners of exchange. Ecological interactions have been previously implied to affect the frequency of HGT [72, 73], but a comprehensive and quantitative view that integrates all previously mentioned factors has not been described yet. In this section, the most common ecological interactions are presented through a review of several case studies from recent literature. Subsequently, the case for how the better understanding of ecological interactions can lead to better predictions about the frequency of exchange between the players on the interaction is made and, finally, the open questions in the field that could be address by integration of genomic and ecological analysis are presented.

Ecological interactions refer to the relationships between species that live together in a community and are categorized based on the effect that one population/species exerts on another one. Well described interactions include: protocooperation, commensalism, neutralism, amensalism and competition [74]. In protocooperation, both organisms involved in the interaction benefit; however, the interaction is not obligatory. This type of interaction is very common in the microbial world when a population can be associated with different partners for a specific cooperation. One of the most clear examples of protocooperation is the cross feeding on food webs (i.e., syntrophy). Examples of protocooperation are the association between *Lactobacillus bulgaricos* and *Streptococcus thermophilus* during yogurt production [75], or the association between methanogens and sulfate reducers with fermenting bacteria in anaerobic sludge [76]. In commensalistic relationships, one organism benefits from the association while the other remains unaffected. One example of commensalism is the relation between purple sulfur bacteria (*Chromatiales* orders) and colorless sulfur oxidizers (*Thiotrichales* order) in microbial mats. In this interaction colorless sulfur oxidizers benefit the growth of purple sulfur bacteria by removing oxygen from the system [77, 78]. In amensalism, the presence of one population inhibits the other, for instance, by the production of acids and antibiotics. A recent study of ruminal fibrolytic bacteria characterized an amensal interaction between *Ruminococcus albus* and *R. flavenciens,* in which the former inhibit the latter by production of bacteriocins [79]. In competition, energy and nutrients are often a limiting factor and therefore the fitness of both populations is decreased. Eventually, competition leads to the exclusion of one of the populations. Numerous studies have demonstrated the effect of competition in natural systems, such as the competition between polyphosphate and polysaccharide accumulating bacteria [80] and competition between sulfate-reducing and methanogenic bacteria [81]. In some cases the fitness of the organisms is not affected by their sharing of the same habitat, and therefore, an interaction is not observed, which is known as neutralism. Neutralism is hard to prove in nature mainly because neutrality may not be stable over long enough periods of time to be easily detectable. However coexistence under neutrality has been described previously under the neutral theory [82-84]; for instance, in laboratory studies of *Lactobacillus* and *Streptococcus,* during growth in a chemostat, where the mixed and the individual cultures had the same apparent fitness [85].

Accordingly, it can be hypothesized that the different ecological interactions among co-occurring populations/species can have an important effect on the frequency of encounter and therefore, can facilitate or impede, depending on the type of interaction, genetic flow between populations (Fig 1.1). For instance, positive (i.e., protocooperation and mutualism) and positive-neutral relations (i.e., commensalism) should favor genetic exchange between the populations, while negative (i.e., competition) and negative-neutral relations, in which one population displaces the other, should foster lower frequency of genetic exchange. Finally, under neutrality exchanges should occur at low frequency. The interactions might have different intensities in situ for different organisms considered while organisms might experience several interactions at the same time or one type after the other. The strength of those interactions might also vary depending of the environmental conditions and the acquired adaptations (HGT or mutation). However, it needs to be pointed out that higher frequency of exchange does not necessarily mean higher probability of fixation, since the later depends also on the adaptive value of the exchange. This implies that in some cases highly adaptive exchanges can occur and get fixed during negative relations or the opposite, i.e., high exchange of non-adaptive, or slightly advantageous, DNA material will not get fixed even when organisms meet often.

In conclusion, a deeper understanding of the mechanisms affecting HGT will require a better understanding of the ecological interactions between populations and the effect of interactions on the frequency of HGT. Through the integration of ecology and population genomics, the most relevant ecological interactions favoring HGT can be identified and their effects be studied. Further, population genomics combined with metagenomics[1] can determine how these interactions may differ between environments (i.e., anaerobic fermentation bioreactors vs. decaying organic matter in the forest), and what genes and functions (i.e., defense mechanisms, metabolism) are more ecologically selected (adaptive) under different environments.

[1] Metagenome refers to the composite genome of all members of a microbial community

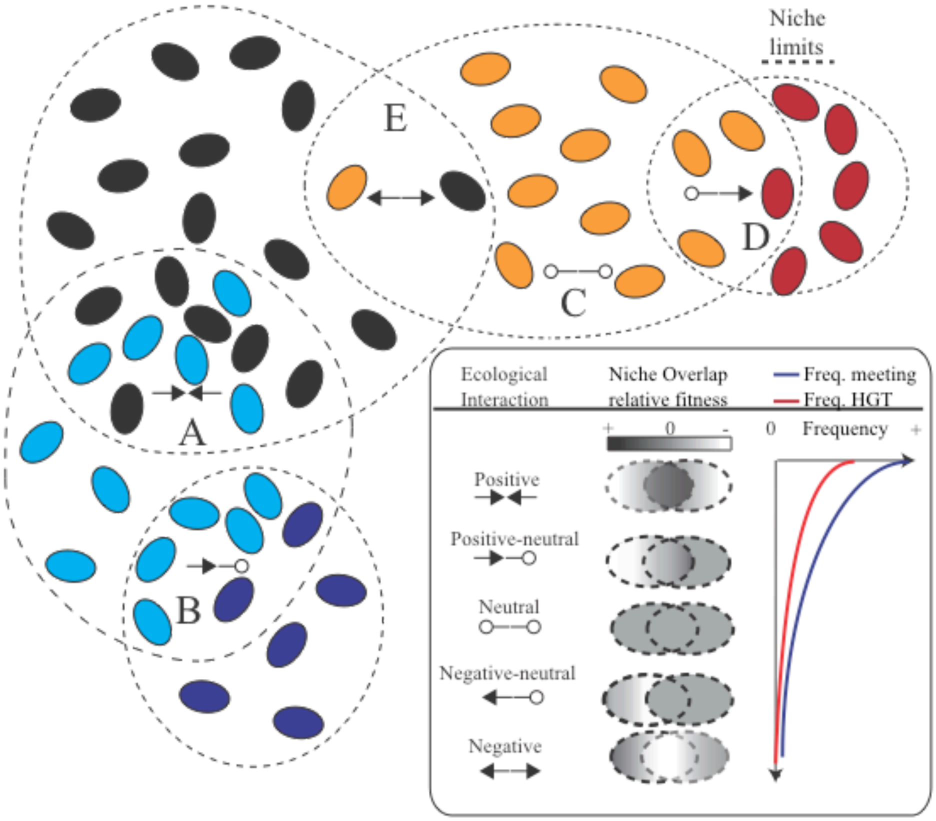


Figure 1.1 Model of the effect of ecological interactions on the frequency of HGT. The figure represents a prokaryotic community; different populations or "species" are represented by different color. Dashed ovals contain the niche range of each population and overlapping ovals denote overlapping niche between the corresponding populations. The inset shows the type of main ecological interaction based on the effect that one species exerts over the other. The niche overlap panel shows the relative fitness of the population in the overlapping and non-overlapping niche range (see scale bar on the top), how frequent populations are expected to meet in the overlapping niche (blue line), and how the latter determines the frequency of genetic exchange (red line). Five main interactions are shown, positive (e.g., protocoperation), positive-neutral (e.g., commensalism), neutral, negative-neutral (e.g., amensalism) and negative (e.g., competition).

1.5 The Importance of HGT for the Models of Prokaryotes Evolution

Genetic exchange can shaped the evolution of prokaryotes in two contrasting ways, as a process that could either maintain populations together (cohesive, mediated by HR) or separate populations by promoting diversification and appearance of new populations. Evidence in support of both outcomes have been reported and models employing the cohesive or diversifying role of HGT have been proposed to explain how genetic coherent populations can exist in nature and evolve in new species. Here, I present the most influential models of prokaryotic evolution at present in which HGT is a fundamental mechanism: the biological species model, the ecological speciation model, and the temporal fragmentation.

1.5.1 The Biological Species Concept

The Biological Species concept defines species in terms of their capability to interbreed [87, 88]. In prokaryotes, HR can have a similar effect to that of interbreeding, frequently replacing small regions of the genome with those from other members of the same species or from closely related species. In this case, a new proposition or species can arise not because of fundamental ecological constrains or geographic separation but rather because the efficiency of recombination decreases between increasingly more divergent DNA sequences [14]. Recombinant organisms are categorized as “sexual” if the rate of mutation “m” (i.e., new polymorphisms brought in by mutation) is lower than the rate of recombination “r” (i.e., polymorphisms removed by recombination during the same time interval). If this scenario is maintained over time, then genetic cohesiveness and discrete populations are expected. The benefits of sexual speciation in prokaryotes have been extrapolated from eukaryotic models of evolution [89]. One of the most interesting evolutionary models used to explain sexuality is the Fisher-Muller model (FM), also known as the adaptive landscape model. Under the FM model, recombination of chromosomes and random mating will produce recombinant genomes with fewer deleterious mutations or conversely promote the propagation of favorable alleles. Laboratory studies in *Helicobacter pylori* have proved that natural transformation increases the rate of adaptation to novel environments, consistent with the expectations of the FM model [68]. Comparative analysis of bacterial genomes also supports the spreading of positively selected genes such as antibiotic [90] resistance and toxin-encoding genes in several pathogenic populations [35-37]. Theoretical modeling has also reinforced the idea that, under realistic conditions, HR can increase the rate of adaptation [34].

An increased number of metagenomic studies have, more recently, revealed that discrete genetic clusters, similar to those expected by high rates of HR, represent a common observation within natural microbial communities [91, 92]. Analysis of large metagenomic datasets from marine (Global Ocean Survey)[91] and freshwater environments (Lake Lanier, GA) uncovered clear genetic discontinuities between co-occurring populations [93]. For instance, a clear separation, in terms of sequence identity, was evident among related populations (*Burkholderiales* order) by recruitment analysis of sequence reads against reference contig sequences in Lake Lanier metagenome. The patterns revealed by recruitment plots (Fig 1.2A), coverage plots (Fig 1.2B), and phylogenetic analysis fig 1.2C), showed genetically distinct populations, with no apparent, abundant intermediate genotypes. These patterns of genetic discontinuity seem to support the notion that distinct genetic populations are a dominant feature in the environment. However, further studies are required to establish whether these patterns of genetic discontinuity are the result HR or that of other processes such as population sweeps caused by periodic natural selection, which can also result in similar discrete populations [94, 95].

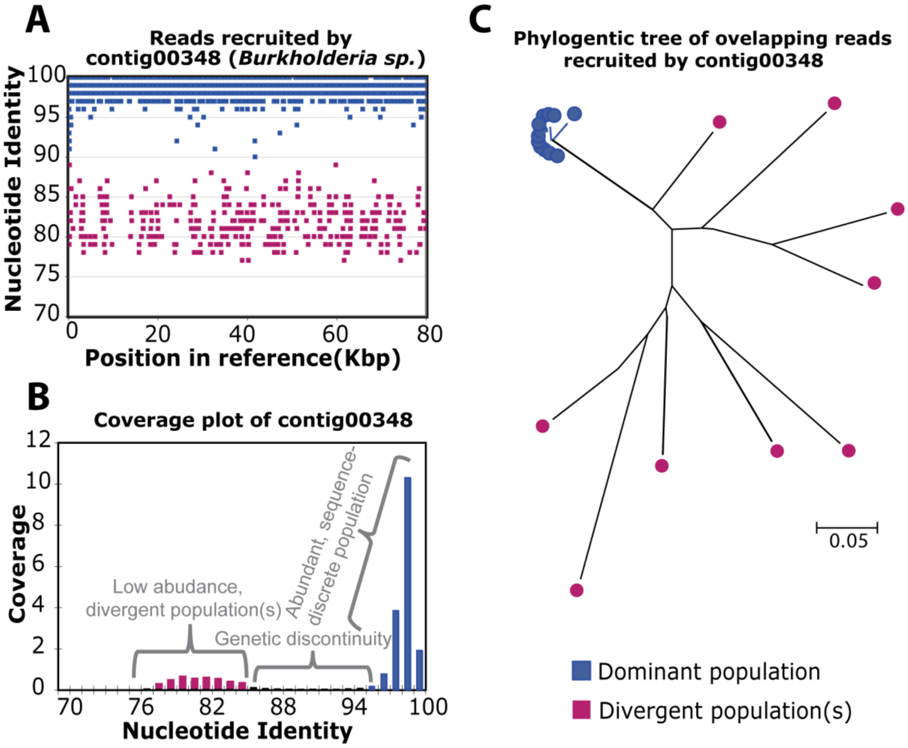


Figure 1.2 Sequence-discrete populations. (A) Fragment recruitment plot of the Lake Lanier (Atlanta, GA) metagenome [96], performed essentially as described previously [91], using as reference a large contig (100 kb) of a *Burkholderia* sp. conting (heterotrophic, *betaproteobacterium*). (B) Coverage plot of the same data as in panel A, performed as described previously [92]. (C) Neighbor-joining phylogenetic tree of all

fully overlapping reads (150 pb length) of the metagenome that mapped on the single-copy transcription termination factor Rho encoded on the contig. Figure is adapted from Caro-Quintero and Konstantinidis, Env. Microbiology 2012 [97].

1.5.2 The Ecological Speciation Models

There are two ecological models that have been advanced to explain how speciation occurs in prokaryotes; both models are based on the acquisition of ecologically relevant gene(s), either by mutation or HGT, which could drive separation of populations into new ecological niches. However, the mechanisms by which this happens are to some extent incompatible between the two models. The first model is the ecotype model. It is based on the periodic selection concept. This concept assumes that the recombination between sub-lineages of a population is very infrequent for the spread of adaptive alleles and therefore, the genetic discontinuity arises between sub-lineages as a result of periodic selection caused by the appearance of advantageous mutation(s) in one of the sub-lineages (but not the other) (*9*). Recent updates of the model incorporating the possibility of HGT suggest that in some cases the acquired genes will bring new features and allow the population to either outcompete the populations within the same or highly overlapping niche or invade a new niche and thus, separate from the ancestral population [62]. One limitation of the ecotype concept is our limited understanding of the environmental conditions (i.e., biochemical, physical and spatial) that define the ecological niche and the niche range of a microorganism. This is both a theoretical and a practical problem because at the micro-scale it is challenging to quantify, identify and untangle the important dimensions of the niche. Most of the original observations of repeated selective population sweeps were made in chemostats that are fairly stable, whereas natural environments are strikingly unstable and diverse [88]. Another limitation of the ecotype model is that it assumes that recombination and exchange are infrequent in nature; however, several recent examples have clearly shown (as described in the HR section above) that rampant HGT, mediated by HR, spread adaptations within populations (see also Table 1.1).

1.5.3 Temporal Fragmentation of Bacterial Speciation

The temporal fragmentation model suggests that, within a niche, specific genes can be maintained while the populations freely recombined at the rest of the genome. However, because of lack of sufficient homology at the region flanking the niche-specific genes, recombination decreases and regional chromosomal isolation (divergence) develops. Therefore, isolation can be established at different chromosomal regions in a population and it is expected that the accumulation of such events will hinder recombination, sometimes at different evolutionary periods, and gradually generate a distinct, nascent lineage [98]. The predicted pattern this model have been identified in different regions of the genome of *Escherichia coli* when compared to that of *Salmonella enterica,* suggesting that genetic exchanges were maintained even after the acquisition of relevant ecological genes for about 10 millions of years. Recent analysis of more recently evolved *Escherichia coli* [99] and *Shewanella baltica* [100] genomes, however, did not find the pattern of the temporal fragmentation model (i.e., accumulation of SNPs flanking ecologically relevant regions). The absence of such pattern could be due to the lack of sufficient evolutionary time elapsed or the different populations considered compared to the original study. On the other hand, it is also possible that the minimum length of DNA required for recombination might be shorter than previously anticipated, e.g., a few base pairs long, [98] and therefore acquisition of a new gene will not have a significant effect on recombination rate at the flanking sides.

1.6 Questions that this Thesis Sought to Address and Thesis Outline

Horizontal gene transfer (HGT) is probably the most important mechanism for functional novelty and adaption in prokaryotes but our understanding of HGT is far from complete. A robust understanding of the rates of genetic exchange for most bacterial species under natural conditions and the influence of the ecological settings on the rates remain elusive, severely limiting our view of the microbial world. Little is known about how ecological interactions affect the frequency of genetic exchange, particularly what kind of relationships might produce effective encounters for genetic exchange to occur. This dissertation describes four studies of HGT in free-living bacteria where I have effectively elucidated some of these ecological interactions and the environmental selective pressure conditions by integrating physiological and ecological data with comparative analyses of whole-genomes of isolates. The effect of ecology on HGT will be presented at three important levels or time-scales: the **species level** (Chapter 2 and 3), the **genus level** (Chapter 4) and the **phylum level** (Chapter 5 and 6**).**

Chapter 2 and 3 show two examples of how ecological settings influence rates of HGT in natural populations. In brief, these chapters discuss the analysis of complete genomic sequences and expressed transcriptomes of several co-occurring *Shewanella baltica* strains recovered from the Baltic Sea. It was found that isolates with more overlapping ecological niches had exchanged a larger fraction of their core and auxiliary genome, up to 20% of the total, in very recent past. The frequency and spatial patterns across the genome of HR suggested that some *S. baltica* strains evolve sexually, fostered by overlapping ecological niches and unbiased exchange of genes. To the best of my knowledge, this represents the first example where sexual speciation in bacteria, fostered by ecology, was unequivocally shown based on whole genome sequences.

Chapter 4 evaluates the frequency of HGT between two different species of *Campylobacter*, *C. coli* and *C. jejuni,* to test whether or not these two distinct species are converging through HGT as suggested by Sheppard and collaborators (Science 2008). Convergence of distinct bacterial species, if true, has major theoretical implications for the species concept and practical consequences for epidemiological studies. In this study, the *Campylobacter* Multi Locus Sequence Typing (MLST) database used previously [9] was re-analyzed in conjunction with available genome sequences of the two species. The analysis convincingly showed that HGT mobilized mostly genes with ecological/selective advantage for the host genome and the two *Campylobacter* species do not converge in their whole genomes.

Chapter 5 focuses on the role of HGT in spreading metabolic capabilities between distantly related organisms of different phyla. In particular, the genome sequences of two members of the newly proposed *Sphaerochaeta* genus (*Spirochaetes* phylum) were analyzed and shown to not only have lost the spiral flagellar genes, a hallmark of *Spirochaetes* but also have acquired more than 40% of their total genes form distantly related organisms, especially members of the *Clostridiales* order (*Firmicutes* phylum). Such a high level of direct inter-phylum genetic exchange is extremely rare among mesophilic organisms and has important implications for the assembly of the prokaryotic Tree of Life.

Chapter 6 aims to extent the analysis of HGT in *Sphaerochaeta* genomes (inter-phylum HGT) to all available complete genome sequences of free-living organisms. In order to obtain a quantitative understanding of inter-phylum HGT, a novel bioinformatics pipeline was developed that determined the uniqueness and significance of HGT events, normalizing for the limitations in the current collection of competed genomes such as the overrepresentation of a few phyla. The pipeline was used to answer questions such as what percent of genomes have undergone inter-phylum HGT and what ecological mechanisms and environmental conditions account for differences between genomes. The results revealed that HGT between distantly related organisms might be more frequently than previously anticipated and that networks of HGT within overlapping ecological niches can assemble large parts of the metabolic functions of the corresponding microbial communities.

Finally, in Chapter 7 I provide a summary of our findings and how they contribute to the better understanding of bacterial evolution, as well as a perspective for future studies.

CHAPTER 2

UNPRECEDENTED LEVELS OF HORIZONTAL GENE TRANSFER AMONG SPATIALLY CO-OCCURRING *SHEWANELLA* BACTERIA FROM THE BALTIC SEA

Reproduced in part with permission from Alejandro Caro-Quintero, Jie Deng, Jennifer Auchtung, Ingrid Brettar, and Manfred Hoefle, and Konstantinos T Konstantinidis.
The ISME Journal. 2010, 5(1), 131-140.
Copyright 2013, International Society for Microbial Ecology.

2.1 Abstract

High-throughput sequencing studies during the last decade have uncovered that bacterial genomes are very diverse and dynamic, resulting primarily from the frequent and promiscuous horizontal gene exchange that characterizes the bacterial domain of life. However, a robust understanding of the rates of genetic exchange for most bacterial species under natural conditions and the influence of the ecological settings on the rates remain elusive, severely limiting our view of the microbial world. Here we analyzed the complete genomic sequences and expressed transcriptomes of several *Shewanella baltica* isolates recovered from different depths in the Baltic Sea and found that isolates from more similar depths had exchanged a larger fraction of their core and auxiliary genome, up to 20% of the total, compared to isolates from more different depths. The exchanged genes appear to be ecologically important and contribute to the successful adaptation of the isolates to the unique physicochemical conditions of the depth. Importantly, the latter genes were exchanged in very recent past, presumably as an effect of isolate's seasonal migration across the water column, and reflected sexual speciation within the same depth. Therefore, our findings reveal that genetic exchange in response to environmental

settings may be surprisingly rapid, which have important broader impacts for understanding bacterial speciation and evolution and for modeling bacterial responses to human-induced environmental impacts.

2.2 Introduction

High-throughput sequencing during the last decade have revealed that bacterial genomes are much more diverse and dynamic than previously anticipated [101-103]. For instance, gene content variation among strains of the same bacterial species may comprise 30-35% of the genes in the genome [101, 104]. This gene diversity and genome fluidity frequently underlies the emergence of new pathogens and the natural attenuation of important environmental pollutants, and hence, has important health and economical consequences [86]. Horizontal gene transfer (HGT) accounts for a substantial fraction, if not the majority, of the bacterial genomic fluidity and diversity [7, 105, 106]. However, a robust understanding of the rates of genetic exchange for most bacterial species under natural conditions and the influence of the ecological settings on the rates remain elusive [86, 107, 108]. An improved understanding of the previous issues has important broader impacts such as for reliable diagnosis of infectious disease agents, successful bioremediation strategies, and robust modeling of bacterial evolution and speciation.

Stratified aquatic systems are characterized by sharp physical, chemical and nutrient gradients and thus, offer unique opportunities for studying the role of the environment in shaping population (and genome) structure and dynamics. One such system, which is among the most stable systems on the planet [e.g., water retention time in the order of 20-30 years [109]] and has been characterized extensively due to its long history of pollutant contamination, is the Baltic Sea [110]. *Shewanella baltica* dominates the pool of heterotrophic nitrate-reducing bacteria isolated from the oxic-anoxic interface of the Baltic Sea [111, 112]. For instance, *S. baltica* organisms (strains) accounted for 32-80% of total cultivable denitrifying bacteria under different growth conditions during our isolation efforts in 1986 [112]. These findings further corroborate the important role of *Shewanella* bacteria in cycling of organic and inorganic materials at redox interfaces [113, 114].

To identify the genetic elements that enable *S. baltica* to adapt to redox gradients and provide novel insights into the mechanisms and rates of genomic adaptation, we performed whole-genome sequence and DNA-DNA microarray comparative analyses of a large collection of isolates from the Baltic Sea (n = 116, Fig. 2.1). Our analyses revealed that *S. baltica* genomic adaptation to environmental settings, mediated by HGT, may be much more rapid and extensive compared to what seen previously in other marine bacteria.

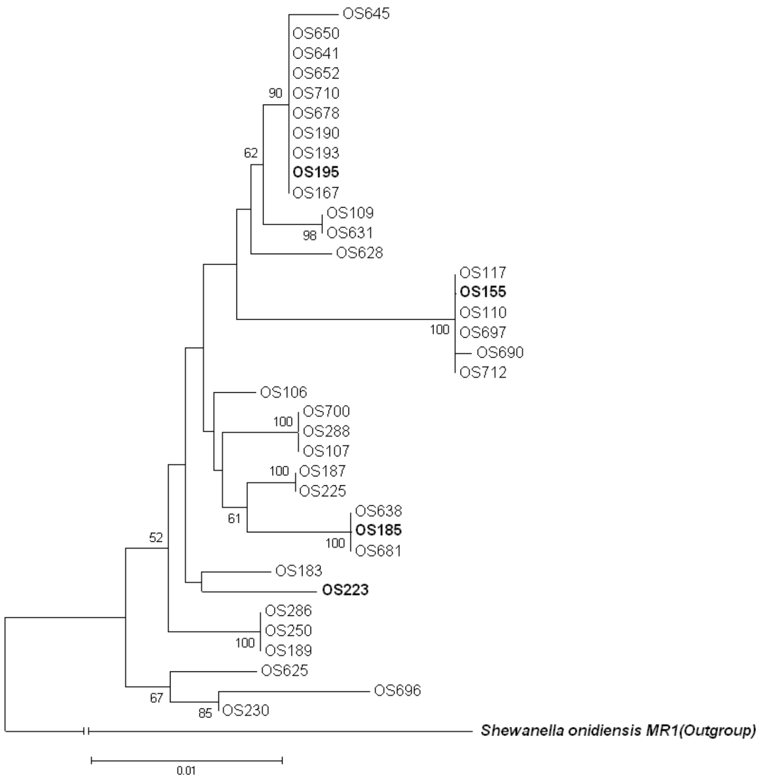


Figure 2.1 Phylogenetic relationships among the *S. baltica* strains used in this study. 36 strains from our collection of total 116 strains, which had the most unique Randomly Amplified Polymorphic DNA (RAPD) fingerprinting profiles [112], were selected for sequencing of their gyrase (*gyrB*) gene. The neighbor joining phylogenetic tree [115] of the 36 strains based on their gyrB gene sequences is shown. The evolutionary distances between the strains were computed using the Maximum Composite Likelihood method, as implemented in the MEGA4 package [116]. Scale bar represents the number of base substitutions per site. Bootstraps values from 500 replicate trees are also shown next to the branches. Strains whose name starts with "OS1" or "OS2" were isolated in 1986; the remaining strains were isolated in 1987.

2.3 Materials and Methods

2.3.1 Organisms Used In This Study

The *S. baltica* strains used in this study were isolated on denitrifying media ($NHNO_3$, $THNO_3$) or anaerobic ZoBell agar. More details on sampling, isolation conditions and genome fingerprinting patterns of each strain are provided in [112]. The complete genome sequences of the four *S. baltica* strains used in the study were obtained from GeneBank [117]. The strains and their GeneBank accession numbers were, OS195 (NC_009997, NC_009998, NC_009999, NC_010000), OS185 (NC_009665, NC_009661), OS155 (NC_009052, NC_009035, NC_009036, NC_009037, NC_009038), OS223 (NC_011663, NC_011664, NC_011665, NC_011668).

2.3.2 Identification of Orthologs

Orthologs among the four *S. baltica* genomes were identified using a reciprocal best-match blastn approach, essentially as described previously [118]. In brief, the sequences of the predicted genes in the genome of strain OS195 were searched, using the blastn algorithm [119], against the genomic sequence of each of the remaining three strains. The best match for each query gene, when better than at least 70% overall nucleotide identity (recalculated to an identity along the entire sequence) and an alignable region covering >70% of the length of the query gene sequence, was extracted using a custom PERL script and searched against the complete gene complement of OS195 to identify reciprocal best matches (RBM). Such RBM conserved genes were denoted as orthologs. Orthologs conserved in all four genome were denoted as core orthologous genes. Genes that found no match better than the previous standards against any of the remaining three genomes were denoted as OS195-specific (strain-specific). Genes conserved in some but not all of the strains were denoted as variable (Table A1, which includes all OS195 genes).

2.3.3 Recombination Analysis

Recombination fragments were detected using a custom-made approach, essentially as described previously [92, 120]. Briefly, the genomic sequence of OS195 was cut *in-silico* in 500 bp-long consecutive sequence fragments. The fragments were subsequently searched against the other *S. baltica* genomes for best matches, using blastn as described above for orthologs. A fragment was flagged as (potentially) recombined in another strain when its best blastn match in the latter strain showed more than 99.5% nucleotide identity while its identity in the other strains was lower <98%, which corresponded to the typical genetic distance between the *S. baltica* strains (i.e., ~96.7%). Such fragments and their adjacent fragments were subsequently visually inspected to determine the presence of recent homologous recombination as shown graphically in Figure 2.2B. The recombined fragments identified this way were further validated by the Genetic Algorithm for Recombination Detection (GARD) [121]. Briefly, all core genes in all genomes were concatenated to provide a whole-genome core gene alignment. The alignment was scanned in 1 or 2 Kbp-long windows by GARD (longer windows are too computationally demanding for GARD) in a pair-wise fashion (i.e., two genomes at a time) and the sequence windows that provided delta AIC values higher than ~10 were flagged as containing recombined segments, as suggested earlier [121]. The recombined fragments identified by GARD were contrasted with those identified by visual inspection of the nucleotide identity patterns (blastn approach). Sequence fragments or genes that showed high nucleotide identity (>98%) between all four genomes encoded typically for highly conserved housekeeping genes such as the rRNA operon genes. Such fragments were excluded from the recombination analysis because it could not be established whether the identity patterns observed were due to recombination or high sequence conservation. Fewer than 100 fragments were excluded from the analysis for the latter reason (from more than 3,000, in total; see Table A1). The number of synonymous substitutions per synonymous site (Ks) for every gene was calculated based on the gene nucleotide codon-based alignment using the codeml module of the PAML package [122].

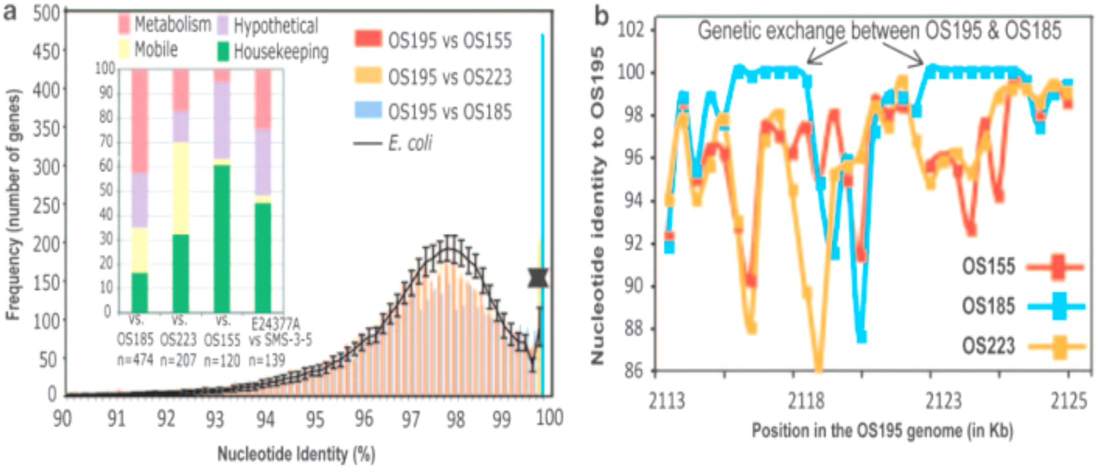


Figure 2.2 Nucleotide identity distribution of orthologous genes in the *S. baltica* genomes. Panel A: All genes in the OS195 genome were compared to their orthologs in strain OS185, OS223, and OS155. For each pairwise comparison (see figure key), the number of orthologs is plotted against their nucleotide identity. The solid line represents

the average of 125 comparison of between *E. coli* genomes with similar ANI (~97%) and number of orthologs genes (~3,500) with the *S. baltica* genomes. Error bars represent one standard deviation from the mean and the "X" represents the value of the most outlier *E. coli* genome pair. The inset in Panel A shows the functional annotation of the 100% nucleotide identity genes identified for each pairwise comparison (for details, see text). An graphical representation of the type of recent genetic exchange events assessed by our analysis is provided in Panel B. Note that the sequences of OS155 and OS223 show consistently lower, and close to the genome average, nucleotide identities to their recombined counterparts in OS195 and OS185.

2.3.4 DNA Microarray Construction and Analysis

Microarray slides were constructed by Biodiscoveries LLC (Ann Arbor, MI, USA), and were consisted of 44-48 bp long, *in-situ* synthesized probes. Probes were designed from the genomic sequences of the four sequenced *S. baltica* strains using the following strategy: for core orthologous genes sharing at least 90% nucleotide sequence identity over 90% of their length, probes were designed only against the ortholog in the OS185 genome; likewise, for the remaining orthologous genes, probes were designed only against the corresponding ortholog gene, using the following preference: OS185> OS195> OS223> OS155. Probes against all genome-specific genes or genes related at a level below the previous standards were also included.

For DNA-DNA microarray studies, genomic DNA was extracted as previously described [123], and sonicated to produce DNA fragments less than 3 kbp in size. DNA samples were labeled with the fluorescent Cy5 dye by incorporation of amino-allyl-dUTP through extension from random primers using *E. coli* DNA polymerase Klenow fragment I, followed by addition of amine-reactive Cy5. Microarray slides were pre-hybridized in buffer containing 0.1% SDS, 5xSSC and 1 mg/mL BSA at 50 ^{0}C for 90 min, and washed with 0.5xSSC and water. Cy5-labeled DNA samples were mixed with the same volume of 2X hybridization buffer (10xSSC, 0.2% SDS, 0.2 mg/mL herring sperm DNA, and 46% formamide), heated at 95 ^{0}C for 5 min and then transferred to 68 ^{0}C. Samples were applied to pre-hybridized slides, which were then incubated at 50 ^{0}C for 18 h before being washed and scanned using an Axon GenePix 4000B scanner (one-channel hybridization). For array data processing and normalization, mean signal intensity from the negative control probes were subtracted from signals of all spots. Subsequently, the median signal from the core probes with 100% identical matches in all four genomes was calculated for each microarray dataset. Based on these calculations, a normalization factor was generated that would bring the median signal of the 100% identity core probes to the same value for each slide. This normalization factor was then applied to all the spots on the slide as proposed recently [124].

For gene expression studies, cells were inoculated into 25 ml of trypticase soy broth and incubated at 22°C with agitation until the cells reached mid-exponential phase (Optical Density at 600 nm = 0.4-0.7). Experiments were repeated in triplicate. Cells were pelleted by centrifugation and resuspended in 350 ml anaerobic HEPES medium

[125]. After 18-19 hours of growth at 22 °C, cells were pelleted anaerobically and resuspended in 10 ml of anaerobic HEPES medium lacking an electron acceptor. 2 ml of this cell resuspension was added to 22.5 ml of HEPES medium containing either 10 mM sodium fumarate, 5 mM sodium nitrate, 10 mM sodium thiosulfate or 10 mM sodium chloride (aerobic culture). All cultures were incubated at 22 °C. The aerobic cultures were aerated by shaking on an orbital shaker at 150 rpm. RNA was extracted from cell pellets using a Qiagen RNeasy kit following the optional protocol for better recovery of low molecular weight RNA. RNA from 3 independent cultures of OS185 and OS195 grown in the presence of oxygen, nitrate, and thiosulfate was used as experimental samples in hybridization experiments. RNA from all four strains grown in each condition (oxygen, nitrate, thiosulfate, and fumarate) was used to construct a reference RNA pool, composed of 45 μg of RNA from each condition for strains OS185, OS195 and OS223 and 15 μg of RNA from each condition for strain OS155. 10 μg of RNA from each experimental condition and a parallel aliquot of reference RNA were reverse transcribed with 9 μg of random primers (Invitrogen). Reactions were incubated at 25 °C for 10 min, 42 °C for 70 min, and 70 °C for 15 min. Remaining RNA was hydrolyzed by adding sodium hydroxide to 33 mM and incubation at 70 °C for 10 min. Labeled cDNA was purified using a QiaQuick MinElute PCR purification column following the manufacturer's protocol with the exception that the sample was eluted in 12 μl of RNase-free water (Qiagen). For each hybridization, 10 μl of labeled experimental cDNA was mixed with an equal volume of labeled reference cDNA and was applied to the oligoarray as described above for DNA-DNA studies. Cy3 and Cy5 signals for each array were normalized to the arithmetic mean of ratios for each array using the GenePix software.

Features that had fewer than 50% of pixels with signal more than two standard deviations above background in both Cy 5 and Cy3 channels were excluded from further analysis. Genes showing significantly increased anaerobic gene expression (nitrate and thiosulfate) relative to aerobic growth were identified using Significance Analysis of Microarrays [126]. Experiments were repeated in triplicate and the mean of the three replicates is reported in Figure 2.3.

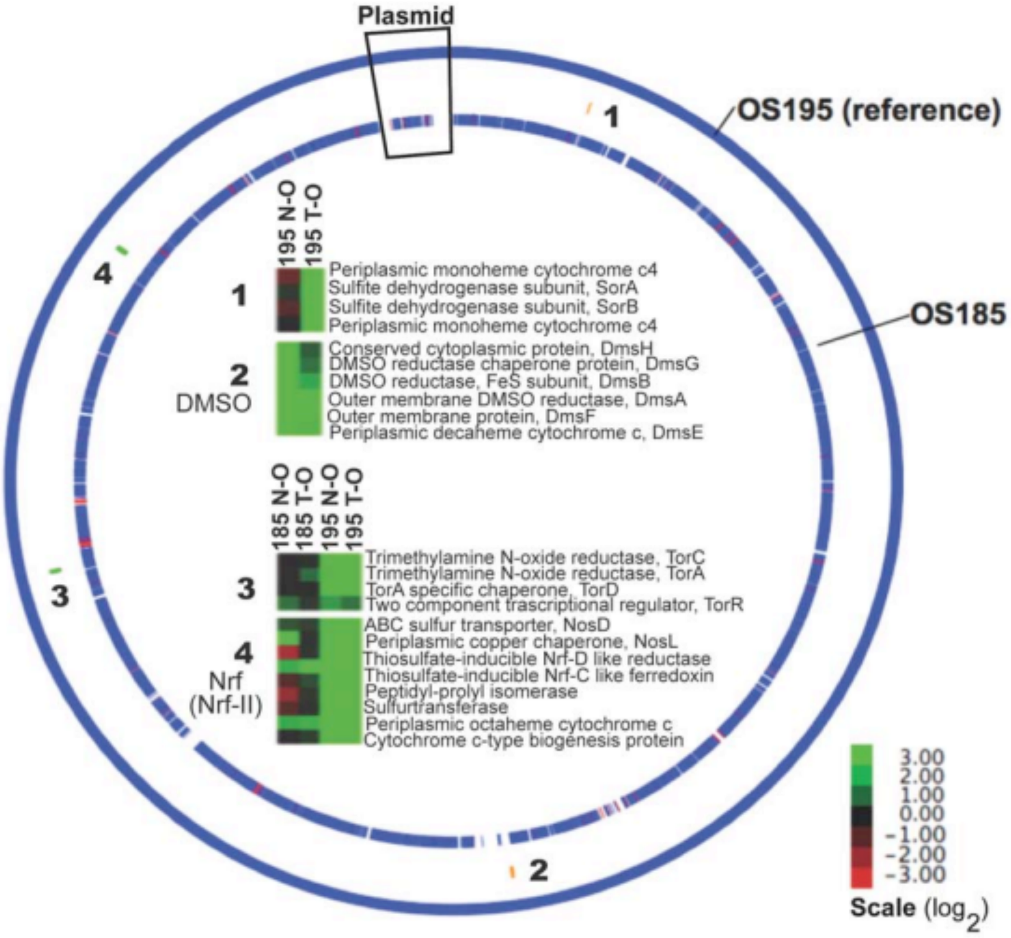

Plasmid
OS195 (reference)
OS185
1
2
3
4
195 N-O
195 T-O
1
Periplasmic monoheme cytochrome c4
Sulfite dehydrogenase subunit, SorA
Sulfite dehydrogenase subunit, SorB
Periplasmic monoheme cytochrome c4
2
DMSO
Conserved cytoplasmic protein, DmsH
DMSO reductase chaperone protein, DmsG
DMSO reductase, FeS subunit, DmsB
Outer membrane DMSO reductase, DmsA
Outer membrane protein, DmsF
Periplasmic decaheme cytochrome c, DmsE
185 N-O
185 T-O
195 N-O
195 T-O
3
Trimethylamine N-oxide reductase, TorC
Trimethylamine N-oxide reductase, TorA
TorA specific chaperone, TorD
Two component trascriptional regulator, TorR
4
Nrf
(Nrf-II)
ABC sulfur transporter, NosD
Periplasmic copper chaperone, NosL
Thiosulfate-inducible Nrf-D like reductase
Thiosulfate-inducible Nrf-C like ferredoxin
Peptidyl-prolyl isomerase
Sulfurtransferase
Periplasmic octaheme cytochrome c
Cytochrome c-type biogenesis protein
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
Scale ($\log_2$)

Figure 2.3 Analysis of gene expression in *S. baltica* OS185 and OS195 strains grown in the presence of different electron acceptors. Genes showing significantly increased anaerobic gene expression [nitrate (N) and thiosulfate (T)] relative to aerobic growth (O) and encoded by genomic islands are shown, with black indicating no change (based on

the average of the three replicates per treatment), bright green indicating up to 32-fold increase in gene transcription, and red indicating decreased transcription. 1 and 2 denote genes that were OS195-specific; 3 and 4 denote genes found in two genomic islands shared by OS195 and OS185. The positions of the genomic islands in the genome are also shown (see Figure 2.6 for details on the outer two circles).

2.4 Results

2.4.1 Unprecedented Levels of Genetic Exchange Among Spatially Co-Occurring *S. baltica* Strains

To unravel the genetic diversity within our *S. baltica* strain collection, four strains that represented the most abundant lineages recovered among the 116 isolates comprising our collection (Fig. 2.1) were fully sequenced. These strains were OS155, OS185, OS223 and OS195 and were recovered from three different depths of the Baltic Sea, 90m, 120m, 120m and 140m, respectively. These depths were characterized by different redox potentials and nutrient availability at the time of isolation. In particular, the 140m depth represented a more anoxic environment, with higher abundance of alternative electron acceptors to oxygen such as nitrate, compared to the (more) oxic environment at 90m depth. The 120m depth was intermediate between these two depths (Fig. 2.4A).

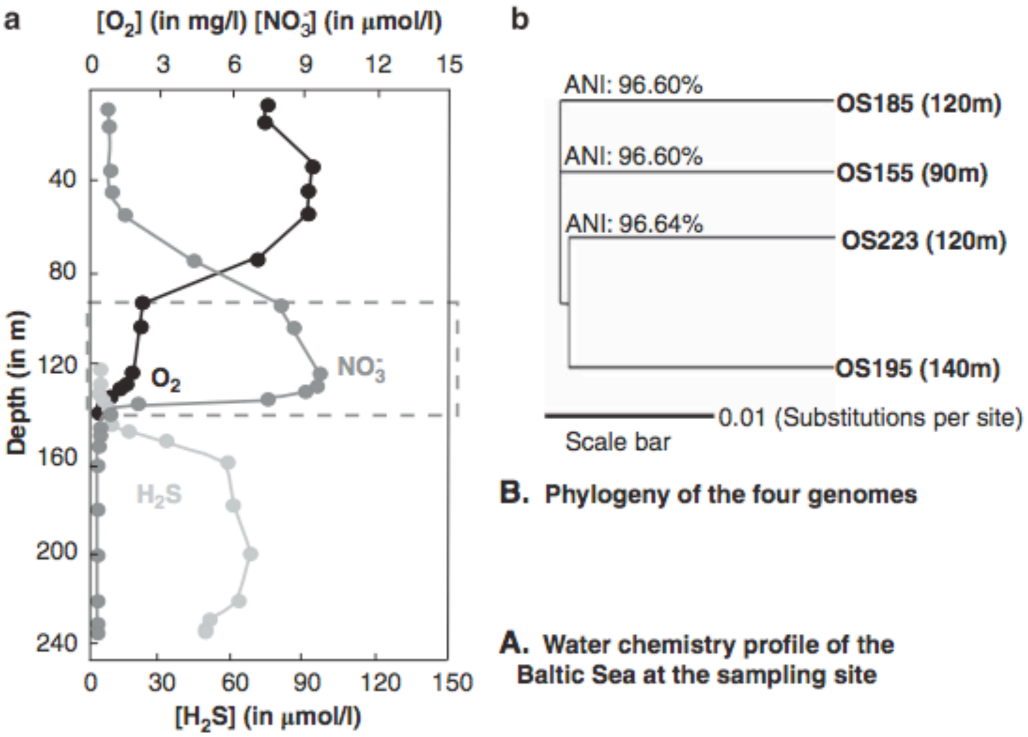


Figure 2.4 The *S. baltica* genomes. Panel A: The water chemistry profile at the site of isolation of the four genomes. Note the appearance of H_2S at around 140m depth is, at least in part, due to the reduction of sulfur compounds including sulfur disproportionation. The whole-genome phylogeny of the genomes based on Maximum Likelihood analysis of the concatenated sequences of all core genes (n = ~2,500) that showed no evidence of recombination, performed as described previously [127], is shown in Panel B. ANI values among the genomes based on the non-recombined core genes are also provided. Panel A is adapted from [111].

The four *S. baltica* genomes showed very similar evolutionary relatedness among each other, e.g., they had identical 16S rRNA gene sequences. To provide for a higher resolution, the genome-aggregate average nucleotide identity (ANI) [101] of all core genes (n = 2,500) with no detectable signal of recombination according to PhiTest analysis [128] was employed. ANI analysis revealed that these four genomes were not only very closely related but also show comparable evolutionary relatedness among each other, with their ANI values being ~96.7% for each pair of genomes compared (Fig. 2.4B). These values are higher than the 95% ANI that corresponds to the 70% DNA-DNA hybridization (DDH) standard frequently used for species demarcation [129]; hence, these genomes belong justifiably to the same species, *S. baltica*.

Despite the comparable evolutionary relatedness among all strains, strains from more similar depths shared, in general, substantially more genes compared to strains from more different depths. For instance, OS195 shared 580 (non-core) genes with OS185 and 350 with OS223, but none of these three strains shared more than 150 genes with OS155 (Fig. 2.5). Remarkably, most (i.e., ~350) of the 580 genes shared between OS195 and OS185 and an additional ~10% of their core genes showed 99.5% to 100% nucleotide identity between OS185 and OS195, contrasting sharply with ~97% identity for the rest genes in the genome and less than 3% high identity (i.e., 99.5% to 100%) core genes among the remaining pairs of genomes, respectively (Table A1). This pattern became more obvious when the frequency of genes was plotted against their nucleotide identity for each pair of genomes compared (nucleotide identity histograms, see Fig 2.2A). Notably, a similar analysis of all pairs of genomes available in GenBank with similar ANI (96.5-97.5%) and genome size (3,500 - 4,500 genes) to the *S. baltica* genome pairs, revealed that the gene nucleotide identity distribution in the OS185 vs. OS195 case was unparalleled and significantly different from any other distribution based on the z-test (p value < 0.001). For instance, among the 125 pairwise comparisons of all available *E. coli* genomes, only *E. coli* strains E24377A and SMS-3-5 had about 150 genes with higher than 99.5% nucleotide identity (still, 4 times fewer genes compared to the OS185 and OS195 case; see Fig. 2.2A). We also observed about 200 genes with >99.5% nucleotide identity between OS195 and OS223, while comparing OS155 against OS195, OS185 or OS223 did not reveal more high identity genes than the average of all genome pairs from Genbank (i.e., n < 100). Because all *S. baltica* genomes show comparable evolutionary distance among each other (Fig. 2.4B), the high identity genes shared between OS185 and OS195 cannot be attributed simply to higher evolutionary relatedness between these two genomes. These findings cannot be explained by preferential deletion of the corresponding genes in OS155 or OS223 either, because the pool of high identity genes included several core genes that showed nucleotide identities in the 95-98% range against their OS155 or OS223 orthologs. Instead, these findings are, most likely, attributed to recent extensive horizontal exchange between OS195 and OS185 or their immediate ancestors.

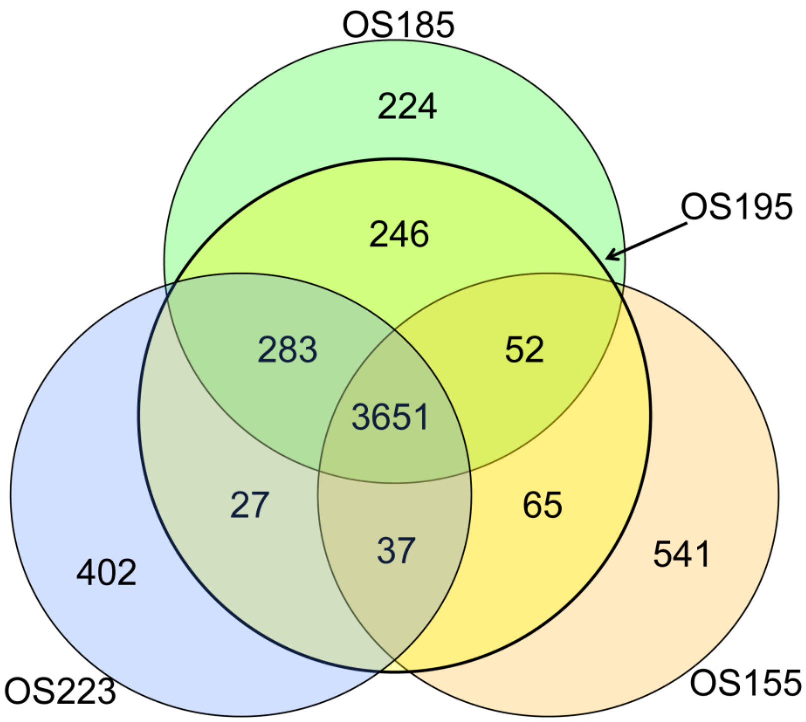


Figure 2.5 Shared and variable *S. baltica* genes. The number of orthologous genes shared between the four *S. baltica* genomes are shown on the venn diagrams. 341 genes were specific to the OS195 genome based on our comparisons (not represented on the diagram but available in Table S1). Orthologous genes were counted only once.

2.4.2 Unconstrained Homologous Recombination Mediates the Genetic Exchange Events

To further validate the previous findings and provide insights into the mechanisms mediating the genetic exchange among the *S. baltica* strains, we examined the functional role of all 100% nucleotide identity genes shared between OS195 and OS185. The genes were assigned to one of the following four categories: (i) genes related to metabolism and regulation, (ii) mobile elements (integrases, transposases and genes contained within prophages, integrons and plasmids), (iii) hypothetical and (iv) housekeeping genes (genes related to central cell functions such as replication and translation), which tend to be more conserved than the genome average at the sequence level [92]. The analysis showed that most of the genes were neither housekeeping nor mobile; instead, most of them encoded for metabolic, transport and regulatory functions related mainly to secondary metabolism. This functional gene distribution contrasted strikingly with that of the OS195 vs. OS155 pair or the *E. coli* genome pairs, which were enriched in housekeeping and hypothetical genes (Fig. 2.2A, inset). Thus, the majority of the exchanged genes do not appear to be the product of a single, specialized vector of horizontal gene transfer such as a bacteriophage or a plasmid.

Further examination of the nucleotide identity patterns of the recently exchanged core genes showed that these genes have been brought into the genome via a homologous recombination mechanism. For instance, the nucleotide identity of the exchanged core genes between OS195 and OS185 against their orthologs in OS155 or OS223 was consistently lower than 100%, and typically in the 95-98% range (for a graphical representation, see Fig. 2.2B). In addition, the majority of the recombined core segments between OS185 and OS195 were randomly distributed in the genome (Fig. 2.6, innermost circle), did not show any strong biases in terms of the function of the genes they contained when compared against the rest of the genome (Fig. 2.7) and were 0.5 to ~10 Kbp long (average ~1.5Kbp; Fig. 2.8). Genes identified as recombined based on such simple sequence comparisons were further validated by GARD, an advanced algorithm for homologous recombination detection [130]. In general, there was a high agreement between the two methods (>80%) in identifying recently recombined fragments (Fig. 2.9). About ten fold more recombined core genes were observed between strains OS195 and OS185 (n=308) than between OS195 and OS233 (n=48) or OS195 and OS155 (n=28), which is consistent with higher genetic flow between OS195 and OS185 compared to the other genome pairs. The majority of the non-core genes shared between OS195 and OS185 showed similar patterns to those described above for core genes, suggesting that they were also brought in the genome via a similar mechanism as the core genes. These patterns are best explained by invoking an unconstrained mechanism for genetic exchange among the *S. baltica* genomes such as transformation or conjugation and homologous recombination as the process through which the exchanged DNA was incorporated into the genome. While the exact mechanism for genetic exchange remains to be elucidated, the genome of *S. baltica* encodes several genes with strong amino acid similarity to known conjugative DNA transfer genes and a complete *recA*-dependent homologous recombination protein complex.

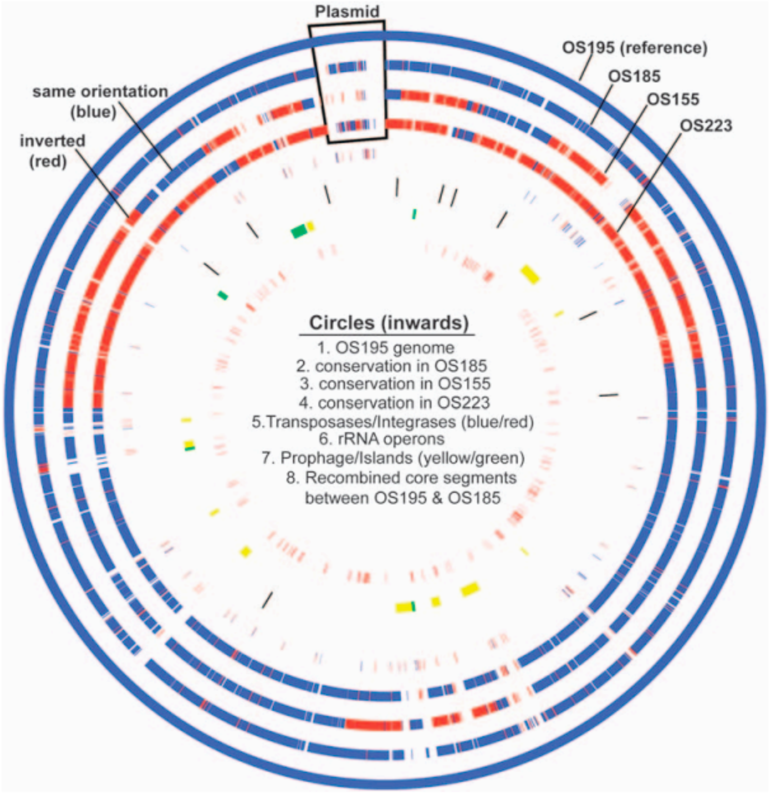


Figure 2.6 Preferential genome-wide and extensive genetic exchange between the *S. baltica* genomes. Circles represent (inwards): the genome of OS195 (#1); the conservation of the OS195 genome in OS185 (#2), OS155 (#3), and OS223 (#4), with red denoting segments of the genome that have been inverted in the latter genomes relative to the OS195; the positions of transposase (blue) and integrase (red) genes in the genome of the OS195 (#5); the position of the rRNA operons (#6); all genomic islands shared between OS195 and OS185, colored either yellow if they corresponded to prophage genomes and prophage remands or green if they encoded probable ecologically important genes (#7); and, the position of the recombined segments between OS195 and OS185 that contained only core genes. Note that the latter segments do not show any spatial bias in the genome, are not typically associated with the mobile genes in the genome and represent a substantial fraction of the core genome.

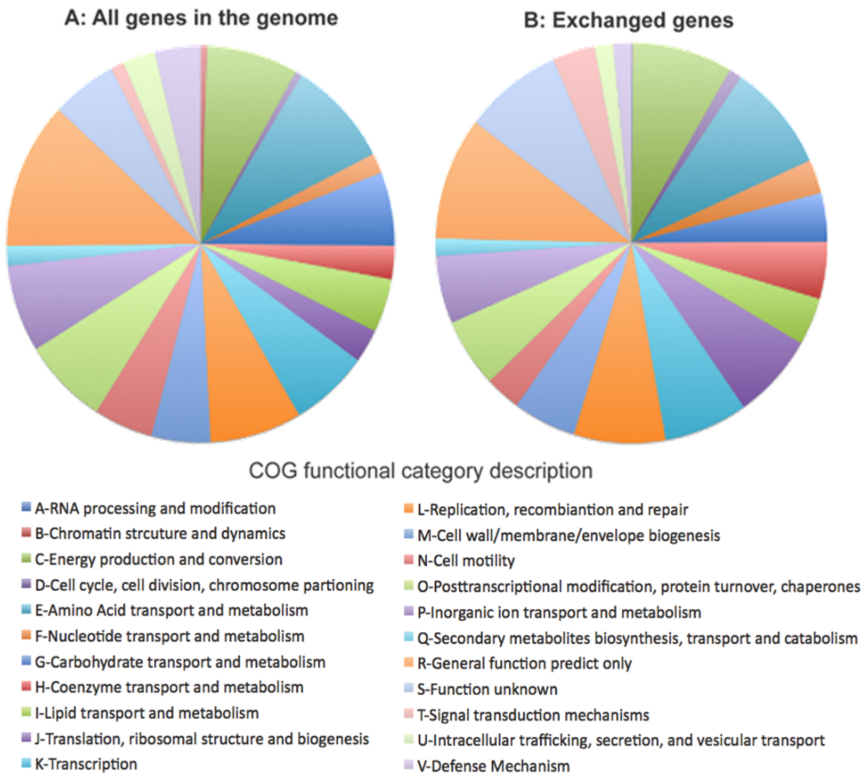


Figure 2.7 Absence of strong functional biases in the genes exchanged between *Shewanella baltica* strains OS195 and OS185. All genes in the genome of *S. baltica* OS195 were assigned to a major gene functional category of the Clusters of Orthologous Groups (COG) database [131], as described previously [132]. The percentages of the total genes in the genome assigned to each category (A) relative to that of only the exchanged genes (B) are shown. Note that the two gene distributions look very similar to each other. Some of the minor differences observed are attributable to category-specific characteristics rather than strong biases in the genes exchanged The description of the categories is also provided (adjusted from the COGs website).

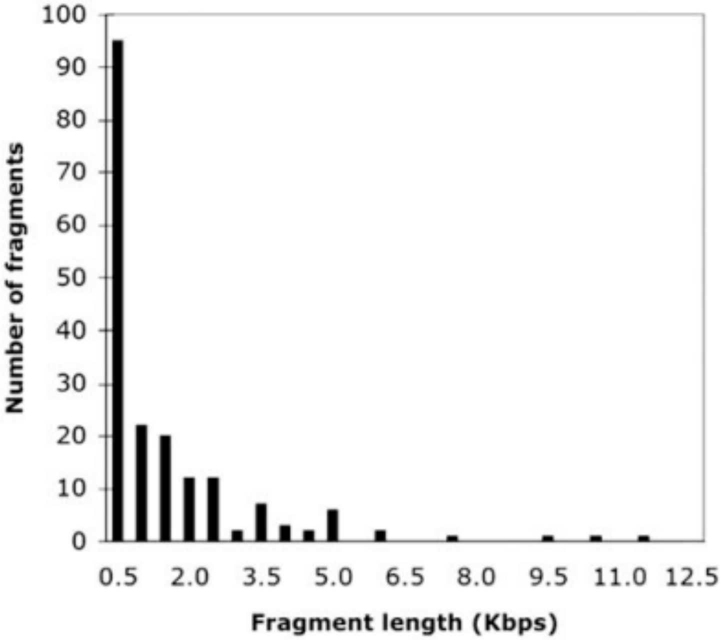


Figure 2.8 Length distribution of the recently recombined fragments between OS185 and OS195. All genetic exchange events between OS195 and OS185 similar to the two events shown in Fig. 2.2B were identified based on visual inspection of the whole-genome alignments (as shown in Fig. 2.2B and described in the material and methods section). The graph shows the length distribution of these recombined fragments.

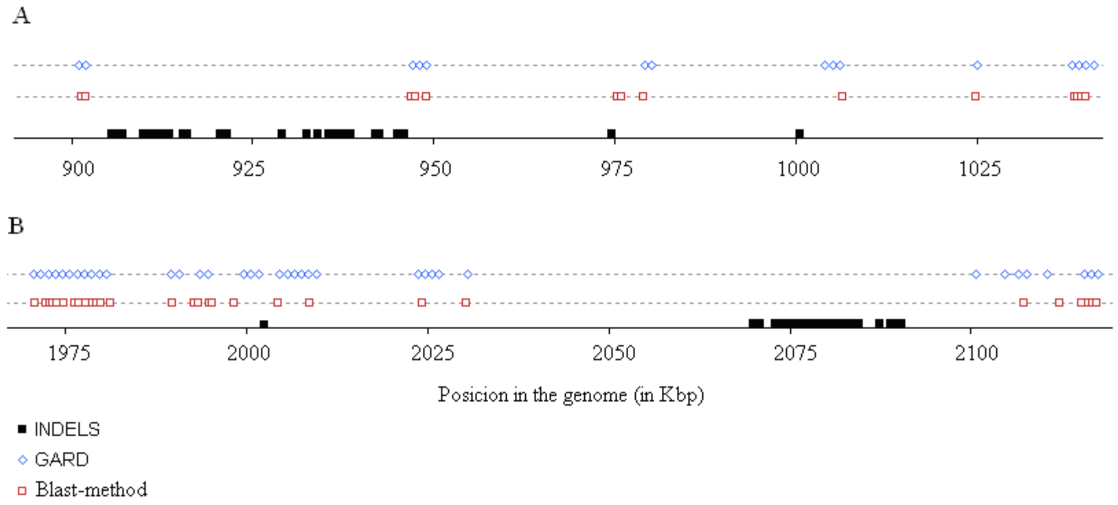


Figure 2.9 Congruence of the blast- and GARD- based methods for detecting recently recombined genes between *S. baltica* strains OS195 and OS185. The graph shows the identified recombined genes based on our blast method (red open squares) and GARD (blue open diamonds) in two representative 150 Kbp long segments of the OS195 genome. Black filled squares represent deletions, insertions or housekeeping genes of high nucleotide identity, i.e., areas of the genome not assessed for recombination. Note the high congruence between the blast- and GARD-based methods in detecting recently recombined genes. At the whole genome level, more than 80% of the total sites identified by the blast method as recombined had significant recombination signal by GARD analysis as well (>90% when recombined segments longer than 2Kbp, which were not considered in the GARD analysis, were removed from the analysis).

Assessing historical, as opposed to recent (e.g., Fig. 2.2B), recombination among the *S. baltica* genomes was severely impeded by the very high nucleotide relatedness of the genomes, multiple (old) recombination events on the same segment of the genome, and the process of amelioration of the newly introduced DNA sequence into the recipient cell [133]. Accordingly, we report here on easily detectable, recent recombination events only.

2.4.3 Clonal or Sexual Divergence?

Even though precise dating of the genetic exchange events cannot be made due to lack of understanding of important population parameters such as the *in-situ* generation time [14], a relative dating was attempted based on the number of generations (*g*). We quantified *g* by dividing the average Ks value (synonymous substitutions per synonymous site) of all core genes with no obvious signal of recent recombination by the mutation rate of bacterial genomes [5.4 x 10 $^{-10}$ substitution/site/generation [134]], as suggested previously [135, 136]. (Synonymous substitutions are thought to be neutral and thus, reflect the intrinsic mutational rate). The distribution of the Ks values of the core genes approximated the normal distribution and was very similar among all pairs of *S. baltica* genomes (6 pairs in total; see Fig. 2.10 for all pairs; Fig. 2.11A for OS195 vs. OS185). The average Ks was ~0.0898, providing for a divergence time since the last common ancestor of all genomes that corresponded to 1.66 x 10^8 generations (±1.03 x 10^7 generations), with 95% confidence. By the same token, and using the average Ks of all recently recombined core genes between OS195 and OS185 (Ks = 0.0015), i.e., the substitutions accumulated since the onset of recombination, we estimated that the recent recombination events identified here took place within the latest ~2.77 x 10^6 generations. Thus, recombination between OS195 and OS185 occurred within the latest ~2% of the total divergence time since the last common ancestor of the *S. baltica* strains (Fig.

2.11B). We also employed the codon usage bias of each gene, essentially as previously described [98], to normalize the Ks values (and derived divergence time estimates) for the different mutational rates of the genes due to the varied selection pressures acting on each gene. The normalized Ks values provided for similar results to those obtained with non-normalized Ks values (data not shown).

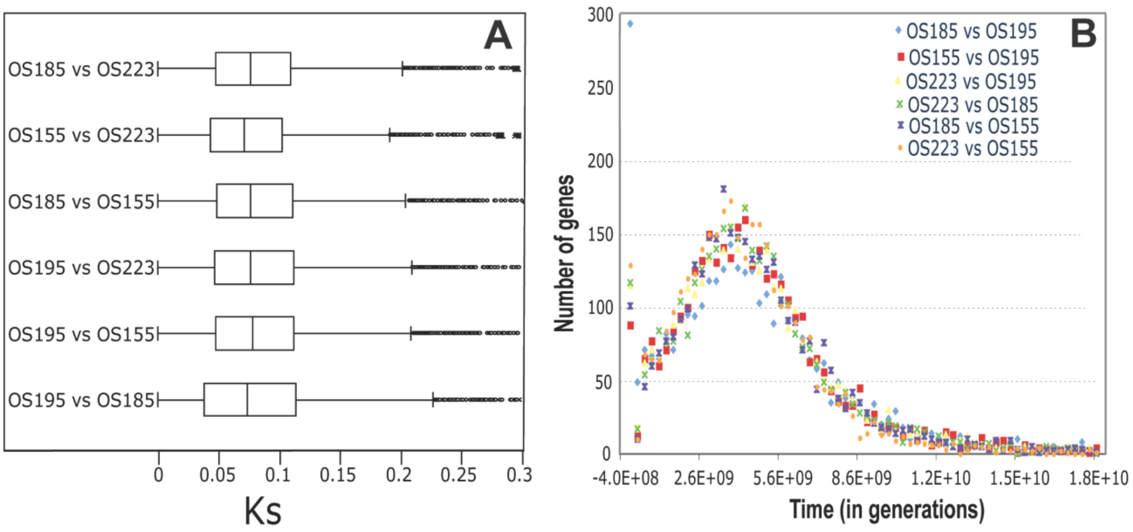


Figure 2.10 Synonymous substitutions among the *S. baltica* genomes. The Ks values (number of synonymous substitutions per synonymous site) were calculated for all core genes (n = 3,500 genes) for every possible pairwise combination of the four *S. baltica*, using the gene nucleotide codon-based alignment and the codeml module of the PAML package [122]. The distribution of the Ks values for every genome pair is shown in Panel A; the vertical line represents the median Ks. Divergence time for each gene (Panel B) was calculated by dividing the Ks value of the gene by the mutation rate of bacterial genomes [5.4 x 10 $^{-10}$ substitution/site/generation [134]].

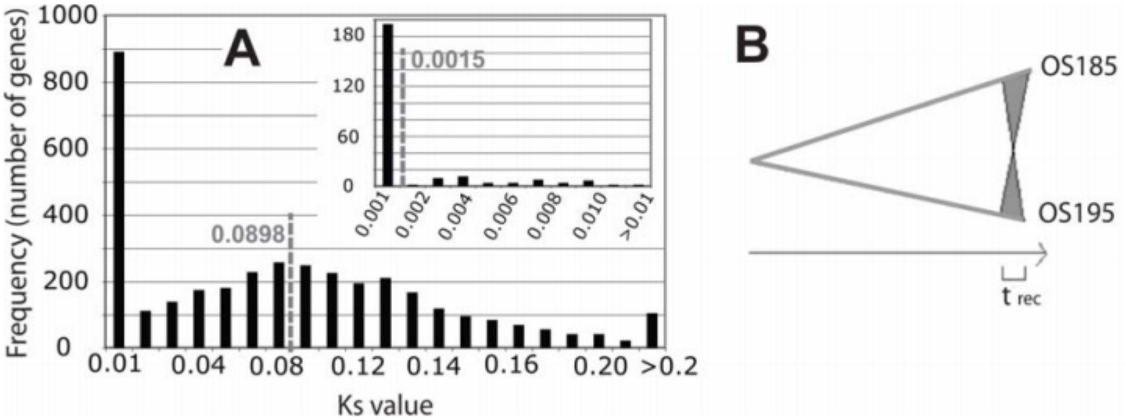


Figure 2.11 Dating recombination events. Panel A shows the distribution of the Ks values of all core genes and the recombined core genes only (inset) for the OS195 vs. OS185 comparison. For the former gene set only genes showing 93% to 98% nucleotide identity were included in the analysis (n = 3550); for the latter one, the analysis was restricted to recombined genes sharing at least 99.5% across their entire length (n = 257). Note the difference in the scale of the x-axes between the main graph and the inset. Panel B represents the period that OS195 and OS185 had been recombining as a fraction of the total divergence time since their last common ancestor. Divergence time was calculated based on the mean Ks value of the non-recombined vs. the recombined core genes as described in the text.

Using a simple strategy based on the Ks values, we also attempted to quantify the relative importance of recombination to mutation. For the time that recombination had been taking place between OS195 and OS185, we assumed that the synonymous substitutions brought in the genome by mutation equal the total length of all core genes (3.5 Mb) multiplied by the number of substitutions observed during this time (i.e., the Ks of recombined genes, which equaled 0.0015). During the same time, recombination purged a total number of synonymous substitutions that equaled the average number of substitutions between two genomes before the onset of recombination (i.e., Ks of non-recombined genes – Ks of recombined genes; or 0.0898 - 0.0015 = 0.0883) multiplied by the total length of the recombined core genes (0.20 Mb for OS195 vs. OS185). Accordingly, the recombination (r) to mutation (m) ratio was ~3.4:1 for OS195 and OS185, indicating sexual evolution [14]. In contrast, and using the same methods and standards, the recombination to mutation ratio for the OS195 vs. OS155 and OS195 vs. OS223 pairs was 1:5 and 3:5, suggesting clonal divergence for these genome pairs.

2.4.4 Are The Exchanged Genes Neutral or Ecologically Important?

DNA-DNA microarray experiments using a *S. baltica* pangenome oligoarray revealed that all OS195-like (n=10) and OS185-like (n=3) strains in our collection examined had consistently greater hybridization signal for probes that corresponded to recombined vs. non-recombined core genes (Fig. 2.12B). In addition, half of these strains, including OS195 and OS185, were isolated from the Gotland Deep sampling station in 1986 and the remaining half in 1987, while the *S. baltica* population was estimated to about 1000 cells per ml of seawater in both sampling years based on most probable number (MPN) estimates using with several liquid media [112]. Therefore, the genetic exchange patterns revealed by the sequenced genomes apply to a large collection of strains and were persistent over a time (1986-1987) in the natural *S. baltica* population.

Our data collectively reveal that the OS195 and OS185 lineages have exchanged recently more than 20% of their genome (core plus variable genes). The factors that have fostered the recent and extensive genetic exchange between OS195 and OS185 lineages are not fully understood but several lines of evidence seem to indicate that at least some of the exchanged genes are ecologically important as opposed to neutral. For instance, the strains of the OS185 lineage and particularly those of the OS195 lineage were isolated from depths (Fig. 2.12A) that were characterized by oxygen depletion and presence of alternative electron acceptors such as nitrate, manganese oxides and sulfur compounds (Fig. 2.4A). To take advantage of the available electron acceptors, the strains possessed in their genomic islands several complete operons that encoded for anaerobic respiratory complexes and associated transport and cytochrome proteins (Fig. 2.12C and 2.13). In fact, the genes shared only by OS195 and OS185 represented either prophage-related (i.e., ephemeral) or genes related, almost exclusively, to anaerobic metabolism and transport (Fig. 2.6, 7th circle). It also appeared that the isolated OS195 strain, which apparently had migrated (sink?) in deeper waters after the recombination event(s) between the OS195 and OS185 lineages, had presumably adapted further to the more anoxic environment of the deeper waters. For instance, its genome encoded additional genomic islands for anaerobic lifestyle, such as a dimethyl sulfoxide reductase (DMSO) containing island (Fig. 2.12C), and OS195-like strains were more abundant and consistently recovered from this depth in both sampling years (Fig. 2.12A).

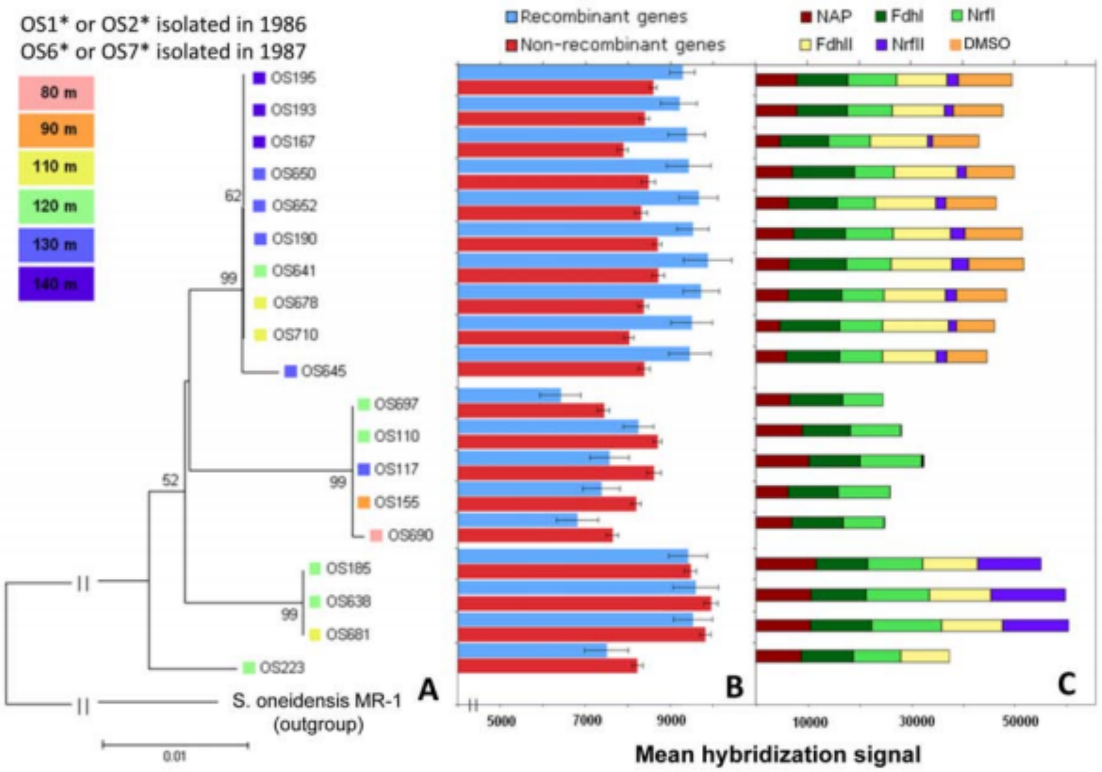


Figure 2.12 The patterns of genetic exchange apply to a large collection of *S.baltica* strains. All strains in the same lineages as the four sequenced strains (Panel A) were hybridized against a pangenome oligonucleotide microarray. The average raw signal of all probes that corresponded to non-recombined OS185 core (red) vs. recombined OS185 core genes with OS195 (blue) are shown (Panel B). Error bars represent one standard deviation from the mean. Note that the latter probes show consistently greater hybridization signal only in the OS195-like strains in agreement with the preferential genetic exchange between the OS185 and OS195 lineages. The hybridization signal of selected ecologically important genes or operons is also shown (Panel C; no signal denotes gene absence). The low signal for the nrf II operon in theOS195lineageis due to a few mismatches between the corresponding probes and the OS195 gene sequences. All operons and their genes are described in detail in TableS2.

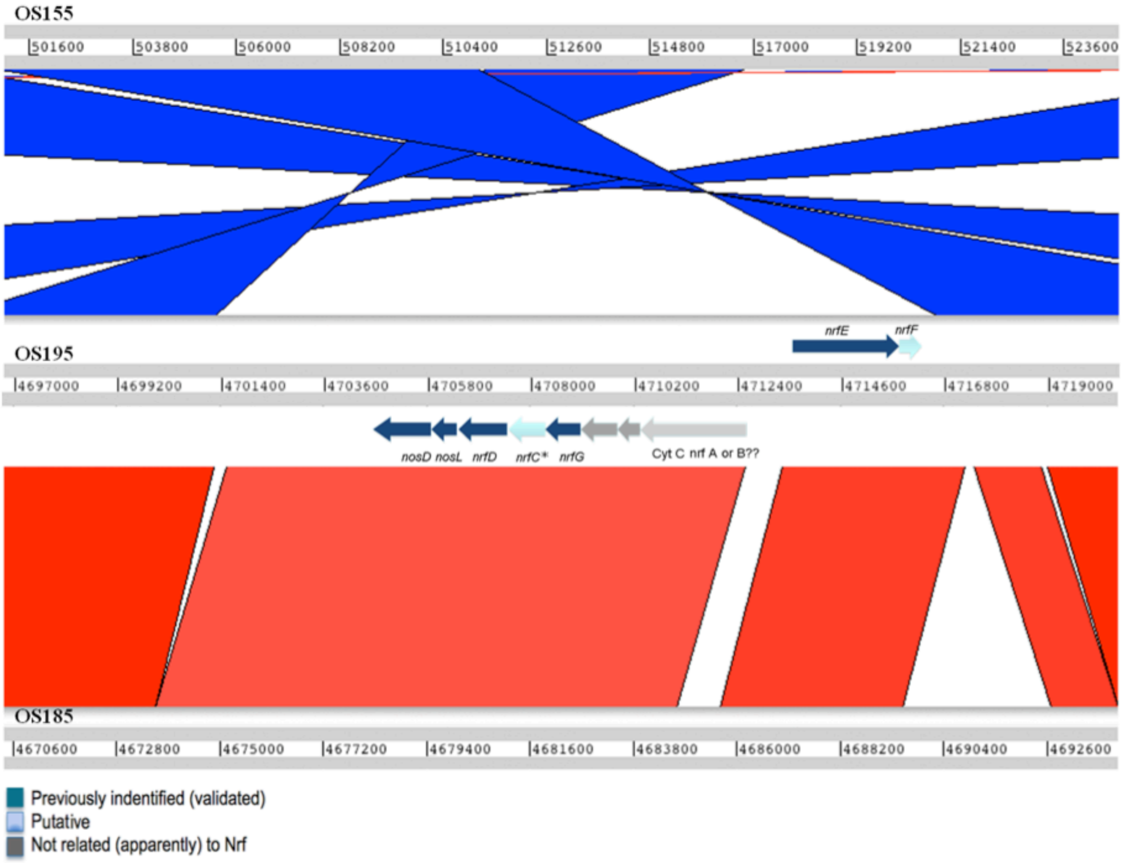


Figure 2.13 An example of an ecologically important genomic island shared between *S. baltica* OS195 and OS185. Graph shows the conservation of an OS195 genomic island (middle) in OS185 (red, bottom) and OS155 (blue, top) genomes using the ACT module of the Artemis package [137]. The island is present in OS185 but clearly absent in OS155. The island encodes, among other metabolic genes, a complete operon that is most similar (30-50% a.a. identity) to previously characterized *nrf* operons [138], which encode for the dissimilatory nitrate reduction to ammonia complex. The genes encoded in the operon of OS195 (or OS185) are also shown and are color-coded according to their role in the complex, which was inferred based on best blast-match searches against the functionally characterized *nrf* operons in GenBank.

While the substrates of the anaerobic genes shared between OS185 and OS195 remain speculative, laboratory microarray analysis revealed that some of these genes were expressed in OS185 and OS195 strains in response to anaerobic growth with nitrate or thiosulfate, indicating that they may be functional. The level of induction of the anaerobic metabolism genes examined typically varied between OS185 and OS195. For instance, the nrf operon, which was shared exclusively between OS185 and OS195 (Fig. 2.13) and encodes for genes putatively involved in the dissimilatory nitrate reduction to ammonia [138], was significantly induced by thiosulfate in both strains but by nitrate only in OS195 (Fig. 2.3). These variations in the level of induction may be due to the artificial batch conditions used in the laboratory compared to the *in-situ* conditions, the experimental noise of the microarray measurements, and/or the varied degrees of ecological/genomic adaptations, which may have altered metabolic and regulatory networks between the two strains.

Consistent with their ecological role, bioinformatics sequence (Table A1) and DNA-DNA microarray (Fig. 2.12C) comparisons suggested that most of the anaerobic metabolism genes shared between OS195 and OS185 were absent from strains of the OS155 lineage, which originated from (more) oxic waters (90-120m vs. 120-140m for strains of the OS195 lineage). Additionally, competition growth experiments suggested that OS155 was outcompeted by OS195 under anaerobic conditions, e.g., OS195 grew twice as rapid and typically to a double as high optical density compared to OS155 in the same anaerobic medium (ZoBell agar) or with thiosulfate as electron acceptor. Some of the potentially ecologically important genes shared between OS195 and OS185 (but not

OS155), but not all (e.g., thiosulfate/nitrate respiration; see Fig. 2.12C), were also present in OS223 (isolated from 120m depth), while the number of genetic exchange events between OS195 and OS223 was higher compared to OS195 and OS155 (48 vs. 28, respectively) but not as high as between OS195 and OS185 (308 events). These findings might indicate that although OS223 was isolated from the same depth as OS185 it might had occupied a slightly different ecological niche in the water column relative to OS185 or OS195, e.g., being associated with sinking particles as opposed to being planktonic (or vise versa) or being transient or allochthonous at the 120-140m depth (see also discussion below). In agreement with the latter hypothesis, only one other OS223-like strain was recovered in our 1986 or 1987 isolation efforts.

Regardless of what the exact ecological niche of the strains or the environmental stimuli that the genes respond to may be, our findings collectively indicate that more anaerobic metabolism genes had been exchanged between strains from more similar (deeper) waters and these genes were apparently important for the successful adaptation of the strains in the deeper, more anoxic, waters. They also reveal that genomic adaptation of the *S. batlica* strains to their immediate environmental conditions, mediated by HGT, may be very fast and lead to sexual divergence (speciation).

2.5 Discussion

To the best of our knowledge, such rapid, extensive and genome-wide adaptation in immediate response to environmental settings, mediated by directed (as opposed to promiscuous) genetic exchange, as the one seen in the OS195 and OS185 or OS223 genomes, has never been observed previously (e.g., Fig. 2.2A). Thus, our findings advance understanding of the speed and mode of bacterial adaptation and underscore the important relationships between ecological setting, biotic interactions, and genetic mechanisms that together shape and sustain microbial population structure. Extensive genetic exchange between co-occurring strains has been previously implied by metagenomic studies of natural populations [92, 139], but the fragmented nature of these datasets did not allow robust estimations of the magnitude of the genetic exchange at the whole-genome level or assessment of its ecological consequences [92, 140]. Recent studies of isolated strains have also reported elevated levels of genetic exchange between pathogenic bacteria such as between distinct *Campylobacter* species [9] or within *Vibrio cholerae* [141]. However, the genes exchanged in these cases are typically limited to a few environmentally selected functions and show strong biases in terms of spatial location in the genome [120]. Accordingly and in contrast with *S. baltica*, genetic exchange is unlikely to lead to sexual speciation and population cohesion in such cases.

The *S. baltica* genomes reveal that genetic exchange, mediated by homologous recombination, could constitute an important mechanism for population cohesion among spatially co-occurring prokaryotes, similar to the role of sexual reproduction in higher eukaryotes. Therefore, our results provide the experimental evidence in support of recent computer simulation studies that suggested that recombination-driven sexual speciation is possible in bacteria [14]. Despite the extensive recombination observed, the *S. baltica* genomes show no evidence in support of the recently proposed fragmented speciation model for bacteria [98]. For instance, the predicted signature of this model, i.e., ecological genomic islands are surrounded by increased levels of nucleotide divergence between ecologically distinct (e.g., OS195 vs. OS155) but not between ecologically coherent (e.g., OS195 vs. OS185) populations, was not observed (Fig. 2.14). The signature was also not observed in comparisons between selected *S. baltica* strains and other closely related (i.e., sharing 80% to 88% ANI to *S. baltica*) but ecologically distinct sequenced *Shewanella* genomes of *Shewanella* sp. MR-4 and MR-7 from the Black Sea, *Shewanella* sp. ANA-3 and *Shewanella oneidensis* MR-1 from freshwater ecosystems in the USA [118]. These results may be due to the fact that the recombined fragments are too small (Fig. 2.8) for recombination to be affected (reduced) by the presence of genomic islands (which would act as barriers to recombination because the sequence is not conserved) among ecologically distinct organisms. Alternatively, the genetic exchange between the incipient ecological distinct species may not be maintained for long enough evolutionary time as previously hypothesized [98] for recombination to create the signature of the model in the *S. baltica* case.

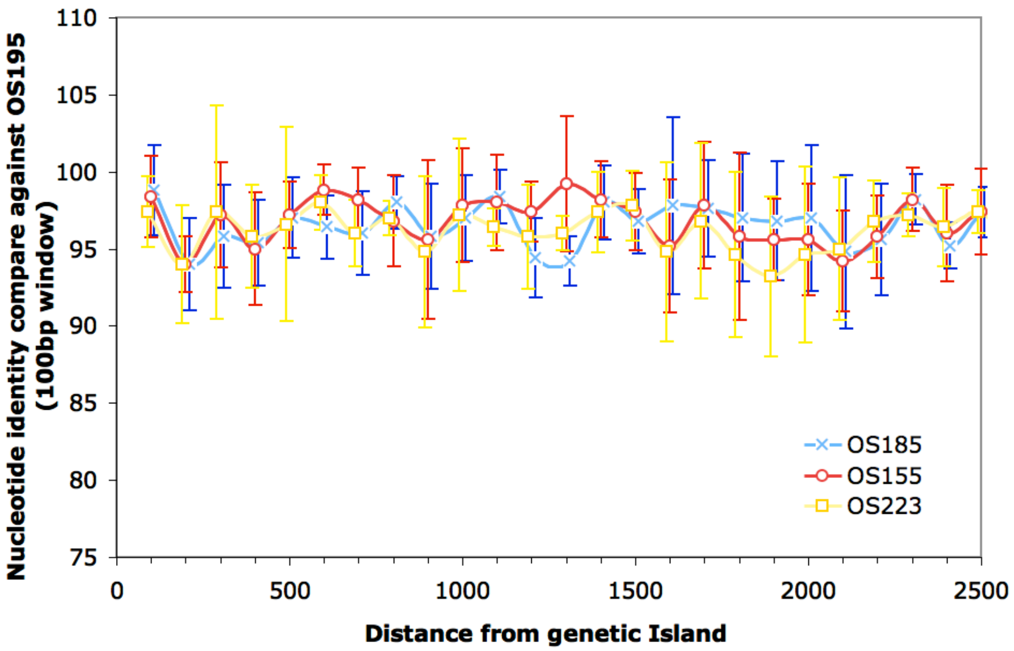


Figure 2.14 Spatial analysis of the nucleotide diversity of the regions surrounding ecological islands. The nucleotide identity of the regions that flank potentially important ecological islands shared only by OS195 and OS185 genomes (y-axis) is plotted against the distance of the region from the ecological island based on the OS195 genome. Five islands were considered in total; errors bars represent one standard deviation from the mean based on the five islands (10 observations were used in total, i.e., one upstream and one downstream for each island). Note that the islands are flanked by similar levels of nucleotide identity in ecologically overlapping (e.g., OS195 vs. OS185) vs. non-overlapping (OS195 vs. OS155 or OS233) genome pairs. A similar pattern was observed in comparisons between selected *S. baltica* strains and other closely related (i.e., sharing 80% to 88% ANI to *S. baltica*) but ecologically distinct sequenced *Shewanella* genomes, such as the *Shewanella* sp. MR-4 and MR-7 from the Black Sea and the *Shewanella* sp.

ANA-3 and *Shewanella oneidensis* MR-1 from freshwater ecosystems in the USA (data not shown).

To what extent the patterns of genetic exchange observed between OS195 and OS185 (Fig. 2.2) and their sister strains (Fig. 2.12) apply to other natural sub-populations of *S. baltica* in the Baltic Sea and what accounts for the reduced genetic flow between OS185 and OS223 (same isolation depth) compared to OS195 (different depth), remain currently unknown. To address these issues, *in-situ* genomic studies (e.g., metagenomics) and sampling of the natural populations over time will be required. However, the OS195 and OS185 example does raise the possibility that bacterial adaptation through genetic exchange may be much more rapid and extensive than previously anticipated and thus, it has broader implications for understanding bacterial evolution and adaptation. Our independent analyses have also ruled out the possibility that the results reported here for OS195 and OS185 are attributable to manmade mixing of the genomic DNA submitted to sequencing or the derived sequences. For instance, if the results were attributable to DNA mixing, we would not have observed a significantly greater hybridization signal with the recombined vs. the non-recombined genes during DNA-DNA microarray experiments (Fig. 2.12). It also appeared that the genomes of OS155 and OS223 had numerous and extensive genomic rearrangements (transposition and inversions) compared to those of OS195 and OS185, while OS185 and OS195 genomes were syntenic in almost their entire length (Fig. 2.6, outer cycles). Whether or not these rearrangements, which could act as barriers to recombination because the sequence is not syntenic, are responsible for the reduced genetic flow between OS223 or OS155 and OS195 relative to OS185 and

OS195 is not clear, but does represent an intriguing hypothesis that warrants further investigations.

In summary, it appears as if the genome of *S. baltica* adapts through continuous internal genome-wide genetic exchange and rearrangement events (Fig. 2.6), in a highly dynamic (electron donors as well as electron acceptors), nutrient rich pelagic environment. This differs fundamentally from what was observed previously in other important marine bacteria such as the *Pelagibacter ubique* [142] and *Prochlorococcus marinus* [143], which have streamlined genomes, developed over eons in rather constant, nutrient poor environments. The latter organisms represent the ultimate marine k-strategist whereas *S. baltica* is very close to the ultimate r-strategist. The patterns observed in *S. baltica* may be broadly applicable to other bacteria that experience frequent environmental fluctuations in the marine environment and elsewhere. Therefore, our findings expand understanding of the rate and mode of bacterial adaptation and underscore the important relationships between ecological setting, biotic interactions, and genetic mechanisms that together shape and sustain microbial population structure.

2.6 Acknowledgments

We thank Profs. James Tiedje and Frank Loeffler for helpful suggestions regarding the manuscript and the Shewanella Federation for supporting work on Shewanella genomics. Contributions of the Joint Genome Institute for the genome sequences used in this study are also acknowledged. Our work is supported by the Department of Energy (contract number DE-FG02-07ER64389).

CHAPTER 3

GENOME SEQUENCING OF FIVE SHEWANELLA BALTICA STRAINS RECOVERED FROM THE OXIC-ANOXIC INTERFACE OF THE BALTIC SEA

Reproduced in part with permission from Alejandro Caro-Quintero, Jennifer Auchtung, Jie Deng, Ingrid Brettar, Manfred Höfle, James M. Tiedje, and Konstantinos T. Konstantinidis. *Journal of bacteriology*. 2012,194(5), 1236-1236.

Copyright 2012, American Society for Microbiology.

3.1 Abstract

Shewanella baltica represents one of the most abundant heterotrophic nitrate-respiring species among those that can be cultivated from the oxic-anoxic interface of the Baltic Sea. We recently described the complete genome sequences of four *S. baltica* strains recovered from the Gotland Deep sampling station in 1986 and 1987 (Caro-Quintero et al., The ISME Journal, 2011). These genomes showed unprecedented high levels intra-species horizontal gene transfer (HGT), driven presumably by adaptation to rapidly changing conditions as the strains migrate seasonally across the water column. Interestingly, two of the strains that were isolated from similar depths were found to evolve sexual. Here we describe the genome sequences of five additional *S. baltica* strains recovered from the same samples (strains OS117, OS183, OS625, and OS678) as well as one recover 10 years later from the same sampling station (strain BA175). These new genomes confirmed and further expanded on our previous observations that *S. baltica* represents a versatile group of fast adapting organisms and that HGT plays a major role during the adaptation process. Collectively, the *S. baltica* genomes represent a valuable resource for assessing the role of environmental settings and fluctuations on genome evolution and adaptation.

3.2 Introduction

The genus *Shewanella baltica* is an important common inhabitant of the stratified water column of the Baltic Sea, playing an important role in cycling of organic matter at low oxic/anoxic water of the central Baltic Sea [144]. Interestingly, *S. baltica* strains ability to use different electron acceptor makes them of great value for potential bioremediation of heavy metals and radioactive waste. Distribution and ecology of *S. baltica* strains is affected by availability of electron acceptors at different depths in the stable stratified water column of the Baltic Sea, such ecological preferences have important implications on genomic adaptations and amount of genetic exchange. Analyses of the first four sequenced genomes of *S. baltica* strains, OS155 (80 m depth), OS195 (140 m depth), OS185 (120 m depth) and OS223 (120 m depth) revealed that strains adapted to more anaerobic environments (OS185, OS195 and OS223) had exchanged genes more frequently than strains from different depth, as evidenced by the patterns of gene sharing and the unprecedented levels of recent homologous recombination [100].

Here we present the 5 additional genomes of *S. baltica* strains OS117 (130 m depth), OS183 (120 m depth), OS625 (80 m depth), OS678 (110 m depth) and BA175 (120 m depth) to expand our understanding of the relative importance of phylogeny and ecology in gene content, genetic exchange and homologous recombination. Selection of these strains was based upon the observations from the four previously sequenced strains and the phylotypes revealed trough MLST (Multi Locus Sequence Typing) and RAPD (Random Amplification of Polymorphic DNA) profiling [145].

3.3 Methods

3.3.1 Nucleotide Sequences Accession Numbers

The following genome sequences were deposited in GenBank: OS183 (NZ_AECY00000000, high-draft status), OS117 (CP002811.1, chromosome; CP002812.1, CP002813.1, and CP002814.1, plasmids), BA175 (CP002767.1, chromosome; CP002768.1 and CP002769.1, plasmids), OS678 (CP002383.1, chromosome; CP002384.1, plasmid), and OS625 (AGEX00000000).

3.3.2 Homologous Recombination Detection

Recombination among the genomes was detected as previously described [100]. Briefly, the sequence of a reference genomes was cut *in-silico* in 500 bp-long consecutive sequence fragments. The fragments were subsequently searched against the other *S. baltica* genomes for best matches, using blastn as described above for orthologs. A fragment was flagged as (potentially) recombined in another strain when its best blastn match in the latter strain showed more than 99.5% nucleotide identity while its identity in the other strains was lower <98%, which corresponded to the typical genetic distance between the *S. baltica* strains (i.e., ~96.7%).

3.4 Results and Discussion

3.4.1 *Shewanella baltica* Strains OS183 and BA175

Sequencing of strains OS183 and BA175 genomes provide a unique opportunity to assess allelic variation, population adaptation and gene conservancy in short periods of time. In brief, both strains belong to the same MLST clade and were isolated from similar depth, but with a 12 years period difference, OS183 was isolated in 1986 and BA175 was isolated in 1998. To address the genomic adaptation, a comparative genomic analysis was done to identify strain specific genes and to quantify allele variation and Single Nucleotide Polymorphisms (SNPs). The analysis revealed that even though the strains BA175 and OS183 are almost identical (99.9 % Average Nucleotide Identity) gene content differences exist between the two strains. In brief, 89 genes were specifically found in BA175, while 114 were found in OS183, most of these genes were hypothetical or mobile elements (data not shown). Interestingly, block of strain-specific genes (16 in BA175 and 10 genes in OS183) were found in the same syntenic location of both genomes. These blocks encode for similar functions, capsular polysaccharide polymerization similar to O-antigen production on enterobacteria; however the genes within the blocks were very divergent from each other to be called orthologs. Further analysis of the translated genes revealed the existence of closer homologs in species of other genus (e.i. *Vibrio cholerae*, *Prosthecochloris aestuarii* DSM 271), which suggests acquisition trough HGT (Table 3.1). Similar cases of acquisition of capsular variants (dTDP-L rhamnose pathway) have been previously described in *Vibrio cholerae,* also a marine organism [146, 147]. Interestingly, a recent assets on O-antigen related genes in several *Vibrio cholerae* serogroups [146] revealed that genetic exchange between *Vibrio sp*. and *Shewanella sp.* may be common and that HGT between the two species has important environmental (i.e., resistance to phage infection) and clinical implication (i.e., emergence of new pandemic serogroups).

Table 3.1 Antigen-O related protein hits in *S. baltica* BA175.

Accession number	**Annotation in *S. baltica* BA175**	**Identity**	**Annotation**
AEG10742.1	dTDP-glucose 4,6-dehydratase	86%	*Vibrio cholerae*
AEG10743.1	glucose-1-phosphate thymidylyltransferase	92%	*Vibrio cholerae*
AEG10744.1	dTDP-4-dehydrorhamnose 3,5-epimerase	83%	*Vibrio cholerae*
AEG10745.1	NAD-dependent epimerase/dehydratase	53%	*Shewanella baltica* OS185
AEG10746.1	hexapeptide repeat-containing transferase	49%	*Enterobacter sp*. 638
AEG10747.1	putative acetyltransferase	52%	*Aeromonas hydrophila*
AEG10748.1	glycosyl transferase family 2	46%	*Geobacter uraniireducens*
AEG10749.1	hypothetical protein Sbal175_1473	30%	*Pseudoalteromonas sp.* SM9913
AEG10750.1	glycosyl transferase group 1	<30%	-
AEG10751.1	hypothetical protein Sbal175_1475	<30%	-
AEG10752.1	glycosyl transferase group 1	45%	*Hippea maritima* DSM 10411
AEG10753.1	GHMP kinase	54%	*Photorhabdus luminescens* subsp. laumondii TTO1
AEG10754.1	Phosphoheptose isomerase	66%	*Prosthecochloris aestuarii* DSM 271
AEG10755.1	Nucleotidyl transferase	45%	*Prosthecochloris aestuarii* DSM 271
AEG10756.1	D,D-heptose 1,7-bisphosphate phosphatase	55%	*Aneurinibacillus thermoaerophilus*
AEG10757.1	undecaprenyl-phosphate alpha-N-acetylglucosaminyl 1-phosphatetransferase	85%	*Shewanella* sp. MR-4
AEG10758.1	phosphoglucosamine mutase	97%	*Shewanella baltica* OS117

Analysis of polymorphic sites detected a total of 3,985 SNPs between the strains. Interestingly, 93 % of the SNPs (3,697) were found within 6 syntenic regions and not randomly distributed, as expected by mutation. These syntenic SNPs patterns are more likely the result of the incorporation of divergent foreign DNA through homologous recombination than the result of random mutation, as suggested by spatial distribution of Ks values similar to what has been previously described for *Streptococcus pneumoniae* [36]. Quantification of Ka/Ks ratio on the recombined segments revealed several genes under positive selection, suggesting a plausible adaptive roll of the genes (Fig 3.1),.

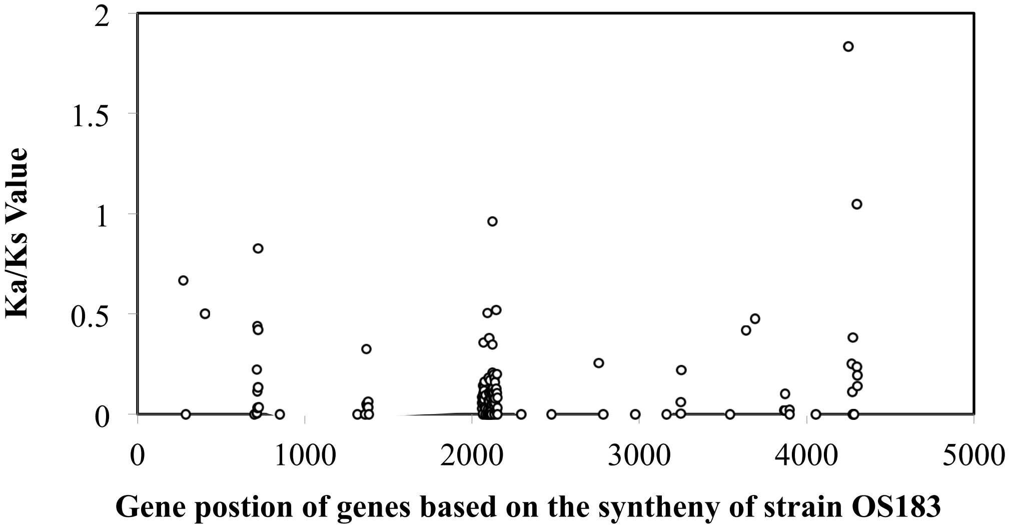


Figure 3.1 Ka/Ks between *S.baltica* strains OS183 and BA175. Substitutions of synonymous and non-synonymous Substitutions are mainly cluster in 6 syntenic regions, suggesting homologous recombination mediated allele substitution. The Ka/Ks values were calculated for all orthologs genes shared between the strains, using the gene nucleotide codon-based alignment and the codeml module of the PAML package [122].

Despite the lack of evidence that BA175 is the direct descendant of OS183, the number of generations between the strains can be used to measure the relative divergence time of the strains. In brief the rate of synonymous substitutions from "non-recombined regions" (Ks= 3.32 × 10 $^{-5}$) is divided by mutation rate for double stranded DNA [134]. A total of 6.02×10^{6} generations between the strains was quantified, the generation per day and per hour (assuming 12 year period of separation) are 14 and 1.7 respectively. These values agree with the doubling time of *Shewanella baltica* under laboratory conditions of 2.14 generation/hr [148]. Nevertheless, it is important to mention that these values do not necessarily reflect the growth rate in the Baltic Sea because of seasonal variation and the fact that *Shewanella baltica* are known to growth in pulses of feast and famine [149], instead of a continuous growth.

3.4.2 *Shewanella baltica* Strains OS625 and OS117

Strains OS625 and OS117 were sequenced to assess the relative roll of phylogenetic affiliation and ecological affiliation in gene content. In brief, strain OS625 belongs to OS195 MLST clade but it was isolated from a more oxic redox zone (80 m). Similarly, strain OS117 belongs to the OS155 MLST clade but was isolated from a more anoxic redox zone (120 m). Comparative genomic analysis was performed to identify specific genes within (i) similar phylogenetic clade but different redox zones and (ii) similar redox zones but different phylogenetic clade.

Strains OS625 and OS195 belong to the same clade, but are not clonal as evidenced by the phylogenetic network analysis (Fig. 3.2, A) and the ANI analysis (99.3%). Comparative genomic analysis between OS625 and OS195 reveal a set of 489 clade specific genes. Similar analysis between OS625 and OS155 (similar redox zone different clade) identified set of 31 share genes, mainly hypothetical proteins and mobile elements. In the other hand, strains OS117 and OS155 (ANI= 99.7%) shared a set of 510 clade specific genes, while OS117 and OS195 (similar redox zone different clade) shared 81 genes. From these 81 genes, 49 are present in all *S. baltica* genomes but OS155, suggesting a deletion in the last (OS155) instead of an ecological relevant island shared between OS117 and OS195. The rest 32 genes are mostly related to hypothetical proteins and mobile elements. In conclusion, our comparative analysis of OS117 and OS625 was dominated by the effect of phylogenetic affiliation of strains (Fig 3.2, A) and did not identify a consistent gene sharing pattern that could suggest adaptation of OS117 or OS625 to a different redox zone. This reveals the biases of dynamic and interconnected environments as the water columns of the Baltic Sea, where upwelling and sinking can bring transient populations not necessarily adapted to the conditions at the depth of isolation. These findings highlight the importance of in depth population genomics or metagenomics to identify dominant vs. transient individuals, which seems a fundamental step to untangled the roll of ecology and adaptation.

3.4.3 *Shewanella baltica* Strain OS678

Finally, strain OS678 isolated from the microaerophilic redox zone, belongs to a MLST clade dominated by strains isolated from the anoxic redox zone (OS195), Interestingly, the patterns of high homologous recombination previously reported between OS185 and OS195 [100], are also observed between OS185 and OS678, supporting the idea that genetic exchange happened between the ancestors of the clades.

Additionally, evidence of extra homologous recombination events in OS678 suggests an ongoing process that might be quantifiable on short periods of time.

3.4.4 The Ecological Pattern of Recombination in *S. baltica*

The new analysis of all sequenced strains revealed that the high inter-clade recombination and gene sharing is not only exclusive of OS195-OS185-OS223, but also observed between the OS155 and OS183 clades (Fig. 3.2, B). Using a similar approach as previously described, 160 core genes were identified as recombined between the OS155 and OS183 clades. Similar to the ecologically relevant anaerobic genes in the OS195-OS185 pair, a set of genes, mostly flagellar genes, were identified between OS155-OS183 clades (Fig 3.2, B). These clades are more abundant just above the chemocline, where motility could be important for maintaining an optimal location in the redox gradient or to reduce the chance predation [150].

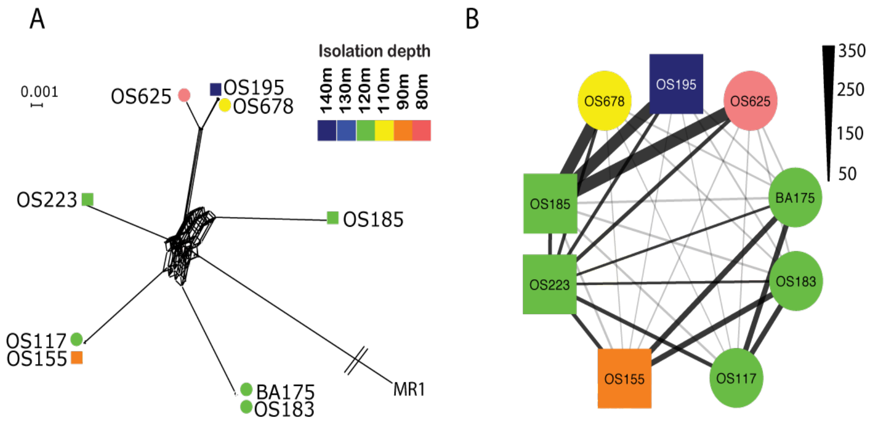


Figure 3.2 Phylogenetic network genomes and homologous recombination events of *S. baltica* sequenced strains. Squares represent previously described genomes, while circles represent the sequenced the recently sequenced genomes. The phylogenetic network of the sequenced *Shewanella baltica* genomes was constructed by using the concatenated alignment of 3,338 shared orthologous genes in SplitsTree 4 [151] (Panel A). The homologous recombination network was constructed using Cytoscape 2.8.1[152]. The network represents the abundance of homologous recombination events between clades, the thicker the line the higher the number of homologous recombined genes detected (Panel B).

3.5 Conclusions

The recently sequenced genomes not only corroborated, but also uncovered new patterns of homologous recombination correlated with ecological constraints. Our findings indicate that HR is more pervasive between ecologically more related populations (e.g. anaerobic adapted or motility adapted), and that HGT is an essential mechanism for the fast adaptability, diversification and genetic versatility observed between *S. baltica* strains.

3.6 Acknowledgments.

We thank the Joint Genome Institute for sequencing and annotating all *S. baltica* genomes. This work was supported in part by the US Department of Energy under Contract No. DE-FG02-07ER64389.

CHAPTER 4

GENOMIC INSIGHTS INTO THE CONVERGENCE AND PATHOGENICITY FACTORS OF *CAMPYLOBACTER JEJUNI* AND *CAMPYLOBACTER COLI* SPECIES

Reproduced in part with permission from Alejandro Caro-Quintero, Gina P. Rodriguez-Castaño, and Konstantinos T. Konstantinidis. *Journal of bacteriology*. 2009,191(18), 5824-5831.

Copyright 2009, American Society for Microbiology.

4.1 Abstract

Whether or not bacteria can cohere together via means of genetic exchange and hence, form distinct species boundaries remains an unsettled issue. A recent report has implied that not only the former may be true but, in fact, the clearly distinct *Campylobacter jejuni* and *Campylobacter coli* species may be converging as a consequence of increased inter-species gene flow, fostered, presumably, by the recent invasion of the same ecological niche (Sheppard et al., Science 2008). We have re-analyzed the *Campylobacter* Multi Locus Sequence Typing (MLST) database used in the previous study and found that the number of inter-species gene transfer events may actually be too infrequent to account for species convergence. For instance, only 1-2% of the 4,507 *Campylobacter* isolates examined appeared to have imported gene alleles from another *Campylobacter* species. Furthermore, by analyzing the available *Campylobacter* genomic sequences, we show that although there seems to be a slightly higher number of exchanged genes between *C. jejuni* and *C. coli* relative to other comparable species (~10% vs. 2-3% of the total genes in the genome, respectively), the function and spatial distribution in the genome of the exchanged genes is far from random, and hence,

inconsistent with the species convergence hypothesis. In fact, the exchanged genes appear to be limited to a few environmentally selected cellular functions. Accordingly, these genes may represent important pathogenic determinants of *Campylobacter* pathogens and convergence of (any) two bacterial species remains to be seen.

4.2 Introduction

High-throughput sequencing studies during the last decade have revealed that bacterial genomes are much more diverse and "fluid" than previously anticipated [102, 107]. This genomic fluidity is primarily attributable to the great pervasiveness and promiscuity of horizontal gene transfer (HGT) in the bacterial world [8, 153]. Nonetheless, evidence for any two distinct bacterial species or lineages merging due to directed (as opposed to promiscuous) inter-species genetic exchange was observed, probably for the first time ever, by the recent study of Sheppard et al [9]. Species convergence, if occurring, has major theoretical implications for the bacterial species concept [reviewed extensively elsewhere [14, 107, 108, 154, 155]] and important practical consequences for accurate identification of bacterial pathogens in the clinic.

Sheppard and colleagues reported that as many as ~18.6% of the unique alleles of housekeeping genes found in *Campylobacter coli* isolates may have been recently imported (through HGT) from a close relative, *Campylobacter jejuni* [9]. The results were based on the analysis of 4507 *Campylobacter* isolates, which have been genotyped at seven genes (loci), available though the *Campylobacter* Multi Locus Sequence Typing

(MLST) database [156]. In brief, the 4507 genotyped isolates contained a total of 2917 unique sequence types (ST). A unique ST represents the concatenated sequence of the seven genes present in the genome of an isolate and contains a unique sequence (allele) for at least one of the seven genes when compared against any other unique ST in the database (different isolates may be characterized by the same ST). The unique STs were assigned to either *C. coli* or *C. jejuni* species using the program STRUCTURE [51]. Neighbor-joining phylogenetic trees of all available unique alleles for each individual gene were subsequently built. Instances where the ST assignment to a species differed from the assignment of an individual gene sequence that constituted the ST were attributed to inter-species transfer of the gene and the number of such instances was reported [9].

Here, we have reevaluated the available *Campylobacter* MLST dataset and show that the predominant STs, i.e., the STs characterizing >98% of the isolates, do not contain imported alleles and hence, do not support the species convergence hypothesis. In agreement with these findings, analyses of the available *Campylobacter* genomic sequences indicated that the inter-species genetic exchange is limited and heavily biased towards a few genes under positive selection. In fact, housekeeping genes (such as those used in MLST) were found to be exchanged between the two species only in (rare) hitchhiking events associated with the horizontal transfer of adaptive genes. Accordingly, a clear species boundary between the *C. jejuni* and *C. coli* species is evident and it is unlikely that this boundary is being eroded, which contrasts with what was hypothesized previously [9].

4.3 Material and Methods

The gene sequences of all isolates analyzed in this study were obtained from the Campylobacter MLST database [156], available through http://pubmlst.org/campylobacter/. The sequence dataset used was identical to that used by Sheppard and colleagues [9]. Assignment of STs to species and identification of imported genes based on neighbor joining phylogenetic trees were performed as described previously [9]. To further validate these tree-based results, a simple blast-based strategy for detecting genes exchanged between *Campylobacter* isolates was also employed. In brief, a gene in a *C. coli* isolate was flagged as (potentially) imported from *C. jejuni* when it showed >95% nucleotide identity to a gene in at least one *C. jejuni* isolate and the average nucleotide identity of the concatenated sequences (i.e., the STs) of the two corresponding isolates was lower than 90%, which corresponded to the typical genetic distance between *C. jejuni* and *C. coli* species (i.e., 86% nucleotide identity). The blast-based method provided very similar results to those obtained with the method employed by Sheppard et al. [9]. The congruence in the results obtained is primarily due to the significantly larger inter-species genetic distance relatively to the intra-species distance (Fig. 4.1), which greatly facilitated the accurate identification of potentially transferred genes, independently of the method employed. Accordingly, ST assignment to species based on STRUCTURE corresponded perfectly to the 90% nucleotide cut-off used in the blast-based method. A few intermediate isolates showing 90-95% nucleotide identity to other isolates (Fig. 4.1) corresponded mainly to the unassigned STs in the previous study [9], and were excluded from counting isolates with imported genes. In the remaining text, the results based on the nucleotide identity (blast-based approach) are preferentially reported because nucleotide identity is a much simpler and more intuitive concept than the concepts associated with phylogenetic trees.

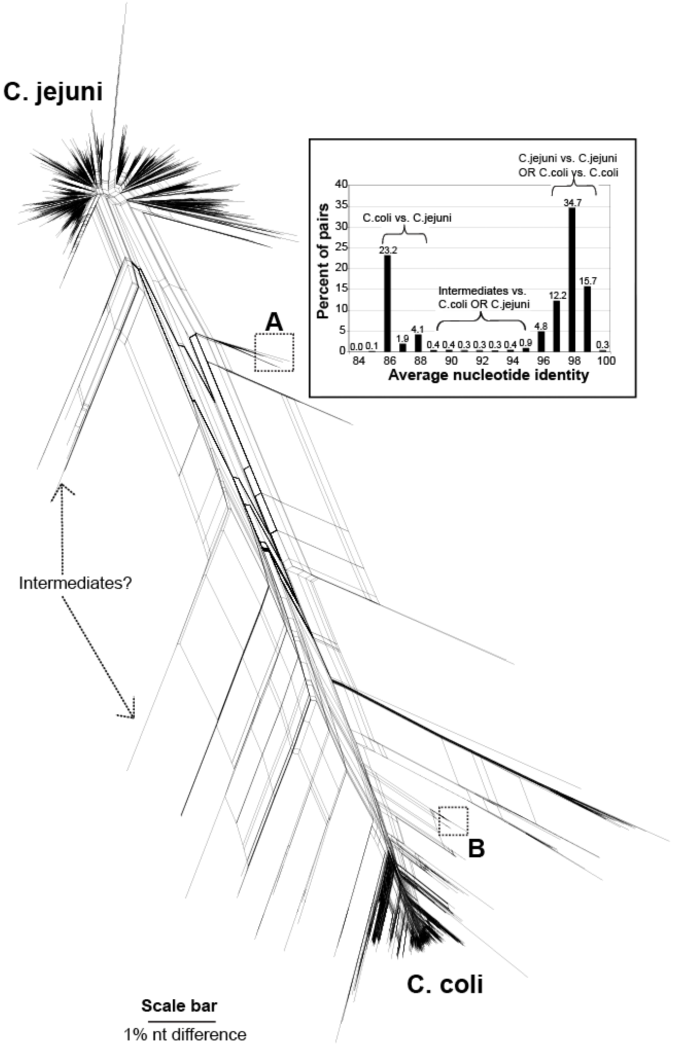


Figure 4.1 Genetic relatedness among the 3693 *C. jejuni* and 814 *C. coli* isolates analyzed in this study. Figure shows the phylogenetic network among all 4507 isolates as calculated by the SplitsTree4 program [151], using default settings and the ST for each isolate as input to the program. Isolates' IDs were omitted for clarity purposes. Horizontal lines between any two branches indicate complex underlying evolutionary scenarios such as the HGT event of one (or more) of the individual genes, as explained

previously [151]. Inset shows the average blast-derived nucleotide identities between all 4507 X 4507 STs. Boxes A and B denote the tight sub-clades with imported *uncA* and *aspA* alleles, respectively (discussed in the text).

The calculation of non-synonymous vs. synonymous amino acid substitution ratio (Dn/Ds) was performed as follows: *C. jejuni* and *C. coli* orthologous protein sequences, when longer than 100 amino acids long, were aligned using the Clustalw algorithm [157]. The corresponding nucleotide sequences of the aligned protein sequences were subsequently aligned, codon by codon, using the pal2nal script, with "remove mismatched codons" enabled, and the protein alignment as the guide [158]. The Dn/Ds ratio for each pair of proteins was calculated on the nucleotide codon-based alignments using the codeml module of the PAML package [122], using the whole sequence, or 30, 40 and 50 amino acid long sliding windows, as proposed previously [159]. Custom PERL scripts were used to automate the Dn/Ds analysis and parse the results of the codeml and blast algorithms. Protein sequences shorter than 100 amino acids long were excluded from the analysis to avoid short spurious open reading frames that may not represent genuine protein-coding regions of the genome [160, 161].

4.4 Results and Discussion

4.4.1 Isolates With Imported Genes Are Extremely Rare

In agreement with the previous study [9], our analysis revealed that 102 of the unique *C. coli* STs (from the 713 total, or 14%) contained alleles potentially imported from *C. jejuni*, and 103 unique *C. jejuni* STs (from the total 2204, or 4.7%) contained imported alleles from *C. coli*. Sheppard and colleagues performed, in addition, ClonalFrame analysis [42] to show that the majority of *C. coli* STs with imported alleles belonged to a sub-clade of the *C. coli* species, which had about 18% of its unique STs with imported alleles from *C. jejuni* [9]. However, when the analysis was performed at the isolate and the individual gene level, as opposed to the ST level, a quantitatively different picture was obtained. The isolates that contained imported alleles were very rare; typically, fewer than 10 isolates per species for each gene evaluated (from the 814 *C. coli* and 3693 *C. jejuni* isolates used in the study, in total; see Table 4.1). Further, these isolates rarely carried imported alleles for more than one of the seven MLST genes used (i.e., 9/102 *C. coli* and 4/103 *C. jejuni* isolates carried imported alleles for two genes; no isolate had three or more imported genes). Only for the *uncA* and *aspA* genes did we observe a substantially larger number of *C. jejuni* and *C. coli* isolates with imported alleles from the other species, 65 and 39, respectively. The great majority, i.e., 56/65 and 33/39, of these isolates, however, clustered together as tight sub-clades within the *C. jejuni* and *C. coli* species, respectively (boxes A & B in Fig. 4.1, respectively). The sub-clades were also evident when the *uncA* and *aspA* gene sequences were omitted from building the reference phylogeny (data not shown). Therefore, the previous imported *uncA* and *aspA* alleles represent, most likely, products of a single HTG event that occurred between the ancestors of specific sub-clades within *C. jejuni* and *C. coli* species and hence, the number of HGT events of the *uncA* and *aspA* genes appears similar to that of the remaining five genes (i.e., n = ~10). The high nucleotide identity (>99%) among the imported *aspA* or *uncA* alleles recovered within the sub-clades is also consistent with a single HGT event.

Table 4.1 *C. coli* and *C. jejuni* isolates with imported gene sequences. The number of isolates of each species (3,693 *C. jejuni*, 814 *C. coli*, in total) whose individual genes were assignable to *C. jejuni* (J) or *C. coli* (C) species based on the phylogenetic approach described previously [9] are shown. Numbers in parenthesis for *uncA* and *aspA* genes denote the number of corresponding isolates found to cluster together in two discernible tight sub-clades of the tree that represents the phylogeny of all isolates (denoted by A and B boxes in figure 4.1, respectively). The complete annotation of genes is as follows: *aspA* - aspartase A; *glnA* - glutamine synthetase; *gltA* - citrate synthase; *glyA* - serine hydroxymethyltransferase; *pgm* – phosphoglucomutase; *tkt* – transketolase; and *uncA* - ATP synthase alpha subunit.

	aspA		glnA		gltA		glyA		pgm		tkt		uncA	
Isolate	**J**	**C**	**J**	**C**	**J**	**C**	**J**	**C**	**J**	**C**	**J**	**C**	**J**	**C**
Jejuni	3683	10	3692	1	3689	4	3690	3	3689	4	3677	16	3628	65 (56)
Coli	39 (33)	775	6	808	10	804	10	804	15	799	18	796	4	810

Our analysis also revealed that the great majority of STs with imported alleles were encountered only in a single isolate. For instance, the 47 *C. jejuni* isolates with imported alleles for at least one gene of the seven MLST genes (excluding the 56 isolates with imported *uncA* alleles, represented by Box A in Fig. 4.1) contained a total of 44 unique STs, i.e., only two STs (ST #352 and #628) were encountered in more than one *C. jejuni* isolate (two and three isolates, respectively; see Table 4.2, which includes all STs with imported alleles and the underlying data for Table 4.1). These results contrasted with an average of ~1.7 isolates per unique ST (3693/2204) for the *C. jejuni* species. In other words, the most predominant STs, i.e., the ST types characterizing two or more isolates, do not typically contain imported alleles.

In summary, assessing HGT at the ST level clearly “inflated”, to a certain degree (see also *uncA* and *aspA* genes above), the extent of HGT between the *Campylobacter* species [9]. We detected a maximum of ~70 inter-species HGT events (assuming that most of the imported *uncA* and *aspA* alleles were exchanged in a single HGT event) in a total of 5698 *C. coli* genes evaluated (number of isolates multiplied by the number of genes available for each isolate), which translates to <2% of the total *C. coli* isolates had exchanged an allele with a *C. jejuni* partner for each gene evaluated (Table 4.1). These results reveal that HGT between the two species may be too infrequent to account, unequivocally, for active species merging.

Table 4.2 *C. coli* and *C. jejuni* sequence types (STs) with imported alleles. All STs with imported alleles for particular genes (column heading), except for those belonging to the clades denoted by the A and B boxes in figure 4.1, are shown. Numbers in parentheses denote the number of isolates characterized by the corresponding ST; absence of parentheses denotes that the ST was encountered only once among the total 4507 isolates evaluated. Blue denotes STs assigned to *C. coli* and yellow STs assigned to *C. jejuni* species.

Gene	aspA	glnA	gltA	glyA	pgm	tkt	uncA
ST	357	1011	1415	57	57	555	647
ST	2194	1087	1420	138	139	1010	648
ST	2501	2470	1553	1362	357	1574	2623
ST	2565	2471	1772	1934	437	1611	2762
ST	2774	2565	1934	2587 (4)	646	2241	352 (2)
ST	2784	2828	2050	2588 (2)	1529	2363	554
ST	348	2775	2051	310	1574	2499	625
ST	625		2055	2775	1623	2502 (2)	628 (3)
ST	714		2064	2799	1638	2503	928
ST	1366		2204		2129	2506	2891
ST	1686		145		2363	2526	
ST	1933		1682		2472	2527	
ST	2124		1722		2499	2528	
ST	2566		1869		2762	2533	
ST	2622				2908	2621	
ST	2775				802	2773	
ST					803	2881	
ST					2775	382	
ST					2803	726	
ST						2039	
ST						2052	
ST						2243	
ST						2446	
ST						2459	
ST						2590	
ST						2600	
ST						2609	
ST						2661	
ST						2664	
ST						2690	
ST						2691	
ST						2781	
ST						2788	

4.4.2 The Possibility of "*In-Silico*" Generated HGT

The fact that most (>95%) STs with imported alleles were encountered only once in the pool of 4,507 genotyped isolates raises the possibility for human-introduced error in sequencing and depositing gene sequences in the *Campylobacter* database [156]. In fact, the majority of the isolates with imported alleles were deposited in the *Campylobacter* database in submissions that included both *C. coli* and *C. jejuni* isolates, which may have promoted (man-made) mix-up of sequences and isolates. Consistent with these interpretations, we indentified several errors or inconsistencies in the *Campylobacter* MLST database. For instance, in a single mixed submission dated 11/17/2006, 36 and 49 isolates identified as *C. coli* and *C. jejuni*, respectively, were submitted. Although 81/85 of these isolates were identified correctly at the species level based on (presumably) their STs, four of them (STs 2467, 2489, 2491, and 2492) were mistakenly identified as *C. coli,* despite the fact that their sequence clearly corresponded to *C. jejuni* and no imported alleles were obvious for these STs. Difficult to detect, human-introduced errors are likely to occur at low frequency during manual handling and depositing of high-volume data to public databases, which is also consistent with the very low number of "questionable" *Campylobacter* isolates identified in our study (Table 4.2). Finally, instances where a foreign allele in a *C. coli* strain was acquired from a species other than *C. jejuni* (identified as the *C. coli* alleles with >10% nucleotide dissimilarity to any other available *C. coli* or *C. jejuni* allele) were rare and at least ten fold less frequent than acquisitions from *C. jejuni* strains. Although this observation is consistent with the hypothesis that *C. coli* recombines preferentially with *C. jejuni* [9], it appears rather unexpected given that *Campylobacter* organisms are members of complex microbial communities in their natural environment(s) [162, 163] (see also results based on genomic comparisons below). Clearly, more research is required to establish more firmly whether or not man-made errors and/or inability to PCR-amplify (and thus, sequence) divergent alleles may have artificially amplify the magnitude of directed genetic exchanged between *C. jejuni* and *C. coli*.

4.4.3 Genomic Insights Into The Inter-Species Gene Transfer

To provide further insights into the extent of interspecies gene transfer, we examined the available *Campylobacter* genomic sequences [164]. We employed a blast-based approach similar to the one described above for MLST data to identify imported genes in the genomic sequences. In brief, all *C. jejuni* RM1221 genes (1838 genes) were searched against the available high-draft *C. coli* RM2228 genome (38 contigs) to indentify orthologs with >97% nucleotide identity that were flanked by loci with substantially lower identity, i.e., identity close to ~86%, which typifies the genome average nucleotide identity between *C. coli* and *C. jejuni* [a high draft genomic sequence typically covers >95% of the genome of the sequenced organism [165]; hence, very few genes, if any, have been presumably missed by our analyses]. Highly conserved genes, i.e., genes typically showing much higher sequence conservation than the genome average such as the ribosomal RNA operon, ribosomal proteins and DNA/RNA polymerases [92], were identified by sequence comparisons against the genomes of *C. upsaliensis* and *C. lari* [164], close relatives of *C. jejuni* and *C. coli*, and were removed from further analyses (33 genes were removed). A total of 117 genes, constituting ~10% of the genes shared between RM1221 and RM2228, passed the criteria above and thus, could (potentially) represent genes exchanged recently between *C. jejuni* and *C. coli* (Fig. 4.2 & Table 4.3). Consistent with these interpretations, phylogenetic analysis of the latter genes and their orthologs (when present) in other sequenced *Campylobacter* species showed that the latter orthologs were considerably divergent, at the sequence level, from their *C. coli* or *C. jejuni* counterparts (data not shown). Thus, the high sequence identity of the 117 genes between *C. jejuni* RM1221 and *C. coli* RM2228 is likely due to HGT rather than high sequence conservation. Similar results were obtained with other *C. jejuni* genomes available (data not shown); RM2228 represents the only sequenced representative of the *C. coli* species currently available.

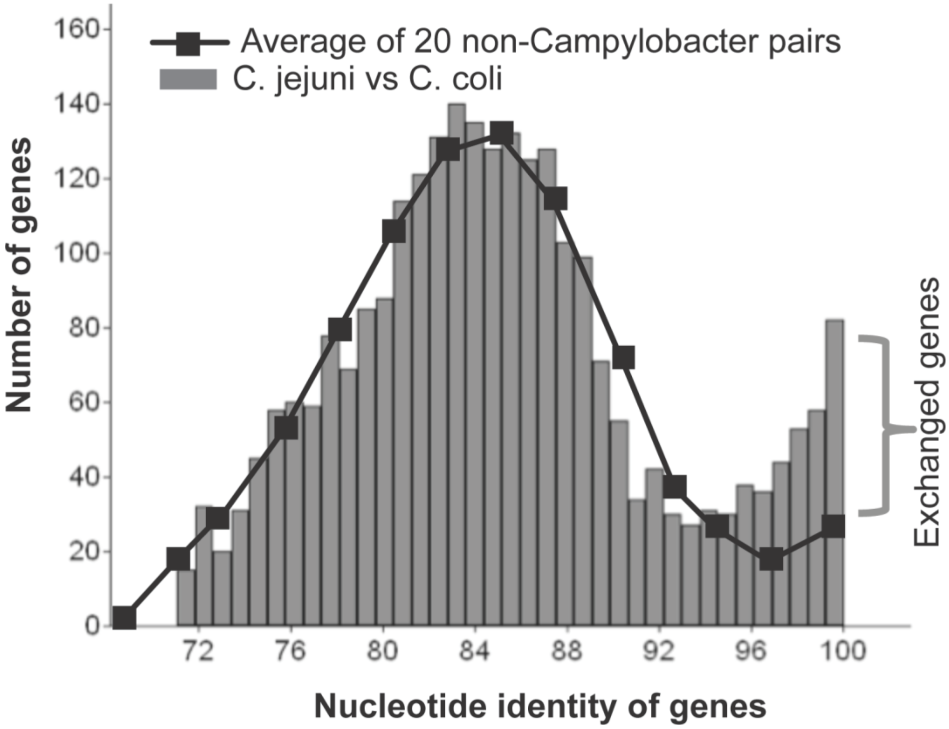


Figure 4.2 Distribution of the nucleotide identities of the genes shared by *C. jejuni* and *C. coli* genomes. The number of genes shared by *C. jejuni* RM1221 and *C. coli* RM2228

genomes (y-axis) is plotted against their nucleotide sequence identity (x-axis). The black line represents the average of 20 pair-wise comparisons of non-*Campylobacter* genomes, performed as described for the *Campylobacter* genomes. These genome pairs show comparable genome sizes and ANI relatedness among themselves to those observed for the two *Campylobacter* genomes and belong to several phylogenetically diverse genera, such as the *Procholorococcus* (*Cyanobacteria*), *Streptococcus* (Gram-positive, *Firmicutes*) and *Neisseria* (Gram-negative, *Proteobacteria*).

Table 4.3 The list of the 117 genes exchanged between *C. jejuni* and *C. coli* genomes. Columns show (from left to right): the gi number of the exchanged *C. jejuni* RM1221 gene (1st column), the annotation of each gene (2nd column), the blastn-derived nucleotide sequence identity of the *C. jejuni* RM1221 gene to its *C. coli* (3rd column) and *C. upsaliensis* (4th column) homolog, the blastp-derived amino acid sequence identity to its *C. upsaliensis* homolog (5th column), and the Dn/Ds ratio between the *C. jejuni* and *C. coli* homologs (6th column). The cut-off used in our nucleotide search was 70% identity; hence, "<70%" on 4th column denotes that either the gene is absent (i.e., no homolog was found) or the nucleotide identity of the homolog (if the latter exists) is below 70%. This cut-off was used because the blastn search (nucleotide level) is not sensitive enough below the 70% identity level. In general, consecutive gi numbers (1st column) indicate that the corresponding genes are adjacent to each other in the genome of *C. jejuni* RM1221.

		C.coli	C.upsaliensis		
gi number	**gene description**	**nuc. Identity**	**nuc. identity**	**aa identity**	**dn/ds**
57236895	RNA pseudouridylate synthase family protein	94.23	77.76	84.58	0.099
57236913	flagellar hook protein FlgE	97.64	73.7	78.41	0.134
57236914	Holliday junction resolvase	97.9	78.78	86.16	
57236921	glutamate synthase, large subunit	93.23	78.19	84.35	0.033
57236928	ExsB	100	73.64	74.55	0.361
57237028	magnesium and cobalt transport protein CorA	99.09	75.71	81.96	0.090
57237029	ABC transporter, periplasmic substrate-binding protein	99.04	<70	Not Found	0.001
57237030	hypothetical protein CJE0828	98.72	<70	60.9	0.073
57237031	hypothetical protein CJE0829	99.13	<70	Not Found	0.094
57237032	ABC transporter, permease protein	98.69	<70	22.48	0.198
57237033	ABC transporter, permease protein	98.46	<70	26.84	0.080
57237034	ABC transporter, ATP-binding protein	99.9	<70	44.95	0.490
57237035	HAD family hydrolase	99.53	<70	Not Found	0.129
57237036	CjaC	96.8	<70	62.7	0.126
57237048	hypothetical protein CJE0033	97.58	<70	Not Found	0.247
57237049	Bcr/CflA subfamily drug resistance transporter	99.67	75.46	77.47	0.299
57237050	hypothetical protein CJE0035	98.81	<70	Not Found	0.031
57237057	flagellar hook protein	92.43	76.43	78.35	0.053
57237058	hypothetical protein CJE0043	97.28	<70	48.49	0.329
57237059	hypothetical protein CJE0044	91.9	<70	47.09	0.169
57237198	trigger factor	94.38	78.17	80.84	0.043
57237300	cysteine desulfurase	93.49	78.07	87.28	0.038
57237346	thiF family protein	96.9	75.32	81.52	0.009
57237354	molybdenum ABC transporter, periplasmic molybdenum-binding protein	97.06	<70	68.27	0.101
57237355	biotin biosynthesis protein BioC	96.8	<70	53.71	0.112
57237357	8-amino-7-oxononanoate synthase	96.2	<70	64.74	0.243
57237358	adenosylmethionine--8-amino-7-oxononanoate transaminase	97.8	72.64	72.58	0.057
57237359	dethiobiotin synthetase	98.68	<70	67.66	0.088
57237366	HAD family hydrolase	96.35	<70	70.59	0.115
57237421	RND efflux system, inner membrane transporter CmeB	96.32	73.34	75.87	0.039
57237476	hypothetical protein CJE0470	99.58	<70	50.95	0.435
57237536	dihydrodipicolinate synthase, putative	94.17	<70	25.44	0.074
57237542	aldehyde dehydrogenase A	99.72	<70	25.22	0.219
57237599	TonB	97.14	<70	36.25	0.070
57237600	hypothetical protein CJE0846	100	<70	Not Found	
57237601	ferric receptor CfrA	93.78	<70	30.23	0.086
57237660	amino acid-binding protein	93.25	<70	28.4	0.233
57237662	UDP-N-acetylglucosamine pyrophosphorylase	95.58	73.91	75.47	0.076
57237716	flagellar hook-associated protein	94.54	71.85	72.64	0.125
57237917	GTP-binding protein LepA	94.16	80.91	91.11	0.003
57237918	hypothetical protein CJE1176	98.11	<70	64.92	0.076
57237919	AcrB/AcrD/AcrF family protein	99.27	72.96	76.99	0.114
57237920	adenylosuccinate lyase	98.51	77.52	80.6	0.087
57237934	succinyl-diaminopimelate desuccinylase	96.26	74.41	79.18	0.046
57237935	amino acid transporter LysE	99.67	<70	53.77	
57237936	NAD-dependent deacetylase	99.57	<70	71.93	0.001
57237937	type II restriction-modification enzyme	99.4	<70	32.27	0.078
57237938	recombination and DNA strand exchange inhibitor protein	93.7	72.64	73.8	0.074
57237940	UDP-N-acetylmuramate--L-alanine ligase	97.69	76.88	79.39	0.314
57237948	acetyltransferase	100	<70	Not Found	0.001
57238043	peptidyl-prolyl cis-trans isomerase B	95.45	76.4	82.5	0.018
57238044	hypothetical protein CJE1306	98.45	79.4	90.64	0.021
57238045	SMR family multidrug efflux pump	98.83	<70	43.4	0.261
57238046	SMR family multidrug efflux pump	97.73	<70	55.84	0.054
57238047	arginyl-tRNA synthetase	94.6	75.08	74.15	0.070
57238061	methyl-accepting chemotaxis protein	99.8	75.51	72.73	0.001
57238062	methyl-accepting chemotaxis protein	98.19	<70	72.77	0.189
57238178	hydrogenase nickel insertion protein HypA	98.26	73.45	67.26	0.069
57238179	hydrogenase expression/formation protein HypE	97.95	74.87	76.54	0.072
57238180	hydrogenase expression/formation protein HypD	98.26	74.86	77.35	0.017
57238182	hydrogenase accessory protein HypB	98.92	74.9	74.4	0.040
57238183	[NiFe] hydrogenase maturation protein HypF	99.09	<70	66.8	0.146
57238184	hypothetical protein CJE0724	96.84	<70	60.08	0.086
57238186	MATE efflux family protein	97.49	73.62	76.15	0.079
57238188	phosphate ABC transporter, ATP-binding protein	95.25	<70	54.96	0.001
57238189	phosphate ABC transporter, permease protein PstA, putative	98.99	<70	37.75	0.091
57238190	phosphate ABC transporter, permease protein PstC, putative	99.89	<70	32.64	0.001
57238191	phosphate ABC transporter, periplasmic phosphate-binding protein, putative	97.89	<70	34.91	0.060
57238251	flagellar protein FliS	97.67	78.82	83.06	0.001
57238252	flagellar capping protein	97.75	<70	62.77	0.464
57238253	flagellar protein FlaG	99.73	<70	57.72	0.001
57238345	hypothetical protein CJE1487	96.05	<70	68.1	
57238346	hypothetical protein CJE1488	83.68	<70	42.75	0.243
57238348	hypothetical protein CJE1490	99.89	<70	Not Found	
57238349	HAD family phosphatase	97.89	<70	Not Found	0.146
57238350	3-oxoacyl-(acyl carrier protein) synthase III	97.27	<70	30.81	0.064
57238351	acyl carrier protein, putative	98.2	<70	Not Found	
57238352	hypothetical protein CJE1494	97.09	<70	41.67	0.146
57238353	hypothetical protein CJE1495	97.78	<70	39.37	
57238354	amino acid adenylation domain-containing protein	96.36	73.54	72.76	0.032
57238355	acetyltransferase	95.18	<70	40.17	0.158
57238356	formyl transferase domain-containing protein	99.44	<70	28.11	0.085
57238364	flagellar protein, putative	99.88	<70	61.73	
57238365	acetyltransferase	99.79	<70	53.42	
57238366	imidazoleglycerol phosphate synthase, cyclase subunit	98.26	<70	36.08	0.072
57238367	imidazole glycerol phosphate synthase subunit HisH	96.7	<70	39.9	0.108
57238368	flagellin modification protein, PseA	98.86	<70	33.24	0.050
57238370	NAD-dependent epimerase/dehydratase family protein	98.04	<70	69.97	0.090
57238371	aminotransferase	98.69	<70	66.4	0.138
57238372	formyltransferase, putative	99.19	71.39	64	
57238373	N-acetylneuraminic acid synthetase (neuB)	98.41	<70	70.66	0.179
57238374	UDP-N-acetylglucosamine 2-epimerase	97.94	<70	67.7	0.069
57238375	nucleotidyltransferase family protein	97.17	73.75	72.89	0.124
57238376	hypothetical protein CJE1519	98.23	<70	60.47	0.190
57238377	posttranslational flagellin modification protein B	98.59	<70	69.26	0.079
57238378	flagellin modification protein A	98.44	76.3	73.44	0.122
57238379	motility accessory factor	89.7	<70	45.29	0.166
57238380	motility accessory factor	98.51	<70	43.67	0.110
57238381	motility accessory factor	90.16	<70	48.89	0.194
57238382	motility accessory factor	97.48	<70	50.47	0.111
57238385	hypothetical protein CJE1531	96.86	<70	49.16	0.134
57238390	phosphatidate cytidylyltransferase	93.8	<70	66.12	0.067
57238431	endoribonuclease L-PSP, putative	93.66	<70	Not Found	0.051
57238435	feoA family protein	99.56	<70	60.81	
57238489	hypothetical protein CJE1639	99.77	76.09	76.39	0.001
57238490	flagellar hook-associated protein FlgK	91.24	76.41	81.74	0.035
57238540	thiS family protein	100	76.47	76.71	
57238541	molybdopterin converting factor, subunit 2	99.77	<70	62.42	0.001
57238575	arsenate reductase	99.53	<70	Not Found	0.292
57238576	arsenical-resistance protein, putative	99.43	<70	Not Found	
57238598	oxidoreductase, FAD-binding, iron-sulfur cluster-binding	98.45	<70	45.84	0.106
57238600	multidrug transporter membrane component/ATP-binding component	96.65	<70	26.07	
57238642	hypothetical protein CJE1803	92.49	<70	49.31	0.197
57238644	hypothetical protein CJE1805	95.43	76.42	81.96	0.093
57238665	Na+/H+ antiporter NhaA	91.88	74.36	78.8	
57238669	FTR1 family iron permease	88.14	<70	Not Found	0.141
57238729	flagellin	90.76	<70	68.83	0.110

The genomic comparisons also revealed that the genome of *C. jejuni* RM1221 possesses more strain-specific genes (~400) than the number of genes it has (potentially) exchanged with *C. coli* RM2228 (~117), and vise versa. The great majority of the former genes appear to have been acquired through HGT from (apparently) non-*C. coli* sources since they were associated with mobile elements and/or were absent in other *Campylobacter* genomes. Although the majority (~70%) of the RM1221-specific genes are contained within the prophage (i.e., ephemeral) parts of the genome, a fraction comparable to that of exchanged genes with *C. coli* represents host-like, as opposed to phage-like, functions and includes several transport, polysaccharide biosynthesis, and metabolism genes, among other functional genes. These findings are consistent with those reported previously in a more comprehensive investigation of the *Campylobacter* gene-content differences [164] and suggest that promiscuous acquisition of genetic material may be as important as, if not more important than, directed genetic exchange in the *C. jejuni* - *C. coli* case.

While the 117 genes represent a slightly higher degree of inter-species genetic exchange in the *C. jejuni* - *C. coli* case relatively to other species (Fig. 4.2, black line), the spatial distribution of these genes in the *C. jejuni* RM1221 genome was not random. Rather, the genes clustered together in a few areas of the genome. For example, 30/117 genes were found in a single large region, located at about 9 o'clock position in the RM1221 genome, while more than half of the 117 exchanged genes were located in just three areas of the RM1221 genome (Fig. 4.3 & Table 4.3). If *C. jejuni* and *C. coli* were indeed converging, as hypothesized previously [9], a much more unbiased (i.e., random) genome-wide distribution of the exchanged genes would have been expected [88].

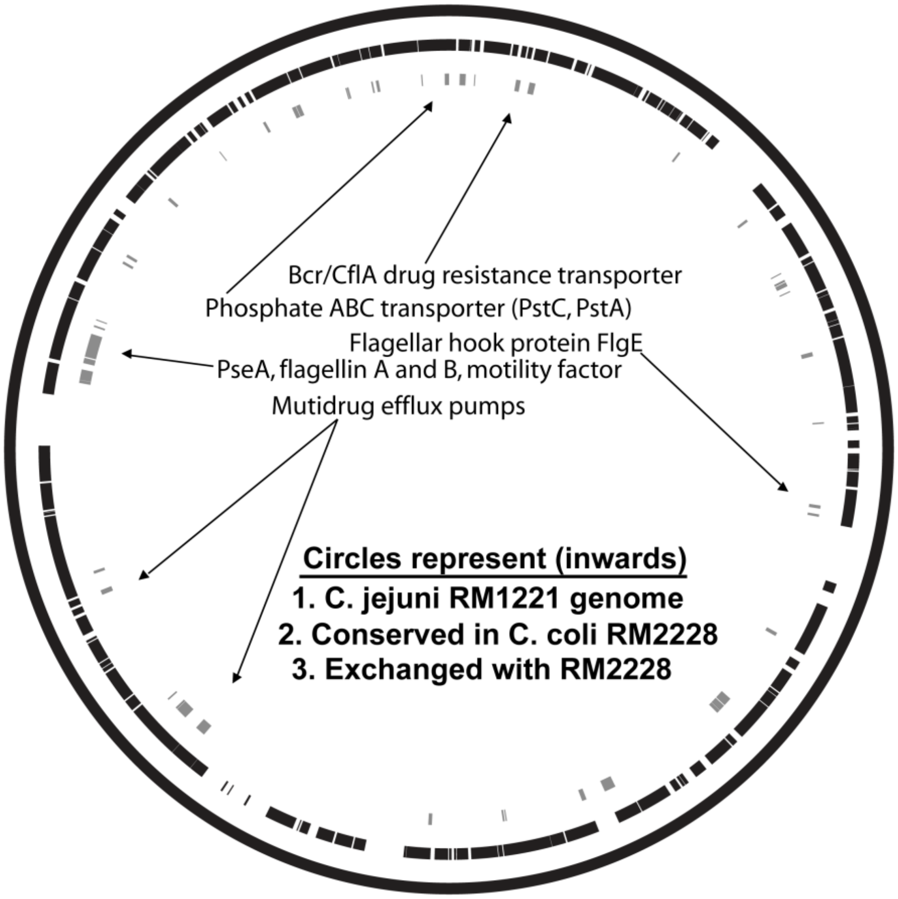

Bcr/CflA drug resistance transporter
Phosphate ABC transporter (PstC, PstA)
Flagellar hook protein FlgE
PseA, flagellin A and B, motility factor
Mutidrug efflux pumps
Circles represent (inwards)
1. C. jejuni RM1221 genome
2. Conserved in C. coli RM2228
3. Exchanged with RM2228

Figure 4.3 Spatial distribution of the exchanged genes in the *Campylobacter* genome. The 117 genes that were identified as exchanged between the *C. jejuni* RM1221 and *C. coli* RM2228 genomes were mapped (gray color, innermost circle) against the genome of *C. jejuni* RM1221 (outermost circle). The parts of *C. jejuni* RM1221 genome shared by *C. coli* RM2228 (middle circle) as well as a few representative examples of exchanged genes (discussed in the text) are also shown on the graph. Circles were drawn using the GenomeViz software [166].

4.4.4 *C. jejuni* and *C. coli* Exchange Ecologically Important Genes

The highly biased spatial distribution of the exchanged genes indicated that unusual (for the genome) evolutionary processes might be acting on these genes. To provide further insights into the latter issue, we examined the functional annotation of the 117 genes more closely. We found that the predicted function of these genes was also far from random when compared to all genes in the RM1221 genome (student's t test < 0.01; Fig. 4.4). In fact, the pool of 117 genes was heavily enriched in hypothetical proteins (21/117), motility accessory factors and flagella genes (16/117), and genes related to metallo beta lactamases, multidrug efflux pumps, two ABC-transport systems, endonucleases III, lipopoligosacharide synthesis, and membrane-associated proteins (Fig. 4.3, Fig. 4.4 & Table 4.3). Thus, the exchanged genes appeared to be functionally limited to motility, drug resistance, transport of nutrients, and genes causing variation in the surface properties of the cell, i.e., accessory genes potentially important for environmental adaptation. Such genes probably enable *Campylobacter* survival and adaptation in the intestinal tract of human and animal species, the presumptive ecological niche of these organisms [162, 163]. For instance, the role of polysaccharide surface antigens in evading phages or the eukaryotic host defense mechanisms has been well documented previously for many pathogenic and environmental bacteria [167, 168]. Hence, environmental selection pressures appear to drive, by and large, the exchange (and, more importantly, the fixation in the population) of genetic material between *C. jejuni* - *C. coli*.

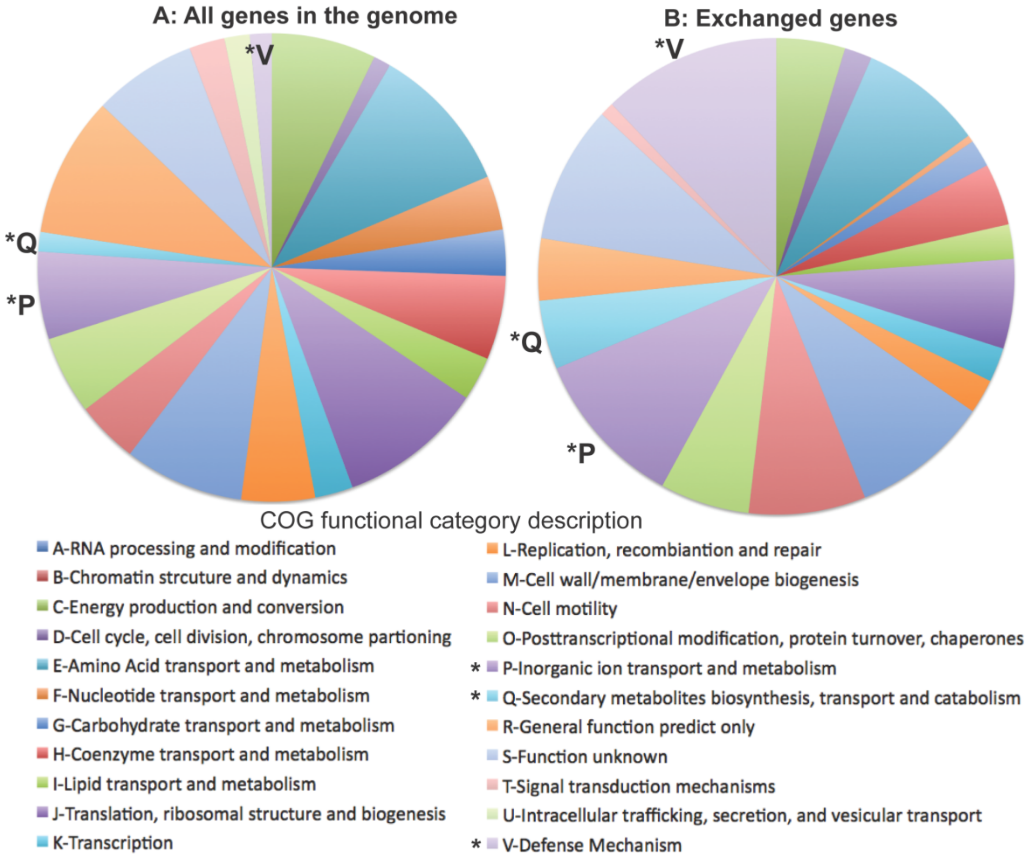


Figure 4.4 Functional biases in the genes exchanged between the *Campylobacter* genomes. All genes in the genome of *C. jejuni* RM1221 were assigned to a major gene functional category of the Cluster of Orthologous Gene (COG) database [131], as described previously [169]. The percentage of the total genes in the genome assigned to each category (Panel A) relative to that of only the exchanged genes (Panel B) are shown. The three most differentially abundant categories in the latter distribution relative to the former are noted on the graph. See legend key for the description of each COG category.

In contrast, housekeeping genes, such as those used in MLST applications, were dramatically depleted from the pool of exchanged genes. A few cases of housekeeping genes exchanged between the two *C. jejuni* and *C. coli* were also noted based on the genomic comparisons (Table 4.3). These cases, however, were, typically, attributable to hitchhiking events associated with the exchange of accessory genes of ecological importance. For instance, *C. coli lepA* (GTP-binding protein) and *purB* (adenylosuccinate lyase) showed >99% identity to their *C. jejuni* orthologs and flanked an AcrB/AcrD/AcrF operon (cation/multidrug efflux pump). The later operon shows >99% to *C. jejuni* and has apparently been transferred into/from the *C. coli* genome through a mobile element mechanism based on its high nucleotide identity and the presence of a phage-like integrase adjacent to the operon (no complete prophage genome was found nearby, nonetheless). Regardless of what the actual mechanism might have been in this case, the most parsimonious scenario is that the *lepA* and *purB* genes were horizontally exchanged together with the multidrug efflux pump.

These findings are in agreement with, and probably explain, the small number of exchanged MLST genes identified by our (Table 4.1) and the previous study [9]. They also suggest that, even though strong environmental pressures and complete niche overlap (if true) could potentially promote the convergence of *C. jejuni* and *C. coli* phenotypes, through selection to acquire or exchange the same environmentally important genes, the two species would, most likely, have remained genomically discrete in their core genome. Consistent with the later interpretation, a recent independent study of the same *Campylobacter* MLST dataset based on coalescent theory suggested that the intra-species genetic flow for housekeeping genes is at least an order of magnitude higher than the inter-species genetic flow in the *C. jejuni* – *C. coli* case [136]. Under such gene flow rates, the two species will continue to diverge from each other in their core genome based on computer simulations [14, 88], which is consistent with our interpretations based on the genomic comparisons. Further, the average nucleotide identity of the transferred genes between *C. coli* and *C. jejuni* is ~97%, which suggests that many of the HGT events between the two lineages occurred long time ago, corresponding presumably to several hundred or thousand years [136, 170]. Thus, if the two *Campylobacter* species were indeed converging in their core genome, there would have been enough evolutionary time elapsed to replace (through genetic exchange) many core alleles, in addition to acquiring the environmentally important genes. The number of core genes replaced, however, was negligible (Table 4.1) despite enough evolutionary time (presumably) available, indicating that the two species are unlikely to be converging.

4.4.5 Several Exchanged Genes May Undergo Adaptive Evolution

The strong bias in exchanged genes toward a few specific cellular functions implied that the corresponding genes might confer a selective advantage to the recipient species. Analysis of non-synonymous vs. synonymous amino acid substitution ratio (Dn/Ds) can provide some clues about the strength of selection acting on protein sequences, with Dn/Ds values higher than one being indicative of positive (adaptive) selection. Analysis of the Dn/Ds ratio among all *C. coli* and *C. jejuni* orthologs showed that the distribution of the Dn/Ds values of the 117 exchanged genes differed significantly (student's t test < 0.01) from that of the remaining genes in the genome. In fact, 54% of the total genes shared between *C. jejuni* and *C. coli* showing Dn/Ds ratio larger than 0.1 were exchanged genes; even though, the latter constituted only ~10% of the total shared genes. The average Dn/Ds ratio of the exchanged genes was twice as large as the average of all shared genes (Fig. 4.5). These data are unlikely to reflect relaxed selection or to be attributable solely to the time-dependency of the Dn/Ds signature [171]. Rather, our findings probably reflect the selective advantage conferred by some of the exchanged genes to the recipient cells. Consistent with the latter hypothesis, when we performed Dn/Ds analysis using a 30 amino acid long sliding window, as proposed recently [159], we found that at least one segment of the sequence of several exchanged accessory genes had Dn/Ds ratio much higher than one (Fig. 4.5, inset). In contrast to accessory genes and as expected, the sequences of the (very few) housekeeping genes exchanged between *C. jejuni* and *C. coli* genomes, such as the *lepA* and *purB* genes mentioned above, showed typically no window with Dn/Ds >1 (Fig. 4.5, inset). Therefore, although sometimes the signature of positive selection (i.e., Dn/Ds >1) was not apparent when considering the whole sequence of a gene, the signature became evident for specific domains of a gene. These results further corroborated the conclusion that several of the exchanged accessory genes may be under positive selection.

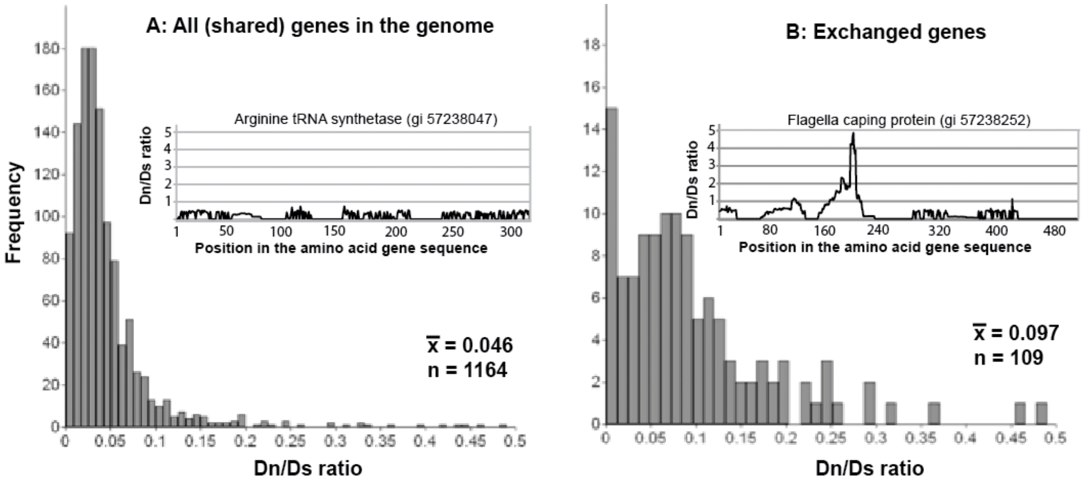


Figure 4.5 Signatures of positive selection of the genes exchanged between the *Campylobacter* genomes. The number of genes shared by *C. jejuni* RM1221 and *C. coli* RM2228 genomes (y-axis) is plotted against their whole-sequence-based synonymous/nonsynonymous amino acid substitution ratio (Dn/Ds) (x-axis). Panel A shows all shared genes while panel B shows only the genes that were exchanged recently between RM1221 and RM2228 genomes. The number of genes used in each panel and their average Dn/Ds ratio are also shown. Dn/Ds analysis was also performed on segments of the sequence of several selected genes using sliding windows. Representative examples of an exchanged accessory gene undergoing (possibly) positive selection (Panel B) and a hitchhiked housekeeping gene (Panel A, no positive selection) are also shown (insets). Note that at least one segment of the sequence of the former gene shows a clear signature of positive selection (i.e., Dn/Ds >> 1), whereas the whole sequence of the latter gene undergoes strong purifying (negative) selection (i.e., Dn/Ds << 1). Several additional exchanged accessory genes showed similar signatures of adaptive evolution; in contrast, virtually no exchanged housekeeping gene showed such signatures (Table 4.3).

4.5 Conclusions and Perspectives

Although evidence for genetic exchange between *C. coli* and *C. jejuni* appear to exist, probably beyond reasonable doubt and the consequences of (possible) human-introduced errors (e.g., *uncA* and *aspA* genes and the results of our genomic comparisons), several independent lines of evidence suggest that the available data are not conclusive about *Campylobacter* species convergence. Further, several reasons may account for preferential acquisition of environmentally favored genes from closely related organisms rather than distantly related ones such as the host-specificity of the vectors of HGT (e.g., phages), similarities in gene regulation and expression (which facilitates the functionality of the exchanged gene in the recipient cell), and in the mechanisms defending invasion of foreign (but not native or similar) DNA. However, an intrinsic preference to recombine with close relatives does not necessarily lead to species convergence, especially in cases where genetic exchanged is likely limited to a few environmentally important functions, like in the *Campylobacter* case. Hence, convergence of (any) two bacterial species remains to be seen.

Our genomic comparisons also provided novel insights into the interplay between environmental selection pressures and genetic exchange in the *Campylobacter* group and identified several environmentally important genes that have been exchanged recently between *C. jejuni* and *C. coli* species. The preferential exchange of the latter genes and their adaptive evolution (Fig. 4.5) indicate that they may contribute substantially to the adaptation, survival and pathogenic potential of *Campylobacter* pathogens and hence, should be targets of further investigation.

4.6 Acknowledgments

We thank Daniel Falush and Samuel Sheppard for providing the list of isolates and their assignment to species used in their study. Our work is supported by the National Science Foundation (award DEB-0516252), and the DOE Genomics: GTL Program (DE-FG-02-02ER63342).

CHAPTER 5

THE CHIMERIC GENOME OF *SPHAEROCHAETA*: NON-SPIRAL SPIROCHETES THAT BREAK WITH THE PREVALENT DOGMA IN SPIROCHETE BIOLOGY

Reproduced in part with permission from A. Caro-Quintero, K. M. Ritalahti, K. D. Cusick, F. E. Löffler, and K. T. Konstantinidis. CITA. mBio. 2012, 3(3).
Copyright 2012 Caro-Quintero et al.

5.1 Abstract

The spirochetes represent one of a few bacterial phyla that are characterized by a unifying diagnostic feature; namely the helical morphology and motility conferred by axial periplasmic flagella. The unique morphology and mode of propulsion also represent major pathogenicity factors of clinical spirochetes. Here we describe the genome sequences of two coccoid isolates of the recently described genus *Sphaerochaeta*, which are members of the *Spirochaetes* phylum based on 16S rRNA gene and whole genome phylogenies. Interestingly, the *Sphaerochaeta* genomes completely lack the motility and associated signal transduction genes present in all sequenced spirochete genomes. Additional analyses revealed that the lack of flagella is associated with a unique, non-rigid cell wall structure, hallmarked by the lack of transpeptidase and transglycosylase genes, which is also unprecedented for spirochetes. The *Sphaerochaeta* genomes are highly enriched in fermentation and carbohydrate metabolism genes relative to other spirochetes, indicating a fermentative lifestyle. Remarkably, most of the enriched genes appear to have been acquired from non-spirochetes, particularly *Clostridia*, in several massive, horizontal gene transfer events (> 40% of the total genes in each genome). Such a high level of direct inter-phylum genetic exchange is extremely rare among mesophilic

organisms and has important implications for the assembly of the prokaryotic Tree of Life.

5.2 Introduction

Spirochetes represent a diverse, deeply-branching phylum of Gram-negative bacteria. Members of this phylum share distinctive morphological features, i.e., spiral shape and axial, periplasmic flagella [172, 173]. These traits enable propulsion through highly viscous media, and thus, are directly associated with the ecological niches spirochetes occupy. For instance, motility mediated by axial flagella represents a major pathogenicity factor that allows strains of the *Treponema*, *Borrelia,* and *Leptospira* genera to invade and colonize host tissues, resulting in important diseases such as Lyme disease and syphilis. Several studies have shown that disruption of the flagellar or the chemotaxis genes that control the periplasmic flagella attenuates spirochete pathogenic potential [174-176].

The focus on clinical isolates has biased our understanding of the ecology, physiology, and diversity of the *Spirochaetes* phylum. Indeed, free-living, non-pathogenic spirochetes are greatly underrepresented in culture collections, while culture-independent studies have revealed that spirochetes are ubiquitous in anoxic environments, implying that they represent key players in anaerobic food webs [177-180]. Consistent with the latter findings, studies of members of the *Spirochaeta* genus demonstrated that environmental isolates possess distinct physiological properties compared to their pathogenic relatives, e.g., they encode a diverse set of saccharolytic enzymes [177], while other members of the genus are alkaliphiles [181] and thermophiles [182]. More recently, screening environmental samples revealed a novel genus of free-living spirochetes, the *Sphaerochaeta* [183]. Phylogenetic analysis of 16S rRNA genes identified this group as a member of the phylum *Spirochaetes*, most closely related to the genus *Spirochaeta*. Interestingly, *Sphaerochaeta pleomorpha* strain Grapes and *Sphaerochaeta globosa* strain Buddy are non-motile and share a spherical morphology during laboratory cultivation [183]. However, currently it remains unclear whether this unusual morphology and the lack of motility represent a distinct stage of the cell cycle and/or responses to culture conditions, or if these distinguishing features have a genetic basis. To elucidate the metabolic properties and evolutionary history of environmental, non-pathogenic spirochetes and to provide insights into the unusual morphological features of *Sphaerochaeta*, we sequenced the genomes of strain Grapes and strain Buddy, representing the type strains of *S. pleomorpha* and *S. globosa*, respectively. Our analyses suggest that *Sphaerochaeta* are unique spirochetes that completely lack the genes of the motility apparatus and have acquired nearly half of their genomes from Gram-positive bacteria, an extremely rare event among mesophilic organisms.

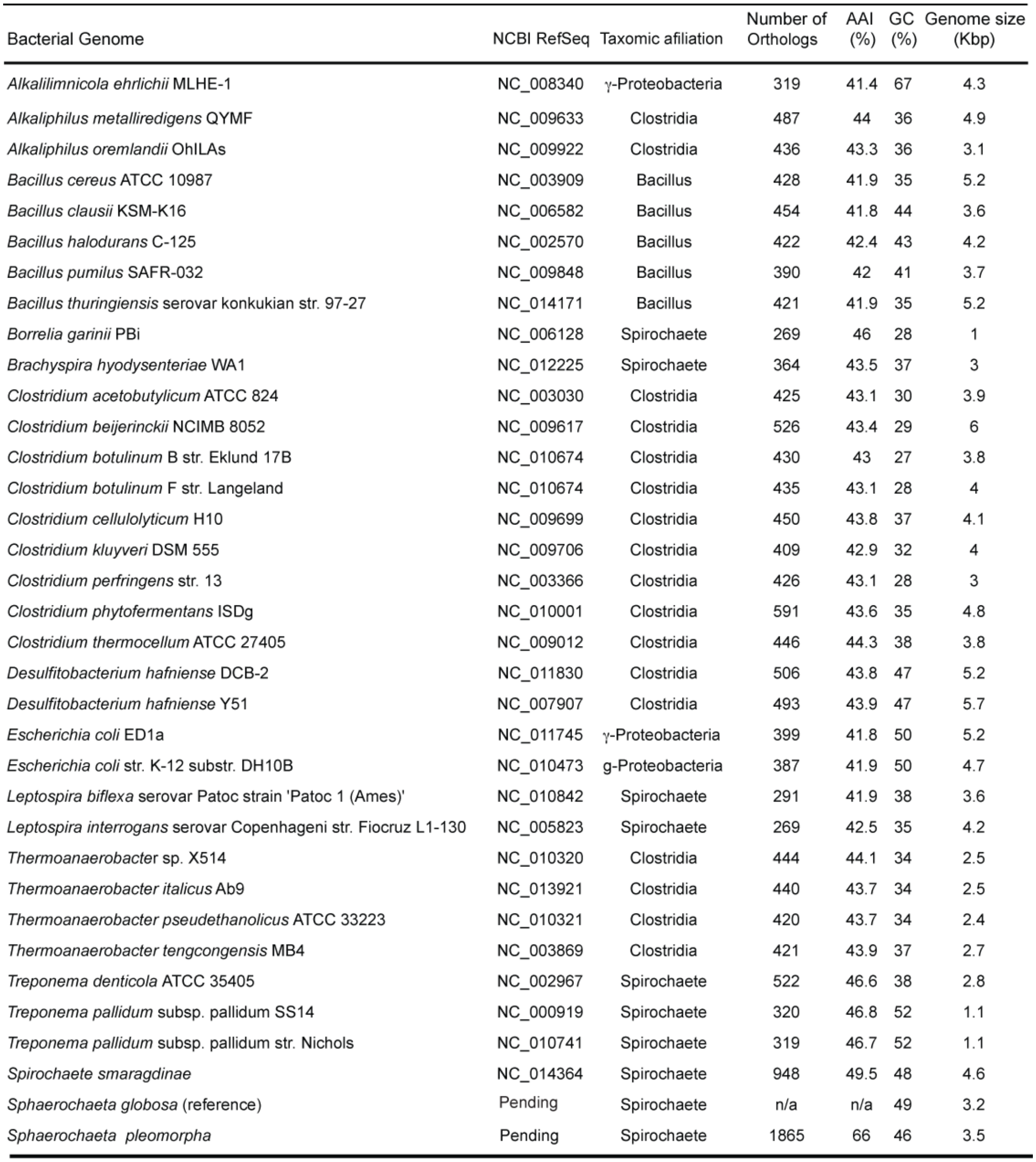
5.3 Materials and Methods

5.3.1 Organisms Used In This Study

The information of the genome sequence of each *Sphaerochaeta* species is provided in Table 5.1. The accession numbers of the genomes are: CP003155 (*S.*

pleomorpha) and CP002541 (*S. globosa*). Details regarding the isolation conditions of type species are available elsewhere [183].

Table 5.1 Bacterial genomes used in the analysis of horizontal gene transfer



Bacterial Genome	NCBI RefSeq	Taxomic afiliation	Number of Orthologs	AAI (%)	GC (%)	Genome size (Kbp)
Alkalilimnicola ehrlichii MLHE-1	NC_008340	γ-Proteobacteria	319	41.4	67	4.3
Alkaliphilus metalliredigens QYMF	NC_009633	Clostridia	487	44	36	4.9
Alkaliphilus oremlandii OhILAs	NC_009922	Clostridia	436	43.3	36	3.1
Bacillus cereus ATCC 10987	NC_003909	Bacillus	428	41.9	35	5.2
Bacillus clausii KSM-K16	NC_006582	Bacillus	454	41.8	44	3.6
Bacillus halodurans C-125	NC_002570	Bacillus	422	42.4	43	4.2
Bacillus pumilus SAFR-032	NC_009848	Bacillus	390	42	41	3.7
Bacillus thuringiensis serovar konkukian str. 97-27	NC_014171	Bacillus	421	41.9	35	5.2
Borrelia garinii PBi	NC_006128	Spirochaete	269	46	28	1
Brachyspira hyodysenteriae WA1	NC_012225	Spirochaete	364	43.5	37	3
Clostridium acetobutylicum ATCC 824	NC_003030	Clostridia	425	43.1	30	3.9
Clostridium beijerinckii NCIMB 8052	NC_009617	Clostridia	526	43.4	29	6
Clostridium botulinum B str. Eklund 17B	NC_010674	Clostridia	430	43	27	3.8
Clostridium botulinum F str. Langeland	NC_010674	Clostridia	435	43.1	28	4
Clostridium cellulolyticum H10	NC_009699	Clostridia	450	43.8	37	4.1
Clostridium kluyveri DSM 555	NC_009706	Clostridia	409	42.9	32	4
Clostridium perfringens str. 13	NC_003366	Clostridia	426	43.1	28	3
Clostridium phytofermentans ISDg	NC_010001	Clostridia	591	43.6	35	4.8
Clostridium thermocellum ATCC 27405	NC_009012	Clostridia	446	44.3	38	3.8
Desulfitobacterium hafniense DCB-2	NC_011830	Clostridia	506	43.8	47	5.2
Desulfitobacterium hafniense Y51	NC_007907	Clostridia	493	43.9	47	5.7
Escherichia coli ED1a	NC_011745	γ-Proteobacteria	399	41.8	50	5.2
Escherichia coli str. K-12 substr. DH10B	NC_010473	g-Proteobacteria	387	41.9	50	4.7
Leptospira biflexa serovar Patoc strain 'Patoc 1 (Ames)'	NC_010842	Spirochaete	291	41.9	38	3.6
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	NC_005823	Spirochaete	269	42.5	35	4.2
Thermoanaerobacter sp. X514	NC_010320	Clostridia	444	44.1	34	2.5
Thermoanaerobacter italicus Ab9	NC_013921	Clostridia	440	43.7	34	2.5
Thermoanaerobacter pseudethanolicus ATCC 33223	NC_010321	Clostridia	420	43.7	34	2.4
Thermoanaerobacter tengcongensis MB4	NC_003869	Clostridia	421	43.9	37	2.7
Treponema denticola ATCC 35405	NC_002967	Spirochaete	522	46.6	38	2.8
Treponema pallidum subsp. pallidum SS14	NC_000919	Spirochaete	320	46.8	52	1.1
Treponema pallidum subsp. pallidum str. Nichols	NC_010741	Spirochaete	319	46.7	52	1.1
Spirochaete smaragdinae	NC_014364	Spirochaete	948	49.5	48	4.6
Sphaerochaeta globosa (reference)	Pending	Spirochaete	n/a	n/a	49	3.2
Sphaerochaeta pleomorpha	Pending	Spirochaete	1865	66	46	3.5

5.3.2 Sequence Analysis and Metabolic Reconstruction

Orthologous proteins between *Sphaerochaeta* and selected publicly available genomes were identified using a reciprocal best-match (RBM) approach and a minimum cut-off for a match of 70% coverage of the query sequence and 30% amino acid identity, as described previously [118]. For phylogenetic analysis, sequence alignments were constructed using the ClustalW software [184] and trees were built using the Neighbor Joining algorithm as implemented in the MEGA 4 package [116]. Central metabolic pathways were reconstructed using Pathway Tools version 14 [185]. The annotation files required as input to the Pathway Tools were prepared from the consensus results of two approaches. First, amino acid sequences of predicted proteins were annotated based on their best BLAST match against NR [186], KEGG [187] and COG [131] databases. Second, the whole genome sequences were submitted to the RAST annotation pipeline [188] to ensure that the previous approach did not miss any important genes, and to assign protein sequences to functions and enzymatic reactions (E.C. numbers). The results of both approaches were used to extract gene names and E.C. numbers. Disagreements between the two approaches were resolved by manual curation.

5.3.3 Horizontal Gene Transfer (HGT) Analysis

For best-match analysis, strain Buddy protein sequences were searched using BLASTP against two databases: i) all completed prokaryotic genomes available in January 2011 (n=1,445) and ii) NR database (release 178). The best match for each query sequence, when better than 70 % coverage of the length of the query protein and 30% amino acid identity, was identified, and the taxonomic affiliation of the genome encoding the best match was extracted from the taxonomy browser of NCBI. HGT events were identified as follows: orthologous protein sequences present in at least one representative genome from the five groups used (i.e., *Sphaerochaeta*, *S. smaragdiane,* other spirochetes, *Clostridiales*, and *E. coli*) were identified and aligned as described above. Phylogenetic trees for each alignment were built in Phylip v3.6, using both Maximum Parsimony and Neighbor Joining algorithms, and bootstrapped 100 times using Seqboot [189]. The topology of the resulting consensus tree was compared to the 16S rRNA gene-based tree topology and conflicting nodes between the two trees, which also had bootstrap support higher than 50, were identified as cases of HGT.

To evaluate how unique the case of inter-phylum gene transfer between *Clostridiales* and *Sphaerochaeta* is, the following approach was used. All available completed bacterial and archaeal genomes (as of January 2011, n=1,445) that showed similar genetic relatedness among them to the relatedness among the *Sphaerochaeta* genomes (i.e., 65 +/- 0.5% gAAI), were assigned to the same group. All protein-coding genes shared between genomes of different groups were subsequently determined using the BLASTP algorithm as described above. The

BLASTP results were analyzed using sets of three genomes at a time, each genome representing one of three distinct groups: i) a reference group, ii) a group from the same phylum as the reference group, and iii) a group from another phylum. The ratio of the number of genes of the reference genome with best matches in the genome of the different phylum vs. the number genes of the reference genome with best matches to the genome of the same phylum was determined for each set and plotted against the gAAI value between the reference genome and the genome of the same phylum (Fig. 5.1). Groups of genomes sharing fewer than forty genes were removed from further analysis to reduce noisy results from very distantly related or small size genomes.

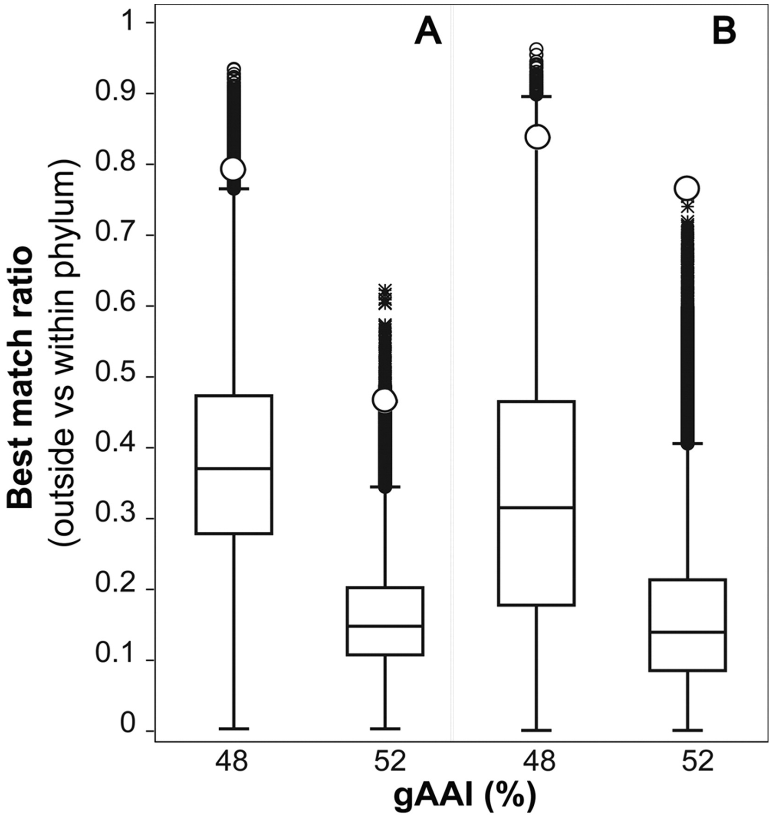


Figure 5.1 Comparisons of the extent of inter-phylum horizontal gene transfer. The ratio of the number of genes of a reference genome with best BLASTP matches in a genome of a different phylum relative to a genome of the same phylum as the reference genome was determined in three-genome comparisons (sets) as described in the text. The graph shows the distribution of the ratios for 150,022 and 86,516 comparisons that included genomes of the same phylum showing ~48% and ~52% gAAI, respectively; the distributions were based on all genes shared among the three genomes in a comparison (A) and all genes in the reference genome (B). Horizontal bars represent the median, the upper and lower box boundaries represent the upper and lower quartiles, and the upper and lower whiskers represent the 99% percentile. Open circles represent the values for the *Sphaerochaeta* – *Clostridiales* case.

5.4 Results

5.4.1 Phylogenetic Affiliation

The *S. pleomorpha* strain Grapes and *S. globosa* strain Buddy complete genomes encode about 3,200 and 3,000 putative protein coding sequences (CDS), have an average % G+C content of 46% and 49%, and a genome size of 3.5 and 3.2 Mbp, respectively (Table 5.1). The two genomes share about 1,850 orthologous genes (i.e., 57-61% of the total genes in the genome, depending on the reference genome), and these genes show, on average, 65% amino acid identity. Therefore, the two genomes represent two divergent species of the *Sphaerochaeta* genus according to current taxonomic standards [129].

Phylogenetic analysis of the concatenated alignment of 43 highly-conserved, single-copy informational genes (Table 5.2) corroborated previous 16S rRNA gene-based findings [183] that identified *Sphaerochaeta* as a distinct lineage of the *Spirochaetes* phylum, most closely related to members of the *Spirochaeta* genus, e.g., *Spirochaeta coccoides* and *Spirochaeta smaragdinae* (Fig. 5.2). The average amino acid identity between *S. smaragdinae* and *S. pleomorpha* or *S. globosa* was 46% (based on 900 shared orthologous genes). This level of genomic relatedness is typically observed between organisms of different families, if not orders [190]; hence, *Sphaerochaeta* and *Spirochaeta* represent distantly related genera of the *Spirochaetes* phylum. Other spirochetal genomes shared fewer orthologous genes with *Sphaerochaeta* (e.g., 300-500), and these genes showed lower amino acid identities compared to *S. smaragdinae* (e.g., 30-45%). No obvious inter- or intra-phylum horizontal gene transfer (HGT) of any of the 43 informational genes was observed when the phylogenetic analysis was expanded to include selected genomes of *Proteobacteria* and Gram-positive bacteria (see below).

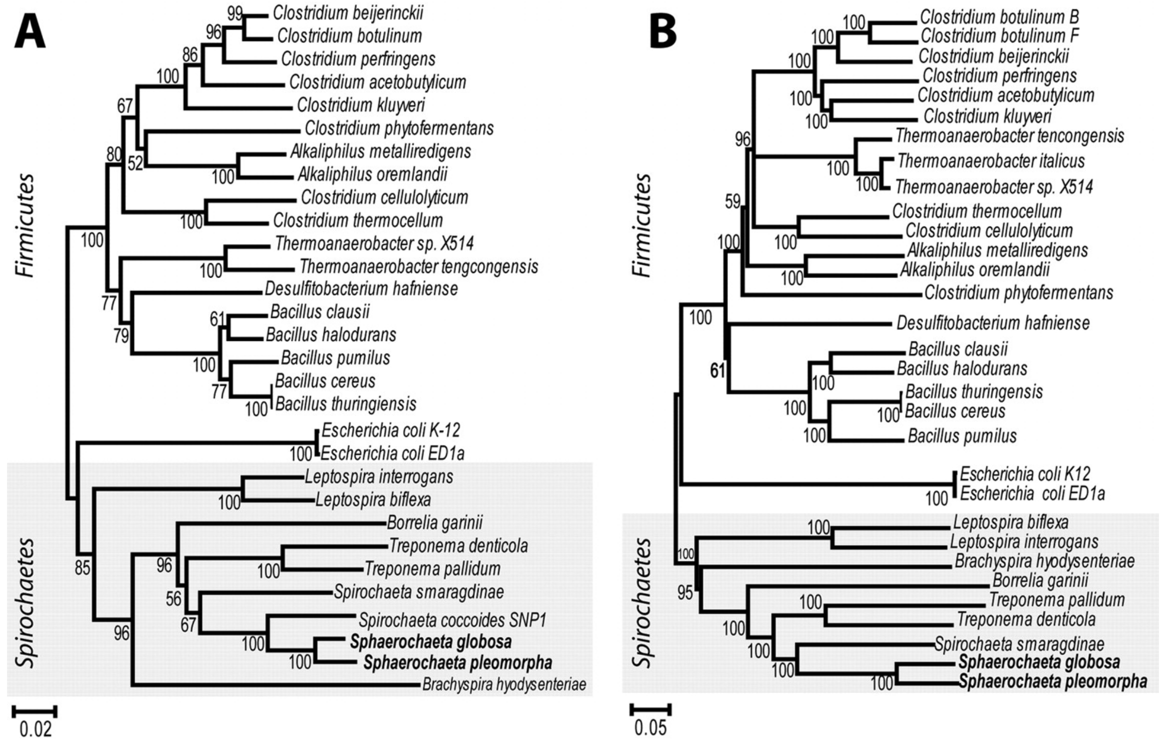


Figure 5.2 Phylogenetic affiliation of *Sphaerochaeta globosa* and *Sphaerochaeta pleomorpha*. Neighbor Joining phylogenetic trees of *Sphaerochaeta* and selected bacterial species based on 16S rRNA gene sequences (A) and the concatenated alignment of 43 single-copy informational gene sequences (B) are shown. Values on the nodes represent bootstrap support from 1,000 replicates. The scale bar represents the number of nucleotide (A) or amino acid (B) substitutions per site.

Table 5.2 List of the 43 informational genes used in the genome phylogeny shown in Figure 5.2B.

	Sphaerochaeta	Sphaerochaeta	Spirochaeta	Spirochete	Spirochete	Spirochete	Spirochete	Spirochete	Spirochete	Spirochete	Bacillus	Bacillus	Bacillus	Bacillus	Bacillus	Bacillus	Clostridiales	Clostridiales
Product	Strain Buddy	Strain Grapes	S.Smaragdinae	Borrelia_garinii	Brachyspira_hyodysenteriae	Leptospira_biflexa_	Leptospira_interrogans_serovar_Copen	Treponema_denticola_	Treponema_pallidum_SS14_	Treponema_pallidum_	Bacillus_cereus	Bacillus_clausii	Bacillus_halodurans	Bacillus_pumilus	Bacillus_subtilus	Bacillus_thuringensis	Alkaliphilus_metalliredigens_QYMF_	Alkaliphilus_oremlandii_OhILAs
enolase	SpiBuddyDRAFT_0104	SpiGrapesDRAFT_1256	gi_170083336_	gi_51598595_	gi_225620799_	gi_189911182_	gi_45657813_	gi_42526461_	gi_189026041_	gi_15639803_	gi_42784284_	gi_56964781_	gi_15616118_	gi_157693805_	gi_157693805_	gi_49481799_	gi_150391308_	gi_158320009_
ribosomal_protein_L13	SpiBuddyDRAFT_0155	SpiGrapesDRAFT_1103	gi_170083342_	gi_51598597_	gi_225621084_	gi_189910877_	gi_45656655_	gi_42526364_	gi_189026248_	gi_15640009_	gi_42779224_	gi_56961965_	gi_15612731_	gi_157690933_	gi_157690933_	gi_49481687_	gi_150392125_	gi_158319578_
ribosomal_protein_S9	SpiBuddyDRAFT_0156	SpiGrapesDRAFT_1102	gi_170083341_	gi_51598596_	gi_225621083_	gi_189910878_	gi_45656656_	gi_42526365_	gi_189026247_	gi_15640008_	gi_42779225_	gi_56961966_	gi_15612732_	gi_157690934_	gi_157690934_	gi_49481688_	gi_150392124_	gi_158319579_
GTP-binding_protein_YchF	SpiBuddyDRAFT_0237	SpiGrapesDRAFT_1023	gi_45658905_	gi_51598496_	gi_225620019_	gi_189911580_	gi_45657485_	gi_42525821_	gi_189025358_	gi_15639118_	gi_42784671_	gi_56965865_	gi_15616613_	gi_157694463_	gi_157694463_	gi_49480471_	gi_150390656_	gi_158320397_
translation_initiation_factor_IF-1	SpiBuddyDRAFT_0254	SpiGrapesDRAFT_997	gi_45656364_	gi_51598430_	gi_225621048_	gi_189911418_	gi_45658681_	gi_42527590_	gi_189025331_	gi_15639091_	gi_42779214_	gi_56961955_	gi_15612721_	gi_157690923_	gi_157690923_	gi_49476717_	gi_150392136_	gi_158319567_
Glycine_hydroxymethyltransferase	SpiBuddyDRAFT_0256	SpiGrapesDRAFT_995	gi_45658020_	gi_51598854_	gi_225621280_	gi_189910524_	gi_45658181_	gi_42528168_	gi_189025562_	gi_15639320_	gi_42784486_	gi_56965621_	gi_15616327_	gi_157694086_	gi_157694086_	gi_49481164_	gi_150390129_	gi_158320729_
GTP-binding_protein_LepA	SpiBuddyDRAFT_0419	SpiGrapesDRAFT_705	gi_45657708_	gi_51598351_	gi_225620814_	gi_189911181_	gi_45657863_	gi_42527398_	gi_189025740_	gi_15639501_	gi_42783446_	gi_56963420_	gi_15613905_	gi_157693051_	gi_157693051_	gi_49481363_	gi_150390802_	gi_158320264_
leucyl-tRNA_synthetase	SpiBuddyDRAFT_0562	SpiGrapesDRAFT_1485	gi_45656193_	gi_51598511_	gi_225621162_	gi_189909823_	gi_45656415_	gi_42527840_	gi_189025813_	gi_15639574_	gi_42783928_	gi_56964635_	gi_15615843_	gi_157693425_	gi_157693425_	gi_49481694_	gi_150390081_	gi_158321607_
ATP-dependent_DNA_helicase_RecG	SpiBuddyDRAFT_0744	SpiGrapesDRAFT_1301	gi_45657206_	gi_51598833_	gi_225621185_	gi_189909907_	gi_45658972_	gi_42528079_	gi_189025912_	gi_15639674_	gi_42782947_	gi_56964070_	gi_15615058_	gi_157692267_	gi_157692267_	gi_49478911_	gi_150390531_	gi_158320474_
adenylate_kinase	SpiBuddyDRAFT_0862	SpiGrapesDRAFT_1812	gi_45657410_	gi_51598674_	gi_225619885_	gi_189911419_	gi_45658682_	gi_42526622_	gi_189025822_	gi_15639583_	gi_42779212_	gi_56961953_	gi_15612718_	gi_157690921_	gi_157690921_	gi_49479237_	gi_150392139_	gi_158319564_
SsrA-binding_protein	SpiBuddyDRAFT_0921	SpiGrapesDRAFT_1734	gi_45657885_	gi_51598296_	gi_225620957_	gi_189910900_	gi_45658263_	gi_42526972_	gi_189025417_	gi_161579587_	gi_42784277_	gi_56964766_	gi_15616114_	gi_157693782_	gi_157693782_	gi_49481124_	gi_150391302_	gi_158320014_
DNA_mismatch_repair_protein_MutS	SpiBuddyDRAFT_0933	SpiGrapesDRAFT_1722	gi_45657900_	gi_51599049_	gi_225620061_	gi_189910358_	gi_45657640_	gi_42528103_	gi_189025561_	gi_15639319_	gi_42782853_	gi_56963952_	gi_15614932_	gi_157692388_	gi_157692388_	gi_49478366_	gi_150390307_	gi_158320586_
transcription termination/antitermination fa	SpiBuddyDRAFT_1011	SpiGrapesDRAFT_2397	gi_45656832_	gi_51598651_	gi_225621220_	gi_189911452_	gi_45656642_	gi_42527927_	gi_189025469_	gi_15639228_	gi_42779177_	gi_56961918_	gi_15612681_	gi_157690884_	gi_157690884_	gi_49476713_	gi_150392173_	gi_158319530_
ribosomal_protein_L1	SpiBuddyDRAFT_1013	SpiGrapesDRAFT_2395	gi_45656834_	gi_51598649_	gi_225621222_	gi_189911450_	gi_45656644_	gi_42527925_	gi_189025471_	gi_15639230_	gi_42779179_	gi_56961920_	gi_15612683_	gi_157690886_	gi_157690886_	gi_49479291_	gi_150392171_	gi_158319532_
ribosomal_protein_L7/L12	SpiBuddyDRAFT_1015	SpiGrapesDRAFT_2393	gi_45656836_	gi_51598647_	gi_225621224_	gi_189911448_	gi_45656646_	gi_42527923_	gi_189025473_	gi_15639232_	gi_42779181_	gi_56961922_	gi_15612685_	gi_157690888_	gi_157690888_	gi_49479289_	gi_150392169_	gi_158319534_
DNA-directed RNA polymerase, beta subu	SpiBuddyDRAFT_1016	SpiGrapesDRAFT_2392	gi_45656837_	gi_51598646_	gi_225621225_	gi_189911447_	gi_45656647_	gi_42527922_	gi_189025474_	gi_15639233_	gi_42779183_	gi_56961924_	gi_15612689_	gi_157690891_	gi_157690891_	gi_49479286_	gi_150392168_	gi_158319535_
DNA-directed RNA polymerase, beta' subu	SpiBuddyDRAFT_1017	SpiGrapesDRAFT_2391	gi_45656838_	gi_51598645_	gi_225621226_	gi_189911446_	gi_45656648_	gi_42527921_	gi_189025475_	gi_15639234_	gi_42779184_	gi_56961925_	gi_15612690_	gi_157690892_	gi_157690892_	gi_49479285_	gi_150392167_	gi_158319536_
ribosomal_protein_S7	SpiBuddyDRAFT_1020	SpiGrapesDRAFT_2389	gi_45656840_	gi_51598643_	gi_225621024_	gi_189911444_	gi_45656650_	gi_42526558_	gi_189025477_	gi_15639236_	gi_42779187_	gi_56961928_	gi_15612693_	gi_157690895_	gi_157690895_	gi_49479282_	gi_150392164_	gi_158319539_
ribosomal_protein_S10	SpiBuddyDRAFT_1022	SpiGrapesDRAFT_2387	gi_45656842_	gi_51598731_	gi_225621027_	gi_189911441_	gi_45658704_	gi_42526278_	gi_189025421_	gi_15639181_	gi_42779190_	gi_56961931_	gi_15612696_	gi_157690899_	gi_157690899_	gi_49479272_	gi_150392161_	gi_158319542_
50S_ribosomal_protein_L3	SpiBuddyDRAFT_1023	SpiGrapesDRAFT_2386	gi_45656843_	gi_51598732_	gi_225621028_	gi_189911440_	gi_45658703_	gi_42526279_	gi_189025422_	gi_15639182_	gi_42779191_	gi_56961932_	gi_15612697_	gi_157690900_	gi_157690900_	gi_49479271_	gi_150392160_	gi_158319543_
ribosomal_protein_L4/L1e	SpiBuddyDRAFT_1024	SpiGrapesDRAFT_2385	gi_45656844_	gi_51598733_	gi_225621029_	gi_189911439_	gi_45658702_	gi_42526280_	gi_189025423_	gi_15639183_	gi_42779192_	gi_56961933_	gi_15612698_	gi_157690901_	gi_157690901_	gi_49476718_	gi_150392159_	gi_158319544_
Ribosomal_protein_L25/L23	SpiBuddyDRAFT_1025	SpiGrapesDRAFT_2384	gi_45656845_	gi_51598735_	gi_225621030_	gi_189911438_	gi_45658701_	gi_42526281_	gi_189025424_	gi_15639184_	gi_42779193_	gi_56961934_	gi_15612699_	gi_157690902_	gi_157690902_	gi_49479267_	gi_150392158_	gi_158319545_
ribosomal_protein_L2	SpiBuddyDRAFT_1026	SpiGrapesDRAFT_2383	gi_45656846_	gi_51598736_	gi_225621031_	gi_189911437_	gi_45658700_	gi_42526282_	gi_189025425_	gi_15639185_	gi_42779194_	gi_56961935_	gi_15612700_	gi_157690903_	gi_157690903_	gi_49479266_	gi_150392157_	gi_158319546_
ribosomal_protein_S19	SpiBuddyDRAFT_1027	SpiGrapesDRAFT_2382	gi_45656847_	gi_51598737_	gi_225621032_	gi_189911436_	gi_45658699_	gi_42526283_	gi_189025426_	gi_15639186_	gi_42779195_	gi_56961936_	gi_15612701_	gi_157690904_	gi_157690904_	gi_49479246_	gi_150392156_	gi_158319547_
ribosomal_protein_L16	SpiBuddyDRAFT_1030	SpiGrapesDRAFT_2379	gi_45656850_	gi_51598740_	gi_225621035_	gi_189911433_	gi_45658696_	gi_42526286_	gi_189025429_	gi_15639189_	gi_42779198_	gi_56961939_	gi_15612704_	gi_157690907_	gi_157690907_	gi_49476719_	gi_150392153_	gi_158319550_
ribosomal_protein_L14	SpiBuddyDRAFT_1033	SpiGrapesDRAFT_2376	gi_45656853_	gi_51598743_	gi_225621038_	gi_189911430_	gi_161621772_	gi_42526289_	gi_189025432_	gi_15639192_	gi_42779201_	gi_56961942_	gi_15612707_	gi_157690910_	gi_157690910_	gi_49481666_	gi_150392150_	gi_158319553_
ribosomal_protein_L5	SpiBuddyDRAFT_1035	SpiGrapesDRAFT_2374	gi_45656855_	gi_51598745_	gi_225621040_	gi_189911428_	gi_45658691_	gi_42526291_	gi_189025434_	gi_15639194_	gi_42779203_	gi_56961944_	gi_15612709_	gi_157690912_	gi_157690912_	gi_49481670_	gi_150392148_	gi_158319555_
ribosomal_protein_S8	SpiBuddyDRAFT_1037	SpiGrapesDRAFT_2372	gi_45656857_	gi_51598747_	gi_225621041_	gi_189911426_	gi_45658689_	gi_42526293_	gi_189025436_	gi_15639196_	gi_42779205_	gi_56961946_	gi_15612711_	gi_157690914_	gi_157690914_	gi_49481672_	gi_150392146_	gi_158319557_
ribosomal_protein_L6	SpiBuddyDRAFT_1038	SpiGrapesDRAFT_2371	gi_45656858_	gi_51598748_	gi_225621042_	gi_189911425_	gi_45658688_	gi_42526294_	gi_189025437_	gi_15639197_	gi_42779206_	gi_56961947_	gi_15612712_	gi_157690915_	gi_157690915_	gi_49481676_	gi_150392145_	gi_158319558_
ribosomal_protein_L18	SpiBuddyDRAFT_1039	SpiGrapesDRAFT_2370	gi_45656859_	gi_51598749_	gi_225621043_	gi_189911424_	gi_45658687_	gi_42526295_	gi_189025438_	gi_15639198_	gi_42779207_	gi_56961948_	gi_15612713_	gi_157690916_	gi_157690916_	gi_49481677_	gi_150392144_	gi_158319559_
ribosomal_protein_S5	SpiBuddyDRAFT_1040	SpiGrapesDRAFT_2369	gi_45656860_	gi_51598750_	gi_225621044_	gi_189911423_	gi_45658686_	gi_42526296_	gi_189025439_	gi_15639199_	gi_42779208_	gi_56961949_	gi_15612714_	gi_157690917_	gi_157690917_	gi_49481680_	gi_150392143_	gi_158319560_
30S_ribosomal_protein_S13	SpiBuddyDRAFT_1045	SpiGrapesDRAFT_2364	gi_45656865_	gi_51598755_	gi_225621049_	gi_189911416_	gi_45658679_	gi_42526301_	gi_189025444_	gi_15639203_	gi_42779216_	gi_56961957_	gi_15612723_	gi_157690925_	gi_157690925_	gi_49481683_	gi_150392134_	gi_158319569_
30S_ribosomal_protein_S11	SpiBuddyDRAFT_1046	SpiGrapesDRAFT_2363	gi_45656866_	gi_51598756_	gi_225621050_	gi_189911415_	gi_45658678_	gi_42526302_	gi_189025445_	gi_15639204_	gi_42779217_	gi_56961958_	gi_15612724_	gi_157690926_	gi_157690926_	gi_49481684_	gi_150392133_	gi_158319570_
DNA-directed RNA polymerase, alpha sub	SpiBuddyDRAFT_1048	SpiGrapesDRAFT_2361	gi_45656868_	gi_51598757_	gi_225621052_	gi_189911413_	gi_45658676_	gi_42526303_	gi_189025446_	gi_15639205_	gi_42779218_	gi_56961959_	gi_15612725_	gi_157690927_	gi_157690927_	gi_49481737_	gi_150392131_	gi_158319572_
ribosome-associated_GTPase_EngA	SpiBuddyDRAFT_1055	SpiGrapesDRAFT_2354	gi_161621771_	gi_51598763_	gi_225620645_	gi_189910901_	gi_45658262_	gi_42525602_	gi_189025914_	gi_15639676_	gi_42780705_	gi_161349991_	gi_15614201_	gi_157692783_	gi_157692783_	gi_49477254_	gi_150390617_	gi_158320432_
excinuclease_ABC,_B_subunit	SpiBuddyDRAFT_1412	SpiGrapesDRAFT_2634	gi_45658200_	gi_51599087_	gi_225619311_	gi_189910370_	gi_45658769_	gi_42526877_	gi_189025350_	gi_15639110_	gi_42784318_	gi_56964823_	gi_15616157_	gi_157693900_	gi_157693900_	gi_49481649_	gi_150391818_	gi_158321280_
ribosomal_protein_L20	SpiBuddyDRAFT_1435	SpiGrapesDRAFT_2661	gi_45658456_	gi_51598449_	gi_225619665_	gi_189911842_	gi_45658303_	gi_42527657_	gi_189026072_	gi_15639834_	gi_42783750_	gi_56964452_	gi_15615700_	gi_157693290_	gi_157693290_	gi_49481332_	gi_150389422_	gi_158321003_
cysteinyl-tRNA_synthetase	SpiBuddyDRAFT_1632	SpiGrapesDRAFT_2012	gi_45658028_	gi_51598852_	gi_225620434_	gi_189911202_	gi_45657880_	gi_42525611_	gi_189025325_	gi_15639085_	gi_42779170_	gi_56961911_	gi_15612674_	gi_157690877_	gi_157690877_	gi_49481691_	gi_150392183_	gi_158319520_
valyl-tRNA_synthetase	SpiBuddyDRAFT_1635	SpiGrapesDRAFT_2009	gi_45658021_	gi_51598989_	gi_225619103_	gi_189909811_	gi_45656371_	gi_42526872_	gi_189026258_	gi_15640019_	gi_42783595_	gi_56964386_	gi_15615600_	gi_157693210_	gi_157693210_	gi_49478642_	gi_150389212_	gi_158321063_
excinuclease_ABC,_C_subunit	SpiBuddyDRAFT_2285	SpiGrapesDRAFT_494	gi_170083653_	gi_51598713_	gi_225619074_	gi_189911545_	gi_45657622_	gi_42527978_	gi_189025703_	gi_15639463_	gi_42783693_	gi_56964432_	gi_15615659_	gi_157693266_	gi_157693266_	gi_49481523_	gi_150391815_	gi_158321277_
ribosome_recycling_factor	SpiBuddyDRAFT_2474	SpiGrapesDRAFT_3009	gi_45658723_	gi_51598384_	gi_225619145_	gi_189912041_	gi_45656744_	gi_42527846_	gi_189025830_	gi_15639592_	gi_42782915_	gi_56964004_	gi_15614987_	gi_157692331_	gi_157692331_	gi_49478397_	gi_150390448_	gi_158320548_
undecaprenyl_diphosphate_synthase	SpiBuddyDRAFT_2475	SpiGrapesDRAFT_3010	gi_45658724_	gi_51598383_	gi_225619144_	gi_189912040_	gi_45656745_	gi_42527845_	gi_189025829_	gi_15639591_	gi_42782914_	gi_56964003_	gi_15614986_	gi_157692332_	gi_157692332_	gi_49478396_	gi_150390446_	gi_158320549_
tRNA (guanine-N1)-methyltransferase	SpiBuddyDRAFT 2654	SpiGrapesDRAFT 2871	gi 45657755	gi 51598952	gi 225620838	gi 189911144	gi 45657431	gi 42526396	gi 189026131	gi 15639893	gi 42782933	gi 56964054	gi 15615042	gi 157692282	gi 157692282	gi 49478409	gi 150390504	gi 158320500

5.4.2 Motility and Chemotaxis

Typical spirochetal flagella are composed of about thirty different proteins [191], and about a dozen additional regulatory or sensory proteins have been demonstrated to directly interact with flagellar proteins, such as the chemotaxis proteins encoded on the *che* operon [172]. To determine whether or not the *Sphaerochaeta* genomes possess motility genes, we queried the sequences of the *Treponema pallidum* flagellar and chemotaxis proteins against the *S. pleomorpha* and *S. globosa* genome sequences. Although the *T. pallidum* protein sequences had clear orthologs in all available spirochetal genomes, none of the chemotaxis and motility related proteins were present in the *S. pleomorpha* or *S. globosa* genomes (Fig. 5.3B). Incomplete sequencing, assembly errors or low sequence similarity did not present plausible explanations for these results since the flagellar genes are typically encoded in three distinct, large gene clusters, each 20-30 kbp long, and it is not likely that such clusters were missed in genome sequencing and annotation. Consistent with these interpretations, all informational genes encoding ribosomal proteins and RNA and DNA polymerases were recovered in the assembled genome sequences. These results were consistent with previous microscopic observations and corroborated that the *Sphaerochaeta*-characteristic spherical morphology is related to the absence of axial flagella [183].

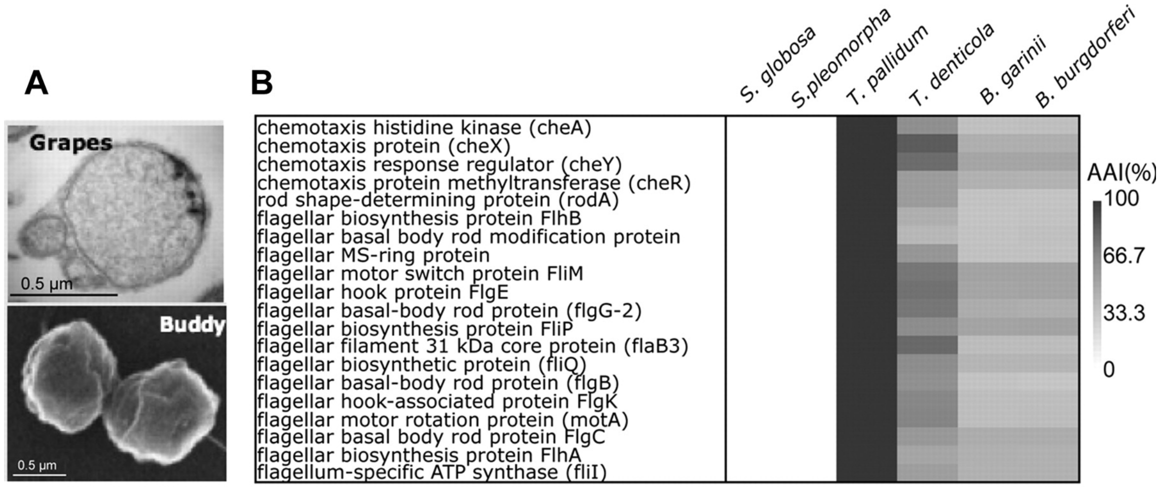


Figure 5.3 Absence of flagellar and chemotaxis genes in *Sphaerochaeta* genomes. Transmission electron micrograph showing the non-spiral shape of *S. globosa* strain Buddy and *S. pleomorpha* strain Grapes cells (A). Heatmap showing the presence/absence and the level of amino acid identity (see scale) of *Treponema pallidum* chemotaxis, flagellar assembly and locomotion gene homologs in selected spirochetal genomes (B).

5.4.3 A Unique Cell Wall Structure

Our analyses revealed additional features in *Sphaerochaeta* that are unusual among spirochetes and Gram-negative bacteria in general, and are probably linked to the lack of axial flagella. Both *Sphaerochaeta* genomes encode all genes required for peptidoglycan biosynthesis, and electron microscopy verified the presence of a cell wall in growing cells [183]; however, the genomes lack genes for penicillin-binding proteins (PBPs). PBPs catalyze the formation of linear glycan chains (transglycosylation) during cell elongation and the transpeptidation of murein glycan chains (Table 5.3), which confers rigidity to the cell wall [192, 193]. Consequently, *Sphaerochaeta* spp. are resistant to β-lactam antibiotics (ampicillin up to 250 μg/mL, which was the highest concentration tested). In Gram-negative bacteria without antibiotic resistance mechanisms, including clinical spirochetes, β-lactam antibiotics block PBP functionality resulting in cell lysis. Often, β-lactam-treated, cell wall-deficient cells can be maintained in isotonic growth media as so-called L-forms with characteristic spherical morphologies [194-196]. While *Sphaerochaeta* spp. cells occur in spherical morphologies (Fig. 5.3A), they possess a cell wall, grow in defined hypertonic and hypotonic media without the addition of osmotic stabilizers [183], and are not L-forms. It is conceivable that a rigid cell wall is required for anchoring of the axial flagella. Thus, the absence of both axial flagella and PBPs presumably explain the atypical spirochete morphology of the *Sphaerochaeta*. The loss of the flagella and PBPs genes occurred likely in the ancestor of the *Sphaerochaeta*, since both members of the genus lack these genes.

Table 5.3 *Sphaerochaeta* genomes lack several universal genes encoding penicillin-binding proteins (PBPs). Four types of penicillin-binding proteins (PBPs) and three low molecular weight proteins (pbp4-pbp6) involved in cell wall biosynthesis are shown. Lack of pbp1 produces unstable cells that lyse easily, absence of pbp2 leads to large, osmotically stable spherical cell forms, lack of pbp3 causes filamentation of cells, and lack of pbp4-6 decreases cell wall rigidity [for a comprehensive review, see [197]]. X denotes the presence of the corresponding gene.

Genes	PBP	Description	*Sphaerochaeta* spp.	*Spirochaeta* spp.	*Leptospira* spp.	*Borrelia* spp.	*Treponema* spp.	*Brachispira* spp.	*Clostridium* spp.
pbpC	pbp 1C	peptidoglycan glycosyltransferase		X	X	X	X	X	
mrcA/B	pbp 1A	penicillin-binding protein, 1A family		X	X	X	X	X	X
mrdA	pbp 2	penicillin-binding protein 2		X	X	X	X	X	X
pbpB	pbp 3	cell division protein FtsI	X	X	X	X	X	X	
Mvin	no pbp	integral membrane protein MviN	X	X	X	X	X	X	X
dacA/C/D	pbp 4/6	D-alanyl-D-alanine carboxypeptidase		X	X	X	X	X	X

5.4.4 Extensive Gene Acquisition From Gram-Positive Bacteria

Searching all *Sphaerochaeta* protein sequences against the non-redundant (nr) protein database of GenBank revealed that ~700 of the protein-encoding genes had best matches to genes of the *Clostridiales*, ~700 to genes of the *Spirochaetes*, and ~100 to genes of the *Bacilli* (Fig. 5.4). Consistent with the best match results, *S. pleomorpha* and *S. globosa* exclusively shared more unique genes with *Clostridia* than with other *Spirochaetes* (~110 vs. ~70 genes, respectively). Both species exclusively shared a substantial number of unique genes with *Bacilli* (*Firmicutes*, 25 genes) and *Escherichia* (*g-Proteobacteria*, 60 and 10 genes for *S. pleomorpha* and *S. globosa*, respectively) (Fig. 5.5B). Functional analysis based on the COG database showed that the spirochete-like genes of *Sphaerochaeta* were mostly associated with informational categories, e.g., transcription and translation, whereas the clostridia-like genes were highly enriched in metabolic functions, e.g., carbohydrate and amino acid metabolism and transport (Fig. 5.4 and 5.6). Several of the carbohydrate and amino acid metabolism genes, such as the multidomain glutamate-synthase (SpiBuddy_0108-0113) and genes related to polysaccharide biosynthesis (SpiBuddy_0254-0259), were found in large gene clusters, indicating that their acquisition likely occurred in single HGT events. Interestingly, many of the clostridia-like genes had high sequence identity to their clostridial homologs (> 70% amino acid identity), even though these genes did not encode informational proteins (e.g., ribosomal proteins and RNA/DNA polymerases). While informational genes tend to show high levels of sequence conservation, much lower sequence conservation was expected for (not horizontally transferred) metabolic genes shared across phyla, revealing that some of the genetic exchange events between *Sphaerochaeta* and *Clostridiales* occurred relatively recently.

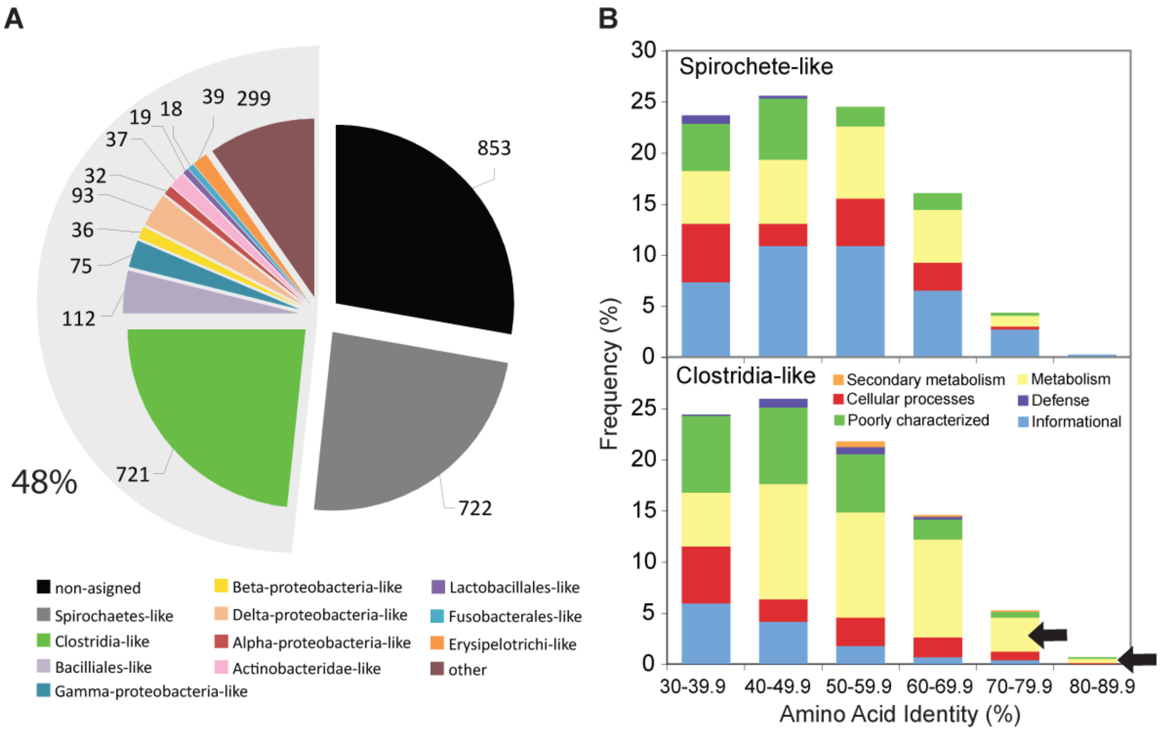


Figure 5.4 Distribution of best BLAST matches of *Sphaerochaeta globosa* protein sequences. Best match analysis against all publicly available complete genomes reveals that the *Sphaerochaeta globosa* genome has as many best matches in *Clostridiales* (clostridia-like) as in *Spirochaetes* (spirochete-like) (A). The histograms show that the spirochete-like genes are enriched in informational functions, while the clostridia-like genes are enriched in metabolic functions (based on assignment of genes to the COG database) (B). Arrows on B highlight the high identity of several clostridia-like metabolic genes (>70% amino acid identity).

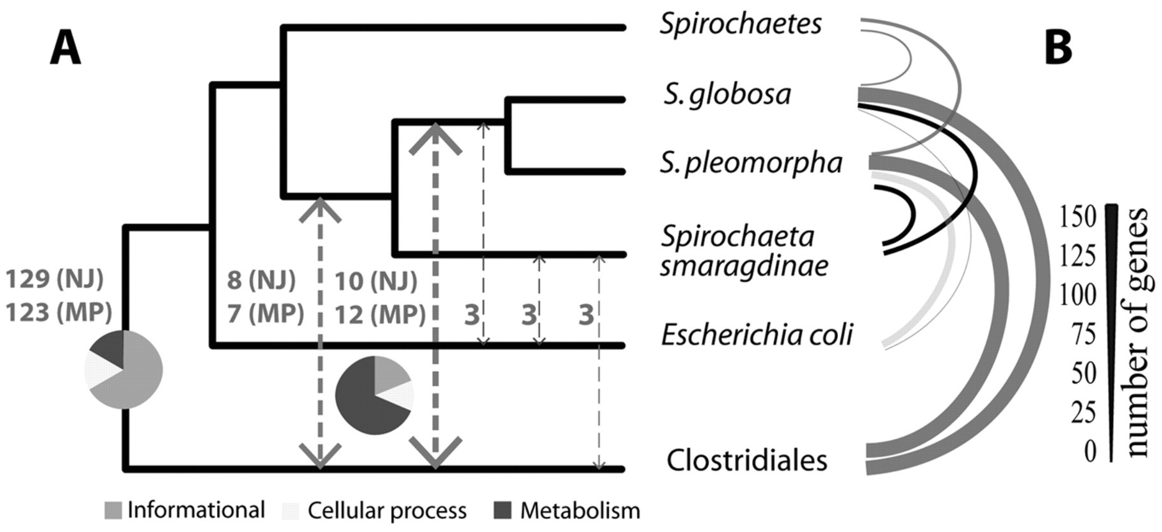


Figure 5.5 Horizontal gene transfer between *Sphaerochaeta* spp. and *Clostridiales*. The cladogram depicts the 16S rRNA gene phylogeny. Arrows connecting branches represent cases of HGT (A); the numbers next to the arrows indicate the number of genes exchanged (out of a total of 178 genes examined). Pie charts show the distribution of the genes in major COG functional categories (see figure key for category designation by color). Orthologous genes shared exclusively between *Sphaerochaeta* and other taxa are graphically represented by arced lines (B). The thickness of the line is proportional to the number of shared genes (see scale bar).

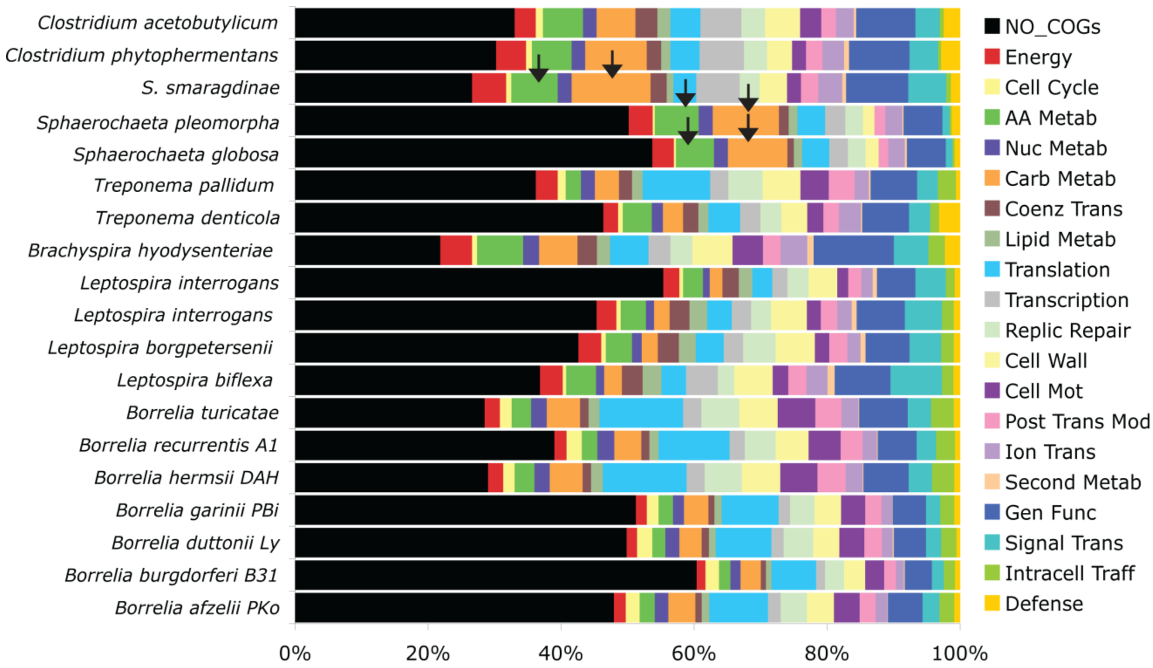


Figure 5.6 Functional characterization of selected spirochetal and clostridial genomes based on the COG database. All genes encoded on the genomes were assigned to the COG database and the graph shows the relative abundance of COGs categories in each genome. Arrows mark the relative enrichment of genes for carbohydrate and amino acid metabolism in *Spirochaeta smaragdinae*, *Sphaerochaeta globosa* and *Sphaerochaeta pleomorpha* genomes.

Homology-based (best-hit) bioinformatic analyses are inherently prone to artifacts including uneven numbers of representative genomes in the database, disparate % G+C content, different rates of evolution, multidomain proteins and gene loss [198, 199]. To provide further insights into the genome fluidity of *Sphaerochaeta* and the inter-phylum HGT events, we performed a detailed phylogenetic analysis of 223 orthologous proteins that had at least one homologous sequence in each of the taxa evaluated (i.e., *Sphaerochaeta* spp., *S. smaragdinae*, other *Spirochaetes*, *E. coli* and *Clostridiales*). We evaluated genetic exchange events based on embedded quartet decomposition analysis [105], using both maximum parsimony (MP) and Neighbor Joining (NJ) methods and 178 trees with at least 50% bootstrap support in all branches. The gene set contributing to the trees was biased towards informational functions; hence, it was not surprising that the most frequent topology obtained [123 trees (MP) and 129 trees (NJ)] was congruent with the 16S rRNA gene-based topology, denoting no inter-phylum genetic exchange. Nonetheless, the analysis also provided trees with topologies consistent with genetic exchange between *Clostridiales* and *Sphaerochaeta*, and identified 19 (MP) and 18 (NJ) genes (i.e., ~10 % of the total trees evaluated) that were most likely subjected to inter-phylum HGT. This gene set was enriched in genes encoding metabolic functions, e.g., carbohydrate metabolism (Fig. 5.5A). About half of the 19 (MP) trees were consistent with genetic exchange between *Clostridiales* and the ancestor of both *S. smaragdinae* and *Sphaerochaeta*, while the other trees were consistent with exchange between the ancestor of *Clostridiales* and *Sphaerochaeta* (more recent events; Fig. 5.5). The phylogenetic distribution of the genes exchanged between *Clostridiales* and *Sphaerochaeta* in other spirochetes and Gram-positive bacteria (e.g., Fig. 5.7) suggested that members of the *Clostridiales* were predominantly the donors (>95% of the genes examined) in these genetic exchange events (unidirectional HGT). These findings corroborated those of the best-match analysis and collectively revealed that, with the exception of informational genes, inter-phylum HGT and gene loss (e.g., flagellar genes) have shaped more than half of the *Sphaerochaeta* genomes through evolutionary time.

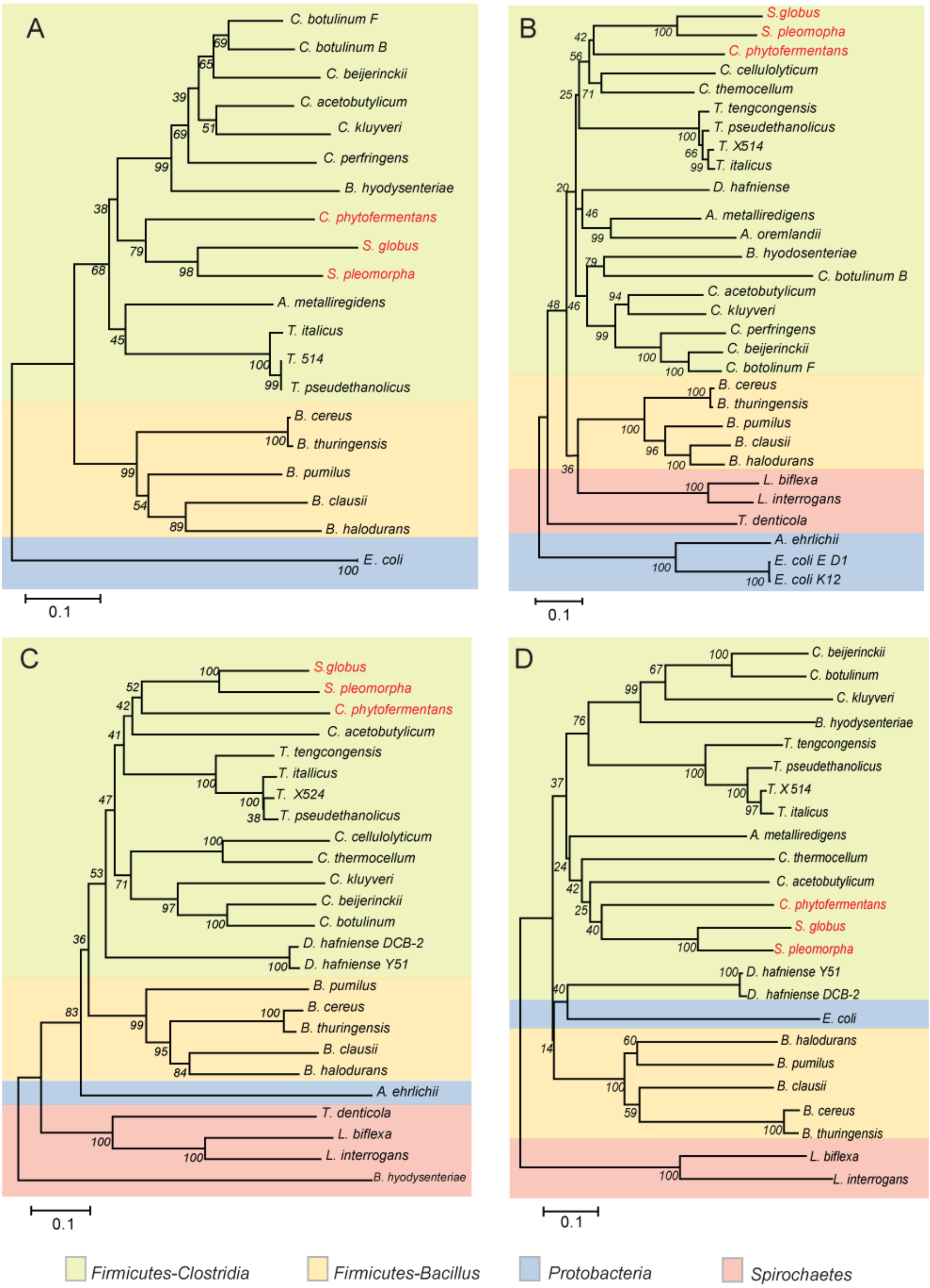


Figure 5.7 Phylogenetic analysis of genes exchanged between the ancestors of *Sphaerochaeta spp.* and *Clostridium phytofermentans*. Neighbor-Joining trees of four horizontally exchanged genes are shown. Values next to the branches denote the bootstrap support from 1,000 replicates. Genes include: phosphoribosyl-amino-

imidazole-succino-carboxamide synthase (A), amidophosphoribosyl transferase (B), arginine biosynthesis bifunctional protein (C), and N-acetyl gamma-glutamyl phosphate reductase (D). The genes in A and B and in C and D are possible expressed as a single cistronic mRNA on the *S. globosa* and *S. pleomorpha* genomes.

5.4.5 How Unique Is The Case of *Sphaerochaeta-Clostridiales* Gene Transfer?

We evaluated how frequently such a high level of inter-phylum gene transfer as that observed between *Clostridiales* and *Sphaerochaeta* genomes occurs within the prokaryotic domain. To this end, the ratio of the number of genes of a reference genome with best matches in a genome of a different phylum vs. the number genes of the reference genome with best matches to a genome of the same phylum was determined. To account for differences in the coverage of phyla with sequenced representatives, the analysis was performed using three genomes at a time (two of the same phylum and one of a different phylum). Further, only genomes of the same phylum that showed similar genetic relatedness among them, measured by the genome-aggregate average amino-acid identity - or gAAI - [190], to that between *Sphaerochaeta* and selected *Spirochaete* genomes, i.e., *Leptospira* (48% gAAI) and *Treponema* (52% gAAI) genomes, were compared. This strategy sidesteps the limitation that the number of genes shared between any two genomes depends on the genetic relatedness among the genomes [Fig. 5.8 and [101]], and thus, can affect estimates of the number of best-match genes and HGT. The compared sets represented 12 different bacterial and three archaeal phyla and 308 and 249 different genomes (150,022 and 86,516 unique three-genome sets) for the 48% and 52% gAAI set comparisons, respectively. The analysis revealed that the extent of genetic exchange between *Sphaerochaeta* and *Clostridiales* is highly uncommon relative to that occurring among other genomes, i.e., upper 99.74 and 99.99 percentiles for the 48% and 52% gAAI sets, respectively. Similar results were obtained when all genes in the genome or only the genes shared between the three genomes, which were enriched in conserved housekeeping functions, were evaluated (Fig. 5.1). Most clostridia-like genes in *Sphaerochaeta* genomes had best matches within a phylogenetically narrow group of clostridia that included fermenters such as *Clostridium saccharolyticum* and *Clostridium phytofermentans* associated with anaerobic organic matter decomposition [200] and species such as *Eubacterium rectale [201]* and *Butyrivibrio proteoclasticus* [202] associated with the animal gut.

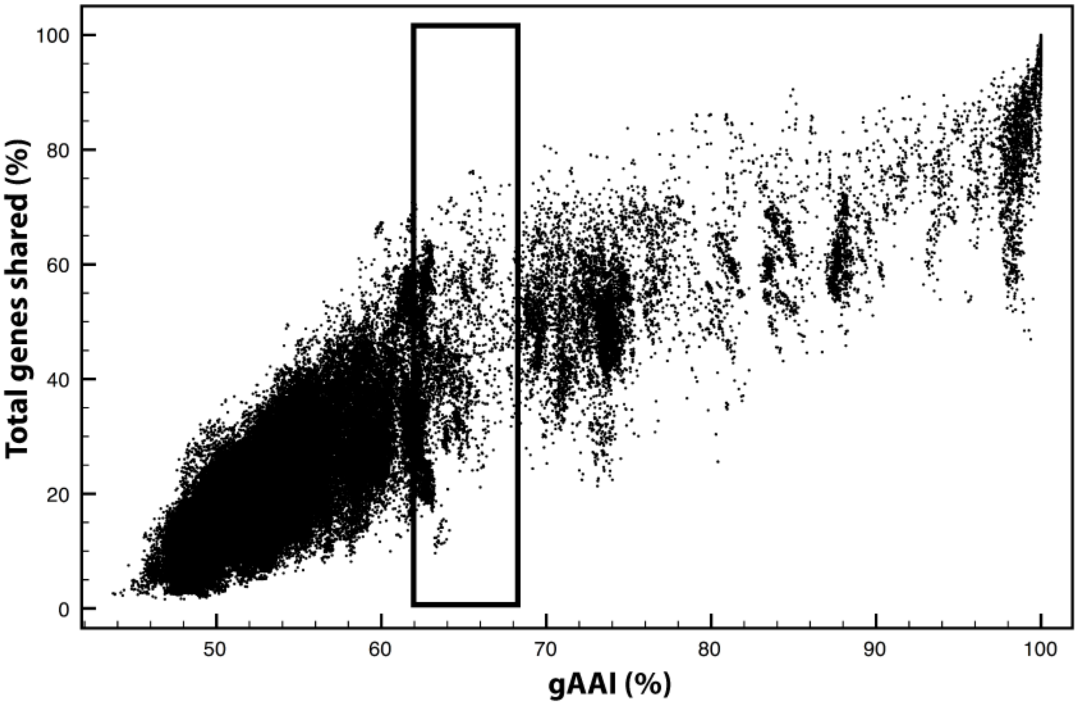


Figure 5.8 Correlation between shared genes and genetic relatedness for 1,445 completed genomes. All vs. all comparisons among the available complete genomes (as of June 2011) were performed and the graph shows the fraction of the total genes in the genome shared by a pair of genomes (y-axis) plotted against the genome-aggregate amino acid identity (gAAI) between the two genomes of the pair (x-axis). Note that there is a significant correlation between the two parameters. To avoid the influence of genetic relatedness on the number of genes shared, and thus, on the estimations of horizontal gene transfer (HGT), we focused on genomes that show similar gAAI values. The boxed region represents the 64 to 66% range, which approximated the 65% gAAI value among the *Sphaerochaeta* genomes and was used to define the groups in the three-group comparisons of HGT (Fig. 5.1).

5.4.6 Metabolic Properties of *Sphaerochaeta*

Metabolic genome reconstruction revealed that most of the central metabolic pathways were shared between *S. pleomorpha* and *S. globosa* (Fig. 5.9). The complete glycolytic and pentose phosphate pathways were present in both genomes. However, only a few, non-specific genes of the tricarboxylic acid cycle (TCA) were found - encoding citrate lyase, 2-oxoglutarate oxidoreductase and succinate dehydrogenase - revealing an incomplete TCA cycle. Another important feature of the two genomes was the absence of key components of respiratory electron transport chains such as c-type cytochromes and the ubiquinol-cytochrome C reductase (cytochrome bc_1 complex), corroborating physiological tests that *Sphaerochaeta* spp. do not respire. Instead, cellular energy (ATP, reducing power) capture in *Sphaerochaeta* relies on fermentation, a feature shared with several other spirochetes lacking respiratory functions, including members of the *Spirochaeta*, *Borrelia,* and *Treponema* genera [203]. In *Sphaerochaeta*, homo-fermentation of lactate and mixed acid fermentation appear to be the dominant fermentation pathways, producing lactate, acetate, formate, ethanol, H_2 and CO_2, consistent with physiological observations. The two *Sphaerochaeta* genomes also encode an assortment of transport proteins for the uptake and utilization of oligo- and mono-saccharides. Genes involved in carbohydrate metabolism and amino acid transport and metabolism are also over-represented relative to other spirochete genomes. In contrast, genes involved in signal transduction, intracellular trafficking, motility, posttranscriptional modification, and cell wall and membrane biogenesis are underrepresented in *Sphaerochaeta* genomes (Fig. 5.6). Consistent with an anaerobic lifestyle [179, 180], several genes related to oxidative stress and protection from reactive oxygen species were found in the *Sphaerochaeta* genomes. Genes encoding alkyl hydroperoxide reductase, superoxide dismutase, manganese superoxide dismutase, glutaredoxin, peroxidase, and catalase indicate that *Sphaerochaeta* spp. are adapted to environments with oxidative stress fluctuations.

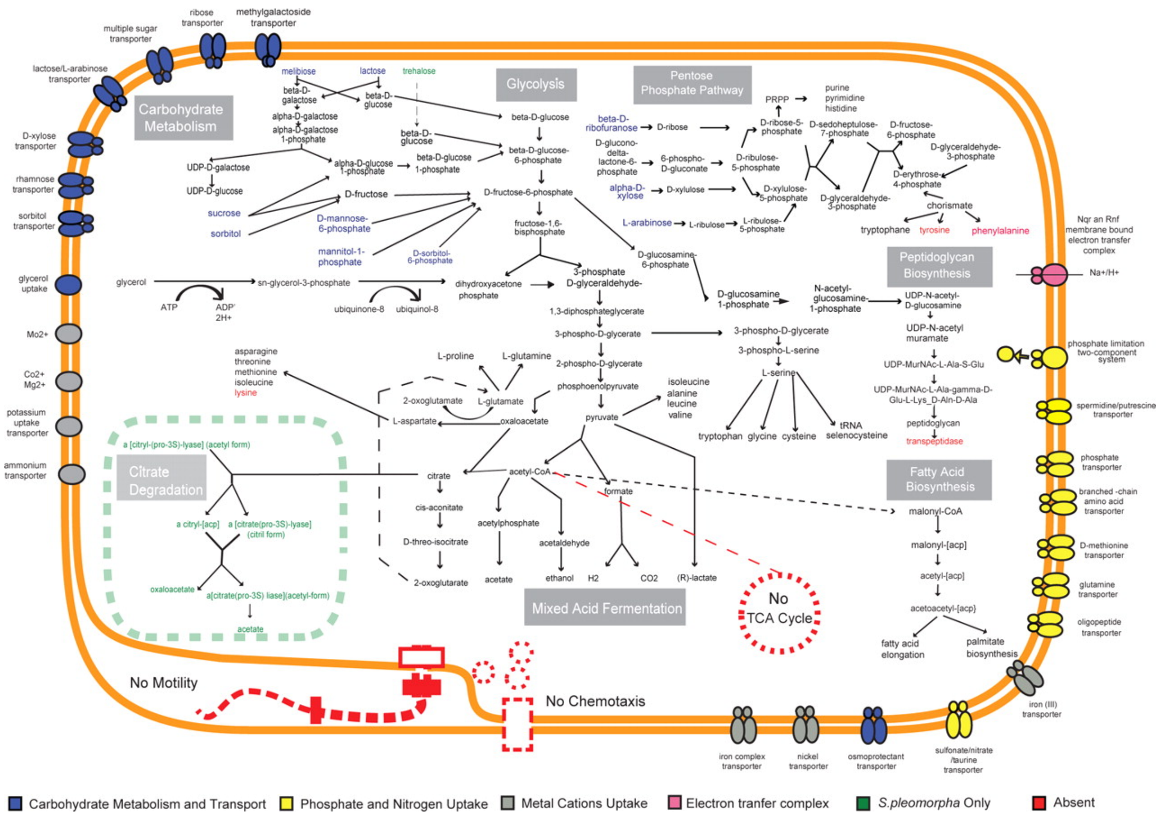


Figure 5.9 Overview of the metabolic pathways encoded in the *Sphaerochaeta pleomorpha* and *Sphaerochaeta globosa* genomes. The graph shows the primary energy generation pathways, diversity of carbohydrate metabolism pathways, biosynthesis genes for amino acids and fatty acids, and cell wall features encoded in both genomes. Pathways not found in the genomes such as those encoding flagellar and two component signal transduction systems related to motility are shown in red. The substrates and pathways found exclusively in *S. pleomorpha* are marked green. Transporters related to carbohydrate metabolism (in blue), metal ion transport and metabolism (in gray), and phosphate and nitrogen uptake (in yellow) are also shown.

Each *Sphaerochaeta* genome encodes about 850 species-specific genes (~25% of the genome), the majority of which represent genes of unknown or poorly characterized functions (Fig. 5.10). Nevertheless, our analyses identified a few genes or pathways that can functionally differentiate the two *Sphaerochaeta* species and might have implications for the habitat distribution of each species. For example, *S. pleomorpha*-specific genes were enriched in sugar metabolism and energy production functions, including genes for trehalose and maltose utilization and the complete (TCA cycle-independent) fermentation pathway for citrate utilization [204] (green-labeled genes in Fig. 5.9). Further, the genome of *S. pleomorpha* uniquely encodes several genes involved in cell wall and capsule formation such as phosphoheptose isomerase (capsular heptose biosynthesis) and the anhydro-N-acetylmuramic acid kinase (peptidoglycan recycling) [205]. These findings revealed that *S. pleomorpha* has both a potential for capsule formation and can use a wider range of carbohydrates than *S. globosa*, which are both consistent with experimental observations [183]. Almost all of the *S. globosa*-specific genes have unknown or poorly characterized functions.

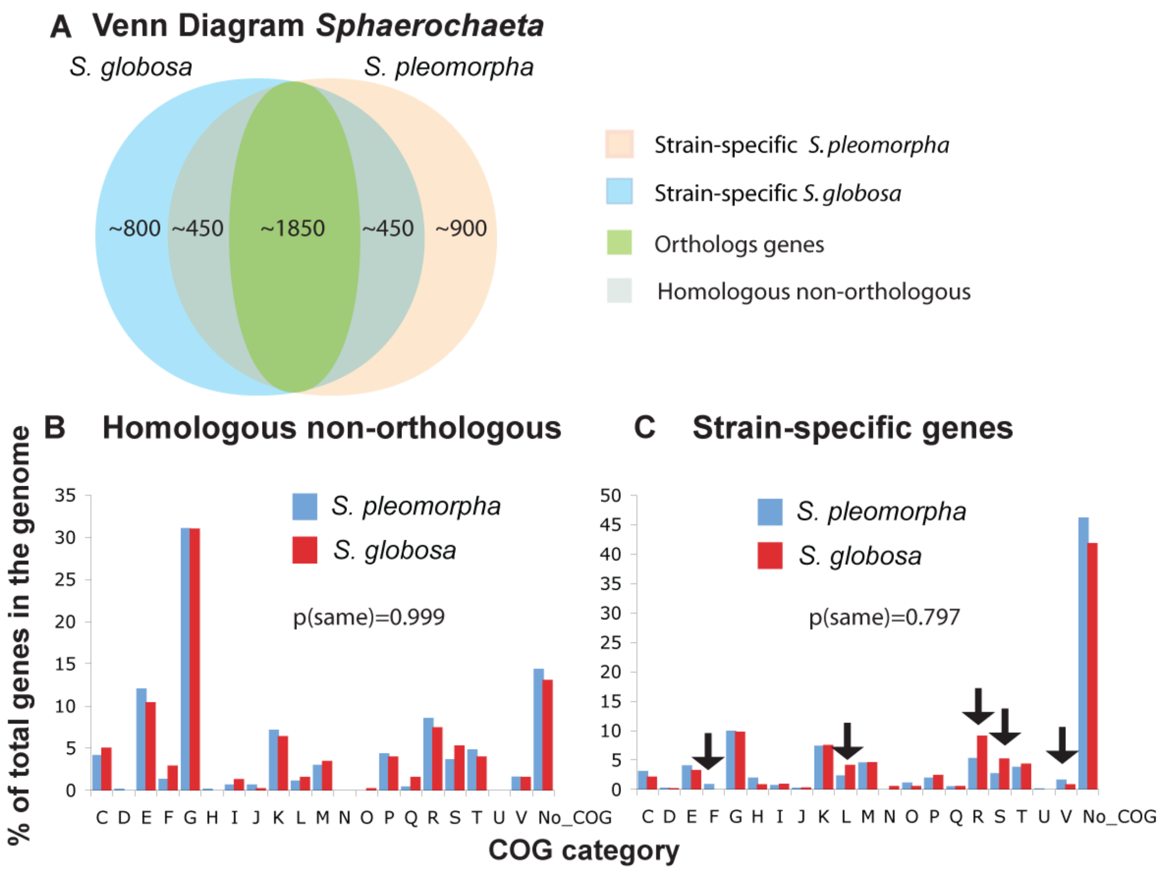


Figure 5.10 Functional comparisons between the *S. globosa* and *S. pleomorpha* genomes. Numbers of shared and strain-specific genes between the *S. globosa* and *S. pleomorpha* genomes are shown in a Venn diagram (A). Distributions of homologous but non-orthologous (i.e., not reciprocal best matches) (B) and strain-specific (C) genes in COG functional categories are also shown. The distributions were significant for strain-specific genes (p <0.05, Student's one-tail t test) but not significant for homologous, non-orthologous genes.

5.4.7 Bioinformatic Predictions In Deeply-Branching Organisms

Sphaerochaeta spp. probably represent a new family or even an order within the *Spirochaetes* phylum based on their divergent genomes and unique morphological and phylogenetic features. Bioinformatic functional predictions, particularly in such deeply-branching organisms, are often limited by weak sequence similarity and/or uncertainty about the actual function of homologous genes or pathways. Nonetheless, bioinformatics remains a powerful tool for hypothesis generation as well as for understanding the phenotypic differences among organisms. For the *Sphaerochaeta*, experimental evidence confirmed all of our bioinformatic predictions. For instance, we have confirmed experimentally [183] the predictions regarding the resistance of *Sphaerochaeta* to β-lactam antibiotics (based on the lack of PBPs), utilization of various oligo- and mono-saccharides, unusual cell wall structure, absence of motility, and tolerance to oxygen. These results revealed that bioinformatic-based inferences about the metabolism and physiology of deep-branching organisms such as the *Sphaerochaeta* can be robust and reliable.

5.4.8 *Sphaerochaeta* and Reductive Dechlorination

Sphaerochaeta commonly co-occur with obligate organohalide respirers of the *Dehalococcoides* genus [180, 183]. The reasons for this association are unclear but may have important practical implications for the bioremediation of chloroorganic pollutants. The *Sphaerochaeta* genomes provided some clues and create new hypotheses with respect to the potential interactions between free-living, non-motile *Sphaerochaeta* spp. and *Dehalococcoides* dechlorinators. For instance, it was previously hypothesized that *Sphaerochaeta* may provide a corrinoid to dechlorinators, an essential cofactor for reductive dechlorination activity [206]. However, the genome analyses revealed that *Sphaerochaeta* genomes encode only the cobalamin salvage pathway, which is not in agreement with the corrinoid hypothesis. Alternative intriguing hypotheses include that the fermentation carried out by *Sphaerochaeta* provides essential substrates (e.g., acetate and H_2) to *Dehalococcoides*, or that *Sphaerochaeta* are helper phenotypes to protect the highly redox-sensitive *Dehalococcoides* cells from oxidants (i.e., oxygen) [207].

5.5 Discussion

Genomic analyses revealed the absence of motility genes, the underrepresentation of sensing/regulatory genes (Fig. 5.3 and Fig. 5.6), the unusual lack of transpeptidase and transglycosylase genes involved in cell wall formation, and explained the resistance of *Sphaerochaeta* to β-lactam antibiotics and their unusual cell morphology. These findings demonstrate that spiral shape and motility are not shared attributes of the *Spirochaetes* phylum, breaking with the prevalent dogma in spirochete biology that "…spirochetes are one of the few major bacterial groups whose natural phylogenetic relationships are evident at the level of phenotypic characteristics" [208]. The reasons that underlie the loss of motility genes in the *Sphaerochaeta* are not clear but the lack of transpeptidase activity (i.e., loss of cell wall rigidity) may have been associated with the loss of axial flagella. Cell wall rigidity is presumably necessary for anchoring the two ends of the axial flagellum; hence, permanent loss of cell wall rigidity is likely detrimental to a properly functioning axial flagellum. It is also possible that habitats such as anoxic sediments enriched in organic matter and/or characterized by a constant influx of nutrients do not select for motility [209, 210] and favor the loss of genes encoding the motility apparatus; *Sphaerochaeta* were obtained from such habitats [183].

The unusual, non-rigid cell wall structure likely imposes additional challenges for *Sphaerochaeta* organisms to maintain cell integrity. A cellular adaptation to maintain membrane integrity, possibly accounting for the lack of a rigid cell wall, is through tight regulation of intracellular osmotic potential. Several genes encoding the biosynthesis of osmoregulating, periplasmic glucans, osmo-protectant ABC transporters, an uptake system for betaine and choline, and potassium homeostasis were found on the genomes of *S. globosa* and *S. pleomorpha*, suggesting fine-tuned responses to osmotic stressors. The importance of these findings for explaining *Sphaerochaeta* spp. survival and ecological success in the environment remains to be experimentally verified.

The loss of motility genes imposes new challenges for the identification of non-motile spirochetes in environmental or clinical samples. Free-living spirochetes are isolated routinely by selective enrichment for spiral motility, using specialized filters and/or solidified media, and by taking advantage of the unique spiral morphology, mode of propulsion, and natural resistance of spirochetes to rifampicin [211]. Therefore, traditional isolation methods have failed to recognize and likely underestimated the abundance and distribution of non-motile spirochetes. New isolation procedures should be adopted to expand our understanding of the ecology and diversity of this clinically and environmentally important bacterial phylum. The genome sequences reported here will greatly assist such efforts; for instance, they have revealed that *Sphaerochaeta* are naturally resistant to β-lactam antibiotics. The *Sphaerochaeta* genomes also provide a long-needed negative control (i.e., lack of axial flagella) to launch new investigations into the flagella-mediated infection process of spirochetes causing life-threatening diseases. Further, the recently determined genome sequence of *Spirochaeta coccoides* (accession number CP002659) also lacks the flagellar, chemotaxis and PBP genes and is more closely related to *Sphaerochaeta* compared to other members of the *Spirochaeta* genus (e.g., *S. smaragdinae*), suggesting that *S. coccoides* is a member of the *Sphaerochaeta* genus.

Our analyses revealed that more than 10% of the core genes and presumably more than 50% of the auxiliary and secondary metabolism genes of *Sphaerochaeta* were acquired from Gram-positive *Firmicutes*. The extensive unidirectional HGT (i.e., *Clostridiales* è *Sphaerochaeta*) implied that the two taxa (or their ancestors) share ecological niche(s) and/or physiological properties. Consistent with these interpretations, ecological overlap was observed previously between *Clostridiales* and both host-associated and free-living spirochetes. For instance, several genes related to carbohydrate metabolism in *Brachyspira hyodosenteriae*, an anaerobic, commensal spirochete, appear to have been acquired from co-occurring members of the *Escherichia* and *Clostridium* genera in the porcine large intestine [203]. Among free-living spirochetes, ecological overlap is likely to occur within anaerobic food webs where spirochetes and clostridia coexist [210, 212]. For example, biomass yield and rates of cellulose degradation by *Clostridium thermocellum* increase when grown in co-culture with *Spirochaeta caldaria* [213]. In agreement with these studies, the genes transferred between *Sphaerochaeta* and *Clostridiales* were heavily biased toward carbohydrate uptake and fermentative metabolism functions. A more comprehensive phylogenetic analysis that included 35 spirochetal and clostridial genomes (Table 5.1) indicated that *Sphaerochaeta* acquired several, but not all, of its clostridia-like genes from the ancestor of the anaerobic cellulolytic bacterium *Clostridium phytofermentans* (Fig. 5.7), which was also consistent with the BLASTP-based results from the three-genome comparisons.

Such a high level of inter-phylum genetic exchange is extremely rare among mesophilic organisms like *Sphaerochaeta* (Fig. 5.1 and in [7]]. This level of HGT has been reported previously only for thermophilic *Thermotoga* spp. (i.e., organisms living under extreme environmental selection pressures) [214]. On the other hand, we did not observe HGT that affected informational proteins such as ribosomal proteins and DNA/RNA polymerases, suggesting that the reconstruction of spirochetal phylogenetic relationships, and in general the construction of the bacterial Tree of Life, can be attained even in cases of extensive genetic exchange of metabolic genes [for a contrasting opinion, see [153]]. In the case of *Sphaerochaeta*, the massive HGT was apparently favored by overlapping ecological niche(s) with *Clostridiales* and/or strong functional interactions within anoxic environments. These findings highlight the importance of both ecology and environment in determining the rates and magnitudes of HGT. Obtaining quantitative insights into the role of the environment and shared ecological niches in HGT will lead to the more educated assembly of the prokaryotic Tree of Life based on measurable and quantifiable properties.

5.6 Acknowledgments

We thank the DOE Joint Genome Institute for sequencing the *Sphaerochaeta* genomes used in this study. This work was supported by the National Science Foundation under Grant No. 0919251.

CHAPTER 6

INTER-PHYLUM HGT HAS SHAPED THE METABOLISM OF SEVERAL MESOPHILIC AND ANAEROBIC BACTERIA

Reproduced in part with permission from: Caro-Quintero, A. and Konstantinidis, K.T. Inter-Phylum HGT Has Shaped the Metabolism of Several Mesophilic and Anaerobic Bacteria. In preparation. All copyrights interests will be exclusively transferred to the publisher upon submission

6.1 Abstract

Genome sequencing during the past two decades has revealed that horizontal gene transfer (HGT) is a major evolutionary process in bacteria but several questions remain. Although it is generally assumed that HGT is more pronounced among closely related organisms relative to distantly related ones, this hypothesis has not been rigorously tested yet, while quantitative data on the number of genes in the genome affected by HGT are lacking for most bacterial species. Here, we devised a novel bioinformatic pipeline to identify gene exchange between bacterial genomes representing different phyla that normalized for many of the known limitations in HGT detection such as the differential representation of phyla in the database. Analysis of all available genomes suggested that organisms with overlapping ecological niches clustered in networks of genetic exchange and such level of exchange was higher among mesophilic anaerobic organisms. Inter-phylum HGT has affected up to ~16% of the total genes and up to 35% of the metabolic genes in some genomes, revealing that HGT among distantly related organisms is much more pronounced than previously thought. Nonetheless, ribosomal proteins were subjected to HGT at least 150 times less frequently than the most promiscuous metabolic

functions (e.g., various dehydrogenases and ABC transport systems), suggesting that the ribosomal protein species Tree may be reliable. All together, our results indicated that the metabolic diversity of microbial communities within most habitats has been largely assembled from preexisting genetic diversity through HGT and that HGT accounts for the functional redundancy commonly observed within communities.

6.2 Introduction

Bacteria are the most ubiquitous organisms of the planet. They catalyze fundamental steps in the geochemical cycles and are players of key ecological relationships (i.e., symbiosis, protocoperation, competition) that determine the diversity and distribution of organisms, including eukaryotes, in most habitats. A key aspect favoring bacteria functional and ecological diversity is their ability to incorporate foreign DNA through horizontal gene transfer (HGT). The ability to incorporate exogenous DNA has been so important in the course of evolution that is believed to be the major process responsible for the large physiological diversity and remarkable adaptability of prokaryotes [1, 2]. In fact, recent analysis of protein families suggests that HGT, and not duplication, has driven protein expansion and functional novelty in bacteria [3]. Genome sequencing of thousands of genomes has expanded our view of the role of HGT in bacterial evolution and allowed the identification of genetic exchange events at different time scales (i.e., from ancestral to recent events), and between organisms of varied evolutionary relatedness (i.e., from close related genomes to very distantly related ones) [4-7].

Genetic exchange between distantly related bacteria is generally thought to be less frequent than between closely related organisms due to ecologic (e.g., less frequent encounters due to different niches) and genetic mechanisms (e.g., defense mechanisms against foreign DNA, lower sequence identity for recombination, and incompatibility in gene regulation). Recently, we have reported massive inter-phylum genetic exchange between mesophilic *Sphaerochaeta* (*Spirochaete*) and clostridia (*Firmicutes*) [215], such an extensive inter-phylum HGT had only been previously documented for organisms living under extreme environmental selection pressures, such as thermophilic [214] and halophilic organisms [216]. These findings in mesophiles indicated that, contrary to what has been previously thought, high levels of HGT among distantly related organisms can also occur within non-extreme environments.

Here we aimed to extent our previous analysis [215] to all available complete genome sequences of free-living organisms to quantitatively evaluate inter-phylum HGT and establish whether or not it is frequent within non-extreme environments. We also evaluated the environmental and ecological conditions that favored massive inter-phylum HGT events and the gene functions that were more frequently transferred. To this end, we developed a novel bioinformatic pipeline that minimized the effect of taxonomic classification and overrepresentation of specific phylogenetic groups to provide unbiased, quantitative estimates of HGT across all taxa evaluated.

6.3 Materials and Methods

6.3.1 Amino Acid And Genome Sequences Used in this Study.

Predicted proteins from completed bacterial and archaeal genome projects were downloaded from NCBI on July 1, 2012 (2,001 genomes) to form an in-house searchable database. To avoid the effect of genome reduction in endosymbiotic organisms, which can bias comparisons of the magnitude of HGT across genomes, only free-living genomes with genome size larger than 2Mbp were used in the analysis (1,356 genomes). The resulting set of genomes represented 28 different phyla. Literature review was performed to identify physiological and ecological information (i.e., source of isolation, optimal growth temperature, respiration) for each genome.

6.3.2 Homolog Identification and Database Normalization.

Orthologous genes for all possible pairs of genomes (1,838,736 pairs) were identified using the reciprocal best match approach [217] and the USEARCH algorithm for its computational efficiency [218]. Only best matches with identity higher than 40% and coverage of the query gene sequence higher than 70% were used in the analysis. For any pair of genomes, the genome-aggregate average amino acid identity (gAAI) was calculated by averaging the identity of shared orthologs as suggested previously [101]. In order to reduce the redundancy (and thus, the size) of the database for faster computations, genomes were clustered in groups that shared higher than 95% gAAI, which corresponds to the frequently used standards to define bacterial species [101]. One genome from each of the resulting groups (n=879) was randomly selected to represent the group, and the gAAI values between these representative genomes from different groups were used as a measurement of genetic divergence.

6.3.3 Quantifying HGT at the Genome-Level.

A genome-wise analysis was carried out to identify pairs of genomes involved in high inter-phylum HGT. To account for the differential representation of taxa in the database, genomes were analyzed in triplets (104,101,468 triplets); each genome triplet included a reference genome (reference), a genome of the same phylum as the reference (insider), and a genome of a different phylum (outsider) (Fig. 6.1). For each triplet, all genes of the reference genome (query) were searched against the insider and the outsider (database) for best matches and the ratio of the number of best matches in the insider vs. total genes with best matches (in the insider or outsider genome) was used to quantify the extent of inter-phylum HGT for each reference genome. Two ratios were calculated; one for reference protein sequences having a match (homolog) in both the insider and outsider genomes (shared proteins), and one for protein sequences with a homolog in either or both, the insider or outsider (all proteins). The ratios for all possible triplets of genomes were determined and sets (ratios) from the same reference genome and similar genetic relatedness (i.e., triplets with gAAI values within ± 1% of a chosen gAAI value) were compared together. For each resulting set of triplets, a mean and standard deviation were calculated. The distribution of ratios was normalized by standardization, by calculating their deviation from the mean in terms of standard deviations.. The triplets with three standard deviations higher than the mean were identified as cases of high inter-phylum HGT (p-value <0.001) and the partners were identified (reference and outsider genomes). Note that the HGT detected by this analysis encompassed both recent and ancestral events because all best-matches with higher that 40% identity were taken into account. The information available about the reference and outsider were further examined to identify the ecological, and functional factors that fostered the HGT.

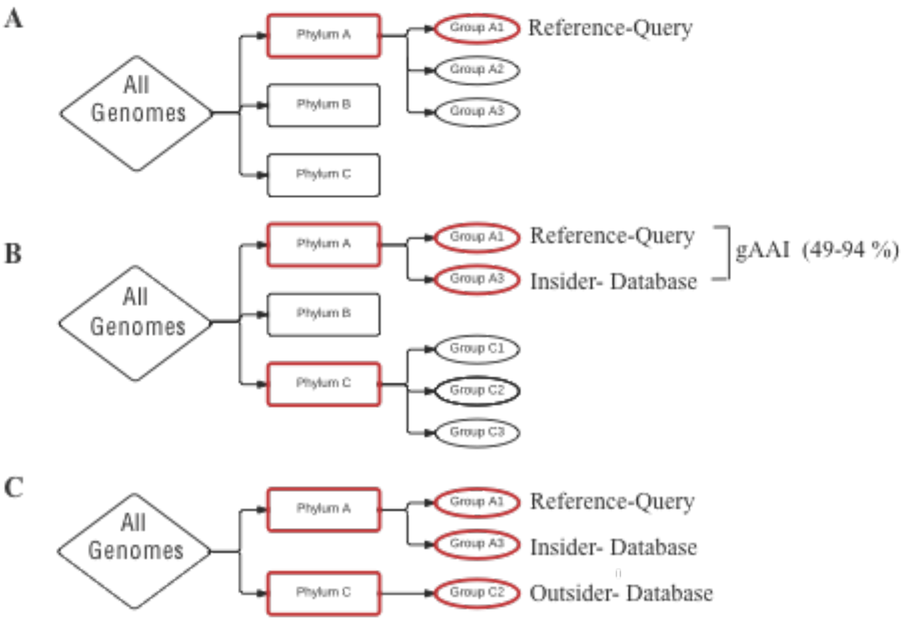


Figure 6.1. A schematic of the approach used to select genome triplets for assessing HGT between bacterial and archaeal phyla. The approach included the following steps: 1) select randomly a reference genome to begin to form a triplet of genomes **(Panel A)**; 2) select a second genome ("insider") representing the same phylum as the reference but from a different group based on gAAI **(Panel B)**; and finally, 3) a genome representing a different phylum ("outsider") is selected **(Panel C)**. The phylogenetic distance between the reference and insider genomes was measured by gAAI; all triplets characterized by similar gAAI values between the reference and insider genomes (-/+1%

from the chosen gAAI values) formed a single set and were analyzed together (compared).

6.3.4 Functional Analysis of Transferred Genes.

The homologs shared between the reference and outsider genomes were evaluated statistically to identify cases of HGT. These genes were also used to determine what functional categories are more commonly transferred across phyla. Two different statistical approaches were employed, one for the homologs present in all genomes of the triplet (shared genes), and one for homologs only shared by the reference and outsider genomes (non-shared genes). For shared genes, all homologs were grouped in sets based on the gAAI values (± 1%) of the corresponding triplets (gAAI between the reference and insider genomes; see above). For each set, the sequence identity between the reference and outsider homologs was subtracted from the identity between the reference and insider homolog (% identity with the insider - % identity with the outsider), and a distribution of these numerical differences was generated. Therefore, one of such distribution was calculated for triplets with the similar gAAI values. Each distribution was fitted to a normal, polynomial, or gamma function and the function that better fitted the observed distribution (Kruskal-Wallis test) was selected. The parameters from the fitted distributions were extracted and used to produce one general model for all gAAIs. This model described the expected probability of finding a homolog shared between the reference and outsider genomes with a specific amino acid identity value, using the identity of the reference genes against the insider homologs to normalize for the different degree of sequence conservation of individual protein families (e.g., ribosomal proteins tend to be more conserved than metabolic proteins). p-values were estimated from the cumulative density distribution of the model (1 – model; Fig. 6.2, A) and HGT events were defined as cases where matches to the outsider had significant higher identity compare to matches to the insider, i.e., p-value < 0.001.

For non-shared homologs, a different method to distinguish cases of HGT from gene loss (in the lineages of the insider genome) was employed. This approach was based on the assumption that the majority of genes identified as orthologs by bidirectional best match searches reflect vertical descent [217], and therefore the variation in amino acid sequence identities among them can be used as a null model to identify cases with sequence identity higher than expected due to HGT. Orthologs from different phyla were identified and assigned, when possible, to the Cluster of Orthologous Groups (COGs) and the mean and standard deviation of the distribution of amino acid sequence identity were calculated. These values were used to evaluate statistically if the identity of matches with the outsider is the expected based on vertical descent or if it is higher than expected (outliers higher than three standard deviations from the mean) and represent case of HGT (p-value < 0.001) (Fig. 6.2, B). For genes that were detected as transferred more than once in the lineage of the outsider or insider genomes, only the case with the highest identity was counted to avoid overestimating the transferred function.

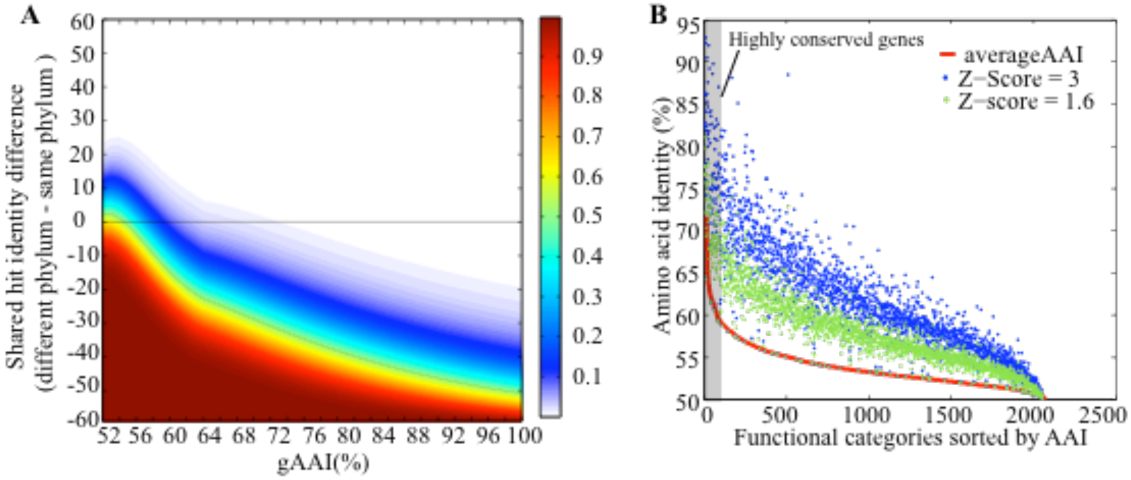


Figure 6.2 Identification of genes exchanged between bacterial and archaeal phyla with statistical confidence. Two different approaches were developed to evaluate HGT signal for shared (reference gene has homologs in the two other genomes of a triplet) and unique (reference gene has homologs only in the outsider). For shared genes, a probabilistic model based on the distribution of amino acid sequence identity difference between the reference–insider match relative to the reference-outsider match was used to detect higher than expected identity of the reference genes with the outsider, which were identified as HGT events (see Material and Methods for details), **(Panel A)**. For unique hits, the distribution of sequence identities was based on homologs assigned to the same (individual) COGs gene family **(Panel B)**. The plot shows the average amino acid identity between reciprocal best-match homologs (orthologs) shared by distantly related organisms, green dots represent 1.6 standard deviations from the average, while blue dots represent 3 standard deviations from the average. The latter threshold was used to identify HGT event.

6.3.5 Networks of HGT.

All pairs of genomes with significant signal of exchange (donor and recipient) were linked in networks that represented the extent of HGT. Networks were constructed using the Cytoscape V 2.8 algorithm [152]. Two networks were evaluated; one based on the significant cases found in the whole-genome level analysis and another based on the individual gene-level analysis. The analysis of both HGT networks was done using the Girvan-Newman greedy algorithm [219, 220] implemented in GLaY [221]. This algorithm clusters the genomes into subnetworks that maximize the amount of connectivity (representing HGT in this case). The organisms/genes in the resulting subnetworks were then examined manually to identify the ecological and/or physiological factors that underlay the high connectivity.

6.3.6 Phylogenetic Reconstruction.

The phylogeny of 879 representative genomes was reconstructed using a similarity matrix built from the gAAI values and the Neighbor Joining algorithm with 1000 bootstraps. The resulting phylogenetic tree was visualized in Cytoscape V2.8 [152] and the putative partners of exchange were connected on the resulting tree using a in-house Perl script.

6.4. Results and Discussion

6.4.1 An Approach to Overcome the Known Limitations in Detecting HGT.

Quantification of HGT among distantly related organisms represents a challenging task, in part because of the lack of complete representation of the prokaryotic diversity and the low number of shared genes between such organisms. For instance, evaluation of the effect of genetic divergence on the proportion of shared genes for all genomes pairs analyzed here (n =1,838,736) revealed that any pair of genomes from different phyla may shared at most 20 % of their total genes in the genome (Fig. 6.3, A). There are currently two commonly used approaches to identify HGT, phylogenetic and homology search methods, primary best-match analysis. Phylogenetic methods are a powerful tool to detect HGT and offer high sensitivity but they are computationally intensive and therefore not suitable for whole-genome analysis of a large number of genomes. An alternative approach is the best-match analysis based on the Smith-Waterman algorithm or its variations [222]. In this approach, gene sequences or their translated peptides are searched against characterized genomes (database) and best-matches to distantly related genomes (when close relatives exist in the database) are identified as putative HGT cases. These approaches are computationally less expensive and can be scaled to large datasets. However, the best-match approach has lower sensitivity compare to the phylogenetic one [199] and can be strongly affected by the genome database used, e.g., several taxa are underrepresented.

In order to implement homology search approaches for the accurate detection of HGT among distantly related genomes, all available genomes were compared in triplets to control for the effect of database representation. Each triplet was composed of two genomes of the same phylum (one reference and other "insider") and the third genome to represent another phylum ("outsider"; see materials and methods for details). The analysis of these triplets showed that the more divergent the reference and insider genomes were, the larger the proportion of best matches of the reference to the outsider genome (Fig. 6.3, B). The high proportion of best matches to the outsider cannot be attributable to gene loss because the same trend was observed when the analysis was restricted to only genes shared by all three genomes in a triplet (Fig. 6.3, B inset) and is presumably attributable to false positive HGT, consistent with previous studies of homology-based approaches [223]. This trend suggests that deep-branching genomes (e.g., relatives from the same phylum with gAAI < 60 %) will always have a substantial amount of genes with best matches in a different phylum, irrespective of the occurrence of HGT (low signal to noise ratio). The results highlighted and quantified the limitation of homology-based approaches with distantly related genomes; the quantification of the limitation provided the basis for an approach to overcome it.

To minimize the number of false positives in the detection of HGT, approaches based on the distribution of best-match ratios (genome-level) and sequence identities of orthologs (gene-level) were used. These approaches identified genes and genomes that have undergone inter-phylum HGT with statistical confidence (see methods for details). At the genome-level, the proportion of best-matches in the outsider was calculated for each triplet, and compared to a distribution build from all triplets with the same reference and genetic relatedness (gAAI). At the gene-level, the method evaluates the individual genes by assessing how uncommon the sequence identity between the reference and the outsider is compared to the expected distribution of identities based in vertical inheritance (null model). The genome-level method represents ancestral to recent HGT because it evaluates the proportion best-matches and not the identity of the hits. In contrast, the gene-level method detect more recent events becuase it relays in the identification of outlier with high identity. Using these approximations significant inter-phylum HGT signal was commonly detected, in 811 out of the total 847 evaluated genomes, which suggests that distant HGT has an important influence in bacterial evolution.

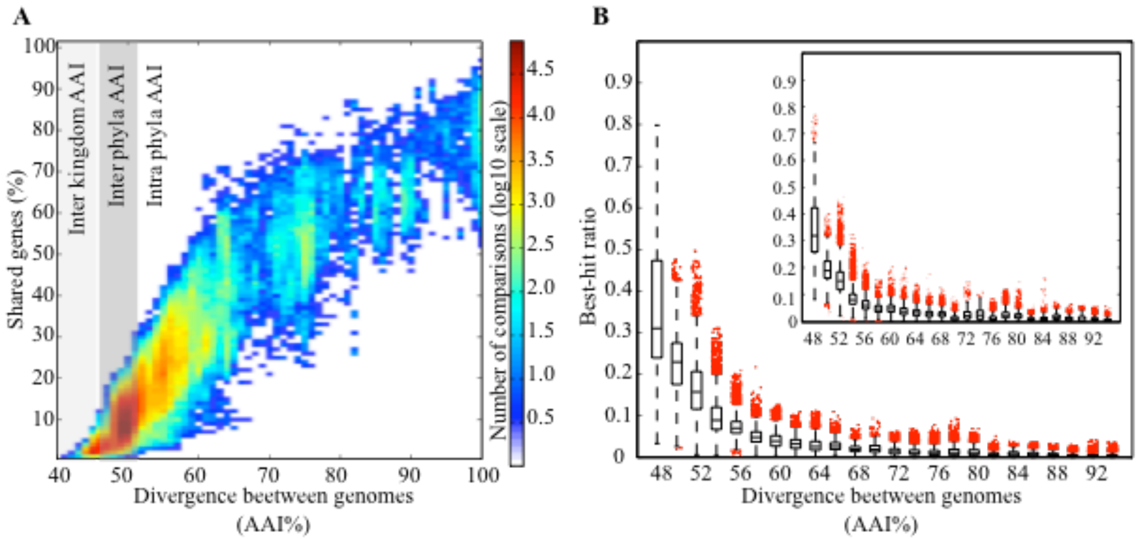


Figure 6.3. Dependence of the number of shared genes and intra- vs. inter-phylum best match on the genetic divergence of the genomes compared. 1,838,736 pairwise whole-genome comparisons were performed and the relationship between genetic divergence and percentage of shared genes for these genomes is represented by a colored density plot **(see scale)**. The shaded areas roughly correspond to the gAAI values between bacteria and archaea (inter-kingdom), between phyla, and within phyla (**Panel A**). The genomes were grouped in triplets, as described in the text, and the genes of the reference genome in the triplet were searched against the other two genomes, one representing the same phylum as the reference and the other representing a different phylum. The ratio of the number of best matches within vs. outside the phylum is plotted against the gAAI values between the two genomes of the same phylum in the triplet (boxplots in **Panel B)**. Each boxplot represents the distribution of ratios from 4,000 randomly drawn triplets per unit of gAAI. Main graph shows the data for reference genes that had a match in both of the other two genomes in the triplet (shared genes); inset shows the genes that had a match in either (but not both) of the genomes. Red points represent the outliers.

6.4.2 Shared Physiology and Ecology Underlie Networks of High HGT.

The influence of ecology and physiology in inter-phylum exchange was evaluated by generating networks that represent the relationships of HGT cases. These networks were made by linking the donors and recipients with statistically significant signal of HGT (p-value <0.001). Two networks were built, one for the genome-level approach and other for the gene-level approach. The genome-level network capture cases of HGT with high genome sharing due to recent and ancestral HGT, while the gene-level network reflects only recent events. Within each network, a community-clustering algorithm [219, 220] was used to cluster the original network into subnetworks that maximize HGT among members (i.e., HGT more abundant among the genomes of the subnetwork than when compare to other genomes or subnetworks).The subnetworks were named as "N" for genome-level and "A" for gene-level analysis, In top of these subnetworks ecology and physiology parameters were mapped to evaluate their correspondence with the observed clustering.

The analysis of the genome-level network revealed that HGT is strongly favored by (shared) ecology and oxygen tolerance. The genome-level network was split by the community- clusteringalgorithm [219, 220], into four subnetworks (N1, N2, N3 and N4; Fig. 6.4, A). Analysis of the available information on the source of isolation of the genomes in a subnetwork showed that subnetwork N3 was clearly enriched (64% of total genomes) in human associated commensals and pathogens.

The latter primarily included members of the *Enterobacteriaceae* (*Proteobacteria* phylum), and the *Streptococcaceae*, *Lactobacillales*, *Listeriaceae* and *Staphylococcacea* (*Firmicutes* phylum). These findings agreed with a previous study that showed higher genetic exchange between human associated bacteria [70], and suggested that the patterns of genetic exchange described previously for closely related organisms are also applicable to distantly related microbes. Subnetwork N2 was enriched in soil and plant associated bacteria (~50%). Most of these exchanges occurred between *Rhizobiales*, *Bradyrhizobiaceae* and *Comamonadaceae* (*Proteobacteria* phylum) and *Streptomycetaceae* and *Micrococcaceae* (*Actinobacteria* phylum). On the other hand, subnetworks N1 and N4 were dominated by aquatic mesophilic and thermophilic organisms (~70%). Mesophilic groups included organisms of the *Chloroflexi* phylum, *Chrorococales* (*Cyanobacteria* phylum), *Flavobacteriacea* (*Bacterioidetes* phylum), and *Alteromonadaceae* (*Proteobacteria*). Meso- and hyper-thermophilic taxa include organisms from the *Deinococcus-Thermus* phylum, *Thermoanaerobacteriales* (*Firmicutes* phylum), representatives from the *Thermotogae* phylum, and archaea from the *Euryarchaeota* and *Crenarcheaota* phyla. Notably, among the evaluated parameters oxygen tolerance appeared to correspond best with the subnetwork clustering. For instance, subnetwork N1 was mainly composed by anaerobic bacteria (80%), while N2, N3 and N4 were dominated by aerobic bacteria (89, 80 and 74%, respectively). This suggests that, among all evaluated environmental parameters, oxygen tolerance plays the most important role in driving HGT within aerobic and anaerobic environment.

The community-clustering algorithm was re-applied to the anaerobic subnetwork N1 (generated in the genome-level approach) to examine in more detail the dynamics of exchange and elucidate more specific ecological interactions between anaerobic mesophiles (Fig. 6.4, B), Four subnetworks (N1.1, N1.2, N1.3, N1.4) were obtained and their structure was analyzed in more detail. Subnetwork N1.2 was the most diverse in terms of phylogeny (encompassing 11 different phyla) but strongly overrepresented by organisms of the *Firmicutes* phylum (57%). Interestingly, elimination of *Firmicutes* from the network reduces the number of transfers (edges) by 97 %, suggesting that *Firmicutes* are the most important partner in HGT for this subnetwork. Further analysis revealed two main physiological groups. The first was composed of aquatic themophilic and hyperthermophilic bacteria (e.g., *Thermoanaerobacterium xylanolyticum* and *Spirochaeta thermophila*) and the second of soil saprophytic fermenters (e.g., *Sphaerochaeta spp* and *Clostridia cellulovorans*) and gut-associated bacteria from insects, humans and ruminants (e.g., *Spirochaeta coccoides, Eubacterium retale* and *Roseburia hominis*). Even though these organisms differ in their source of isolation and optimal growth temperature, they are all characterized by saccharolytic and fermentative lifestyles. Therefore, subnetwork N1.2 showed that organic matter degradation genes are relevant across several ecological niches rich in organic matter content and have been commonly transferred from/to *Firmicutes* multiple times.

Analysis of subnetwork N1.1 revealed the importance of strong ecological interactions (i.e., protocooperation) in favoring genetic exchange. The three main groups that made up the network were either syntrophs, or had representatives reported to be partners of syntrophic interactions, and included the sulfate reducing bacteria (SRB) and syntrophic bacteria from the *Firmicutes* phylum (e.g., *Desulfotomaculum spp*.), the *Proteobacteria* phylum (e.g., *Syntrophus spp*.) and methanogenic archaea of the *Euryarcheota* phylum (eg., *Methanocella spp*.). These groups not only are assigned to different phyla, but also have drastically different ecologies. Therefore, the unexpected high frequency of HGT among these groups indicates that syntrophic associations play a key role for HGT. These results were consistent with previous phylogenetic approaches that showed high gene sharing between syntrophic organisms [73]. Additionally, it has been suggested that HGT is responsible for similar codon usage bias between *Pelotomaculum thermopropionicum* and other syntrophic organisms [224] and that syntrophic interactions between *Desulfovibrio vulgaris* and *Methanosarcina barkeri* had evolved as the result of ancestral HGT [225]. In conclusion, syntrophic metabolism represents a clear example of how tight ecological relationships (i.e., physiological dependence and physical contact) have favored the transfer of genetic material between distantly related organisms.

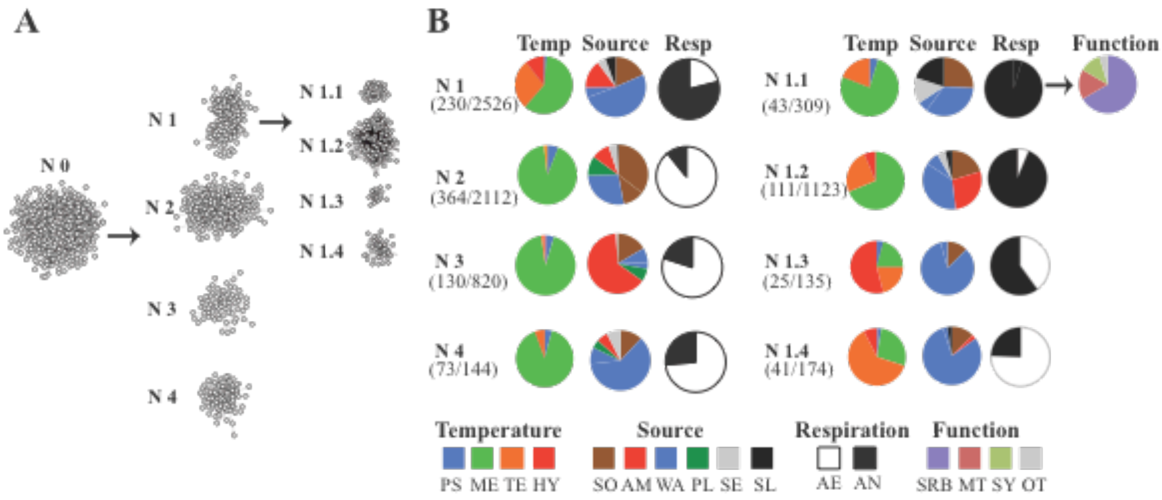


Figure 6.4. The effect of shared physiology and ecology on the structure of HGT networks. A network representing all inter-phylum HGT events was obtained as described in the main text and was divided into subnetworks using the community-clustering algorithm (GLaY) [219, 220] that maximizes the connectivity between network nodes. Four subnetworks were obtained (N1, N2, N3, N4). Network N1 encompassed the highest number of anaerobic representatives and was further subdivided using GLaY. Four subnetworks were obtained (N1.1, N1.2, N1.3, N1.4**; Panel A)**. The optimal growth temperature (Temp), source of isolation (Source) and type of respiration (Resp), was extracted from the literature for all genomes in each subnetworks **(Panel B)** and categorized as follows. I) For optimal growth temperature category: psycrophilic (PS), mesophilic (ME), thermophilic (TE), and hyperthermophilic (HY). II) For source of isolation: soil (SO), animal associated (AM), aquatic (WA), plant (PL), sediment (SE), and sludge-bioreactor (SL). III) For respiration: aerobic (AE) and anaerobic (AN). The data revealed that the organisms grouped in Network N1.1 had predominantly syntrophic interactions among themselves and were categorized further by their metabolic function (Function) to sulfate reducing bacteria (SRB), methanogens (MT), general syntrophic-

secondary fermenting bacteria (SY) or other functions (OT). Note that respiration type separates more clearly subnetwork N1 from N2 and N3 and N4 than the other categories, also important subdivision of N1 creates two subnetworks that clearly match syntrophic (N1.1) and fermentative metabolism (N1.2).

Along the same lines, gene-level network analysis showed that oxygen tolerance explain best the clustering of genomes in the three largest sub-networks A1 (119 genomes), A2 (89 genomes) and A3 (82 genomes) (Fig. 6.5, A). For instance, subnetwork A1 was mainly composed by aerobic organisms while sub-networks A2 and A3 were mainly composed by sulfate reducing and syntrophic bacteria, and fermenting bacteria. Analysis of the frequency of genes transferred within the sub-networks showed that metabolic functions in the networks composed of (primarily) anaerobic, A2 and A3, have been exchanged twice as frequent compared to aerobic metabolic genes (sub-network A1; Fig 6.6, B). Further, inter-phylum exchange within a sub-network was 6 to 37 times larger compared to between the sub-networks, confirming that the network analysis was robust (Fig 6.5,B). Exchange between sub-networks A1 and A3 was the lowest while A2 and A3 (both encompassing mostly anaerobic organisms) showed the highest frequency of exchange. Although the exact reasons for the higher frequency of HGT within anaerobic vs. aerobic networks remain speculative, it is reasonable to hypothesize that within anaerobic environments there is more niche overlap and/or physical proximity among organisms due to physiological dependence, which apparently favors HGT. For instance, aerobic microorganisms can frequently oxidize substrates to water and carbon dioxide without any significant cooperation with other organisms while anaerobic microorganisms often depend to a greater extent on associations with different partners. As an example, the complete conversion of cellulose to methane and carbon dioxide requires the concerted action of at least four different metabolic groups of organisms, including primary fermenters, secondary fermenters, and methanogenic archaea [226].

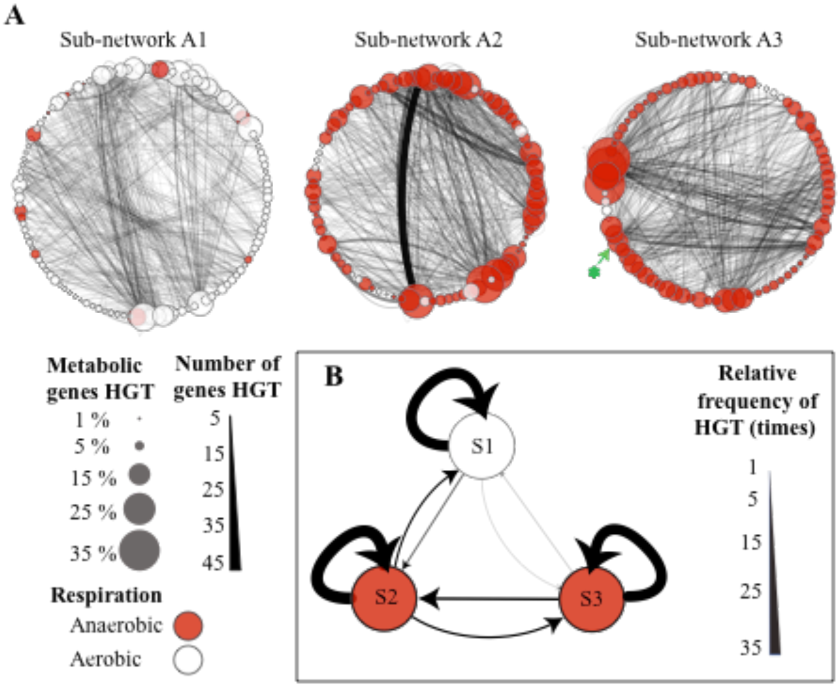


Figure 6.5. Cases of extensive inter-phylum HGT. A network representing all cases of HGT was obtained by linking genomes that had exchanged more than three genes. Nodes represent the genomes and the lines represent the cases of HGT. The network was divided into sub-networks using the community-clustering algorithm (GLaY) [219, 220] that maximizes the connectivity between nodes. Three sub-networks were obtained (A1, A2, A3; **Panel A**). The number of genes exchanged between the genomes is represented by the thickness of the lines (see scale at the bottom left). The percentage of the total

metabolic genes in the genome transferred is represented by the size of each node and the colors of the node represent aerobic (white) and anaerobic organisms (red). The amount of exchange within and between the networks was calculated by selecting randomly 40 genomes, with 1000 replicates, and taking the average of the number of exchanges detected in all replicates. The relative value was calculated by dividing all resulting average frequencies by the lowest inter-network frequency (see figure key; **Panel B)**.

6.4.3 Genomes Shaped by Extensive Inter-Phylum Genetic Exchange.

To establish whether or not the large inter-phylum exchange previously observed in *Sphaerochaeta* [215] represents a unique case, the proportion of genes in the genome that have signal of inter-phylum exchange was quantified for every reference genome (Fig. 6.6, A). The results showed that *Sphaerochaeta* ranked in the higher 97% percentile, with 6 % of the total genes in the genome showing signal of HGT and 15% of all metabolic genes. Thus, *Sphaerochaeta* is not the only mesophile characterized by large genetic exchange; in fact, 24 out of the top 37 cases of extreme inter-phylum HGT also involved mesophiles (Table 6.1). Collectively, these findings revealed that inter-phylum HGT is more pronounced than previously anticipated, accounting for up to 16 % of the total genes and 35 % of the metabolic genes in some genomes. It should be able also mentioned that our method identified only HGT events with high confidence (p-value < 0.01); thus, the previous results most likely represent an underestimation of the magnitude of HGT. For instance, using a less stringent cut-off (best-match with more than 40% a.a. identity over

70% length of the query protein), manual inspection of the results, and phylogenetic analysis of selected genes, we calculated previously that *Sphaerochaeta* genomes have exchanged up to 40 % of the total genes with *Firmicutes*) [215].

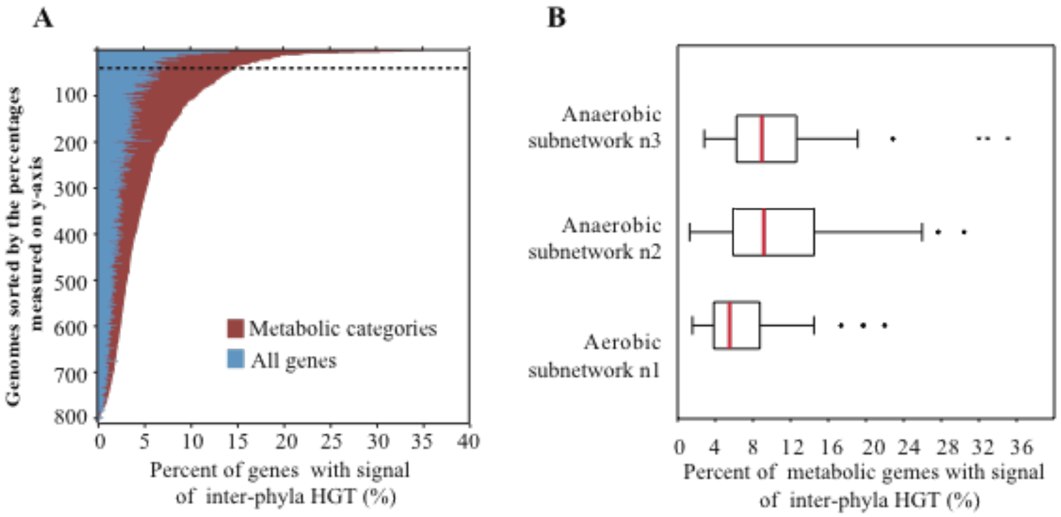


Figure 6.6. Frequency of inter-phylum HGT per genome and gene. Each bar represents one genome; the red portions of the bar represent the proportion of metabolic genes exchanged (i.e., the number of metabolic genes exchanged divided by the total number of metabolic genes in the genome); the blue portion represents the proportion of all genes exchanged (e.g., the number of genes exchanged divided by the total number of genes in the genome). Genomes are sorted by the number of genes exchanged. The dashed line represents the *Sphaerochaeta-Clostridia* case reported previously [215] **(Panel A**). The box plots represent the distribution of the percentages of metabolic genes that have significant signal of HGT shown in panel A based on the subnetworks A1, A2 and A3 from the gene-based analysis (**Panel B**). The red line denotes the median, the left and right box boundaries represent the lower and upper quartiles and the whisker delimit the 97% percentile of the data, dots represent outliers. Note that the median of anaerobic networks n2 and n3 is almost twice as high as that of aerobic network n1.

Table 6.1. Organisms with the highest percentage of gene acquired from organisms of different phyla. Organisms are ranked by the number of genes (as a fraction of the total genes in the genome) with signal of HGT.

Genome name	Optimal growth temperature	Oxygen Tolerance	Metabolic categories (%)	Total genome (%)
Ilyobacter polytropus DSM 2926	mesophilic	anaerobic	35.1	16.2
Leptotrichia buccalis C 1013 b	mesophilic	anaerobic	32.9	11.1
Sebaldella termitidis ATCC 33386	mesophilic	anaerobic	32.0	11.0
Desulfurispirillum indicum S5	mesophilic	anaerobic	30.4	14.2
Thermodesulfatator indicus DSM 15286	thermophilic	anaerobic	27.7	12.6
Deferribacter desulfuricans SSM1	thermophilic	anaerobic	26.0	10.6
Fusobacterium nucleatum ATCC 25586	mesophilic	aerobic	22.9	11.2
Thermodesulfovibrio yellowstonii	thermophilic	aerobic	22.2	11.2
Candidatus Solibacter usitatus Ellin6076	mesophilic	aerobic	22.0	5.9
Geobacter sulfurreducens KN400	mesophilic	anaerobic	21.7	8.6
Candidatus Nitrospira defluvii	mesophilic	anaerobic	20.3	8.7
Thermaerobacter marianensis DSM 12885	hyperthermopilic	aerobic	19.7	8.5
Rubrobacter xylanophilus DSM 9941	thermophilic	aerobic	19.7	9.5
Rhodothermus marinus DSM 4252	thermophilic	aerobic	19.7	7.2
Calditerrivibrio nitroreducens	thermophilic	anaerobic	19.4	8.6
Eggerthella lenta DSM 2243 9	mesophilic	anaerobic	19.1	6.7
Denitrovibrio acetiphilus DSM 12809	mesophilic	anaerobic	18.8	6.8
Geobacter uraniireducens Rf4	mesophilic	anaerobic	18.8	7.0
Slackia heliotrinireducens DSM 20476	mesophilic	anaerobic	18.8	6.6
Desulfotomaculum kuznetsovii DSM 6115	mesophilic	anaerobic	18.7	7.5
Heliobacterium modesticaldum Ice1	thermophilic	anaerobic	17.9	6.1
Ammonifex degensii KC4	thermophilic	anaerobic	17.5	7.2
Anaerobaculum mobile DSM 13181	thermophilic	anaerobic	17.5	8.8
Gemmatimonas aurantiaca T 27	mesophilic	aerobic	17.4	5.9
Treponema primitia ZAS 2	mesophilic	anaerobic	17.3	5.1
Eggerthella YY7918	mesophilic	anaerobic	17.1	6.2
Treponema brennaborense DSM 12168	mesophilic	anaerobic	16.7	6.3
Granulicella mallensis MP5ACTX8	mesophilic	aerobic	16.6	6.4
Treponema succinifaciens DSM 2489	mesophilic	anaerobic	16.4	5.1
Flexistipes sinusarabici DSM 4947	thermophilic	anaerobic	16.4	6.7
Geobacter metallireducens GS 15	mesophilic	anaerobic	16.2	6.9
Clostridium clariflavum DSM 19732	thermophilic	anaerobic	15.6	4.8
Desulfurivibrio alkaliphilus AHT2	mesophilic	anaerobic	15.5	6.1
Thermosediminibacter oceani	thermophilic	anaerobic	15.3	7.4
Desulfobulbus propionicus DSM 2032	mesophilic	anaerobic	15.1	5.3
Pelobacter carbinolicus DSM 2380	mesophilic	anaerobic	15.1	6.7
Sphaerochaeta pleomorpha Grapes **	mesophilic	anaerobic	15.0	5.9

6.4.4 Gene Functional Categories More Frequently Exchanged.

The genes that were recently transferred across phyla were examined to determine functional biases in HGT. Metabolic genes were among the most commonly exchanged genes, making up 60 % of all detected HGT events and 70 of the top 100 most frequently exchanged individual functions (Fig. 6.7, A). The general functional categories more ubiquitously transferred were those related to lipid transport and metabolism, energy production and conversion, amino acid transport and metabolism, and carbohydrate transport and metabolism. The specific functions most frequently exchanged included short dehydrogenases with different specificities (COG1028; 3.8% of all cases), NAD-dependent aldehyde dehydrogenases (COG1012; 2.2% of all cases), predicted oxydoreductases, ABC-type polar amino acid transport system (COG1126; 1.8% of all cases), and Acetyl-CoA acetyltransferase (COG0183; 1.7% of all cases). In contrast, informational functions were the least frequently transferred (12% of all cases); only four informational functions were found among the 100 most transferred functions (i.e., peptide chain release factor RF-3, threonyl-tRNA synthetase, methionine aminopeptidase and methionyl-tRNA synthethase) and none of these categories were related to ribosomal proteins or DNA/RNA polymerases,

The highly conserved genes currently used as phylogenetic markers to reconstruct the Tree of Life [227] were transferred between phyla at extremely low frequencies. Six genes were found to be transferred (Fig 6.7, A, inset) and their frequency was at least 151 times lower compared to the top six most transferred functions. The different abundance of the two sets of genes in the genome, i.e., metabolic and highly conserved, did not account for these results. For instance, the six most transferred metabolic functions were enriched 5 to 16 times in the set of transferred genes relative to their average abundance in the genome, while the six highly conserved genes were 2 to 20 times less abundant in the transferred gene set (Table 6.2). For instance, we found that Arginyl-tRNA synthetase (COG0018) was transferred between *Salinispora tropica* and *Sorangium cellulosum* (all cases are provided in Table 6.3). The low frequency of exchange of informational genes is thought to be related to the high connectivity of their expressed proteins [228] and suggested that phylogenetic reconstruction based on these genes is largely impervious to HGT, at least for the part of the Tree can be robustly resolved by these genes (i.e., within phylum but not phylum-level relationships).

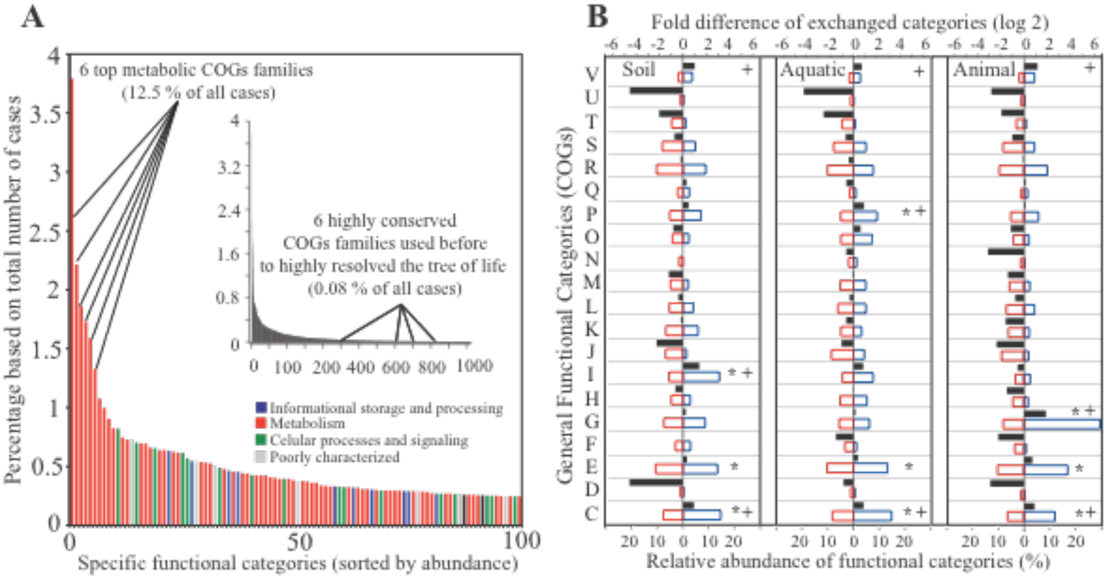


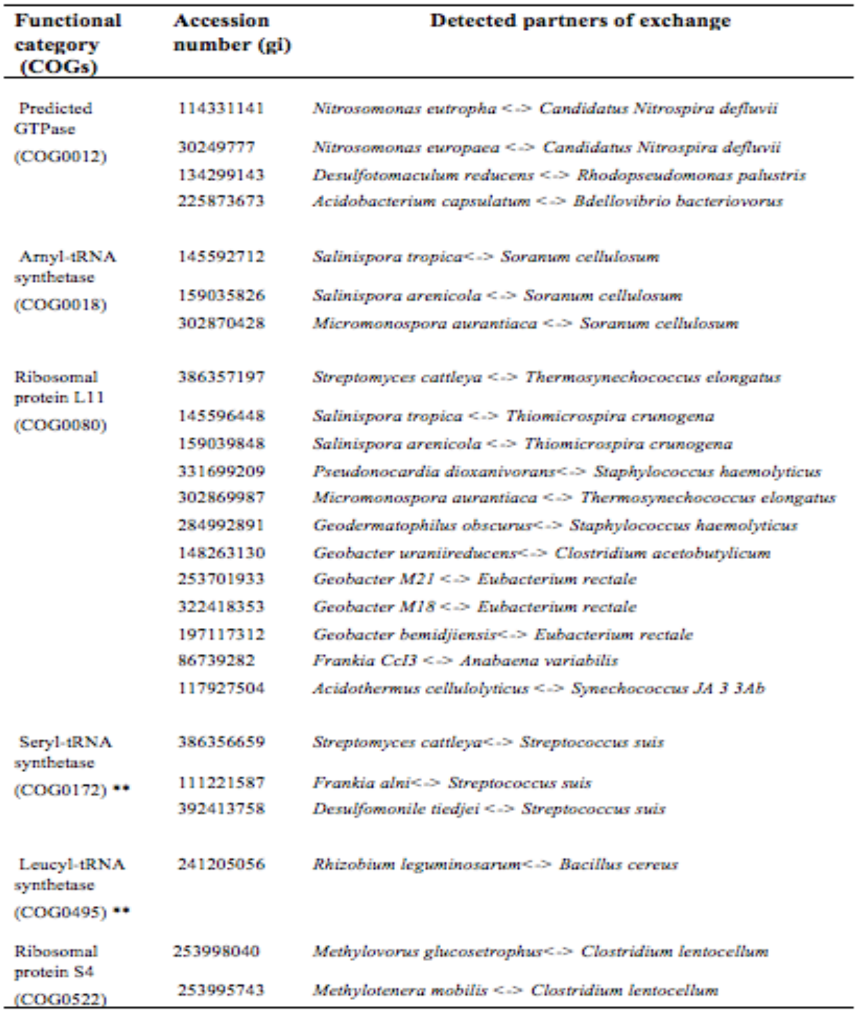
Figure 6.7. Frequency of functional genes transferred across bacterial and archaeal phyla. The top one hundred proteins families (COGs) most frequently transferred across bacterial and archaeal phyla are shown **(Panel A).** Individual COGs are colored based on the major functional category they are assigned to (Figure key). The genomes engaged in the HGT events detected were assigned to one of three major habitats on Earth and the functional enrichment of transferred genes within each habitat is also shown **(Panel B)**. Red bars represent the relative frequency of the COGs categories in the average genome (description of categories is provided in Table A.2). Blue bars represent the relative frequency based on genes exchanged. Black bars represent the fold difference between the previous two frequencies (enrichment). Symbols denote the categories most frequently exchanged (*) and with the higher fold increase (+).

Table 6.2. Comparison of the frequency of inter-phylum HGT between the most transferred metabolic categories and conserved housekeeping genes used to resolved the Tree of Life.

Functional group (COGs)	Functional Classification	Frequency in HGT genes (%)	Frequency in the genome (%)	Ratio (HGT / genome)
COG0080	Informational	0.039	0.059	0.654
COG0012	Informational	0.013	0.059	0.219
COG0018	Informational	0.010	0.056	0.173
COG0172	Informational	0.010	0.061	0.160
COG0522	Informational	0.006	0.061	0.105
COG0495	Informational	0.003	0.055	0.058
Total		0.081	0.351	
COG1028	Metabolic	3.793	0.731	5.185
COG1012	Metabolic	2.215	0.361	6.139
COG0667	Metabolic	1.860	0.147	12.681
COG1126	Metabolic	1.731	0.145	11.965
COG0183	Metabolic	1.587	0.186	8.557
COG0129	Metabolic	1.328	0.082	16.213
Total		12.515	1.651	
Ratio (Metabolic/Informational)		155.4	4.7	

Table 6.3. Detected cases of inter-phyla HGT of highly conserved housekeeping genes.

Notably, the functions with higher frequency of exchange (e.g., NAD-dependent aldehyde dehydrogenases) were also those that have been transfer between organisms from a larger number of different phyla (Fig. 6.8). Therefore, the functions that have been exchanged more frequently are also more promiscuous in terms of the phylogenetic diversity of the partners involved. Thus, it appears that genes assigned to these functional categories might play important roles in metabolic adaptation to several different habitats and organisms.



Functional category (COGs)	Accession number (gi)	Detected partners of exchange
Predicted GTPase (COG0012)	114331141	*Nitrosomonas eutropha <-> Candidatus Nitrospira defluvii*
	30249777	*Nitrosomonas europaea <-> Candidatus Nitrospira defluvii*
	134299143	*Desulfotomaculum reducens <-> Rhodopseudomonas palustris*
	225873673	*Acidobacterium capsulatum <-> Bdellovibrio bacteriovorus*
Arnyl-tRNA synthetase (COG0018)	145592712	*Salinispora tropica<-> Soranum cellulosum*
	159035826	*Salinispora arenicola <-> Soranum cellulosum*
	302870428	*Micromonospora aurantiaca <-> Soranum cellulosum*
Ribosomal protein L11 (COG0080)	386357197	*Streptomyces cattleya <-> Thermosynechococcus elongatus*
	145596448	*Salinispora tropica <-> Thiomicrospira crunogena*
	159039848	*Salinispora arenicola <-> Thiomicrospira crunogena*
	331699209	*Pseudonocardia dioxanivorans<-> Staphylococcus haemolyticus*
	302869987	*Micromonospora aurantiaca <-> Thermosynechococcus elongatus*
	284992891	*Geodermatophilus obscurus<-> Staphylococcus haemolyticus*
	148263130	*Geobacter uraniireducens<-> Clostridium acetobutylicum*
	253701933	*Geobacter M21 <-> Eubacterium rectale*
	322418353	*Geobacter M18 <-> Eubacterium rectale*
	197117312	*Geobacter bemidjiensis<-> Eubacterium rectale*
	86739282	*Frankia CcI3 <-> Anabaena variabilis*
	117927504	*Acidothermus cellulolyticus <-> Synechococcus JA 3 3Ab*
Seryl-tRNA synthetase (COG0172) **	386356659	*Streptomyces cattleya<-> Streptococcus suis*
	111221587	*Frankia alni<-> Streptococcus suis*
	392413758	*Desulfomonile tiedjei <-> Streptococcus suis*
Leucyl-tRNA synthetase (COG0495) **	241205056	*Rhizobium leguminosarum<-> Bacillus cereus*
Ribosomal protein S4 (COG0522)	253998040	*Methylovorus glucosetrophus<-> Clostridium lentocellum*
	253995743	*Methylotenera mobilis <-> Clostridium lentocellum*

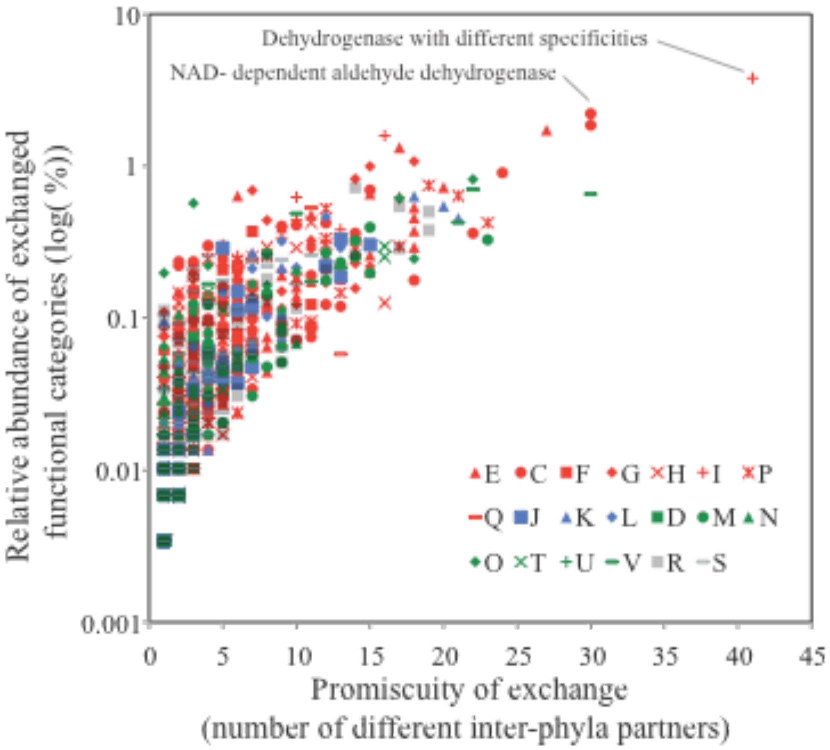


Figure 6.8. Relationship between frequency of HGT and promiscuity. All exchanged genes (p-value < 0.001) were assigned to an individual COG (Table A.2) and the relative abundance of the COG (y-axis) is plotted against the number of genomes (promiscuity) that exchanged the genes assigned to the COG (x-axis). Red symbols represent metabolic categories, green symbols represent cellular processes and signaling, blue symbols represent informational storage and processing, and gray symbols represent poorly characterized functions. Note that the higher the frequency of exchange the higher usually the promiscuity of the exchanged (i.e., more different genomes exchanged the corresponding genes/COG). For instance, the "NAD-dependent aldehyde dehydrogenase" one of the most transferred categories has been exchanged across 30 different pairs of phyla.

The functional biases in exchanged genes within soil, aquatic and animal-associated organisms were examined more closely to elucidate what functions are selected within each corresponding environment. Categories enriched in each environment included: lipid transport and metabolism (I) was most abundantly exchanged in soil, inorganic ion transport and metabolism (P) in aquatic habitats, and carbohydrate transport and metabolism (G) among animal-associated bacteria. These three categories were also found to be among the ones with the highest fold increase in the transferred gene set compared to genome average (Fig. 6B, black bar). These results suggested that the functions exchanged across phyla do not represent random collections of genes but rather reflected the acquisition of ecologically important functions for the corresponding organisms within their habitat(s) (fixed HGT events).

6.4.5 The Role of Inter-Phylum HGT in Bacterial Adaptation.

To examine the importance of genetic exchange between distantly related organisms for adaptation and ecology, the genome pairs with the highest number of exchanged genes were further analyzed, focusing on transferred regions with two or more syntenic genes (Table A.3). As expected, the analysis of exchanged genes between specific genomes reflected the general trends mentioned above for the complete genome set (e.g., Fig 6.7, B). Here, three examples that clearly demonstrate the importance of inter-phylum HGT for acquiring metabolic capabilities essential for the ecological niches of the recipient organism are highlighted.

One of the most interesting cases of inter-phylum HGT is between the syntrophic bacteria *Pelotomaculum thermopropionicum* (*Firmicutes*) and *Syntrophobacter fumaroxidans* (*Proteobacteria*). Three large, syntenic regions, encoding mostly genes involved in the electron transport chain for ATP production and active transport of nitrate or sulfonate, were identified between representatives of these taxa with significantly higher amino acid identity than expected by vertical decent. The identity of these regions, ranged from 82% to 62% with an average of 67%; this level of identity is significantly higher than average identity of the ribosomal proteins, (61%). Further, the genes in syntenic region 2 and 3 (Table A.3) appeared to be involved in reverse electron transport during syntrophic propionate metabolism and be fundamental for the establishment of successful syntrophic relationships [76]. Propionate is an important intermediate in the conversion of complex organic matter under anaerobic conditions and its oxidation to acetate requires the presence of a methanogenic partner to maintain low hydrogen partial pressure [229]. These results not only show clear evidence of genetic exchange between distantly related organisms but also, more importantly, suggest that overlapping ecology within anoxic environments had favored the exchange of key adaptive genes.

Another notable case was *Listeria ivanovvi* (*Firmicutes*) and *Sebaldella termiditis* (*Fusobacteria*), where 11 syntenic regions were exchanged, encoding genes associated to carbohydrate metabolism and transport (Table A.3). The largest region, syntenic region 4, encodes for 16 genes involved in propanediol utilization pathway. This represents potentially an important ecological function since these organisms have been associated with the ruminant and the termite gut (*L. ivannovi* and *S. termiditis*, respectively) and propanediol is thought to be important in these anoxic environments [230]. Propanediol is a major product of the anaerobic degradation of common plant sugars (e.g., rhamnose and fucose); however, its degradation is highly toxic and bacteria need micro-compartments (carboxysomes) to enclose the highly reactive intermediates of the degradation [231]. Consistent with this, several carboxysome structural proteins were also exchanged between these genomes (e.g., gi numbers 347548556 and 269119660) relatively recently, as reflected by the high amino acid identities, ranging from 57% to 85%. These findings suggest that the capabilities for degradation of plant sugars under anaerobic conditions have been transferred between phyla multiple times, and might have been fundamental for adaptation to the animal gut environment.

Noteworthy cases of gene transfer between oral-associated bacteria, *Streptococcus gordonni* (*Firmicutes*) and *Leptotricha buccalis* (*Fusobacteria*), were also observed, where nine syntenic regions (Table A.3), mainly related to carbohydrate transport and metabolism, were exchanged. Among these regions, an operon of seven genes related to the degradation of lactose through the tagatose 6-phosphate pathway, with amino acid identities ranging from 63 to 82%, was observed. Lactose is an important component of the human diet and it has been suggested that lactose catabolism can influence the ecological balance of oral bacteria and colonization of oral cavities and soft tissues [232, 233].

As expected, the main mechanism underlying these inter-phylum HGT events was non-homologous recombination based on several lines of evidence. Several transferred genes were flanked by transposases and integrases as exemplified by the HGT event between *Desulfuruspirillum indicum* (*Chrysiogenetes*) and *Marinobacter aquaeolei* (*Proteobacteria*), where a cation efflux pump gene was recently exchanged (97% amino acid identity), flanked by transposases and integrase genes (99.3% amino acid identity) (Table S2). Additionally, syntenic phage-related proteins (~50 genes) were shared among aquatic bacteria, *Candidatus Nitrospira defluvii* (*Nitrospira*) and *Janthinobacterium sp*. strain Marseille (*Proteobacteria*), with high identity (85% average amino acid identity), indicating recent genetic exchange.

6.5 Conclusions and Perspectives

Exchange between distant related organisms representing different bacterial and/or archaeal phyla is thought to be very infrequent [234]; however, our analysis revealed that inter-phylum exchanges had occurred in almost all of the evaluated genomes. Analysis of networks of HGT revealed that lifestyle and ecology drive most of the HGT events, especially the ones involving a large number of genes exchanged (massive HGT) and metabolic genes. This analysis also revealed that metabolic genes are exchange twice as frequent among anaerobic organisms compared to aerobic ones.

Extensive HGT among thermophiles, pathogens and cyanobacteria has been described previously, e.g., "highways" of HGT [7, 235], and was attributed to substantial ecological overlap among the partner genomes. Along the same lines, a recent study of intra-phylum HGT showed that very recent gene transfer (reflected by 99% nucleotide sequence identity) is clearly structured by ecology, where the highest frequency of HGT was observed among organisms recovered from the same site of the human body [70]. None of these previous studies, however, described cases of such extensive inter-phylum HGT as those described by the network analysis presented here or evaluated the environmental and ecological parameters that account for the "highways" of HGT. In contrast to what was previously reported, our results showed that the most extensive genetic exchange occurs among mesophilic organisms with saccharolytic and fermenting metabolisms, mainly associated to anoxic environments characterized by high plant organic matter concentration (e.g., termite gut, ruminant gut and anaerobic sludge). The differences between our findings and those reported previously might be related to the normalization of the database (mostly overrepresented by human pathogens) and the fact that our method evaluated recent as well as more ancient HGT events (e.g., amino acid sequence identity < 60 %).

It is also important to point out that, due to the still limited representation of the total natural microbial diversity by genome sequences, many more cases of extensive inter-phylum HGT evade detection currently. Advancements in DNA sequencing and single-cell technologies have exponentially lowered the cost of genome sequencing and, as a consequence, the pace at which natural diversity is being characterized is continuously increasing. To keep up with this trend, faster methods for HGT detection are needed and the simple strategy presented here, which is based on comparisons in genome triplets and the statistical significance of the identity of a match, provides means for fast HGT detection. In addition, our strategy provides a standardized framework to compare rates of HGT between organisms, identify the putative partners of exchange, and assess the functions exchanged.

Extensive HGT within anaerobic mesophilic environments was first described between *Sphaerochaeta spp* and *Clostridia* [215]. In total, 37 cases with more extensive HGT than that observed in *Sphaerochaeta* were detected in the present study; 28 of the

37 involved also anaerobic mesophilic organisms like *Sphaerochaeta*. Inspection of the individual genes exchanged suggested that the ability to engage in syntrophic metabolism, degrade toxic intermediates of plant organic matter, and metabolize sugars in the oral cavity, have been exchanged across phyla several times during the relative recent evolutionary history. It thus appears that inter-phylum HGT has not only affected a substantial part of the genome in almost every bacterium but also it has been fundamental for the adaptation of these organisms to their perspective ecological niche(s). These data suggest that members of some communities essentially share their metabolism through a network of HGT, while preserving phylogenetic distinctiveness at housekeeping genes, and that barriers to genetic exchange among distantly related organisms may not be as strong as previously thought. Therefore, although members of microbial communities appear to share metabolic genes and pathways as a somewhat “common good”, they remain distinct and phylogenetically tractable at their highly conserved genes.

6.6 Acknowledgments

This work was supported in part by the National Science Foundation under grant 0919251.

CHAPTER 7

SUMMARY AND PERSPECTIVES

7.1. HR as a Mechanism of Genetic Coherence within and between Species

The case of spatially co-occurring *Shewanella baltica* isolates has expanded our understanding of the rate and mode of bacterial evolution. The comparative analyses of *S. baltica* genomes revealed a unique case of unconstrained gene exchange between strains sharing similar ecology, where no spatial (syntenic) or functional biases were observed (Chapters 2 and 3). Such patterns of recombination can serve as a homogenization force to purge polymorphisms within populations and maintain genetic cohesiveness according to the Biological Species Concept. Thus, the *S. baltica* genomes analyzed here appear to evolve sexually, mediated by homologous recombination, similar to reproduction in higher eukaryotes. However, it remains unclear whether the *S. baltica* case represents a rare example or the norm. More populations and habitats must be analyzed before a more complete understanding of the influence of the environment on evolutionary processes such as recombination can emerge.

On the other hand, an intrinsic preference to recombine with close relatives does not necessarily lead to population cohesion or species convergence, especially in cases where genetic exchange is limited to a few environmentally important functions, like in the *Campylobater* case (Chapter 4). Initially, MLST analysis indicated that *C. coli* and *C. jejuni* species were converging (merging) due to high levels of HR, resulting from expansion of the ecological niche of *C. coli* into that of *C. jejuni* [9]. Our reanalysis of the MLST data and additional genomic comparisons showed that, even though higher levels of HR were indeed observed between *C. coli* and *C. jejuni* compared to other bacterial species, the recombined genes were constrained to a few parts of the genome and represented mostly environmentally-selected functions such as antibiotic resistance and flagella biosynthesis and were mainly mediated by non-homologous recombination. These results suggests that the two distinct species were unlikely to be converging via HR [37]. A more recent study of 42 strains of *C. coli* and 43 strains of *C. jejuni* confirmed that the patterns of HR observed between the two species did not support convergence or “despeciation” [236].

7.2 HGT Between Distantly Related Organisms Can Be Massive and Spread Metabolic Adaptations

The large genetic exchange observed between *Sphaerochaeta* and *Clostridiales*, two distinct bacterial phyla, is unprecedented among mesophilic organisms (Chapter 5). Such high inter-phylum HGT had been previously described in organisms living under extreme conditions, like thermophilic [214] and halophilic organisms [216]. HGT in the *Sphaerochaeta*-*Clostridiales* case was favored by overlapping ecological niche(s) and/or strong functional interactions within anaerobic food webs. The latter was evident by the fact that transferred genes were heavily biased toward carbohydrate uptake and fermentative metabolism functions, including complete operons. These findings reveal that, contrary to previous observations [214, 216], high genetic exchange might also occur been distantly related genomes that live in non-extreme environments.

Even though the fixation of genes exchanged between distant organisms is believed to be very infrequent due to their deleterious effects and the incompatibility conferred by molecular mechanisms (e.g., defense mechanisms against foreign DNA, incompatible codon usage and transcription regulation), our comparative analyses of all available genomes revealed that large genetic exchange across phyla is more common than previously anticipated and can account for up to one third of all metabolic genes in the genome of certain organisms (Chapter 6). Thus, inter-phylum genetic exchange has contributed significantly to the adaptation of the recipient genomes. The partners of inter-phylum genetic exchange revealed the existence of several networks of high HGT that are driven by ecological and physiological factors. Interestingly, exchange of metabolic genes appeared to be more frequent among anaerobic organisms based on these networks. Nonetheless, universal genes, e.g., ribosomal proteins, DNA polymerase, were exchanged across phyla at least 150 times less frequently than most metabolic genes, suggesting that reconstruction of the species phylogeny and the bacterial Tree based on the former genes is reliable.

7.3 Future and Perspectives

The analysis of *S. baltica* genome sequences represents a clear example of how frequent HR can contribute to population genetic cohesiveness. Nevertheless, these genome sequences represent only a single snapshot in the evolution of the "species", preventing a more accurate estimate of the rate of HR and its effect on populations structure. Advancing our understanding of these dynamics requires a continuous monitoring of genetic events within populations. Experimental evolution studies (i.e., mesocosms) provide means to evaluate rates of HR in recombinogenic bacteria (e.g., *S. baltica*, *Vibrio cholerae*) while controlling for environmental fluctuations. Monitoring of these systems through time-series metagenomics and single-cell genomics would allow to robust estimate population genetic parameters and HGT as well as to determine the genes exchanged and their selective advantages, if not neutral. Such research efforts could elucidate the modes and tempo of population adaption and the importance of HR in the maintenance of genetic coherence.

The analysis of inter-phyla HGT, as shown in Chapter 6, suggests that a large proportion of metabolic genes have been exchanged between organisms characterized by similar life styles (i.e., fermentative, syntrophic), revealing that HGT has been an important processes in the optimization of the metabolic capabilities of bacteria. Nevertheless, most of the detected inter-phyla exchanges are unlikely to be recent based on the percentage identity of the exchanged genes and the different source of isolation of the putative partners. In order to evaluate how significant HGT is between distantly related genomes (also applicable to closely related ones) in the short term adaptation of bacteria, future studies should aim to recover the genomic diversity of organisms co-existing in the same habitat (i.e., comprehensive sampling of termite gut microbes), and to follow these communities through time. Metagenomic technologies present an opportunity to sample this genetic diversity; however, the fragmented nature of the technology (i.e., short DNA sequences) makes the disentangling of population diversity and detecting of HGT challenging. Furthermore, in typical metagenomics studies the sample is homogenized during the DNA extraction, destroying the microscale interactions between organisms (i.e., syntrophisms) that might be relevant to link ecology and frequency of HGT (see Chapter 1 Fig 1.1). It will be important to develop of new technologies that allow the study of microbial communities as the microscale level, bypassing the need to isolate the organisms in the laboratory. Microfluidics devices and single cell sequencing [237, 238] provide means to perform such microscale studies. The picture to emerge from such studies will advance our understanding of the role of HGT in the evolution of bacteria, how and what genes spread through populations/species, and how selection acts to fix HGT events.

Finally, the biological interpretation of the detected patterns of genetic exchange will be informative only if a good understanding of the gene functions and physiology of the organisms is available. Even though an increasingly larger fraction of the extant genetic diversity on the planet has been characterized due to improvements in sequencing technologies, little is known about the function and relevance of thousands of available gene sequences. Most of the genes with functional annotation are either classified in broad functional categories or are wrongly classified; many more genes have only hypothetical functions assigned to them. For example, 1/4 of the genes in *Sphaerochaeta* spp. have hypothetical function (Chapter 5). Closing such a large gap between information and function will required the collaborative effort of bioinformaticians and microbiologists to decipher the function of (at least) the abundant and ubiquitous uncharacterized genes and new high throughput methods to functionally characterized gene sequences. Characterization of the gene functions would help elucidating the physiological role of the corresponding organisms in the environments. Such information for unculturable organisms is currently challenging, but emerging technologies such as nanometer-scale secondary-ion mass spectrometry (NanoSIMS) and transcriptomics, can potentially allow the monitoring of microbial activities *in-situ*. These efforts will provide an important framework not only to better interpret HGT patterns but also to better study ecology and evolution of Bacteria in general.

APPENDIX A

TABLES

Table A.1 Exchanged genes between *S. baltica* OS195 and the other strains

OS195 gi #	Gene Annotation	Recombinant	OS185	OS155	OS223
160873241	hypothetical protein Sbal195 0115	*OS223*	97.17	92.14	93.71
160873439	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein	*OS223*	90.78	97.45	99.09
160873464	hypothetical protein Sbal195 0339	*OS223*	90.6	91.86	99.88
160873466	histidine ammonia-lyase	*OS223*	97.21	96.7	99.87
160873648	outer membrane efflux protein	*OS223*	97.35	96.83	98.84
160873649	secretion protein HlyD family protein	*OS223*	98.05	97.74	99.79
160873851	periplasmic serine protease DegS	*OS223*	98.06	98.34	100
160873867	hypothetical protein Sbal195 0745	*OS223*	97.27	96.86	99.18
160874169	transport system permease protein	*OS223*	97.44	97.25	98.82
160874395	GTP-binding protein LepA	*OS223*	98.72	98.72	100
160874571	polyketide-type polyunsaturated fatty acid synthase PfaA	*OS223*		94.3	94.5
160874611	lipid-A-disaccharide synthase	*OS223*	96.83	96.74	97.74
160874670	hypothetical protein Sbal195 1553	*OS223*	97.47	97.11	99.76
160874671	beta-lactamase	*OS223*	95.78	97.49	97.81
160874793	nuclease SbcCD, D subunit	*OS223*	96.95	97.29	100
160874988	cystathionine beta-lyase	*OS223*	97.33	97.42	100
160874989	integral membrane sensor signal transduction histidine kinase	*OS223*	97.83	97.33	99.86
160875277	hypothetical protein Sbal195 2164	*OS223*	99.8	97.96	100
160875335	response regulator receiver modulated metal dependent phosphohydrolase	*OS223*	98.41	91.09	99.3
160875337	phage integrase family protein	*OS223*	95.45	92.9	99.92
160875520	ribonucleotide-diphosphate reductase subunit beta	*OS223*	96.9	98.54	99.66
160875771	inosine kinase	*OS223*	96.78	98.01	98.47
160875772	ferrochelatase	*OS223*	95.69	97.65	99.8
160875794	hypothetical protein Sbal195 2682	*OS223*	99.56	88.56	99.89
160876081	AMP-dependent synthetase and ligase	*OS223*	96.89	95.76	99.95
160876162	flagellar biosynthesis regulator FlhF	*OS223*	96.17	96.17	98.55
160876255	hypothetical protein Sbal195 3149	*OS223*	96.22	96.87	100
160876399	hypothetical protein Sbal195 3293	*OS223*	96.58	98.72	100
160876725	23S rRNA methyluridine methyltransferase	*OS223*	96.59	96.93	100
160876793	hypothetical protein Sbal195 3688	*OS223*	97.6	97.82	100
160876859	putative manganese transporter	*OS223*	96.77	98.58	99.53
160876947	peptidase S9B dipeptidylpeptidase IV subunit	*OS223*	94.86	96.63	98.1
160877253	Na+/H+ antiporter NhaC	*OS223*	96.72	96.65	97.86

160877271	AraC family transcriptional regulator	*OS223*	96.18	98.16	100
160877272	PEBP family protein	*OS223*	98.15	98.7	100
160877392	O-succinylbenzoate synthase	*OS223*	96.37	95.55	99.82
160877565	formamidopyrimidine-DNA glycosylase	*OS223*	97.79	96.69	98.9
160877567	SNARE associated Golgi protein	*OS223*	98.8		100
160873305	DNA-binding transcriptional repressor FabR	*OS185*	100	97.4	96.26
160873401	hypothetical protein Sbal195 0275	*OS185*	100	97.12	97.59
160873457	3-oxoacyl-(acyl carrier protein) synthase II	*OS185*	98.33	94.84	95.71
160873505	hypothetical protein Sbal195 0380	*OS185*	100	97.11	96.88
160873507	orotate phosphoribosyltransferase	*OS185*	100	95.79	97.35
160873509	peptidase S9 prolyl oligopeptidase	*OS185*	100	91.25	98.34
160873524	isopropylmalate isomerase large subunit	*OS185*	99.93	96.8	96.94
160873527	glycerol kinase	*OS185*	97.85	96.23	96.97
160873538	UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase	*OS185*	100	97.34	97.43
160873539	UDP-N-acetylmuramate--alanine ligase	*OS185*	100	97.07	97.21
160873546	preprotein translocase subunit SecA	*OS185*	99.96	98.46	98.31
160873552	twin-arginine translocation protein, TatB subunit	*OS185*	100	95.35	96.32
160873554	2-polyprenylphenol 6-hydroxylase	*OS185*	100	96.97	98
160873567	TonB-dependent receptor	*OS185*	97.37	95.2	96.2
160873611	hypothetical protein Sbal195 0489	*OS185*	99.87	97.34	98.01
160873613	band 7 protein	*OS185*	100	96.69	95.51
160873614	band 7 protein	*OS185*	100	97.12	97.34
160873622	regulatory protein CsrD	*OS185*	99.95	98.05	98.2
160873626	MSHA biogenesis protein MshK	*OS185*	100	97.51	95.95
160873627	pilus (MSHA type) biogenesis protein MshL	*OS185*	98.63	96.37	96.43
160873628	MSHA biogenesis protein MshM	*OS185*	100	98.02	98.35
160873629	TPR repeat-containing protein	*OS185*	97.94	91.96	91.96
160873630	type II secretion system protein E	*OS185*	99.89	96.88	97.39
160873631	type II secretion system protein	*OS185*	100	95.09	97.95
160873632	hypothetical protein Sbal195 0510	*OS185*	100	97.98	91.31
160873633	MSHA pilin protein MshB	*OS185*	100	95.79	95.14
160873639	hypothetical protein Sbal195 0517	*OS185*	98.55	94.54	94.71
160873641	rod shape-determining protein MreC	*OS185*	100	96.95	97.33
160873643	maf protein	*OS185*	100	98.01	97.51
160873815	diguanylate cyclase with PAS/PAC sensor	*OS185*	99.9	98.8	98.7
160873829	protein of unknown function RIO1	*OS185*	99.77	99.19	99.07
160873844	ABC transporter related	*OS185*	99.64	96.88	96.88
160873871	TonB-dependent siderophore receptor	*OS185*	99.18	97.58	97.63
160873872	putative hydroxylase	*OS185*	100	94.33	97.23
160873928	tRNA-dihydrouridine synthase A	*OS185*	99.51	96.67	96.37
160873931	enoyl-CoA hydratase/isomerase	*OS185*	100	98.24	98.37
160873955	ABC transporter related	*OS185*	98.9	94.68	95.12
160873957	peptidase M50	*OS185*	100	98.51	98.16
160873975	diguanylate cyclase	*OS185*	98.88		96.78
160873990	amino acid carrier protein	*OS185*	98.38	95.51	95.14

160873997	curlin-associated protein	*OS185*	99.93	94.77	99.53
160873998	curlin-associated protein	*OS185*	100	96.43	95.71
160873999	hypothetical protein Sbal195 0878	*OS185*	100	98.44	97.27
160874027	hypothetical protein Sbal195 0906	*OS185*	100	97.42	98.66
160874104	ABC transporter related	*OS185*	100	93.8	94.35
160874105	polar amino acid ABC transporter, inner membrane subunit	*OS185*	100	95.01	95.01
160874111	GCN5-related N-acetyltransferase	*OS185*	100	98.17	96.95
160874207	D-isomer specific 2-hydroxyacid dehydrogenase NAD-binding	*OS185*	99.79	95.56	95.35
160874215	ATPase central domain-containing protein	*OS185*	100	96.83	97.49
160874253	hypothetical protein Sbal195 1133	*OS185*	100	98.49	97.42
160874255	diguanylate cyclase	*OS185*	100	97.85	98.01
160874259	von Willebrand factor type A	*OS185*	97.2	96.58	96.68
160874261	hypothetical protein Sbal195 1141	*OS185*	99.17	92.29	96.42
160874339	methyl-accepting chemotaxis sensory transducer	*OS185*	98.92	94.06	97.12
160874389	hypothetical protein Sbal195 1270	*OS185*	100	97.05	96.15
160874390	L-aspartate oxidase	*OS185*	99.94	97.96	97.96
160874457	uroporphyrin-III C/tetrapyrrole methyltransferase	*OS185*	99.37	97.1	97.35
160874486	multi anti extrusion protein MatE	*OS185*	99.04	97.57	97.42
160874500	1-deoxy-D-xylulose-5-phosphate synthase	*OS185*	97.97	97.22	97.59
160874508	hypothetical protein Sbal195 1390	*OS185*	100	97.98	96.28
160874510	putative RNA 2'-O-ribose methyltransferase	*OS185*	99.91	98.25	97.42
160874530	diguanylate cyclase	*OS185*	99.7	93.91	94.21
160874582	hypothetical protein Sbal195 1465	*OS185*	100	98.28	98.28
160874589	pseudouridine synthase	*OS185*	100	98.15	96.64
160874614	tRNA(Ile)-lysidine synthetase	*OS185*	97.7	94.17	94.77
160874615	diguanylate cyclase	*OS185*	100	98.09	98.32
160874623	fructokinase	*OS185*	98.04	97.06	97.6
160874634	transcriptional regulator, TyrR	*OS185*	99.61	97.53	97.92
160874638	outer membrane protein W	*OS185*	100	93.52	93.67
160874639	short-chain dehydrogenase/reductase SDR	*OS185*	99.87	98.79	99.46
160874641	homoserine O-succinyltransferase	*OS185*	99.47	96.92	97.45
160874644	acyl-CoA dehydrogenase domain-containing protein	*OS185*	98.45	97.06	96.46
160874645	enoyl-CoA hydratase	*OS185*	100	98.19	97.67
160874646	enoyl-CoA hydratase/isomerase	*OS185*	99.91	96.96	96.35
160874656	FAD linked oxidase domain-containing protein	*OS185*	98.96		96.18
160874658	hypothetical protein Sbal195 1541	*OS185*	98.09	98.41	86.76
160874659	hypothetical protein Sbal195 1542	*OS185*	99.29	99.52	93.7
160874669	acetyl-CoA hydrolase/transferase	*OS185*	98.76	97.13	97.59
160874716	AraC family transcriptional regulator	*OS185*	99.28	95.1	92.83
160874730	decaheme cytochrome c	*OS185*	99.27	99.13	
160874735	ferrous iron transport protein B	*OS185*	98.47	96.3	96.34
160874746	ATP-dependent protease La	*OS185*	99.41	98.39	98.35
160874748	PpiC-type peptidyl-prolyl cis-trans isomerase	*OS185*	100	97.64	97.91

160874749	TOBE domain-containing protein	*OS185*	100	97.18	94.84
160874750	trans-2-enoyl-CoA reductase	*OS185*	100	97.34	98.17
160874752	oligopeptide/dipeptide ABC transporter, ATPase subunit	*OS185*	100	97.22	97.62
160874753	binding-protein-dependent transport systems inner membrane component	*OS185*	99.89	97.53	96.75
160874754	binding-protein-dependent transport systems inner membrane component	*OS185*	100	96.12	97.09
160874755	extracellular solute-binding protein	*OS185*	100	97.53	97.59
160874756	Fis family transcriptional regulator	*OS185*	99.63	97.71	97.62
160874765	histone deacetylase superfamily protein	*OS185*	100	96.81	97.91
160874769	ATP-dependent DNA helicase DinG	*OS185*	100	97.88	97.59
160874770	DNA polymerase II	*OS185*	98.23	96.91	96.87
160874771	porin	*OS185*	98.84	95.63	95.3
160874775	transposase, IS4 family protein	*OS185*	100	94.35	96.98
160874777	transposase IS4 family protein	*OS185*	100	93.77	94.16
160874794	SMC domain-containing protein	*OS185*	99.8	96.5	99.9
160874799	DEAD/DEAH box helicase domain-containing protein	*OS185*	99.85	97.27	98.01
160874873	malonyl CoA-acyl carrier protein transacylase	*OS185*	100	98.06	97.41
160874879	6-phosphogluconate dehydrogenase NAD-binding	*OS185*	100	97.83	97.15
160874881	acyl-CoA dehydrogenase domain-containing protein	*OS185*	98.66	95.64	95.92
160874882	TonB-dependent siderophore receptor	*OS185*	99.78	96.91	94.75
160874885	peptidase M28	*OS185*	100	95.23	95.17
160874886	UMP phosphatase	*OS185*	100	97.19	96.79
160874887	phosphoribosylglycinamide formyltransferase	*OS185*	100	98.6	98.91
160874891	Na+/H+ antiporter NhaC	*OS185*	99.77	98.26	98.33
160874942	peptidase M48 Ste24p	*OS185*	99.93	98.1	98.1
160874960	hypothetical protein Sbal195 1845	*OS185*	99.66	96.79	95.99
160875042	two component transcriptional regulator	*OS185*	99.85	95.91	95.6
160875050	hypothetical protein Sbal195 1935	*OS185*	98.81	96.63	97.25
160875167	hypothetical protein Sbal195 2053	*OS185*	98.48	97.74	97.86
160875208	exodeoxyribonuclease V, gamma subunit	*OS185*	98.3	93.86	95.06
160875209	transglutaminase domain-containing protein	*OS185*	98.87	94.97	
160875210	hypothetical protein Sbal195 2097	*OS185*	99.91	95.9	97.36
160875212	diguanylate phosphodiesterase	*OS185*	99.57	95.89	96.88
160875215	putative sulfite oxidase subunit YedY	*OS185*	99.9	96.71	96.52
160875280	SecC motif-containing protein	*OS185*	100	96.21	96.97
160875281	hypothetical protein Sbal195 2168	*OS185*	100	98.09	97.74
160875326	hypothetical protein Sbal195 2213	*OS185*	100	95.82	95.24
160875328	heavy metal translocating P-type ATPase	*OS185*	99.25	95.67	95.12
160875330	cytochrome c oxidase, cbb3-type, subunit III	*OS185*	100	96.49	97.11
160875334	hypothetical protein Sbal195 2221	*OS185*	99.59	89.3	88.27
160875458	integral membrane sensor hybrid histidine kinase	*OS185*	99.11	95.35	97.29
160875530	TonB-dependent receptor plug	*OS185*	100	78.75	99.14

160875531	Holliday junction DNA helicase B	*OS185*	95.42	93.93	96.02
160875536	diguanylate cyclase	*OS185*	98.35	97.74	97.2
160875537	putative methyltransferase	*OS185*	100	97.81	97.54
160875538	putative methyltransferase	*OS185*	100	97.49	98.29
160875539	hypothetical protein Sbal195 2427	*OS185*	100	98.79	98.79
160875540	gonadoliberin III-related protein	*OS185*	100	96.81	97.62
160875541	alpha-L-glutamate ligase-like protein	*OS185*	100	97.88	97.78
160875542	response regulator receiver protein	*OS185*	100	96.4	96.56
160875543	LysR family transcriptional regulator	*OS185*	100	97.58	97.03
160875544	protein-glutamate O-methyltransferase	*OS185*	96.45	95.29	94.09
160875549	UBA/THIF-type NAD/FAD binding protein	*OS185*	98.37	94.68	94.35
160875550	thiamine-phosphate pyrophosphorylase	*OS185*	100	94.24	94.08
160875565	glucan 1,4-alpha-glucosidase	*OS185*	100	96.24	99.03
160875570	DNA-directed DNA polymerase	*OS185*	99.92	97.96	97.18
160875571	cupin 4 family protein	*OS185*	100	98.88	97.93
160875572	DNA polymerase III, epsilon subunit	*OS185*	100	97.77	97.29
160875573	LacI family transcription regulator	*OS185*	100	97.24	96.86
160875575	methyl-accepting chemotaxis sensory transducer	*OS185*	99.94	97.19	97.07
160875616	serine O-acetyltransferase	*OS185*	99.88	98.66	98.18
160875617	RNA methyltransferase	*OS185*	100	97.67	98.08
160875619	LolC/E family lipoprotein releasing system, transmembrane protein	*OS185*	99.92	98.24	97.92
160875624	hypothetical protein Sbal195 2512	*OS185*	99.3	98	97.5
160875716	glycoside hydrolase family protein	*OS185*	98.14	97.34	98.58
160875721	DNA topoisomerase I	*OS185*	99.58	97.65	97.92
160875722	succinylarginine dihydrolase	*OS185*	100	96.7	97
160875830	zinc carboxypeptidase-related protein	*OS185*	99.9	96.5	98.06
160875837	hypothetical protein Sbal195 2725	*OS185*	98.2	97.25	97.33
160875838	hypothetical protein Sbal195 2726	*OS185*	99.75	96.6	94.97
160875841	4-hydroxyphenylpyruvate dioxygenase	*OS185*	100	97.6	97.6
160875865	DSBA oxidoreductase	*OS185*	100	96.94	96.94
160875866	NAD-dependent epimerase/dehydratase	*OS185*	100	97.92	98.15
160875867	metal dependent phosphohydrolase	*OS185*	99.92	83.17	97.18
160875868	hypothetical protein Sbal195 2757	*OS185*	100	98.14	96.27
160875870	cell division protein ZipA	*OS185*	99.81	93.13	94.22
160875879	formate/nitrite transporter	*OS185*	100	96.07	97.08
160875902	hypothetical protein Sbal195 2791	*OS185*	99.92	97.83	97.58
160875906	hypothetical protein Sbal195 2795	*OS185*	99.48	97.1	97.52
160875933	aldehyde oxidase and xanthine dehydrogenase molybdopterin binding	*OS185*	99.38	97.15	97.82
160875934	2Fe-2S iron-sulfur cluster binding domain-containing protein	*OS185*	100	98.27	96.92
160875935	hypothetical protein Sbal195 2824	*OS185*	100	96.88	97.15
160875980	acriflavin resistance protein	*OS185*	98.43	97.04	96.73
160875981	hypothetical protein Sbal195 2871	*OS185*	100	99.53	100
160875983	GCN5-related N-acetyltransferase	*OS185*	100	98.17	98.78
160876182	flagellar protein FliS	*OS185*	100	83.94	84.67
160876184	flagellar hook-associated 2 domain-containing protein	*OS185*	99.78	71.52	71.38

160876196	flagellar hook-associated protein FlgL	*OS185*	98.1	96.7	98.27
160876206	flagellar basal body rod protein FlgB	*OS185*	100	98.02	100
160876289	alanine racemase domain-containing protein	*OS185*	99.86	97.42	97.85
160876290	pyrroline-5-carboxylate reductase	*OS185*	100	98.78	98.53
160876292	hypothetical protein Sbal195 3186	*OS185*	100	98.67	97.67
160876307	hypothetical protein Sbal195 3201	*OS185*	100	99.42	97.68
160876308	hypothetical protein Sbal195 3202	*OS185*	100	95.24	96.97
160876439	thioesterase superfamily protein	*OS185*	100	96.06	95.6
160876440	diguanylate cyclase/phosphodiesterase	*OS185*	98.46	96.99	97.05
160876464	DNA repair protein RadA	*OS185*	100	97.88	95.31
160876467	phosphoserine phosphatase SerB	*OS185*	97.15	95.31	94.39
160876471	thymidine phosphorylase	*OS185*	99.02	97.45	97.07
160876503	hypothetical protein Sbal195 3397	*OS185*	99.84	95.35	95.01
160876628	methyl-accepting chemotaxis sensory transducer	*OS185*	100	94.49	93.85
160876702	peptidylprolyl isomerase FKBP-type	*OS185*	100	99.1	99.48
160876734	flavocytochrome c	*OS185*	99.94	98.27	99.05
160876735	D-isomer specific 2-hydroxyacid dehydrogenase NAD-binding	*OS185*	100	98.28	97.78
160876738	TonB-dependent receptor	*OS185*	98.42	97.16	96.8
160876807	hypothetical protein Sbal195 3703	*OS185*	98.05	96.62	97.89
160876826	hypothetical protein Sbal195 3722	*OS185*	100	99.69	99.69
160876827	major facilitator transporter	*OS185*	100	92.16	92.44
160876846	methyl-accepting chemotaxis sensory transducer	*OS185*	96.89	95.12	95.28
160876850	pseudouridine synthase	*OS185*	99.48	97.01	97.15
160876856	glutathione S-transferase domain-containing protein	*OS185*	100	97.79	98.58
160876857	TonB-dependent receptor	*OS185*	99.88	85.1	99.92
160876879	glycerophosphoryl diester phosphodiesterase	*OS185*	99.3	96.87	96.32
160876954	hypothetical protein Sbal195 3850	*OS185*	100	98.07	97.37
160876957	sodium:dicarboxylate symporter	*OS185*	100	98.3	97.83
160876988	glutamine amidotransferase of anthranilate synthase	*OS185*	100	97.09	96.41
160876991	cytochrome c1	*OS185*	100	93.28	93.85
160876992	cytochrome b/b6 domain-containing protein	*OS185*	100	92.17	91.76
160877008	MscS mechanosensitive ion channel	*OS185*	99.35	97.39	98.38
160877059	major facilitator transporter	*OS185*	99.92	95.98	97.4
160877060	glyceraldehyde-3-phosphate dehydrogenase	*OS185*	100	93.79	93.89
160877126	protein of unknown function DUF853 NPT hydrolase putative	*OS185*	100	97.16	96.17
160877128	TRAP transporter solute receptor TAXI family protein	*OS185*	100	97.81	96.76
160877129	TRAP transporter, 4TM/12TM fusion protein	*OS185*	100	97.25	97.64
160877130	hypothetical protein Sbal195 4026	*OS185*	100	96.95	97.18
160877131	diguanylate cyclase with PAS/PAC sensor	*OS185*	100	97.15	98.07
160877133	thioredoxin	*OS185*	100	97.27	98.05
160877134	anion transporter	*OS185*	100	96.86	96.86
160877135	major facilitator transporter	*OS185*	100	98.37	98.71

160877176	LysR family transcriptional regulator	OS185	97.36	91.15	90.72
160877197	adenylate cyclase	OS185	98.26	96.52	96.93
160877202	outer membrane adhesin like proteiin	OS185	98.62		95.95
160877210	ATP-dependent DNA helicase Rep	OS185	99.32	99.13	98.1
160877248	diguanylate cyclase/phosphodiesterase with PAS/PAC and GAF sensor(s)	OS185	98.56	95.65	97.15
160877249	branched-chain amino acid aminotransferase	OS185	99.82	97.99	98.53
160877305	type IV pilus secretin PilQ	OS185	99.71	95.66	98.39
160877306	pilus assembly protein PilP	OS185	100	96.12	94.57
160877307	pilus assembly protein PilO	OS185	100	95.67	92.82
160877309	type IV pilus assembly protein PilM	OS185	100	94.54	98.33
160877310	1A family penicillin-binding protein	OS185	98.46	98.54	96.84
160877508	DNA-directed DNA polymerase	OS185	100	72.15	71.28
160877562	molybdopterin-guanine dinucleotide biosynthesis protein B	OS185	97.68	97.41	96.69
160877571	hypothetical protein Sbal195 4470	OS185	100	99.39	97.97
160874115	MORN repeat-containing protein	OS155	96.41	98.57	97.32
160874503	flagellar motor protein PomA	OS155	98.05	99.74	97.92
160874920	arginine decarboxylase	OS155	97.6	98.75	97.49
160874931	hypothetical protein Sbal195 1816	OS155	98.62	98.95	98.04
160875437	alanine dehydrogenase	OS155	97.76	99.46	95.97
160875712	FAD linked oxidase domain-containing protein	OS155	93.89	98.29	96.61
160875713	phosphoenolpyruvate synthase	OS155	95.49	99.24	98.69
160875947	hypothetical protein Sbal195 2837	OS155	98.2	97.71	96.94
160876240	methyl-accepting chemotaxis sensory transducer	OS155	97.5	98.2	97.01
160876381	tRNA pseudouridine synthase D TruD	OS155	97.09	98.68	97.44
160876387	nucleoside triphosphate pyrophosphohydrolase	OS155	95.29	97.7	93.09
160876412	putative ABC transporter ATP-binding protein	OS155	97	99.22	97.24
160876430	hydrophobe/amphiphile efflux-1 (HAE1) family protein	OS155	97.09	98.51	97.37
160876555	DNA polymerase III, delta subunit	OS155	96.61	99.9	98.16
160876875	ribokinase	OS155	95.5	100	96.05
160876930	ABC transporter related	OS155	97.27	99.78	96.72
160876983	arginine N-succinyltransferase	OS155	98.73	99.71	96.96
160877033	MscS mechanosensitive ion channel	OS155	97.39	98.06	98.06
160877206	OmpA/MotB domain-containing protein	OS155	84.56	99.84	84.56
160877583	rhodanese domain-containing protein	OS155	96.73	100	98.69
160877606	UDP-N-acetylglucosamine pyrophosphorylase	OS155	96.02	97.4	97.76
160873400	methyl-accepting chemotaxis sensory transducer	*OS185	99.49	98.31	97.55
160873402	2OG-Fe(II) oxygenase	*OS185	99.9		98.68
160873403	hypothetical protein Sbal195 0277	*OS185	98.23	98.48	98.74
160873506	ribonuclease PH	*OS185	99.86	97.76	97.9
160873508	GTP cyclohydrolase I	*OS185	100	99.08	97.39
160873510	nucleoid occlusion protein	*OS185	100	97.98	98.48
160873511	deoxyuridine 5'-triphosphate	*OS185	100	99.13	98.47

	nucleotidohydrolase				
160873512	phosphopantothenoylcysteine decarboxylase/phosphopantothenate--cysteine ligase	**OS185*	99.02	95.26	97.71
160873525	isopropylmalate isomerase small subunit	**OS185*	99.01	97.03	100
160873548	diguanylate cyclase	**OS185*	98.98	97.75	97.8
160873549	TatD-related deoxyribonuclease	**OS185*	100	98.51	98.76
160873550	hypothetical protein Sbal195 0425	**OS185*	100	99.05	97.92
160873551	Sec-independent protein translocase, TatC subunit	**OS185*	100	99.07	98.94
160873556	ubiquinone/menaquinone biosynthesis methyltransferase	**OS185*	98.15	97.88	97.88
160873565	putative manganese-dependent inorganic pyrophosphatase	**OS185*	99.78	98.26	98.15
160873609	hypothetical protein Sbal195 0487	**OS185*	100	98.59	98.59
160873610	uridine phosphorylase	**OS185*	100	98.81	98.81
160873612	hypothetical protein Sbal195 0490	**OS185*	99.58	98.51	96.6
160873640	cell shape determining protein MreB	**OS185*	100	98.95	99.43
160873642	rod shape-determining protein MreD	**OS185*	100	97.75	98.77
160873644	ribonuclease G	**OS185*	97.96	98.23	97.75
160873774	MltD domain-containing protein	**OS185*	99.17	98.21	98.21
160873929	hypothetical protein Sbal195 0807	**OS185*	100	98.97	99.66
160873930	phage shock protein C, PspC	**OS185*	100	99	100
160873958	hypothetical protein Sbal195 0836	**OS185*	100	99.47	99.21
160874000	pantoate--beta-alanine ligase	**OS185*	99.88	96.57	97.04
160874001	3-methyl-2-oxobutanoate hydroxymethyltransferase	**OS185*	99.5	98.49	98.24
160874002	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	**OS185*	99.8	97.55	97.55
160874003	poly(A) polymerase	**OS185*	97.41	97.08	97.15
160874028	hypothetical protein Sbal195 0907	**OS185*	100	99.04	98.46
160874081	uroporphyrin-III C-methyltransferase	**OS185*	98.94	97.16	96.69
160874097	dihydropteridine reductase	**OS185*	100	98.47	98.01
160874106	extracellular solute-binding protein	**OS185*	99.73	96.05	93.23
160874192	hypothetical protein Sbal195 1072	**OS185*	99.74	99.08	97.9
160874193	hypothetical protein Sbal195 1073	**OS185*	100	98.73	96.84
160874216	sulfate ABC transporter, ATPase subunit	**OS185*	96.2	97.08	96.2
160874221	hypothetical protein Sbal195 1101	**OS185*	99.81	96.46	90.47
160874223	integral membrane sensor signal transduction histidine kinase	**OS185*	96.3	94.88	94.96
160874232	NADPH-dependent FMN reductase	**OS185*	96.42	95.12	95.12
160874254	hypothetical protein Sbal195 1134	**OS185*	100	98.47	98.01
160874312	rhodanese domain-containing protein	**OS185*	98.02	96.3	96.15
160874387	transcriptional activator NhaR	**OS185*	97.86	97.01	98.61
160874505	thiamine biosynthesis protein ThiI	**OS185*	100	98.83	98.08
160874506	hypothetical protein Sbal195 1388	**OS185*	100	100	98.35
160874507	DNA-binding transcriptional activator GcvA	**OS185*	100	98.25	98.03
160874509	hypothetical protein Sbal195 1391	**OS185*	100	98.47	99.75
160874529	hypothetical protein Sbal195 1411	**OS185*	100	98.47	97.77

160874586	hypothetical protein Sbal195 1469	**OS185*	99.74	98.19	99.74
160874588	hypothetical protein Sbal195 1471	**OS185*	100	97.43	98.86
160874616	potassium efflux system protein	**OS185*	98.77	97.64	97.48
160874640	hypothetical protein Sbal195 1523	**OS185*	100	99.05	98.33
160874661	transcriptional regulator, CadC	**OS185*	97.98	92.29	91.94
160874736	glutaminyl-tRNA synthetase	**OS185*	98.2	97.85	98.08
160874751	ABC transporter related	**OS185*	100	98.73	97.58
160874764	hypothetical protein Sbal195 1647	**OS185*	100	99.66	99.66
160874766	hypothetical protein Sbal195 1649	**OS185*	99.71	97.83	98.55
160874774	transposase, IS4 family protein	**OS185*	100	93.55	99.5
160874810	hypothetical protein Sbal195 1693	**OS185*	99.74	97.14	99.22
160874821	DNA internalization-related competence protein ComEC/Rec2	**OS185*	95.6	97.15	97.49
160874872	3-oxoacyl-(acyl carrier protein) synthase III	**OS185*	99.06	98.12	98.54
160874878	thioesterase superfamily protein	**OS185*	98.75	98.12	97.71
160875174	methyltransferase type 11	**OS185*	100	98.37	99.05
160875216	putative sulfite oxidase subunit YedZ	**OS185*	99.85	98.65	96.71
160875217	lactoylglutathione lyase	**OS185*	99.76	98.78	99.03
160875279	diguanylate cyclase/phosphodiesterase with PAS/PAC sensor(s)	**OS185*	99.91	96.66	100
160875282	Smr protein/MutS2	**OS185*	94.59	94.75	94.42
160875329	hypothetical protein Sbal195 2216	**OS185*	99.79	92.29	92.71
160875332	cytochrome c oxidase, cbb3-type, subunit II	**OS185*	100	98.72	98.09
160875333	cytochrome c oxidase, cbb3-type, subunit I	**OS185*	99.93	98.75	98.75
160875551	thiamine biosynthesis protein ThiC	**OS185*	99.53		96.94
160875569	peptidase S24/S26 domain-containing protein	**OS185*	99.76	98.1	98.1
160875574	hypothetical protein Sbal195 2462	**OS185*	100	99.18	98.49
160875600	Bcr/CflA subfamily drug resistance transporter	**OS185*	98.2	99.1	97.88
160875603	acetolactate synthase 3 regulatory subunit	**OS185*	99.6	97.78	97.78
160875620	lipoprotein releasing system, ATP-binding protein	**OS185*	100	99.43	98.28
160875676	paraquat-inducible protein A	**OS185*	99.84	98.1	97.31
160875677	paraquat-inducible protein A	**OS185*	99.84	97.25	98.22
160875688	uridine kinase	**OS185*	99.84	97.97	98.59
160875700	isocitrate dehydrogenase, NADP-dependent	**OS185*	98.52	97.35	97.3
160875720	hypothetical protein Sbal195 2608	**OS185*	100	98.08	99.36
160875724	sodium:dicarboxylate symporter	**OS185*	97.35	95.77	97.43
160875752	Ion transport protein	**OS185*	99.77	97.54	97.66
160875775	heat shock protein 90	**OS185*	99.43	97.39	97.65
160875802	DNA polymerase III, subunits gamma and tau	**OS185*	97.8	95.59	95.92
160875832	hypothetical protein Sbal195 2720	**OS185*	99.63	95.97	100
160875839	LysR family transcriptional regulator	**OS185*	100	97.6	98.25
160875840	homogentisate 12-dioxygenase	**OS185*	99.83	98.11	97.85
160875842	hexapaptide repeat-containing transferase	**OS185*	99.83	96.92	94.87
160875863	Na+/solute symporter	**OS185*	97.64	96.66	97.35
160875864	hypothetical protein Sbal195 2753	**OS185*	100	99.63	99.63
160875901	putative periplasmic protease	**OS185*	99.8	98.03	99.12
160875936	hypothetical protein Sbal195 2825	**OS185*	99.41	97.74	98.33

160875959	phosphohistidine phosphatase, SixA	**OS185*	100	98.09	97.88
160875960	peptidase M16 domain-containing protein	**OS185*	98.64	97.74	97.71
160875984	peptidase S8 and S53 subtilisin kexin sedolisin	**OS185*	99.63	99.23	98.46
160875990	preprotein translocase subunit SecD	**OS185*	100	98.81	98.65
160875992	queuine tRNA-ribosyltransferase	**OS185*	99.29	98.67	98.76
160876193	transposase, putative	**OS185*	99.52	88.57	94.69
160876203	flagellar hook protein FlgE	**OS185*	99.34	79.25	98.68
160876253	histidyl-tRNA synthetase	**OS185*	99.14	98.59	98.52
160876288	twitching motility protein	**OS185*	99.9	98.17	98.75
160876291	protein of unknown function YGGT	**OS185*	100	99.09	99.45
160876309	hypothetical protein Sbal195 3203	**OS185*	100	98.9	98.53
160876465	type IV pilus assembly PilZ	**OS185*	97.36	87.32	97.15
160876472	deoxyribose-phosphate aldolase	**OS185*	100	99.35	97.15
160876701	endonuclease/exonuclease/phosphatase	**OS185*	98.85	97.74	98.55
160876703	WD-40 repeat-containing protein	**OS185*	98.74	97.79	98
160876733	hypothetical protein Sbal195 3627	**OS185*	99.76	97.08	99.27
160876750	branched-chain amino acid transport system II carrier protein	**OS185*	99.72	98.23	98.23
160876774	putative diguanylate cyclase	**OS185*	98.53	98.13	96.36
160876825	peptidyl-prolyl cis-trans isomerase cyclophilin type	**OS185*	100	98.29	98.8
160876828	nucleotide-binding protein	**OS185*	100	98.35	98.77
160876939	nitrate/nitrite sensor protein NarQ	**OS185*	98.53	93.14	93.19
160876955	hypothetical protein Sbal195 3851	**OS185*	100	79.8	100
160876956	hypothetical protein Sbal195 3852	**OS185*	100		97.45
160876989	ClpXP protease specificity-enhancing factor	**OS185*	100	98.17	100
160876990	stringent starvation protein A	**OS185*	100	99.84	99.37
160877006	phosphatidylserine decarboxylase	**OS185*	99.89	98.29	98.29
160877007	hypothetical protein Sbal195 3903	**OS185*	99.77	97.05	97.39
160877061	redox-active disulfide protein 2	**OS185*	100	97.89	97.89
160877132	peptidylprolyl isomerase FKBP-type	**OS185*	100	98.51	98.51
160877150	bifunctional aconitate hydratase 2/2-methylisocitrate dehydratase	**OS185*	98.39	97.79	97.72
160877189	HupE/UreJ protein	**OS185*	98.79	97.24	98.1
160877198	porphobilinogen deaminase	**OS185*	98.5	98.18	98.29
160877308	fimbrial assembly family protein	**OS185*	99.83	95.56	96.41
160877416	cytochrome c oxidase subunit III	**OS185*	99.89	98.06	97.6
160877570	NAD-dependent epimerase/dehydratase	**OS185*	99.8	96.83	97.52

Table A.2 Description of the COG general functional categories. Adapted from the COG website: http://www.ncbi.nlm.nih.gov/COG/

Category	Description	General category
A	RNA processing and modification	Information processes and signaling
B	Chromatin Structure and dynamics	Information processes and signaling
C	Energy production and conversion	Metabolism
D	Cell cycle control and mitosis	Cellular processes and signaling
E	Amino Acid metabolism and transport	Metabolism
F	Nucleotide metabolism and transport	Metabolism
G	Carbohydrate metabolism and transport	Metabolism
H	Coenzyme metabolism	Metabolism
I	Lipid metabolism	Metabolism
J	Translation	Information processes and signaling
K	Transcription	Information processes and signaling
L	Replication and repair	Information processes and signaling
M	Cell wall/membrane/envelop biogenesis	Cellular processes and signaling
N	Cell motility	Cellular processes and signaling
O	Post-translational modification, protein turnover	Cellular processes and signaling
P	Inorganic ion transport and metabolism	Metabolism
Q	Secondary Structure	Metabolism
T	Signal Transduction	Cellular processes and signaling
U	Intracellular trafficking and secretion	Cellular processes and signaling
Y	Nuclear structure	Cellular processes and signaling
Z	Cytoskeleton	Cellular processes and signaling
R	General Functional Prediction only	Poorly Characterized
S	Function Unknown	Poorly Characterized

Table A.3 Larger cases of genetic exchange across phyla based on probabilistic models

Pelotomaculum thermopropionicum --Syntrophobacter fumaroxidans

Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_147676911	gi_116751364	YP 001211126.1 ABC-type nitrate/sulfonate/bicarbonate transport system, ATPase component	62
1	gi_147676910	gi_116751363	YP 001211125.1 ABC-type nitrate/sulfonate/bicarbonate transport system, periplasmic components	61.9
1	gi_147676909	gi_116751362	YP 001211124.1 ABC-type nitrate/sulfonate/bicarbonate transport system, permease component	65.6
1	gi_147676908	gi_116751361	YP 001211123.1 hypothetical protein PTH 0573	40.8 *
1	gi_147676907	gi_116751360	YP 001211122.1 ABC-type nitrate/sulfonate/bicarbonate transport system, periplasmic components	54.4 *
1	gi_147676906	gi_116751359	YP 001211121.1 hypothetical protein PTH 0571	49.3 *
1	gi_147676905	gi_116751358	YP 001211120.1 permease	64.2
1	gi_147676904	gi_116751357	YP 001211119.1 hypothetical protein PTH 0569	64.6
2	gi_147678107	gi_116751351	YP 001212322.1 transcriptional regulator	68.5
2	gi_147678106	gi_116751350	YP 001212321.1 acyl CoA:acetate/3-ketoacid CoA transferase	79.2
2	gi_147678105	gi_116751349	YP 001212320.1 aromatic ring hydroxylase	81.5
2	gi_147676350	gi_116751348	YP 001212319.1 acyl-CoA dehydrogenases	81.5
2	gi_147676849	gi_116751347	YP 001211064.1 electron transfer flavoprotein	68.3
2	gi_147676352	gi_116751346	YP 001210567.1 electron transfer flavoprotein, alpha subunit	61
2	gi_147676353	gi_116751344	YP 001210568.1 dehydrogenases	67.1
2	gi_147678100	gi_116751343	YP 001212315.1 ferredoxin-like protein	72.9
2	gi_147676354	gi_116751343	YP 001210569.1 ferredoxin-like protein	71.9
2	gi_147678099	gi_116751342	YP 001212314.1 sugar phosphate permease	72.8
3	gi_147678347	gi_116748291	YP 001212562.1 NADH:ubiquinone oxidoreductase, 24 kD subunit	75.2
3	gi_147678346	gi_116748290	YP 001212561.1 NADH:ubiquinone oxidoreductase, NADH-binding 51 kD subunit	81.9
3	gi_147678345	gi_116748289	YP 001212560.1 hydrogenase subunit	81.8
3	gi_147677315	gi_116748288	YP 001211530.1 hypothetical protein PTH 0980	70.9
3	gi_147677319	gi_116748287	YP 001211534.1 thiamine biosynthesis protein ThiH	73.8

Desulfurivibrio alkaliphilus --Thermodesulfatator indicus

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_297569850	gi_337286693	YP 003691194.1 ATP synthase F1, epsilon	64.1

			subunit	
1	gi_297569851	gi_337286692	YP 003691195.1 ATP synthase F1, beta subunit	81.1
1	gi_297569852	gi_337286691	YP 003691196.1 ATP synthase F1, gamma subunit	53.6
1	gi_297569853	gi_337286690	YP 003691197.1 ATP synthase F1, alpha subunit	71.1
2	gi_297568705	gi_337287265	YP 003690049.1 acetolactate synthase, small subunit	66
2	gi_297568704	gi_337287264	YP 003690048.1 acetolactate synthase, large subunit, biosynthetic type	66.1
3	gi_297570015	gi_337287397	YP 003691359.1 flavodoxin/nitric oxide synthase	64.3
3	gi_297570016	gi_337287396	YP 003691360.1 desulfoferrodoxin	75
4	gi_297568804	gi_337287522	YP 003690148.1 CO dehydrogenase/acetyl-CoA synthase complex, beta subunit	67.7
4	gi_297568803	gi_337287521	YP 003690147.1 CO dehydrogenase/acetyl-CoA synthase delta subunit, TIM barrel	65.2
5	gi_297569271	gi_337286233	YP 003690615.1 ATP-dependent protease La	61
5	gi_297569272	gi_337286232	YP 003690616.1 ATP-dependent Clp protease, ATP-binding subunit ClpX	66.7
5	gi_297569273	gi_337286231	YP 003690617.1 ATP-dependent Clp protease, proteolytic subunit ClpP	69.4
6	gi_297569689	gi_337285563	YP 003691033.1 flagellar biosynthesis protein FlhA	61.4
6	gi_297569688	gi_337285562	YP 003691032.1 flagellar biosynthetic protein FlhB	45.9
6	gi_297569686	gi_337285560	YP 003691030.1 flagellar biosynthetic protein FliQ	50.6
6	gi_297569685	gi_337285559	YP 003691029.1 flagellar biosynthetic protein FliP	60.5
7	gi_297568282	gi_337285778	YP 003689626.1 sulfite reductase, dissimilatory-type alpha subunit	65.3
7	gi_297568283	gi_337285777	YP 003689627.1 sulfite reductase, dissimilatory-type beta subunit	67.7
8	gi_297569325	gi_337286362	YP 003690669.1 ATP phosphoribosyltransferase	70.1
8	gi_297569326	gi_337286361	YP 003690670.1 Phosphoribosyl-AMP cyclohydrolase	67.5
8	gi_297568921	gi_337286359	YP 003690265.1 3-deoxy-D-manno-octulosonate cytidylyltransferase	55.2
9	gi_297570151	gi_337286467	YP 003691495.1 ornithine carbamoyltransferase	62.8
9	gi_297569521	gi_337286466	YP 003690865.1 thiamine biosynthesis protein ThiC	64.6

Streptococcus gordonii Challis substr CH1 -- Leptotrichia buccalis C 1013 b

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity

				(%)
1	gi_157149908	gi_257125329	YP 001450422.1 acetoin dehydrogenase	72.1
1	gi_157151664	gi_257125330	YP 001450421.1 acetoin dehydrogenase	78.2
1	gi_157151137	gi_257125331	YP 001450420.1 dihydrolipoamide acetyltransferase	62.5
1	gi_157150243	gi_257125332	YP 001450419.1 dihydrolipoamide dehydrogenase	65.3
1	gi_157149679	gi_257125333	YP 001450418.1 lipoate protein ligase A	65
2	gi_157150143	gi_257125371	YP 001450805.1 galactose-6-phosphate isomerase subunit LacA	66
2	gi_157149701	gi_257125372	YP 001450797.1 galactose-6-phosphate isomerase subunit LacB	78.9
2	gi_157151561	gi_257125373	YP 001450796.1 tagatose-6-phosphate kinase	62.8
2	gi_157151000	gi_257125374	YP 001450795.1 tagatose 1,6-diphosphate aldolase	71.4
2	gi_157150563	gi_257125375	YP 001450793.1 PTS system lactose-specific transporter subunit IIA	65.7
2	gi_157151244	gi_257125376	YP 001450792.1 PTS system lactose-specific transporter subunit IIBC	80.5
2	gi_157150880	gi_257125377	YP 001450791.1 6-phospho-beta-galactosidase	82
3	gi_157151415	gi_257125430	YP 001450823.1 F0F1 ATP synthase subunit alpha	60
3	gi_157151073	gi_257125432	YP 001450821.1 F0F1 ATP synthase subunit beta	70.4
4	gi_157150337	gi_257125543	YP 001449457.1 V-type ATP synthase subunit A	66.6
4	gi_157149878	gi_257125544	YP 001449458.1 V-type ATP synthase subunit B	73.2
5	gi_157150912	gi_257125927	YP 001449690.1 malate dehydrogenase	68.4
5	gi_157150902	gi_257125929	YP 001449344.1 tRNA-specific 2-thiouridylase MnmA	64.6
7	gi_157150310	gi_257126555	YP 001450452.1 putative lipoprotein	68.9
7	gi_157150275	gi_257126556	YP 001450451.1 tat translocated dye-type peroxidase family protein	64.2
7	gi_157149693	gi_257126557	YP 001450450.1 FTR1 family iron permease	52
7	gi_157150071	gi_257126558	YP 001450449.1 Sec-independent protein translocase TatC	59.4
7	gi_157151040	gi_257126559	YP 001450448.1 twin arginine-targeting protein translocase	62.5
8	gi_157149993	gi_257126077	YP 001450429.1 ATP-dependent protease ATP-binding subunit ClpX	60.6
8	gi_157151545	gi_257126078	YP 001450909.1 ATP-dependent Clp protease proteolytic subunit	59.6
8	gi_157149990	gi_257126963	YP 001449596.1 dihydroorotate dehydrogenase 1A	78.1
8	gi_157149754	gi_257126964	YP 001450542.1 NAD-dependent deacetylase	62.7

9	gi_157151254	gi_257125263	YP 001451012.1 integral membrane protein	78.2
9	gi_157151094	gi_257125264	YP 001449935.1 glycerol kinase	59
10	gi_157150100	gi_257125243	YP 001450958.1 PTS system mannose/fructose/sorbose family transporter subunit IID	68
10	gi_157150304	gi_257125244	YP 001450957.1 phosphotransferase system enzyme II	63.1
10	gi_157151038	gi_257125245	YP 001450956.1 phosphotransferase system enzyme II	61.8

Desulfurispirillum indicum S5 -- Marinobacter aquaeolei VT8

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_317050217	gi_120553820	YP 004111333.1 transposase IS204/IS1001/IS1096/IS1165 family protein	99.3
1	gi_317050216	gi_120553821	YP 004111332.1 lipoprotein signal peptidase	98.8
1	gi_317050206	gi_120553822	YP 004111322.1 cation efflux protein	97
1	gi_317050205	gi_120553826	YP 004111321.1 Cd(II)/Pb(II)-responsive transcriptional regulator	90.4
1	gi_317050214	gi_120553826	YP 004111330.1 Cd(II)/Pb(II)-responsive transcriptional regulator	97.8
1	gi_317050213	gi_120553909	YP 004111329.1 integron integrase	52.5
1	gi_317050211	gi_120553989	YP 004111327.1 small multidrug resistance protein	68
2	gi_317050253	gi_120553460	YP 004111369.1 nitrogen regulatory protein P-II	64.3
2	gi_317050254	gi_120554275	YP 004111370.1 general secretion pathway protein G	65.2
3	gi_317051135	gi_120555535	YP 004112251.1 sulfate adenylyltransferase small subunit	77.7
3	gi_317051136	gi_120555646	YP 004112252.1 sulfate adenylyltransferase large subunit	63.9
4	gi_317051301	gi_120553293	YP 004112417.1 TRAP dicarboxylate transporter subunit DctM	80
4	gi_317051300	gi_120553294	YP 004112416.1 tripartite ATP-independent periplasmic transporter subunit DctQ	61.3
4	gi_317051299	gi_120553295	YP 004112415.1 family 7 extracellular solute-binding protein	69.3
4	gi_317051296	gi_120553973	YP 004112412.1 ABC transporter-like protein	60.5
4	gi_317051303	gi_120554460	YP 004112419.1 binding-protein-dependent transporter inner membrane component	68
4	gi_317051304	gi_120554461	YP 004112420.1 ABC transporter-like protein	55.6
5	gi_317051351	gi_120554670	YP 004112467.1 Agmatine deiminase	52.1
5	gi_317051350	gi_120554671	YP 004112466.1 nitrilase/cyanide hydratase -- apolipoprotein N-acyltransferase	62
5	gi_317051352	gi_120554979	YP 004112468.1 TRAP transporter, 4TM/12TM	62.8

			fusion protein	
5	gi_317051353	gi_120554980	YP 004112469.1 TAXI family TRAP transporter solute receptor	63.1
7	gi_317052328	gi_120556164	YP 004113444.1 phosphonate ABC transporter periplasmic phosphonate-binding protein	64.2
7	gi_317052327	gi_120556165	YP 004113443.1 phosphonate ABC transporter ATPase subunit	69.8
7	gi_317052326	gi_120556166	YP 004113442.1 phosphonate ABC transporter inner membrane subunit	66.9
7	gi_317052325	gi_120556167	YP 004113441.1 phosphonate ABC transporter inner membrane subunit	65.8

Caldicellulosiruptor hydrothermalis 108 --Thermotoga thermarum DSM 5069

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_312128371	gi_338730006	YP 003991766.1 3-isopropylmalate dehydrogenase	72.8
1	gi_312128370	gi_338730005	YP 003991765.1 3-isopropylmalate dehydratase, small subunit	74.4
1	gi_312128369	gi_338730004	YP 003991764.1 3-isopropylmalate dehydratase, large subunit	83.6
1	gi_312128368	gi_338730295	YP 003991973.1 pyridoxine biosynthesis protein	64.6
2	gi_312128334	gi_338730008	YP 003991968.1 oligopeptide/dipeptide ABC transporter ATPase subunit	68.8
2	gi_312128333	gi_338730007	YP 003992157.1 tryptophan synthase subunit alpha	60.6
3	gi_312128165	gi_338729930	YP 003992156.1 tryptophan synthase subunit beta	74.5
3	gi_312128164	gi_338730159	YP 003992155.1 phosphoribosylanthranilate isomerase	62
4	gi_312127781	gi_338731040	YP 003992154.1 indole-3-glycerol-phosphate synthase	70.9
4	gi_312127780	gi_338730055	YP 003992153.1 anthranilate phosphoribosyltransferase	81.4
5	gi_312127526	gi_338730088	YP 003992152.1 glutamine amidotransferase of anthranilate synthase	74.7
5	gi_312127524	gi_338730086	YP 003992151.1 chorismate binding-like protein	67.9
6	gi_312127472	gi_338730808	YP 003992346.1 histidinol dehydrogenase	62.8
6	gi_312127471	gi_338730807	YP 003992345.1 ATP phosphoribosyltransferase	65.4
7	gi_312127099	gi_338731576	YP 003993207.1 isocitrate dehydrogenase (nad(+))	69.8
7	gi_312127094	gi_338731090	YP 003993245.1 acetolactate synthase, large subunit, biosynthetic type	63.3

8	gi_312126892	gi_338730292	YP 003993244.1 acetolactate synthase, small subunit	60.5
8	gi_312126891	gi_338730293	YP 003993243.1 ketol-acid reductoisomerase	66.3
8	gi_312126890	gi_338730294	YP 003993242.1 2-isopropylmalate synthase	64.6

Candidatus Nitrospira defluvii -- Janthinobacterium Marseille

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_302035457	gi_152981820	YP 003795779.1 hypothetical protein NIDE0063	61.5
1	gi_302035458	gi_152981934	YP 003795780.1 mercuric resistance operon regulatory protein	67.7
1	gi_302035459	gi_152982220	YP 003795781.1 mercury ion transport protein	69.2
1	gi_302035460	gi_152982938	YP 003795782.1 periplasmic mercury ion binding protein	71.9
1	gi_302035462	gi_152982873	YP 003795784.1 hypothetical protein NIDE0068	74.2
1	gi_302035463	gi_152982221	YP 003795785.1 hypothetical protein NIDE0069	95.8
1	gi_302035464	gi_152981666	YP 003795786.1 putative site-specific recombinase, resolvase family (phage related)	94.4
1	gi_302035465	gi_152982797	YP 003795787.1 hypothetical protein NIDE0071	91.3
1	gi_302035466	gi_152983289	YP 003795788.1 hypothetical protein NIDE0072	82.4
1	gi_302035471	gi_152983290	YP 003795793.1 hypothetical protein NIDE0079	71.3
1	gi_302035472	gi_152982677	YP 003795794.1 hypothetical protein NIDE0080	80.9
1	gi_302035473	gi_152982207	YP 003795795.1 hypothetical protein NIDE0081	90.4
1	gi_302035474	gi_152982323	YP 003795796.1 hypothetical protein NIDE0082	84.4
1	gi_302035475	gi_152982824	YP 003795797.1 hypothetical protein NIDE0083	85.2
1	gi_302035476	gi_152982461	YP 003795798.1 hypothetical protein NIDE0084	75.5
1	gi_302035477	gi_152982378	YP 003795799.1 hypothetical protein NIDE0085	83.3
1	gi_302035478	gi_152982760	YP 003795800.1 hypothetical protein NIDE0086	67.4
1	gi_302035479	gi_152981706	YP 003795801.1 putative DNA primase'	87.9
1	gi_302035480	gi_152982142	YP 003795802.1 putative polynucleotidyl transferase	90.2
1	gi_302035481	gi_152983291	YP 003795803.1 hypothetical protein NIDE0090	79.6
1	gi_302035483	gi_152982005	YP 003795805.1 site-specific DNA-methyltransferase N-4/N-6 (phage related)	85.1
1	gi_302035484	gi_152981982	YP 003795806.1 site-specific DNA-methyltransferase N-4/N-6 (phage related)	92
1	gi_302035485	gi_152983294	YP 003795807.1 hypothetical protein NIDE0094	82.4
1	gi_302035486	gi_152982304	YP 003795808.1 hypothetical protein NIDE0095	85.7
1	gi_302035487	gi_152982162	YP 003795809.1 hypothetical protein NIDE0097	68.2
1	gi_302035489	gi_152982161	YP 003795811.1 hypothetical protein NIDE0099	96.6
1	gi_302035490	gi_152981093	YP 003795812.1 phage terminase large subunit	95.3
1	gi_302035491	gi_152982062	YP 003795813.1 hypothetical protein NIDE0101	95.7
1	gi_302035492	gi_152982876	YP 003795814.1 hypothetical protein NIDE0102	93.1
1	gi_302035493	gi_152982972	YP 003795815.1 hypothetical protein NIDE0103	93.2
1	gi_302035494	gi_152981081	YP 003795816.1 phage portal protein, lambda family	87.3
1	gi_302035495	gi_152981533	YP 003795817.1 putative phage minor capsid protein C	73.8

1	gi_302035496	gi_152982282	YP 003795818.1 hypothetical protein NIDE0106	80
1	gi_302035497	gi_152982830	YP 003795819.1 hypothetical protein NIDE0107	91
1	gi_302035498	gi_152982165	YP 003795820.1 hypothetical protein NIDE0108	80
1	gi_302035499	gi_152982164	YP 003795821.1 hypothetical protein NIDE0109	92.9
1	gi_302035500	gi_152982980	YP 003795822.1 hypothetical protein NIDE0110	71.8
1	gi_302035501	gi_152982163	YP 003795823.1 hypothetical protein NIDE0111	98.4
1	gi_302035502	gi_152982160	YP 003795824.1 hypothetical protein NIDE0112	95.5
1	gi_302035503	gi_152982159	YP 003795825.1 hypothetical protein NIDE0113	98.1
1	gi_302035504	gi_152982158	YP 003795826.1 putative phage tail length tape measure protein	91.3
1	gi_302035505	gi_152982157	YP 003795827.1 hypothetical protein NIDE0115	96.9
1	gi_302035506	gi_152982156	YP 003795828.1 hypothetical protein NIDE0116	87.8
1	gi_302035507	gi_152982127	YP 003795829.1 hypothetical protein NIDE0117	87.8
1	gi_302035509	gi_152982262	YP 003795831.1 hypothetical protein NIDE0119	89
1	gi_302035510	gi_152982615	YP 003795832.1 hypothetical protein NIDE0120	97.5
1	gi_302035511	gi_152982541	YP 003795833.1 hypothetical protein NIDE0121	87.1
1	gi_302035512	gi_152982982	YP 003795834.1 hypothetical protein NIDE0122	91
2	gi_302036778	gi_152979893	YP 003797100.1 chorismate synthase	74.2
2	gi_302036779	gi_152980654	YP 003797101.1 ribonuclease H	68.8
3	gi_302038815	gi_152981067	YP 003799137.1 multidrug efflux system subunit C	60.7
3	gi_302038816	gi_152981117	YP 003799138.1 multidrug efflux system subunit B	63.2

Clostridium saccharolyticum WM1 -- Sphaerochaeta pleomorpha Grapes

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_302385696	gi_374314595	YP 003821518.1 binding-protein-dependent transport system inner membrane protein	72.3
1	gi_302385695	gi_374314596	YP 003821517.1 binding-protein-dependent transport system inner membrane protein	69.1
1	gi_302385694	gi_374314597	YP 003821516.1 extracellular solute-binding protein	64.4
2	gi_302386292	gi_374314822	YP 003822114.1 ABC transporter	72.1
2	gi_302386293	gi_374314823	YP 003822115.1 inner-membrane translocator	75.9
2	gi_302386294	gi_374314824	YP 003822116.1 LacI family transcriptional regulator	72.2
3	gi_302387219	gi_374314977	YP 003823041.1 short-chain dehydrogenase/reductase SDR	76
3	gi_302387813	gi_374314978	YP 003823635.1 4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase	59.6
4	gi_302385761	gi_374315043	YP 003821583.1 L-fucose isomerase-like protein	63.5
4	gi_302385109	gi_374315044	YP 003820931.1 class II aldolase/adducin family protein	62.8

5	gi_302387893	gi_374315132	YP 003823715.1 protein-tyrosine phosphatase	76.4
5	gi_302387095	gi_374315133	YP 003822917.1 redox-active disulfide protein 2	50.4
5	gi_302387097	gi_374315134	YP 003822919.1 permease	69.9
6	gi_302388266	gi_374315140	YP 003824088.1 ABC transporter	56.6
6	gi_302386838	gi_374315141	YP 003822660.1 inner-membrane translocator	70.4
6	gi_302386840	gi_374315143	YP 003822662.1 ABC transporter	63.4
6	gi_302386841	gi_374315144	YP 003822663.1 basic membrane lipoprotein	64.6
7	gi_302384518	gi_374315235	YP 003820340.1 flavodoxin/nitric oxide synthase	75.8
7	gi_302387044	gi_374315237	YP 003822866.1 arsenical-resistance protein	69.4
7	gi_302387889	gi_374315238	YP 003823711.1 ArsR family transcriptional regulator	60
8	gi_302387949	gi_374315291	YP 003823771.1 tryptophan synthase subunit beta	77.2
8	gi_302387950	gi_374315292	YP 003823772.1 tryptophan synthase subunit alpha	60.8
9	gi_302385599	gi_374315380	YP 003821421.1 binding-protein-dependent transport system inner membrane protein	66.9
9	gi_302385598	gi_374315381	YP 003821420.1 extracellular solute-binding protein	64.2
10	gi_302387979	gi_374315440	YP 003823801.1 dihydroxy-acid dehydratase	65.7
10	gi_302387980	gi_374315441	YP 003823802.1 3-isopropylmalate dehydrogenase	62.8
10	gi_302386582	gi_374315442	YP 003822404.1 3-isopropylmalate dehydratase small subunit	70.2
10	gi_302386583	gi_374315443	YP 003822405.1 3-isopropylmalate dehydratase large subunit	71.1
10	gi_302386585	gi_374315446	YP 003822407.1 ketol-acid reductoisomerase	67
11	gi_302386734	gi_374315727	YP 003822556.1 polar amino acid ABC transporter inner membrane subunit	71.9
11	gi_302386735	gi_374315728	YP 003822557.1 family 3 extracellular solute-binding protein	61.5
12	gi_302387418	gi_374315759	YP 003823240.1 malate/L-lactate dehydrogenase	62.8
12	gi_302387311	gi_374315763	YP 003823133.1 ABC transporter	66.1
12	gi_302387310	gi_374315764	YP 003823132.1 ABC transporter	59.1
12	gi_302387309	gi_374315765	YP 003823131.1 inner-membrane translocator	59.7
12	gi_302387308	gi_374315766	YP 003823130.1 inner-membrane translocator	78.5
12	gi_302387307	gi_374315767	YP 003823129.1 extracellular ligand-binding receptor	73.8
12	gi_302385731	gi_374315788	YP 003821553.1 sodium ion-translocating decarboxylase subunit beta	60.9
12	gi_302384784	gi_374315790	YP 003820606.1 dCMP deaminase	62.3

13	gi_302384774	gi_374315940	YP 003820596.1 xylose isomerase domain-containing protein TIM barrel	65.6
13	gi_302384775	gi_374315941	YP 003820597.1 binding-protein-dependent transport system inner membrane protein	72.4
13	gi_302384776	gi_374315942	YP 003820598.1 binding-protein-dependent transport system inner membrane protein	67.6
13	gi_302384777	gi_374315943	YP 003820599.1 extracellular solute-binding protein	68.6
14	gi_302384523	gi_374316702	YP 003820345.1 ABC transporter	67.1
14	gi_302384524	gi_374316703	YP 003820346.1 inner-membrane translocator	67.2
14	gi_302384525	gi_374316704	YP 003820347.1 LacI family transcriptional regulator	72.6
15	gi_302385244	gi_374317120	YP 003821066.1 extracellular solute-binding protein	62.7
15	gi_302385245	gi_374317121	YP 003821067.1 tripartite AtP-independent periplasmic transporter subunit DctQ	68.2
15	gi_302385246	gi_374317122	YP 003821068.1 TRAP dicarboxylate transporter subunit DctM	81.9
16	gi_302386148	gi_374317162	YP 003821970.1 phage major capsid protein, HK97 family	62.8
16	gi_302386147	gi_374317163	YP 003821969.1 peptidase S14 ClpP	53.7
16	gi_302386146	gi_374317164	YP 003821968.1 phage portal protein, HK97 family	66.2

Deferribacter desulfuricans SSM1 -- Geobacter uraniireducens Rf4

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_291280213	gi_148265082	YP 003497048.1 acetyl-CoA C-acetyltransferase	66.8
1	gi_291280212	gi_148265081	YP 003497047.1 3-hydroxybutyryl-CoA dehydrogenase	65.3
1	gi_291280211	gi_148265080	YP 003497046.1 3-hydroxybutyryl-CoA dehydratase	62.8
1	gi_291280210	gi_148265079	YP 003497045.1 butyryl-CoA dehydrogenase	74.3
1	gi_291280209	gi_148263663	YP 003497044.1 iron-sulfur cluster-binding protein	65.8
1	gi_291280208	gi_148265077	YP 003497043.1 electron transfer flavoprotein subunit beta	69.3
1	gi_291280207	gi_148265076	YP 003497042.1 electron transfer flavoprotein subunit alpha	72.5
1	gi_291280192	gi_148265419	YP 003497027.1 acetate kinase	70.2
2	gi_291279999	gi_148264216	YP 003496834.1 cytochrome bd oxidase subunit II	65.7
2	gi_291279998	gi_148264217	YP 003496833.1 cytochrome bd oxidase subunit I	69.8
3	gi_291279856	gi_148265390	YP 003496691.1 nitrogen regulatory protein P-II	72.3
3	gi_291279855	gi_148264278	YP 003496690.1 glutamine synthetase type I	70.4

4	gi_291279849	gi_148263653	YP 003496684.1 long-chain fatty-acid-CoA ligase	65.4
4	gi_291279848	gi_148263654	YP 003496683.1 3-hydroxyacyl-CoA dehydrogenase/enoyl-CoA hydratase	66.8
4	gi_291279847	gi_148263655	YP 003496682.1 3-ketoacyl-CoA thiolase	74.6
4	gi_291279846	gi_148263656	YP 003496681.1 acyl-CoA dehydrogenase	76.2
5	gi_291279843	gi_148264890	YP 003496678.1 HNH endonuclease	69.2
5	gi_291279842	gi_148262944	YP 003496677.1 phosphoenolpyruvate carboxykinase (ATP)	62
6	gi_291279569	gi_148264363	YP 003496404.1 2-isopropylmalate synthase	64.5
6	gi_291279568	gi_148264364	YP 003496403.1 aspartate kinase monofunctional class	63.8
7	gi_291279489	gi_148264234	YP 003496324.1 riboflavin synthase beta chain	61.8
7	gi_291279488	gi_148264235	YP 003496323.1 riboflavin biosynthesis bifunctional protein RibBA	67.6
8	gi_291279312	gi_148264247	YP 003496147.1 malate dehydrogenase	75
8	gi_291279311	gi_148264248	YP 003496146.1 isocitrate dehydrogenase NADP-dependent	67.2
8	gi_291279310	gi_148263996	YP 003496145.1 aconitate hydratase	71.6
9	gi_291279213	gi_148263639	YP 003496048.1 citrate synthase	65.3
9	gi_291279211	gi_148262430	YP 003496046.1 porphobilinogen synthase	70.8
10	gi_291278972	gi_148263636	YP 003495807.1 acyl-CoA synthase	61.7
10	gi_291278971	gi_148266340	YP 003495806.1 pyruvate:ferredoxin oxidoreductase	66.5
11	gi_291278510	gi_148262626	YP 003495345.1 Ni-Fe hydrogenase small subunit	66.9
11	gi_291278509	gi_148262625	YP 003495344.1 Ni-Fe hydrogenase large subunit	73.4

Listeria ivanovii PAM 55 --Sebaldella termitidis ATCC 33386

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_347547968	gi_269118910	YP 004854296.1 putative NADP-specific glutamate dehydrogenase	65.1
1	gi_347549798	gi_269118929	YP 004856126.1 putative phosphate ABC transporter ATP binding protein	64
2	gi_347548523	gi_269119662	YP 004854851.1 putative PduU protein	60.5
2	gi_347548524	gi_269119663	YP 004854852.1 putative PduV protein	44.1
2	gi_347548529	gi_269119665	YP 004854857.1 putative propanediol utilization protein PduA	76.5

2	gi_347548530	gi_269119652	YP 004854858.1 putative propanediol utilization protein PduB	75.9
2	gi_347548531	gi_269119653	YP 004854859.1 putative propanediol dehydratase subunit alpha	76.7
2	gi_347548532	gi_269119654	YP 004854860.1 putative diol dehydrase subunit gamma	58.5
2	gi_347548533	gi_269119655	YP 004854861.1 putative diol dehydrase subunit gamma PddC	54.7
2	gi_347548534	gi_269119656	YP 004854862.1 putative diol dehydratase-reactivating factor large subunit	67.3
2	gi_347548535	gi_269119657	YP 004854863.1 putative diol dehydratase-reactivating factor small chain	41.7
2	gi_347548537	gi_269119665	YP 004854865.1 putative carboxysome structural protein	82.8
2	gi_347548543	gi_269119659	YP 004854871.1 putative ethanolamine utilization protein EutE	55.4
2	gi_347548556	gi_269119660	YP 004854884.1 putative carboxysome structural protein	56.6
2	gi_347548557	gi_269119666	YP 004854885.1 putative acetaldehyde dehydrogenase / alcohol dehydrogenase	60.7
2	gi_347548558	gi_269119661	YP 004854886.1 putative carboxysome structural protein	85.7
2	gi_347548560	gi_269119668	YP 004854888.1 putative PduL protein	51.5
2	gi_347548562	gi_269119670	YP 004854890.1 putative carbon dioxide concentrating mechanism protein	62.8
3	gi_347547746	gi_269121938	YP 004854074.1 putative phospho-beta-glucosidase	67.2
3	gi_347547927	gi_269121939	YP 004854255.1 putative 6-phospho-beta-glucosidase	61.1
3	gi_347547940	gi_269121939	YP 004854268.1 putative 6-phospho-beta-glucosidase	67.3
3	gi_347550094	gi_269121938	YP 004856422.1 putative beta-glucosidase	68.4
4	gi_347547782	gi_269121842	YP 004854110.1 putative oxidoreductase	71.3
4	gi_347549403	gi_269121832	YP 004855731.1 putative oxidoreductase	70.9
5	gi_347547708	gi_269121624	YP 004854036.1 DeoR family transcriptional regulator	69
5	gi_347547709	gi_269121623	YP 004854037.1 putative N-acetylmannosamine-6-phosphate epimerase	80.7
5	gi_347547710	gi_269121621	YP 004854038.1 putative mannose-specific PTS system enzyme IIB	64.7
5	gi_347547711	gi_269121620	YP 004854039.1 putative mannose-specific PTS system enzyme IIC	84.3
5	gi_347547712	gi_269121619	YP 004854040.1 putative mannose-specific PTS system enzyme IID	78.3
5	gi_347547713	gi_269121618	YP 004854041.1 putative mannose-specific PTS system enzyme IIA	61.1
6	gi_347549949	gi_269121095	YP 004856277.1 putative phosphotriesterase	70.2
6	gi_347549950	gi_269121096	YP 004856278.1 putative PTS enzyme IIC component	67.9

7	gi_347548252	gi_269120483	YP 004854580.1 putative amino acid ABC transporter ATP-binding protein	66.1
7	gi_347549641	gi_269120483	YP 004855969.1 putative amino acid ABC transporter ATP binding protein	61.2
8	gi_347550146	gi_269120141	YP 004856474.1 hypothetical protein	61.7
8	gi_347550147	gi_269120140	YP 004856475.1 putative alcohol dehydrogenase	74.9
8	gi_347550148	gi_269120139	YP 004856476.1 putative sugar ABC transporter permease	69.6
8	gi_347550149	gi_269120138	YP 004856477.1 putative sugar ABC transporter permease	65.1
9	gi_347548281	gi_269119824	YP 004854609.1 putative PTS system, beta-glucoside enzyme IIB component	67.9
9	gi_347548282	gi_269119823	YP 004854610.1 putative PTS system, Lichenan-specific enzyme IIC component	71.8
9	gi_347548284	gi_269119821	YP 004854612.1 putative oxidoreductase	62.3
10	gi_347548555	gi_269119678	YP 004854883.1 putative carboxysome structural protein EutL	70
10	gi_347548564	gi_269119679	YP 004854892.1 putative ethanolamine utilization protein EutH	73.8
11	gi_347548553	gi_269119676	YP 004854881.1 eutB gene product	71.8
11	gi_347548552	gi_269119675	YP 004854880.1 eutA gene product	51.1

REFERENCES

1. McDaniel, L.D., et al., *High frequency of horizontal gene transfer in the oceans.* Science, 2010. **330**(6000): p. 50.
2. Ochman, H., J.G. Lawrence, and E.A. Groisman, *Lateral gene transfer and the nature of bacterial innovation.* Nature, 2000. **405**(6784): p. 299-304.
3. Treangen, T.J. and E.P. Rocha, *Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes.* PLoS Genet, 2011. **7**(1): p. e1001284.
4. Gogarten, J.P., W.F. Doolittle, and J.G. Lawrence, *Prokaryotic evolution in light of gene transfer.* Mol Biol Evol, 2002. **19**(12): p. 2226-38.
5. Zhaxybayeva, O., et al., *Intertwined evolutionary histories of marine Synechococcus and Prochlorococcus marinus.* Genome Biol Evol, 2009. **1**: p. 325-39.
6. Ochman, H., E. Lerat, and V. Daubin, *Examining bacterial species under the specter of gene transfer and exchange.* Proc Natl Acad Sci U S A, 2005. **102 Suppl 1**: p. 6595-9.
7. Beiko, R.G., T.J. Harlow, and M.A. Ragan, *Highways of gene sharing in prokaryotes.* Proc Natl Acad Sci U S A, 2005. **102**(40): p. 14332-7.
8. Lawrence, J.G., *Gene transfer in bacteria: speciation without species?* Theor Popul Biol, 2002. **61**(4): p. 449-60.
9. Sheppard, S.K., et al., *Convergence of Campylobacter species: implications for bacterial evolution.* Science, 2008. **320**(5873): p. 237-9.
10. Doolittle, W.F. and O. Zhaxybayeva, *On the origin of prokaryotic species.* Genome Res, 2009. **19**(5): p. 744-56.
11. Ereshefsky, M., *Microbiology and the species problem.* Biology & Philosophy, 2010. **25**(4): p. 553-568.
12. Nesbo, C.L., M. Dlutek, and W.F. Doolittle, *Recombination in Thermotoga: implications for species concepts and biogeography.* Genetics, 2006. **172**(2): p. 759-69.
13. Feil, E.J., et al., *Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 182-7.
14. Fraser, C., W.P. Hanage, and B.G. Spratt, *Recombination and the nature of bacterial speciation.* Science, 2007. **315**(5811): p. 476-80.
15. Kristensen, D.M., et al., *New dimensions of the virus world discovered through metagenomics.* Trends Microbiol, 2010. **18**(1): p. 11-9.
16. Jensen, E.C., et al., *Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa.* Appl Environ Microbiol, 1998. **64**(2): p. 575-80.
17. Evans, T.J., et al., *Characterization of a broad-host-range flagellum-dependent phage that mediates high-efficiency generalized transduction in, and between, Serratia and Pantoea.* Microbiology, 2010. **156**(Pt 1): p. 240-7.

18. Vogelmann, J., et al., *Conjugal plasmid transfer in Streptomyces resembles bacterial chromosome segregation by FtsK/SpoIIIE.* EMBO J, 2011. **30**(11): p. 2246-54.
19. Stachel, S.E. and E.W. Nester, *The genetic and transcriptional organization of the vir region of the A6 Ti plasmid of Agrobacterium tumefaciens.* EMBO J, 1986. **5**(7): p. 1445-54.
20. Smith, E.F. and C.O. Townsend, *A Plant-Tumor of Bacterial Origin.* Science, 1907. **25**(643): p. 671-3.
21. Rawlings, D.E. and E. Tietze, *Comparative biology of IncQ and IncQ-like plasmids.* Microbiol Mol Biol Rev, 2001. **65**(4): p. 481-96, table of contents.
22. Dubnau, D., *DNA uptake in bacteria.* Annu Rev Microbiol, 1999. **53**: p. 217-44.
23. Chen, I. and D. Dubnau, *DNA uptake during bacterial transformation.* Nat Rev Microbiol, 2004. **2**(3): p. 241-9.
24. Meibom, K.L., et al., *Chitin induces natural competence in Vibrio cholerae.* Science, 2005. **310**(5755): p. 1824-7.
25. Vulic, M., R.E. Lenski, and M. Radman, *Mutation, recombination, and incipient speciation of bacteria in the laboratory.* Proc Natl Acad Sci U S A, 1999. **96**(13): p. 7348-51.
26. Roberts, M.S. and F.M. Cohan, *The effect of DNA sequence divergence on sexual isolation in Bacillus.* Genetics, 1993. **134**(2): p. 401-8.
27. Zawadzki, P., M.S. Roberts, and F.M. Cohan, *The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust.* Genetics, 1995. **140**(3): p. 917-32.
28. Majewski, J. and F.M. Cohan, *DNA sequence similarity requirements for interspecific recombination in Bacillus.* Genetics, 1999. **153**(4): p. 1525-33.
29. Majewski, J., et al., *Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation.* J Bacteriol, 2000. **182**(4): p. 1016-23.
30. Thaler, D.S. and M.O. Noordewier, *MEPS parameters and graph analysis for the use of recombination to construct ordered sets of overlapping clones.* Genomics, 1992. **13**(4): p. 1065-74.
31. Budroni, S., et al., *Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination.* Proc Natl Acad Sci U S A, 2011. **108**(11): p. 4494-9.
32. Ray, J.L., et al., *Sexual isolation in Acinetobacter baylyi is locus-specific and varies 10,000-fold over the genome.* Genetics, 2009. **182**(4): p. 1165-81.
33. Rocha, E.P., E. Cornet, and B. Michel, *Comparative and evolutionary analysis of the bacterial homologous recombination systems.* PLoS Genet, 2005. **1**(2): p. e15.
34. Levin, B.R. and O.E. Cornejo, *The population and evolutionary dynamics of homologous gene recombination in bacterial populations.* PLoS Genet, 2009. **5**(8): p. e1000601.
35. Lefebure, T. and M.J. Stanhope, *Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition.* Genome Biol, 2007. **8**(5): p. R71.
36. Croucher, N.J., et al., *Rapid pneumococcal evolution in response to clinical interventions.* Science, 2011. **331**(6016): p. 430-4.

37. Caro-Quintero, A., G.P. Rodriguez-Castano, and K.T. Konstantinidis, *Genomic insights into the convergence and pathogenicity factors of Campylobacter jejuni and Campylobacter coli species.* J Bacteriol, 2009. **191**(18): p. 5824-31.
38. Orsi, R.H., et al., *Recombination and positive selection contribute to evolution of Listeria monocytogenes inlA.* Microbiology, 2007. **153**(Pt 8): p. 2666-78.
39. Vos, M. and X. Didelot, *A comparison of homologous recombination rates in bacteria and archaea.* ISME J, 2009. **3**(2): p. 199-208.
40. Vergin, K.L., et al., *High intraspecific recombination rate in a native population of Candidatus pelagibacter ubique (SAR11).* Environ Microbiol, 2007. **9**(10): p. 2430-40.
41. McVean, G., P. Awadalla, and P. Fearnhead, *A coalescent-based method for detecting and estimating recombination from gene sequences.* Genetics, 2002. **160**(3): p. 1231-41.
42. Didelot, X. and D. Falush, *Inference of bacterial microevolution using multilocus sequence data.* Genetics, 2007. **175**(3): p. 1251-66.
43. Octavia, S. and R. Lan, *Frequent recombination and low level of clonality within Salmonella enterica subspecies I.* Microbiology, 2006. **152**(Pt 4): p. 1099-108.
44. Hanage, W.P., et al., *Hyper-recombination, diversity, and antibiotic resistance in pneumococcus.* Science, 2009. **324**(5933): p. 1454-7.
45. Corander, J., et al., *Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations.* BMC Bioinformatics, 2008. **9**: p. 539.
46. Tang, J., et al., *Identifying currents in the gene pool for bacterial populations using an integrative approach.* PLoS Comput Biol, 2009. **5**(8): p. e1000455.
47. Morelli, G., et al., *Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families.* PLoS Genet, 2010. **6**(7): p. e1001036.
48. Whitaker, R.J., D.W. Grogan, and J.W. Taylor, *Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus.* Mol Biol Evol, 2005. **22**(12): p. 2354-61.
49. Papke, R.T., et al., *Frequent recombination in a saltern population of Halorubrum.* Science, 2004. **306**(5703): p. 1928-9.
50. Hanage, W.P., C. Fraser, and B.G. Spratt, *Fuzzy species among recombinogenic bacteria.* BMC Biol, 2005. **3**: p. 6.
51. Falush, D., M. Stephens, and J.K. Pritchard, *Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.* Genetics, 2003. **164**(4): p. 1567-87.
52. Juhas, M., et al., *Genomic islands: tools of bacterial horizontal gene transfer and evolution.* FEMS Microbiol Rev, 2009. **33**(2): p. 376-93.
53. Ubeda, C., et al., *A pathogenicity island replicon in Staphylococcus aureus replicates as an unstable plasmid.* Proc Natl Acad Sci U S A, 2007. **104**(36): p. 14182-8.
54. Schluter, A., et al., *Genomics of IncP-1 antibiotic resistance plasmids isolated from wastewater treatment plants provides evidence for a widely accessible drug resistance gene pool.* FEMS Microbiol Rev, 2007. **31**(4): p. 449-77.

55. Rodriguez-Minguela, C.M., et al., *Worldwide prevalence of class 2 integrases outside the clinical setting is associated with human impact.* Appl Environ Microbiol, 2009. **75**(15): p. 5100-10.
56. Ajiboye, R.M., et al., *Global spread of mobile antimicrobial drug resistance determinants in human and animal Escherichia coli and Salmonella strains causing community-acquired infections.* Clin Infect Dis, 2009. **49**(3): p. 365-71.
57. Stokes, H.W. and M.R. Gillings, *Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens.* FEMS Microbiol Rev, 2011. **35**(5): p. 790-819.
58. Mohd-Zain, Z., et al., *Transferable antibiotic resistance elements in Haemophilus influenzae share a common evolutionary origin with a diverse family of syntenic genomic islands.* J Bacteriol, 2004. **186**(23): p. 8114-22.
59. Juhas, M., et al., *Sequence and functional analyses of Haemophilus spp. genomic islands.* Genome Biol, 2007. **8**(11): p. R237.
60. Dimopoulou, I.D., et al., *Diversity of antibiotic resistance integrative and conjugative elements among haemophili.* J Med Microbiol, 2007. **56**(Pt 6): p. 838-46.
61. Aziz, R.K., M. Breitbart, and R.A. Edwards, *Transposases are the most abundant, most ubiquitous genes in nature.* Nucleic Acids Res, 2010. **38**(13): p. 4207-17.
62. Wiedenbeck, J. and F.M. Cohan, *Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches.* FEMS Microbiol Rev, 2011. **35**(5): p. 957-76.
63. Budroni, S., et al., *Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination.* Proc Natl Acad Sci U S A, 2011. **108**(11): p. 4494-9.
64. van der Ploeg, J.R., *Analysis of CRISPR in Streptococcus mutans suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages.* Microbiology, 2009. **155**(Pt 6): p. 1966-76.
65. Bhaya, D., M. Davison, and R. Barrangou, *CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation.* Annu Rev Genet, 2011. **45**: p. 273-97.
66. Marraffini, L.A. and E.J. Sontheimer, *CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA.* Science, 2008. **322**(5909): p. 1843-5.
67. Perron, G.G., et al., *Bacterial recombination promotes the evolution of multi-drug-resistance in functionally diverse populations.* Proc Biol Sci, 2011.
68. Baltrus, D.A., K. Guillemin, and P.C. Phillips, *Natural transformation increases the rate of adaptation in the human pathogen Helicobacter pylori.* Evolution, 2008. **62**(1): p. 39-49.
69. MacLean, R.C., et al., *The population genetics of antibiotic resistance: integrating molecular mechanisms and treatment contexts.* Nat Rev Genet, 2010. **11**(6): p. 405-14.
70. Smillie, C.S., et al., *Ecology drives a global network of gene exchange connecting the human microbiome.* Nature, 2011.
71. Tuller, T., et al., *Association between translation efficiency and horizontal gene transfer within microbial communities.* Nucleic Acids Res, 2011. **39**(11): p. 4743-55.

72. Aravind, L., et al., *Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles.* Trends Genet, 1998. **14**(11): p. 442-4.

73. Cordero, O.X. and P. Hogeweg, *The impact of long-distance horizontal gene transfer on prokaryotic genome size.* Proc Natl Acad Sci U S A, 2009. **106**(51): p. 21748-53.

74. Atlas, R.M. and R. Bartha, *Microbial ecology: fundamentals and applications.* 1986.

75. Driessen, F., F. Kingma, and J. Stadhouders, *Evidence that Lactobacillus bulgaricus in yogurt is stimulated by carbon dioxide produced by Streptococcus thermophilus.* Netherlands Milk and Dairy Journal, 1982. **36**.

76. Sieber, J.R., M.J. McInerney, and R.P. Gunsalus, *Genomic insights into syntrophy: the paradigm for anaerobic metabolic cooperation.* Annu Rev Microbiol, 2012. **66**: p. 429-52.

77. Visscher, P.T., et al., *Competition between Anoxygenic Phototrophic Bacteria and Colorless Sulfur Bacteria in a Microbial Mat.* Fems Microbiology Ecology, 1992. **101**(1): p. 51-58.

78. Visscher, P.T., R.A. Prins, and H. Vangemerden, *Rates of Sulfate Reduction and Thiosulfate Consumption in a Marine Microbial Mat.* Fems Microbiology Ecology, 1992. **86**(4): p. 283-293.

79. Odenyo, A.A., et al., *The Use of 16s Ribosomal-Rna-Targeted Oligonucleotide Probes to Study Competition between Ruminal Fibrolytic Bacteria - Development of Probes for Ruminococcus Species and Evidence for Bacteriocin Production.* Appl Environ Microbiol, 1994. **60**(10): p. 3688-3696.

80. Cech, J.S. and P. Hartman, *Competition between Polyphosphate and Polysaccharide Accumulating Bacteria in Enhanced Biological Phosphate Removal Systems.* Water Res, 1993. **27**(7): p. 1219-1225.

81. Robinson, J.A. and J.M. Tiedje, *Competition between Sulfate-Reducing and Methanogenic Bacteria for H-2 under Resting and Growing Conditions.* Arch Microbiol, 1984. **137**(1): p. 26-32.

82. Segura, A.M., et al., *Emergent neutrality drives phytoplankton species coexistence.* Proc Biol Sci, 2011. **278**(1716): p. 2355-61.

83. Fraser, C., W.P. Hanage, and B.G. Spratt, *Neutral microepidemic evolution of bacterial pathogens.* Proc Natl Acad Sci U S A, 2005. **102**(6): p. 1968-73.

84. Ofiteru, I.D., et al., *Combined niche and neutral effects in a microbial wastewater treatment community.* Proc Natl Acad Sci U S A, 2010. **107**(35): p. 15345-50.

85. Meers, J. and H. Jannasch, *Growth of bacteria in mixed cultures.* Critical Reviews in Microbiology, 1973. **2**(2): p. 139-184.

86. Handelsman, J., et al., *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*2007, Washington, DC: The National Academies Press.

87. Mayr, E., *Systematics and the origin of species from the viewpoint of a zoologist*. Columbia biological series ...1942, New York,: Columbia University Press. xiv, 334 p. incl. illus. (incl. maps) tables, diagrs.

88. Fraser, C., et al., *The bacterial species challenge: making sense of genetic and ecological diversity.* Science, 2009. **323**(5915): p. 741-6.

89. Vos, M., *Why do bacteria engage in homologous recombination?* Trends Microbiol, 2009. **17**(6): p. 226-32.

90. Handelsman, J., *Metagenomics: application of genomics to uncultured microorganisms.* Microbiol Mol Biol Rev, 2004. **68**(4): p. 669-85.

91. Rusch, D.B., et al., *The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific.* PLoS Biol, 2007. **5**(3): p. e77.

92. Konstantinidis, K.T. and E.F. DeLong, *Genomic patterns of recombination, clonal divergence and environment in marine microbial populations.* ISME J, 2008. **2**(10): p. 1052-65.

93. Oh, S., et al., *Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem.* Appl Environ Microbiol, 2011. **77**(17): p. 6000-11.

94. Acinas, S.G., et al., *Fine-scale phylogenetic architecture of a complex bacterial community.* Nature, 2004. **430**(6999): p. 551-4.

95. Cohan, F.M., *Bacterial species and speciation.* Syst Biol, 2001. **50**(4): p. 513-24.

96. Oh, S., et al., *Metagenomic insights into the evolution, function and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem.* Appl Environ Microbiol, 2011.

97. Caro-Quintero, A. and K.T. Konstantinidis, *Bacterial species may exist, metagenomics reveal.* Environ Microbiol, 2012. **14**(2): p. 347-55.

98. Retchless, A.C. and J.G. Lawrence, *Temporal fragmentation of speciation in bacteria.* Science, 2007. **317**(5841): p. 1093-6.

99. Luo, C., et al., *Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species.* Proc Natl Acad Sci U S A, 2011. **108**(17): p. 7200-5.

100. Caro-Quintero, A., et al., *Unprecedented levels of horizontal gene transfer among spatially co-occurring Shewanella bacteria from the Baltic Sea.* ISME J, 2011. **5**(1): p. 131-40.

101. Konstantinidis, K.T. and J.M. Tiedje, *Genomic insights that advance the species definition for prokaryotes.* Proc Natl Acad Sci U S A, 2005. **102**(7): p. 2567-72.

102. Welch, R.A., et al., *Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 17020-4.

103. Lawrence, J.G. and H. Ochman, *Reconciling the many faces of lateral gene transfer.* Trends Microbiol, 2002. **10**(1): p. 1-4.

104. Tettelin, H., et al., *Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".* Proc Natl Acad Sci U S A, 2005. **102**(39): p. 13950-5.

105. Zhaxybayeva, O., et al., *Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events.* Genome Res, 2006. **16**(9): p. 1099-108.

106. Lang, A.S. and J.T. Beatty, *Importance of widespread gene transfer agent genes in alpha-proteobacteria.* Trends Microbiol, 2007. **15**(2): p. 54-62.

107. Konstantinidis, K.T., A. Ramette, and J.M. Tiedje, *The bacterial species definition in the genomic era.* Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1475): p. 1929-40.

108. Gevers, D., et al., *Opinion: Re-evaluating prokaryotic species.* Nat Rev Microbiol, 2005. **3**(9): p. 733-9.

109. Neumann, T., *The fate of river-borne nitrogen in the Baltic Sea e An example for the River Oder.* Estuar Coast Shelf Sci, 2006. **73**: p. 1-7.

110. Backer, H., et al., *HELCOM Baltic Sea Action Plan - A regional programme of measures for the marine environment based on the Ecosystem Approach.* Mar Pollut Bull, 2009.

111. Brettar, I., E.R. Moore, and M.G. Hofle, *Phylogeny and Abundance of Novel Denitrifying Bacteria Isolated from the Water Column of the Central Baltic Sea.* Microb Ecol, 2001. **42**(3): p. 295-305.

112. Ziemke, F., I. Brettar, and M.G. Hofle, *Stability and diveristy of the genetic structure of a Shewanella putrefaciens population in the water column of the central Baltic.* Aquatic Microbial Ecology, 1997. **13**: p. 63-74.

113. Fredrickson, J.K., et al., *Towards environmental systems biology of Shewanella.* Nat Rev Microbiol, 2008. **6**(8): p. 592-603.

114. Myers, C.R. and K.H. Nealson, *Bacterial Manganese Reduction and Growth with Manganese Oxide as the Sole Electron Acceptor.* Science, 1988. **240**(4857): p. 1319-1321.

115. Saitou, N. and M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): p. 406-25.

116. Tamura, K., et al., *MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.* Mol Biol Evol, 2007. **24**(8): p. 1596-9.

117. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2009. **37**(Database issue): p. D26-31.

118. Konstantinidis, K.T., et al., *Comparative systems biology across an evolutionary gradient within the Shewanella genus.* Proc Natl Acad Sci U S A, 2009. **106**(37): p. 15909-14.

119. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucl. Acids. Res., 1997. **25**(17): p. 3389-3402.

120. Caro-Quintero, A., G.P. Rodriguez-Castano, and K.T. Konstantinidis, *Genomic insights into the convergence and pathogenicity factors of Campylobacter jejuni and Campylobacter coli species.* J Bacteriol, 2009.

121. Kosakovsky Pond, S.L., et al., *Automated phylogenetic detection of recombination using a genetic algorithm.* Mol Biol Evol, 2006. **23**(10): p. 1891-901.

122. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.

123. Richards, E., M. Reichardt, and S. Rogers, *Preparation of Genomic DNA from Plant Tissue*, in *Current Protocols in Molecular Biology* M. Ausubel, et al., Editors. 1994, John Wiley: Hoboken, NJ. p. 2.3.1-2.3.7.

124. Oh, S., et al., *Evaluating the performance of oligonucleotide microarrays for bacterial strains of increasing genetic divergence to the reference strain.* Appl Environ Microbiol, 2010: p. In press.

125. Cruz-Garcia, C., et al., *Respiratory nitrate ammonification by Shewanella oneidensis MR-1.* J Bacteriol, 2007. **189**(2): p. 656-62.

126. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

127. Konstantinidis, K.T., A. Ramette, and J.M. Tiedje, *Toward a more robust assessment of intraspecies diversity, using fewer genetic markers.* Appl Environ Microbiol, 2006. **72**(11): p. 7286-93.

128. Bruen, T.C., H. Philippe, and D. Bryant, *A simple and robust statistical test for detecting the presence of recombination.* Genetics, 2006. **172**(4): p. 2665-81.

129. Goris, J., et al., *DNA-DNA hybridization values and their relationship to whole-genome sequence similarities.* Int J Syst Evol Microbiol, 2007. **57**(Pt 1): p. 81-91.

130. Kosakovsky Pond, S.L., et al., *GARD: a genetic algorithm for recombination detection.* Bioinformatics, 2006. **22**(24): p. 3096-8.

131. Tatusov, R., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**(1): p. 41.

132. Konstantinidis, K.T. and J.M. Tiedje, *Trends between gene content and genome size in prokaryotic species with larger genomes.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 3160-5.

133. Lawrence, J.G. and H. Ochman, *Amelioration of bacterial genomes: rates of change and exchange.* J Mol Evol, 1997. **44**(4): p. 383-97.

134. Drake, J.W., et al., *Rates of spontaneous mutation.* Genetics, 1998. **148**(4): p. 1667-86.

135. Jarvik, T., et al., *Short-term signatures of evolutionary change in the Salmonella enterica serovar typhimurium 14028 genome.* J Bacteriol, 2010. **192**(2): p. 560-7.

136. Wilson, D.J., et al., *Rapid evolution and the importance of recombination to the gastroenteric pathogen Campylobacter jejuni.* Mol Biol Evol, 2009. **26**(2): p. 385-97.

137. Carver, T., et al., *Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database.* Bioinformatics, 2008. **24**(23): p. 2672-6.

138. Hussain, H., et al., *A seven-gene operon essential for formate-dependent nitrite reduction to ammonia by enteric bacteria.* Mol Microbiol, 1994. **12**(1): p. 153-63.

139. Tyson, G.W., et al., *Community structure and metabolism through reconstruction of microbial genomes from the environment.* Nature, 2004. **428**(6978): p. 37-43.

140. Eppley, J.M., et al., *Genetic exchange across a species boundary in the archaeal genus ferroplasma.* Genetics, 2007. **177**(1): p. 407-16.

141. Chun, J., et al., *Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae.* Proc Natl Acad Sci U S A, 2009.

142. Giovannoni, S.J., et al., *Genome streamlining in a cosmopolitan oceanic bacterium.* Science, 2005. **309**(5738): p. 1242-5.

143. Coleman, M.L., et al., *Genomic islands and the ecology and evolution of Prochlorococcus.* Science, 2006. **311**(5768): p. 1768-70.

144. Brettar, I. and M.G. Hofle, *Nitrous-Oxide Producing Heterotrophic Bacteria from the Water Column of the Central Baltic - Abundance and Molecular-Identification.* Marine Ecology-Progress Series, 1993. **94**(3): p. 253-265.

145. Ziemke, F., I. Brettar, and M.G. Hofle, *Stability and diversity of the genetic structure of a Shewanella putrefaciens population in the water column of the central Baltic.* Aquatic Microbial Ecology, 1997. **13**(1): p. 63-74.

146. Li, Q., M. Hobbs, and P.R. Reeves, *The variation of dTDP-L-rhamnose pathway genes in Vibrio cholerae.* Microbiology, 2003. **149**(Pt 9): p. 2463-74.

147. Aydanian, A., et al., *Genetic diversity of O-antigen biosynthesis regions in Vibrio cholerae.* Appl Environ Microbiol, 2011. **77**(7): p. 2247-53.

148. Kato, C. and Y. Nogi, *Correlation between phylogenetic structure and function: examples from deep-sea Shewanella.* FEMS Microbiol Ecol, 2001. **35**(3): p. 223-230.

149. Lauro, F.M., et al., *The genomic basis of trophic strategy in marine bacteria.* Proc Natl Acad Sci U S A, 2009. **106**(37): p. 15527-33.

150. Matz, C. and S. Kjelleberg, *Off the hook--how bacteria survive protozoan grazing.* Trends Microbiol, 2005. **13**(7): p. 302-7.

151. Huson, D.H. and D. Bryant, *Application of phylogenetic networks in evolutionary studies.* Mol Biol Evol, 2006. **23**(2): p. 254-67.

152. Smoot, M.E., et al., *Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics, 2011. **27**(3): p. 431-2.

153. Doolittle, W.F. and E. Bapteste, *Pattern pluralism and the Tree of Life hypothesis.* Proc Natl Acad Sci U S A, 2007. **104**(7): p. 2043-9.

154. Ward, D., A., *A Macrobiological Perspective on Microbial Species.* Microbe Magazine, 2006(June, 2006).

155. Rossello-Mora, R. and R. Amann, *The species concept for prokaryotes.* FEMS Microbiol Rev, 2001. **25**(1): p. 39-67.

156. Dingle, K.E., et al., *Multilocus sequence typing system for Campylobacter jejuni.* J Clin Microbiol, 2001. **39**(1): p. 14-23.

157. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.

158. Suyama, M., D. Torrents, and P. Bork, *PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W609-12.

159. Schmid, K. and Z. Yang, *The trouble with sliding windows and the selective pressure in BRCA1.* PLoS ONE, 2008. **3**(11): p. e3746.

160. Ochman, H., *Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes.* Trends Genet, 2002. **18**(7): p. 335-7.

161. Lawrence, J., *When ELFs are ORFs, but don't act like them.* Trends Genet, 2003. **19**(3): p. 131-2.

162. McCarthy, N.D., et al., *Host-associated genetic import in Campylobacter jejuni.* Emerg Infect Dis, 2007. **13**(2): p. 267-72.

163. Coker, A.O., et al., *Human campylobacteriosis in developing countries.* Emerg Infect Dis, 2002. **8**(3): p. 237-44.

164. Fouts, D.E., et al., *Major structural differences and novel potential virulence mechanisms from the genomes of multiple campylobacter species.* PLoS Biol, 2005. **3**(1): p. e15.

165. Branscomb, E. and P. Predki, *On the high value of low standards.* J Bacteriol, 2002. **184**(23): p. 6406-9; discussion 6409.

166. Ghai, R., T. Hain, and T. Chakraborty, *GenomeViz: visualizing microbial genomes.* BMC Bioinformatics, 2004. **5**: p. 198.

167. Palenik, B., et al., *Genome sequence of Synechococcus CC9311: Insights into adaptation to a coastal environment.* Proc Natl Acad Sci U S A, 2006. **103**(36): p. 13555-9.

168. Liu, M., et al., *Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage.* Science, 2002. **295**(5562): p. 2091-4.

169. Konstantinidis, K.T. and J.M. Tiedje, *Trends between gene content and genome size in prokaryotic species with larger genomes.* PNAS, 2004. **101**(9): p. 3160-3165.

170. Lawrence, J.G. and H. Ochman, *Molecular archaeology of the Escherichia coli genome.* Proc Natl Acad Sci U S A, 1998. **95**(16): p. 9413-7.

171. Rocha, E.P., et al., *Comparisons of dN/dS are time dependent for closely related bacterial genomes.* J Theor Biol, 2006. **239**(2): p. 226-35.

172. Charon, N.W. and S.F. Goldstein, *Genetics of motility and chemotaxis of a fascinating group of bacteria: the spirochetes.* Annu Rev Genet, 2002. **36**: p. 47-73.

173. Paster, B.J. and F.E. Dewhirst, *Phylogenetic foundation of spirochetes.* J Mol Microbiol Biotechnol, 2000. **2**(4): p. 341-4.

174. Rosey, E.L., M.J. Kennedy, and R.J. Yancey, Jr., *Dual flaA1 flaB1 mutant of Serpulina hyodysenteriae expressing periplasmic flagella is severely attenuated in a murine model of swine dysentery.* Infect Immun, 1996. **64**(10): p. 4154-62.

175. Lux, R., et al., *Motility and chemotaxis in tissue penetration of oral epithelial cell layers by Treponema denticola.* Infect Immun, 2001. **69**(10): p. 6276-83.

176. Sadziene, A., et al., *A flagella-less mutant of Borrelia burgdorferi. Structural, molecular, and in vitro functional characterization.* J Clin Invest, 1991. **88**(1): p. 82-92.

177. Leschine, S., Paster, B., Canale-Parola, E., *Free-living saccharolytic spirochetes: The genus "Spirochaeta"*, in *The Prokaryotes*2006, Springer New York. p. 195-210.

178. Magot, M., et al., *Spirochaeta smaragdinae sp. nov., a new mesophilic strictly anaerobic spirochete from an oil field.* FEMS Microbiol Lett, 1997. **155**(2): p. 185-91.

179. Franzmann, P.D. and S.J. Dobson, *Cell wall-less, free-living spirochetes in Antarctica.* FEMS Microbiol Lett, 1992. **76**(3): p. 289-92.

180. Ritalahti, K.M. and F.E. Löffler, *Populations implicated in anaerobic reductive dechlorination of 1,2-dichloropropane in highly enriched bacterial communities.* Appl Environ Microbiol, 2004. **70**(7): p. 4088-95.

181. Zhilina, T.N., et al., *Spirochaeta alkalica sp. nov., Spirochaeta africana sp. nov., and Spirochaeta asiatica sp. nov., alkaliphilic anaerobes from the Continental Soda Lakes in Central Asia and the East African Rift.* Int J Syst Bacteriol, 1996. **46**(1): p. 305-12.

182. Janssen, P.H. and H.W. Morgan, *Glucose catabolism by Spirochaeta thermophila RI 19.B1.* J Bacteriol, 1992. **174**(8): p. 2449-53.

183. Ritalahti, K.M., et al., *Isolation of Sphaerochaeta (gen. nov.), free-living, spherical spirochetes, and characterization of Sphaerochaeta pleomorpha (sp.*

nov.) and Sphaerochaeta globosa (sp. nov.). Int J Syst Evol Microbiol, 2011: p. In press.

184. Thompson, J.D., T.J. Gibson, and D.G. Higgins, *Multiple sequence alignment using ClustalW and ClustalX.* Curr Protoc Bioinformatics, 2002. **Chapter 2**: p. Unit 2 3.

185. Paley, S.M. and P.D. Karp, *The Pathway Tools cellular overview diagram and Omics Viewer.* Nucleic Acids Res, 2006. **34**(13): p. 3771-8.

186. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2007. **35**(Database issue): p. D21-5.

187. Moriya, Y., et al., *KAAS: an automatic genome annotation and pathway reconstruction server.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W182-5.

188. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC Genomics, 2008. **9**: p. 75.

189. Felsenstein, J., *PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author*, in *Department of Genome Sciences, University of Washington, Seattle*2005.

190. Konstantinidis, K.T. and J.M. Tiedje, *Towards a genome-based taxonomy for prokaryotes.* J Bacteriol, 2005. **187**(18): p. 6258-64.

191. Liu, R. and H. Ochman, *Stepwise formation of the bacterial flagellar system.* Proc Natl Acad Sci U S A, 2007. **104**(17): p. 7116-21.

192. Mattei, P.J., D. Neves, and A. Dessen, *Bridging cell wall biosynthesis and bacterial morphogenesis.* Current opinion in structural biology, 2010. **20**(6): p. 749-55.

193. Sauvage, E., et al., *The penicillin-binding proteins: structure and role in peptidoglycan biosynthesis.* FEMS microbiology reviews, 2008. **32**(2): p. 234-58.

194. Allan, E.J., C. Hoischen, and J. Gumpert, *Bacterial L-forms.* Adv Appl Microbiol, 2009. **68**: p. 1-39.

195. Mursic, V.P., et al., *Formation and cultivation of Borrelia burgdorferi spheroplast-L-form variants.* Infection, 1996. **24**(3): p. 218-26.

196. Wise, E.M., Jr. and J.T. Park, *Penicillin: its basic site of action as an inhibitor of a peptide cross-linking reaction in cell wall mucopeptide synthesis.* Proc Natl Acad Sci U S A, 1965. **54**(1): p. 75-81.

197. Scheffers, D.J. and M.G. Pinho, *Bacterial cell wall synthesis: new insights from localization studies.* Microbiol Mol Biol Rev, 2005. **69**(4): p. 585-607.

198. Eisen, J.A., *Horizontal gene transfer among microbial genomes: new insights from complete genome analysis.* Curr Opin Genet Dev, 2000. **10**(6): p. 606-11.

199. Koski, L.B. and G.B. Golding, *The closest BLAST hit is often not the nearest neighbor.* J Mol Evol, 2001. **52**(6): p. 540-2.

200. Warnick, T.A., B.A. Methe, and S.B. Leschine, *Clostridium phytofermentans sp. nov., a cellulolytic mesophile from forest soil.* Int J Syst Evol Microbiol, 2002. **52**(Pt 4): p. 1155-60.

201. Mahowald, M.A., et al., *Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla.* Proc Natl Acad Sci U S A, 2009. **106**(14): p. 5859-64.

202. Moon, C.D., et al., *Reclassification of Clostridium proteoclasticum as Butyrivibrio proteoclasticus comb. nov., a butyrate-producing ruminal bacterium.* Int J Syst Evol Microbiol, 2008. **58**(Pt 9): p. 2041-5.

203. Bellgard, M.I., et al., *Genome sequence of the pathogenic intestinal spirochete Brachyspira hyodysenteriae reveals adaptations to its lifestyle in the porcine large intestine.* PLoS One, 2009. **4**(3): p. e4641.

204. Bott, M., *Anaerobic citrate metabolism and its regulation in enterobacteria.* Arch Microbiol, 1997. **167**(2/3): p. 78-88.

205. Uehara, T., et al., *Recycling of the anhydro-N-acetylmuramic acid derived from cell wall murein involves a two-step conversion to N-acetylglucosamine-phosphate.* J Bacteriol, 2005. **187**(11): p. 3643-9.

206. He, J., et al., *Influence of vitamin B12 and cocultures on the growth of Dehalococcoides isolates in defined medium.* Appl Environ Microbiol, 2007. **73**(9): p. 2847-53.

207. Morris, J.J., et al., *Facilitation of robust growth of Prochlorococcus colonies and dilute liquid cultures by "helper" heterotrophic bacteria.* Appl Environ Microbiol, 2008. **74**(14): p. 4530-4.

208. Paster, B.J., et al., *Phylogenetic analysis of the spirochetes.* J Bacteriol, 1991. **173**(19): p. 6101-9.

209. Kimsey, R.B. and A. Spielman, *Motility of Lyme disease spirochetes in fluids as viscous as the extracellular matrix.* J Infect Dis, 1990. **162**(5): p. 1205-8.

210. Canale-Parola, E., *Motility and chemotaxis of spirochetes.* Annu Rev Microbiol, 1978. **32**: p. 69-99.

211. Breznak, J.A. and E. Canale-Parola, *Morphology and physiology of Spirochaeta aurantia strains isolated from aquatic habitats.* Arch Microbiol, 1975. **105**(1): p. 1-12.

212. Harwood, C.S. and E. Canale-Parola, *Ecology of spirochetes.* Annu Rev Microbiol, 1984. **38**: p. 161-92.

213. Leschine, S.B., *Cellulose degradation in anaerobic environments.* Annu Rev Microbiol, 1995. **49**: p. 399-426.

214. Zhaxybayeva, O., et al., *On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales.* Proc Natl Acad Sci U S A, 2009. **106**(14): p. 5865-70.

215. Caro-Quintero, A., et al., *The chimeric genome of Sphaerochaeta: nonspiral spirochetes that break with the prevalent dogma in spirochete biology.* MBio, 2012. **3**(3).

216. Nelson-Sathi, S., et al., *Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea.* Proc Natl Acad Sci U S A, 2012. **109**(50): p. 20537-42.

217. Wolf, Y.I. and E.V. Koonin, *A Tight Link between Orthologs and Bidirectional Best Hits in Bacterial and Archaeal Genomes.* Genome Biol Evol, 2012. **4**(12): p. 1286-94.

218. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-1.

219. Newman, M.E.J. and M. Girvan, *Finding and evaluating community structure in networks.* Physical Review E, 2004. **69**(2).

220. Clauset, A., M.E.J. Newman, and C. Moore, *Finding community structure in very large networks.* Physical Review E, 2004. **70**(6).

221. Su, G., et al., *GLay: community structure analysis of biological networks.* Bioinformatics, 2010. **26**(24): p. 3135-7.

222. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

223. Nelson, K.E., et al., *Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima.* Nature, 1999. **399**(6734): p. 323-9.

224. Kosaka, T., et al., *The genome of Pelotomaculum thermopropionicum reveals niche-associated evolution in anaerobic microbiota.* Genome Res, 2008. **18**(3): p. 442-8.

225. Scholten, J.C., et al., *Evolution of the syntrophic interaction between Desulfovibrio vulgaris and Methanosarcina barkeri: Involvement of an ancient horizontal gene transfer.* Biochem Biophys Res Commun, 2007. **352**(1): p. 48-54.

226. Schink, B. and A.J. Stams, *Syntrophism among prokaryotes.* Prokaryotes, 2006. **2**: p. 309-335.

227. Ciccarelli, F.D., et al., *Toward automatic reconstruction of a highly resolved tree of life.* Science, 2006. **311**(5765): p. 1283-7.

228. Cohen, O., U. Gophna, and T. Pupko, *The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer.* Mol Biol Evol, 2011. **28**(4): p. 1481-9.

229. de Bok, F.A.M., C.M. Plugge, and A.J.M. Stams, *Interspecies electron transfer in methanogenic propionate degrading consortia.* Water Res, 2004. **38**(6): p. 1368-1375.

230. Obradors, N., et al., *Anaerobic metabolism of the L-rhamnose fermentation product 1,2-propanediol in Salmonella typhimurium.* J Bacteriol, 1988. **170**(5): p. 2159-62.

231. Sampson, E.M. and T.A. Bobik, *Microcompartments for B-12-dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate.* J Bacteriol, 2008. **190**(8): p. 2966-2971.

232. Chen, Y.Y., et al., *Pathways for lactose/galactose catabolism by Streptococcus salivarius.* FEMS Microbiol Lett, 2002. **209**(1): p. 75-9.

233. Jagusztyn-Krynicka, E.K., et al., *Streptococcus mutans serotype c tagatose 6-phosphate pathway gene cluster.* J Bacteriol, 1992. **174**(19): p. 6152-8.

234. Kurland, C.G., B. Canback, and O.G. Berg, *Horizontal gene transfer: a critical view.* Proc Natl Acad Sci U S A, 2003. **100**(17): p. 9658-62.

235. Doolittle, W.F., *You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes.* Trends Genet, 1998. **14**(8): p. 307-11.

236. Lefebure, T., et al., *Evolutionary dynamics of complete Campylobacter pan-genomes and the bacterial species concept.* Genome Biol Evol, 2010. **2**: p. 646-55.

237. Stepanauskas, R., *Single cell genomics: an individual look at microbes.* Curr Opin Microbiol, 2012. **15**(5): p. 613-20.

238. Ishoey, T., et al., *Genomic sequencing of single microbial cells from environmental samples.* Curr Opin Microbiol, 2008. **11**(3): p. 198-204.