# Precise correlation and metagenomic binning uncovers fine microbial community structure

by

W. Evan Durno

B.A. Mathematics, minor Statistics, The University of British Columbia, 2012

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2017

# Abstract

Bacteria and Archaea represent the invisible majority of living things on Earth with an estimated numerical abundance exceeding $10^{30}$ cells. This estimate surpasses the number of grains of sand on Earth and stars in the known universe. Interdependent microbial communities drive fluxes of matter and energy underlying biogeochemical processes, and provide essential ecosystem functions and services that help create the operating conditions for life. Despite their abundance and functional imperative, the vast majority of microorganisms remain uncultivated in laboratory settings, and therefore remain extremely difficult to study. Recent advances in high-throughput sequencing are opening a multi-omic (DNA and RNA) window to the structure and function of microbial communities providing new insights into coupled biogeochemical cycling and the metabolic problem solving power of otherwise uncultivated microbial dark matter (MDM). These technological advances have created bottlenecks with respect to information processing, and innovative bioinformatics solutions are required to analyze immense biological data sets. This is particularly apparent when dealing with metagenome assembly, population genome binning, and network analysis.

This work investigates combined use of single-cell amplifed genomes (SAGs) and metagenomes to more precisely construct population genome bins and evaluates the use of covariance matrix regularization methods to identify putative metabolic interdependencies at the population and community levels of organization. Applying dimensional reduction with principal components and a Gaussian mixture model to k-mer statistics from SAGs and metagenomes is shown to bin more precisely, and has been implemented as a novel pipeline, SAG Extrapolator (SAGEX). Also, correlation networks derived from small subunit ribosomal RNA gene sequences are shown to be more precisely inferred through regularization with factor analysis models applied via Gaussian copula. SAGEX and regularized correlation are applied toward 368 SAGs and 91 metagenomes, postulating populations metabolic capabilities via binning, and constraining interpretations via correlation. The application describes coupled biogeochemical cycling in low-oxygen waters. Use of SAGEX leverages SAGs deep taxonomic descriptions and metagenomes breadth, produces precise population genome bins, and enables metabolic reconstruction and analysis of population dynamics over time. Regularizing correlation networks overcomes a known analytic bottleneck based in precision limitations.

# Lay summary

Prokaryotic microorganisms, including bacteria and archaea, work together to transform the environment on local and global scales. Global scale influences occur through a combination of two important factors: 1) microbial life, while for the most part invisible, exists on truly massive scales, and 2) each microbial cell derives energy from local environmental transformations. The cumulative effects are substantial enough to drive global biogeochemical cycles over billions of years. Despite playing these integral roles, the majority of microorganisms remain uncultivated, rendering them similar to astronomical dark matter. High-throughput genome sequencing approaches are now shining light onto uncultivated microbial diversity and function, creating a number of bioinformatics challenges related to microbial genome assembly, taxonomic binning, and community metabolic network reconstruction. This thesis contributes toward precise taxonomic binning and correlation network methods, improving our capacity to understand the metabolic linkages between uncultivated microorganisms and biogeochemical cycles in natural and engineered ecosystems.

# Preface

While the majority of the work for this thesis was done by the author, W. Evan Durno, his reading on oxygen minimum zone microbial ecology was mostly directed by Alyse K. Hawley, but also by Steven J. Hallam. Both Alyse K. Hawley and Steven J. Hallam made essential contributions toward assisting the author's interpretation of all data and analyses. All data were taken from a pre-existing project studying the Saanich Inlet oxygen minimum zone.

For the metagenomic binning project (see chapter 2), essential classifier algorithms were largely inspired by the work of Dodsworth et al. [76]. The project motivations were initially imagined by Steven J. Hallam and Jody J. Wright prior to writing Hawley et al. [126]. Connor Morgan-Lang wrote the SAGEX interface. The author designed and wrote the primary software pipeline, and ran the methods comparison experiment. W. Evan Durno, Steven J. Hallam, and Alyse K. Hawley designed experiments which were implemented by W. Evan Durno, Alyse K. Hawley, or Connor Morgan-Lang.

For the SSU rRNA correlation project (see chapter 3), microbial counts were generated and taxonomically annotated by Monica Torres-Beltran. Kai He and Jessica Ngo assisted in the univariate model goodness-of-fit survey under the direction of the author. The author designed this project, implemented nearly all software, and ran all experiments. Bo Chang's reading group on graphical models with guidance from Harry Joe and Ruben Zamar was invaluable for this project. Ed Gabbott provided the GeForce GTX 980 Ti GPU, which was put to good use.

# Table of Contents

## Appendices

# List of Tables

# List of Figures

# Glossary

## Bioinformatics

**ACE** Abundance-based coverage estimator

**BLAST** Basic Local Alignment Search Tool

**BWA** Burrows-Wheeler Aligner

**CONCOCT** Clustering cONtigs with COverage and ComposiTion

**DNA** Deoxyribonucleic acid

**EBPR** Enhanced biological phosphorus removal

**ESOM** Emergent Self Organizing Map

**HGT** Horizontal gene transfer

**IMG** Integrated Microbial Genomes system

**LCA** Lowest common ancestor

**MEGAN** MEtaGenome ANalyzer

**MGA** Marine Group A or Marinimicrobia

**MSA** Multiple sequence alignment

**NCBI** National Center for Biotechnology Information

**OMZ** Oxygen minimum zone

**ORF** Open reading frame

**OTU** Operational taxonomic unit

**PCR** Polymerase chain reaction

**QIIME** Quantitative Insights Into Microbial Ecology (software)

**RNA** Ribonucleic acid

**RPKM** Reads per kilobase mapped

**SAGEX** SAG EXtrapolator

**SAG** Single-cell amplified genome

**SOM** Self Organizing Map

**SSU rRNA** Small subunit of the ribosomal RNA gene

## Math and Statistics

**AIC** Akaike information criterion

**CCA** Canonical Correspondence Analysis

**CMP** Conway-Maxwell-Poisson (distribution)

**FDR** False discovery rate

**FMWC** Floor model with copula

**HMM** Hidden Markov Model

**LNM** Logistic normal multinomial

**MLE** Maximum likelihood estimate

**MM** Method of moment estimate

**NBGC** Negative binomial with Gaussian copula

**PCA** Principal component analysis

**PC** Principal component

**VST** Variance stabilizing transform

## Computation

**ACM** Association for Computing Machinery

**BFGS** Broyden-Fletcher-Goldfarb-Shanno numerical optimization algorithm

**BWT** BurrowsWheeler transform

**CPU** Central processing unit

**CUDA** Compute unified device architecture

**GPU** Graphics processing unit

**LAPACK** Linear Algebra PACKage

**MC** Monté Carlo (integration)

**RAM** Random access memory

# Acknowledgments

This degree would not have been possible without support from my family, especially Agnes. Neither would I have succeeded without the scholarly guidance provided by Alyse K. Hawley. I'd like to thank Harry Joe and Ruben Zamar for actively guiding me toward resources which gave me excellent results. I would like to thank the Hallam Lab students and staff for their guidance. I'd like to thank Steven J. Hallam, not just for research guidance and resources, but also for guidance in writing which made this document something I can be proud of.

# Chapter 1

# Introduction

Prokaryotic life is estimated to account for $4 - 6 \times 10^{30}$ cells, and to sequester $350 - 550$Pg of carbon, comparable to that of all plant life [269]. Because microbes make their living transforming energy and matter, their cumulative transformative effect is truly massive, having great participation in Earth's biogeochemical cycles [85]. Microbial forces not only contribute toward long-term effects in biogeochemical cycling, determining much of the Earth's chemical history [148, 236] and future [128, 229], but are also applicable in more immediate time-scales [31, 99, 184, 198, 202, 219, 259]. Understanding how these microbial forces work in ecological contexts requires sequencing genetic material directly from the environment (see section 1.1). For studying microbial life's ecological machine, this work's inferential mechanism is a two-step, bioinformatic process. First, metabolic capabilities are attributed to taxa. This can be done in a variety of ways (see section 1.2). Second, the breadth of ecological interpretations is constrained through a correlative analysis (see section 1.3). This work makes contributions toward such inferential mechanisms in chapter 2 and chapter 3, and efforts are evaluated with precision-recall curves (see subsection 1.6.5) and through application toward studying denitrification in Saanich Inlet (see section 1.4). Requisite math, statistics, and computational concepts are described in section 1.5, section 1.6, and section 1.7, respectively.

## 1.1 Metagenomics and the network perspective

The greatest challenge to understanding microbial forces is caused by the lab itself, because the vast majority of taxa will not grow in the lab. For example, less than 1% of marine microbes will grow on standard agar plates [58, 89, 197]. This means that results concluded from lab-grown microbes and communities are prone to *cultivation bias*. Fortunately, modern *cultivation-independent* methods, often developed involving *genetic sequencing*, exist to exchange the concerns due to the lab bench for the uncertainties of observational experiments. The exchange to reduce bias at the cost of experimental constraints is worthwhile, especially when both perspectives can be combined. The key difference that makes cultivation-independent methods work is that samples are sequenced soon after being taken directly from the environment with a significantly reduced opportunity for bias-enducing effects to afflict the sample. Genetic sequencing operates like a quickly-taken snapshot, most-closely representing the actual community in the environment as it was at sampling. This work largely focuses on three cultivation-independent data types: **metagenomes**, single-cell amplified genomes (**SAGs**), and small subunit of the ribosomal RNA gene (**SSU rRNA**).

### 1.1.1 Data types

A metagenome is a sample of DNA taken directly from the environment [260]. It has the advantage of describing DNA as it existed in the environment, without bias-enducing, requisite cultivation steps. The DNA itself is not immediately useful and requires further processing. Modern high-throughput sequencing through the Illumina platform [26] translates the DNA into millions or billions of short strings, often 100-200 base pairs (bp) in length. Each string of A,T,C,G characters is sampled randomly from the initial DNA and may overlap. Overlaps are valuable because the small per-base-pair error rate (often below 1%) accumulates opportunities for errors in the many reads. If certain DNA was sequenced more often, it is said to have more *(sequencing) depth*. Important alternatives to metagenomes include *metatranscriptomes* and *metaproteomes*. A metatranscriptomes is sequenced from reverse-transcribed, environmentally sampled RNA. A metaproteome is a set of amino acid sequences from environmentally sampled protein.

After sequencing, metagenomes exist in many *reads*, which are small sub-sequences. To learn what the DNA encodes, a *short-read aligner* (such as BWA (Burrows-Wheeler Aligner) [160]) can be employed to perform look-ups in a database of known functions or taxa (such as NCBI's (National Center for Biotechnology Information) RefSeq-nr [254, 268], GreenGenes [71], or SILVA [220]). It also is possible to attempt a reconstruction of the underlying genomes which produced the reads through *assembly* (see subsection 1.2.1), but the process tends to produces merely longer sub-sequences and can introduce errors. After assembly, several genes may share a single sequence. While some genes may identify genetic function or taxonomy, many sequences cannot have their taxonomy known definitively. This presents a fundamental challenge in Microbial Ecology, because linking function to taxa is such an important goal.

If taxonomy is the focus of a study, all sequencing power can be directed toward sequencing SSU rRNA genes. This is done by amplifying SSU rRNA genes prior to sequencing. After sequencing and processing (see subsection 1.2.3), sequences counts can be used as a proxy for microbial abundances. The analysis of SSU rRNA data is the subject of chapter 3 and further described in section 1.3.

Genetic function and taxa can be more confidently linked with SAGs [29, 249, 253], which are the genomes of individual cells. SAGs are produced by first sorting (perhaps with a microfluidic device [106]), amplifying with multiple displacement amplification (MDA), then sequencing. While contamination is possible, software can be used for quality assurance [207], though it is not yet proven to provide perfect results. The result is a relatively confident description of functional and taxonomic links. SAG collections are growing large [225]. For example, in chapter 2, 368 SAGs bioinformatically studied. However, SAGs remain focused descriptions of relatively few organisms and cannot meet the descriptive breadth achievable with metagenomes and SSU rRNA data. Therefore a combined approach is often motivated.

### 1.1.2 Information flow and the network abstraction

The *Central Dogma of Molecular Biology* [64] is an abstraction which describes how information encoded in the genetic alphabet ({A,T,C,G}) is translated to RNA and then to protein. This direction of informational flow is assumable for this work, but does not hold in general. Effectively, it provides a permanent caveat for all genomic analyses: genetic potential does not ensure expression. So if a gene is found, it does not certainly get used. It might be argued that in

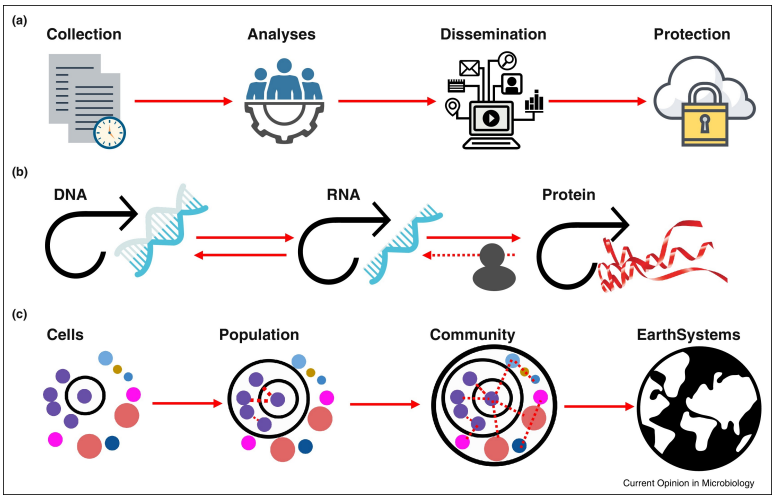



**Figure 1.1:** The Central Dogma compared to Information Theory and microbial communities. Image credit: [120]

resource-starved environments, microbial life can hardly spare wasted genetic material, but ultimately transcriptomics (RNA) is more reliable, and proteomics even more so. While describing a likely future of cloud-based bioinformatic analysis, Hahn et al. [120] described the information flow in the Central Dogma (Figure 1.1 (b)) and related information warehousing (Figure 1.1 (a)) with Claude Shannon's Theory of Communication [239], and also likened microbes to information processors acting in larger networks (Figure 1.1 (c)), metabolically processing their world as a communal machine. This work adopts such a network perspective through genetically-described correlation networks and compares it to previous interpretations [125, 167] (see subsection 3.4.4).

Working toward a better understanding of microbial-mediated environmental transformation, it is pragmatic digest data and communicate understanding with ecological models. Often models are conceptual [125, 198], other times they are differential [20, 27, 179, 187], but all these models are coherently abstractable to entities and their interactions. Taking entities as nodes and interactions as edges, there is a network representation of biological interaction [22, 137]. These network representations have faced some controversy in their early years [4, 78], but the scientific community has since adopted a necessary respect for the possible breadth of interpretation [267]. Modern genomic data now provides many opportunities to survey general community behaviour on a grand scale. For example, Figure 1.2 is a co-occurence network generated from globally-sampled data. These networks' abstractions allows large-scale analysis and interpretation [87, 98].

The idealization of conceptual and differential ecological models into network abstractions provides a perspective of deeply complex entities that is seductively simple. For example, it is easy to forget that a correlation network only comments on data covariation, and might be perceived as similar to a metabolic network. The perspectives are dependent but not equivalent. Worse yet, the automatic generation of networks through correlation may require great familiarity with an inferential model to properly interpret results, especially in cases of model failures which result in artefactual expressions. The greatest challenge in SSU rRNA correlation methods is confident edge calling. Popular methods have been shown to have low precision in successfully detecting

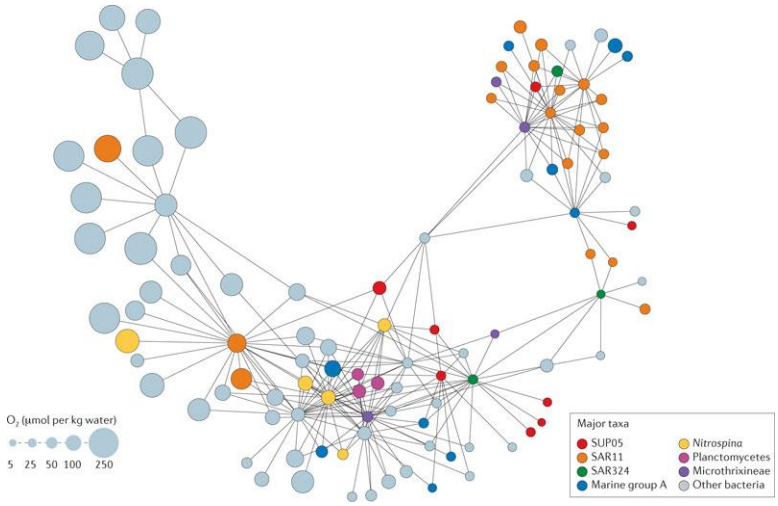



**Figure 1.2:** Microbial co-occurrences produced from Saanich Inlet, Hawaii Ocean Time-series, Microbial ecology of expanding oxygen minimum zones, and eastern tropical South Pacific OMZs' SSU rRNA data. Image credit: [271]

correlations [267]. Doubt in network edges invalidates claims of interaction, preventing confident estimation of correlation networks. Low edge-calling precision denies network application outside of any graph summary statistics which might be robust to this. This work shows how edge-calling precision can be achieved in chapter 3.

This work further concretizes its network abstraction by describing its nodes (chosen taxa) genomically. The task of assigning genomic sequences and their functions to taxonomies is a popular bioinformatic problem in cultivation-independent analysis and thus Microbial Ecology. In section 1.2, methods for attributing taxa to genomic function are described, where the concept of *metagenomic binning* is developed. The process is naturally tedious and erroneous, motivating pragmatism and thus clear descriptions of confidence are desired. In chapter 2, a fundamental pragmatism-quality exchange is explored through precision-recall analyses amongst select binning strategies, and new strategies are described which allow descriptive frontiers to be expanded. Ultimately, a careful comparison and measurement of binning methodologies describes how pragmatism and precision may be exchanged, and binning is generally described as more accurate near when attributing taxa nearer to the phylogenetic root.

This work adopts the network model perspective of oxygen minimum zone (OMZ) microbial ecology, and the majority of its emphasis is on precise network inference. This thesis targets both the nodes and edges of these networks. First, concerns about binning [240] are met by showing that SAG-guided binning can improve precision, thus better describing network's nodes. Second, this work addresses concerns about imprecise correlation detection [267] by using the right statistical model, thus better describing the network's edges. Ultimately, the data-driven argument made with correlation networks is made more credible through better nodes and better edges.

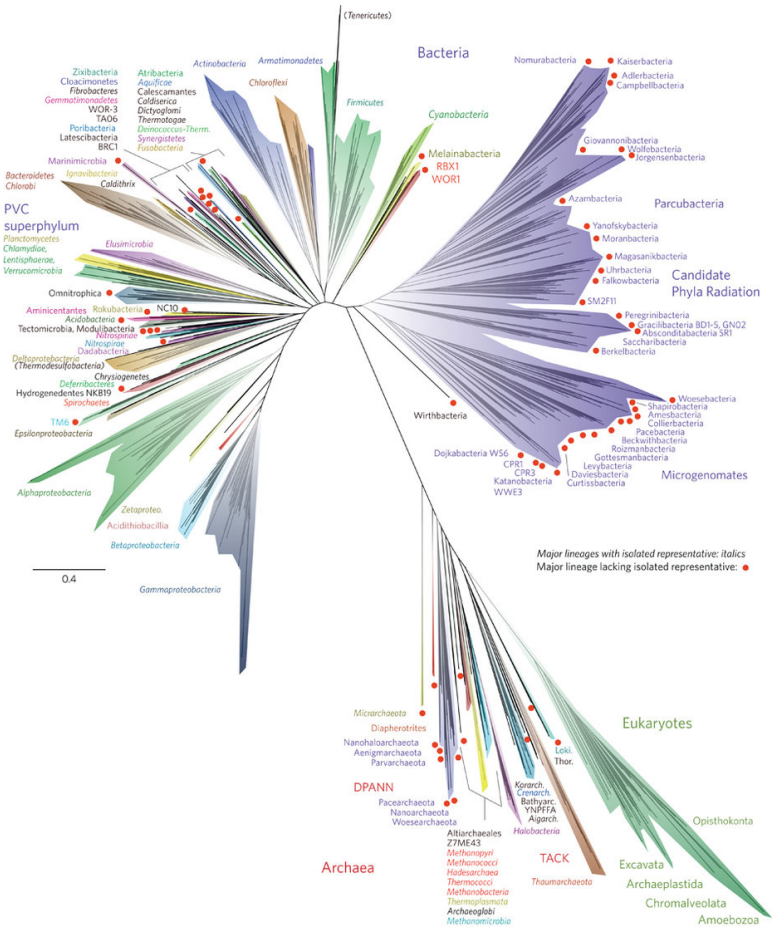



**Figure 1.3:** *Tree of Life* phylogeny estimated with binning. Image credit: [133]

## 1.2 Taxonomic estimation

The task of understanding how microbes cycle energy and nutrients is inseparable from evolutionary dependencies. For the sake of this work, it is sufficient to define a **phylogeny** as a particular evolutionary history or tree. It is also sufficient to define a **taxon** (plural: *taxa*) as a clade or entire branch of a phylogenetic tree, and **taxonomy** is the science of defining or classifying organisms into taxa. An example of an estimated phylogenetic tree is in Figure 1.3.

### 1.2.1 Assembly

Metagenomic sequences may be assigned a sense of phylogenetic similarity by attempting to reconstruct the originating genetic sequence. If in fact several metagenomic sequences do originate from a single strand of DNA, then they certainly share a single taxonomic source. A sequence assembled from overlapping DNA is a **contig** (as in *contiguous*). Reconstructing originating genetic material is difficult, because genetic sequencing technology is only able to read DNA in pieces. For example, the popular Illumina [26] platform may translate DNA into 150 base pair (bp)-length reads in a summary file several gigabytes in size. Therefore modern assembly is computational

problem with necessary approximations and algorithmic solutions [195]. It is important to note that the assembly of metagenomes increases the risk of joining sequences of different taxonomy (**chimeras**) [35], and also begs philosophical questions like *what is a genome?*

The intuitive algorithm of searching for similarity at sequences' ends, overlapping, and *walking through the genome* [194] is the overlap-layout-consensus (OLC) algorithm, and while still relevant, is computationally intractable for many sequences. To generate larger contigs from many small reads, is the **de Bruijn graph** approach [215, 280]. De Bruijn graphs are abstract data structures, concretized in both obvious and clever ways. These data structures can still be large, perhaps motivating distribution over a computer network [244]. A succinct de Bruijn graph [33] is modern reinterpretation inspired by the Burrows-Wheeler Transform (BWT) [43], and decreases memory requirements drastically [159].

### 1.2.2 Phylogenetic estimation

#### Bootstrapped phylogenetic tree estimation

A phylogenetic tree can be estimated for a set of genetic sequences. The genetic sequences must be similar enough to compare (perhaps encoding the same protein), and need to exist across the breadth of the tree–otherwise individual clades would be estimable and incomparable. Genes which are known exist ubiquitously are **marker genes**. An example of an important marker gene is the SSU rRNA gene [155], because it has both identifiable, conserved regions and known hyper-variable regions. A common approach to phylogenetic estimation is to first bioinformatically process marker gene sequences with multiple sequence alignment (MSA) [72, 80], which lines-up genes as closely as possible. Then, sequences may be process through an evolutionary model to produce a phylogenetic tree [248]. The comparison of ubiquitous marker genes makes this possible [207, 240].

There is variance in the estimation of phylogenetic trees. It is fair to ask "*is this clade real?*" or "*is my result merely due to chance?*" Statistical methods largely exist to extract signals from noisy data, and despite the complexity of phylogenetic estimation, there is a method for describing clade confidence: the **bootstrap**. Bootstrapping phylogenetic trees [84, 88, 132] requires randomly re-sampling data and estimating a tree for each sample. If a clade exists in many estimates, it is likely not a false inference of sampling's natural variation. It is important to note that phylogenetic tree estimates have a typical error profile: confidence increases near the root [204]. For example, it is easier to attribute phylum than a species.

#### Phylogenetics for statisticians

Having described estimation methods, it is fair to ask *what was being estimated?* Different underlying models are used per context depending on how useful they are, and often go unstated. A framework which forces explicit model definition is in **phylogenetic regression** [178], where hypothesis testing can be used to compare the likelihood of different phylogenetic structures. Using the tree to constrain Brownian motion or Ornstein-Uhlenbeck drift, a covariance structure can be defined (via precision matrices) [224] as illustrated in Figure 1.4.

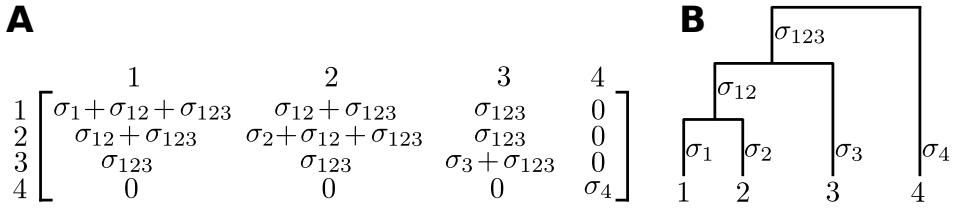



**Figure 1.4:** A covariance sturcture (A) as implied by a Phylogenetic estimate (B)

### 1.2.3 QIIME, SSU rRNA data processing

QIMME (Quantitative Insights Into Microbial Ecology, pronounced 'chime') [51] is a bioinformatic pipeline used to ready raw genomic sequencing data for statistical analysis. In this work, QIIME is used to prepare SSU rRNA sequence data for multivariate regression analysis. An early step in this QIIME workflow is demultiplexing, the sorting of genetic sequences into originating samples based on prepended nucleic tags on every sequence. Second, QIIME employs UCLUST [82] to approximately cluster SSU rRNA sequences in linear time [81]. For this work, SSU rRNA gene sequences must be at least 97% identical to share a cluster (percent-identity is an adjustable parameter). Third, cluster representatives are aligned to a taxonomy database. This work uses GreenGenes [71], but SILVA [220] is also popular. After counting the number of sequences recruited to each cluster, data are summarized in a tabular format. An example of this sort of data follows.

$$\begin{array}{c} \\ sample_1 \\ sample_2 \\ sample_3 \\ \vdots \\ sample_n \end{array} \begin{array}{c} \begin{array}{ccccc} taxon_1 & taxon_2 & taxon_3 & \cdots & taxon_p \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & \dots & 0 \\ 12 & 4 & 521 & \dots & 91 \\ 1642 & 1373 & 1209 & \dots & 1031 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 3 & 2 & \dots & 2 \end{bmatrix} \end{array}$$

While these data provides valuable descriptions of a community's taxa [59], it comes with some statistical challenges. *First*, it suffers a common problem in Bioinformatics: it often has thousands of dimensions (taxa) and relatively few samples. For example, using SSU rRNA data to survey for inter-taxa correlations has a high error rate [267]. *Second*, sequencing depth is variable per sample, which causes statistical dependence between taxa, obfuscating meaningful signals–biological interpretation is hidden behind methodological complications. This statistical complications are addressed in chapter 3.

### 1.2.4 Metagenomic binning

In subsection 1.1.1, metagenomes were described as having vast descriptive potential while often suffering a decoupling of function and taxonomy. Knowing which organisms are performing which functions is a central concern of any ecological field, including microbial ecology. Attempts at linking taxonomy and function are often made through a combination genomic sequence

Not all genetic material encodes marker genes or can be further assembled. Further strategies

fall into the broad category of *metagenomic binning*. For this work, it is sufficient to define **metagenomic binning** or **binning** as any attempt to decide that genetic sequences are phylogenetically similar. Binning produces **bins**, which are collections of sequences. A *binner* is a tool which assists or automates binning. A binner which produces bins with a taxonomic label is a *classifying* binner, otherwise it is a *clustering* binner. Binning strategies are often automated and work with assembly software [93, 234, 261].

Metagenomic binning has been reinterpreted over time [174, 180, 234], so generally applicable definitions will now be provided. A general trend has been toward pragmatic strategies, further from precise methods. Earliest software specifically for metagenomic binning includes TETRA [256] and Phylopythia [182]. TETRA is best described as a clustering method whereas Phylopythia is a classifier. Self-Organizing Maps (SOMs) were used as a clustering binner [1], despite originally being designed for data visualization and dimensional reduction [147]. A popular emergent SOM (ESOM; designed for cases with more nodes than data) software source is available from http://databionic-esom.sourceforge.net/ [263]. Another clustering method is MaxBin 2.0 [276], and another classifying method is MEGAN's (Metagenome Analyzer) LCA (lowest common ancestor) algorithm [135]. Many binners use tetranucleotide frequencies [255]. By 2012, both a variety of classifying and clustering binners had come into existence [174]. Recently the field has been re-interpreting binning as exclusively clustering methods [234]. Clustering methods have been used to inform on symbiotic relationships [40], genome isolations [65], and the tree of life [133]. A classic example of a binning application is an extraction of genomes from an acidophilic biofilm [261].

Since there are so many binning softwares, it is convenient to organize them by strategy. Some binners may only require a metagenome to operate (consider ESOMs [1, 147, 263]), and will leverage *kmer* (pronounced 'Kay-mer') and *coverage* statistics. A kmer statistic is a vector of counts, per genetic sequence, how many times the sequence contains each unique subsequence of length $k$ ($k = 4$ is popular, and also called a *tetranucleotie frequency*). So a kmer statistic will describe how many times a sequence contains the substrings AAAA, AAAT, AAAC, AAAG, AATA, and so on. A tetranucleotide statistic is a vector of $256 = 4^4$ counts. A coverage statistic is a description of how much DNA from a sample belongs to a sequence. This is necessary because assembly algorithns will collapse many reads into a single sequence. RPKM (Reads per kilobase mapped) [191] is a popular coverage statistic. Other binners will also use additional information, such as a marker gene data base [275]. Marker genes are conserved genes with a known phylogenetic distribution. An example of one such binner is MaxBin2.0 [276]. Other binners may further use aggregate genome collections such as the Integrated Microbial Genomes (IMG) [176] system and NCBI's [268] RefSeq [254]. Phylopythia is one such binner [111, 182, 208]. Use of additional data resources is important for quality binning, but also narrows a binner's applicable scope.

Evaluation of binning products is a topic of debate amongst microbial ecologists. A popular approach has been to study marker gene distribution within bins [5, 225, 240]. This approach has been automated [207] and will likely be included in future genomic standards [90, 91]. This work shows that choice of binning strategy is an important factor in determining binning quality.

A modern concern for binning is that error and incompleteness is inconsistently described and understood [55, 175, 240]. In chapter 2 error profiles for binning strategies are analyzed for trends, contributing toward a better understanding of binning error.

## 1.3 SSU rRNA correlation

Characterizations of ecological diversity are long-sought and have various perspectives. While $\alpha$, $\beta$, and $\gamma$-diversity concepts [13] emphasize diversity at the sample level or higher, and have employed statistics like Chao1 [56], ACE (abundance-based coverage estimator) [134], Sorensen, Jaccard, and Bray-Curtis [36] amongst others, genomic sequencing has allowed exploration of a high-resolution frontier. Correlating sequence counts describes taxa interactions and holds weight in ecological models [87]. Sophisticated modern correlation approaches in microbial ecology include Local Similarity Analysis [79, 231, 278], random matrix theory-based [70], hypergeometric sampling-based [54], maximal information coefficient [223], SparCC [95], and others [157]. Simple techniques such as co-presence and mutual exclusion [162], Pearson, and Spearman correlation are also popular.

An important recent discovery is that modern correlation techniques in microbial ecology have poor precision-recall exchanges [267], meaning that graph edges are often incorrectly attributed. This is a clear concern, motivating further research, because it casts doubt over the results of correlation network-based methods. This work contributes to solving this problem in chapter 3. While low edge-calling precision should increase spurious edges in graphs, it might not disturb certain graph structures in larger graphs. The current applied potential of existing methods can be seen in studies of the soil microbiome relationships with carbon dioxide ($CO_2$) [281], human gut microbiome [14] relationships with genetics [107] and disease [109], and poultry [201]. Considering these achievements highlights the fact that precise edge-calling is a frontier for Microbial Ecology. Precision in individual correlation inference is valuable, because it would confidently inform on interactions between individual pairs of taxa.

Relevant multivariate regression approaches exist for SSU rRNA data, particularly Canonical Correspondence Analysis (CCA) [258]. The fundamental challenge in analyzing SSU rRNA data is high-dimensionality. There are far more measured taxa than samples taken. The solution for CCA is similar to the approaches in chapter 3, both are a form of dimensional reduction. In CCA, dimensional reduction is accomplished by regressing against the primary axis of variation (an eigenvector). In chapter 3, precision is increased by constraining the breadth of possible correlation matrices. In this way, CCA is the spiritual predecessor of methods proposed in chapter 3.

Modern correlation networks in microbial ecology are usually generated with SSU rRNA data. SSU rRNA counts are an attempt to indirectly count taxa abundances per sample. The technical aspect of generating SSU rRNA data is elaborated upon in an example in subsection 1.1.1 and subsection 1.2.3. The effective data is thus a list of non-negative integer counts per microbial taxa for each sample. Often many samples are gathered, and the resulting data product is a non-negative integer matrix. This count matrix may be further processed before network production with a previously mentioned correlation tool. SSU rRNA data is often sampled along with environmental or experimental data (*regressors*), which may be used to infer meaningful changes in community structure.

The indirect observation of taxa counts through SSU rRNA data carries certain caveats. The primary caveat is that amplification and sequencing can induce a general positive correlative effect between all taxa. Effectively, every sample's counts will be shifted up or down together, thus obfuscating the true values. A popular solution has been to convert each sample into a list of proportions or *relative abundances*. Unfortunately, ratios inherently induce a general

negative correlative effect between all taxa. While converting to proportions has enabled plenty of insight, proportions' inherent negative correlative dependencies can cause obfuscation as well, thus better corrections are motivated. The best modern solution has been to estimate the effect as an unmeasured regressor in a count model [185]. While unmeasured, this regressor is not treated as a mixed effect, because as estimate (generated from the count matrix) is instead used as an observed regressor. Another approach is to apply variance stabilizing transforms (VSTs) McMurdie and Holmes [185].

### 1.3.1 Compositional vs. mixed effect perspective

The *compositional* and *mixed effect* perspectives of SSU rRNA modelling are the products of independent thought, but have elegantly produced a false dichotomy. Under the compositional perspective, it is imagined that taxa's SSU rRNA counts compete for sequencing depth, and thus tend to be negatively correlated. Under the mixed effect perspective, it is imagined that taxa's SSU rRNA counts are all driven up or down together with overall sequencing depth, and thus tend to be positively correlated. The idea that counts might bias toward positive or negative correlation with each other is the false dichotomy, because it is only imposed by the models used to understand this data. Of course, it is possible that either perspective might be more or less relevant dependent on unknown circumstances. The compositional perspective has been realized in the study of correlation networks [95, 277], while the mixed effect perspective is the result of univariate regression surveys [168, 185, 226, 228]. In chapter 3, this work explores SSU rRNA data with multivariate regression, which brings these two perspectives together in a surprisingly meaningful way (see subsection 3.3.2).

While the compositional effect merely indicates a tendency toward negative correlation, the mixed effect perspective borrows from the broad category of it mixed effect linear models [183]. In regression analyses (see subsection 1.6.2), a mixed effect is an unobserved random variable which obfuscates statistical signals, but can be dealt with through modelling the effect. In RNA seq and SSU rRNA regression, the concept is generalized to an unobserved random effect shared over all measurements in a single observation. The effect is usually attributed to sequencing depth effects. For example, if $(Y_1, Y_2)$ are taxa counts, $X$ is environmental measurements, $(\varepsilon_1, \varepsilon_2)$ are observational errors, and $M$ is the mixed effect, then the following equations share $M$ as a mixed effect. These kinds of equations are decribed in section 1.5.2.

$$Y_1 = \mu_1 + X\beta_1 + M\gamma_1 + \varepsilon_1, Y_2 = \mu_2 + X\beta_2 + M\gamma_2 + \varepsilon_2$$

## 1.4 Use case: Saanich Inlet OMZ

### 1.4.1 Oceanic nitrogen loss

An oxygen minimum zone (OMZ) is a subsurfance body of anoxic water ($< 20\mu$M $O_2$/kg). Under this definition, OMZs account for 7% of the ocean by volume [209]. Oceanic oxygen ($O_2$) concentrations have recently decreased rapidly in the north east Pacific OMZ [143], and OMZs are generally expected to expand as the planet warms [250]. OMZs play a disproportionately large role in oceanic denitrification, and because nitrogen is a limiting nutrient, excessive nitrogen

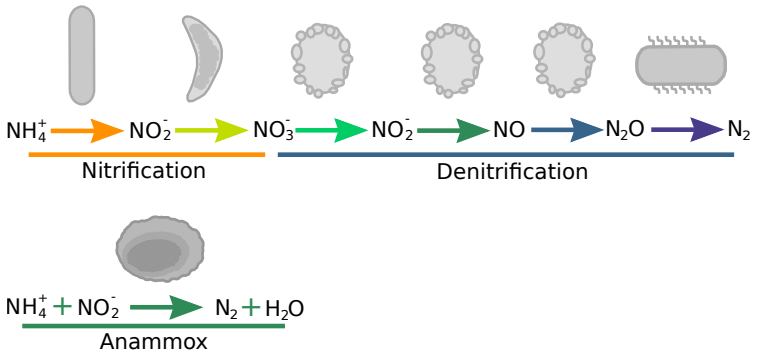



**Figure 1.5:** Nitrogen loss examples. Taxa will be argued throughout this work.

loss limits the ocean's ability to sequester atmospheric $CO_2$ [61]. These facts combined with the clear correlation between $CO_2$ and global warming [229] motivate understanding. It is important that global warming's tipping points [158] be identified. The microbial component of the global nitrogen cycle is both substantial and not fully understood [50, 154, 216, 262]. It is clear that the anoxic conditions favour chemolithotrophic energy metabolism [77], and these alternative metabolic strategies produce microbial ecological networks resulting in oceanic nitrogen loss [271].

Nitrogen gas ($N_2$) is abundant in the atmosphere, very stable, and nearly biologically inaccessible. When $N_2$ is produced, it has effectively left the biotic realm of the nitrogen cycle. Biologically accessible nitrogen is produced during nitrogen fixation ($N_2 \rightarrow NH_3$), an energetically costly transformation. Diazotrophs are a group of nitrogen fixing cyanobacteria. Microbially mediated nitrification causes accumulations of nitrate ($NO_3^-$) in the ocean. Because OMZs favour chemolithotrophy, nitrate and ammonium are used as an energetic resource. **Anammox** (anaerobic ammonium oxidation) and **denitrification** are the primary avenues of oceanic nitrogen loss [19] as exemplified in Figure 1.5, with anammox accounting for 30–50% of oceanic $N_2$ production [73].

### 1.4.2 Saanich Inlet

Saanich Inlet is a seasonally anoxic fjord on the coast of Vancouver Island, British Columbia, Canada [12]. Most of the year, the fjord has an anoxic basin. In the later summer and early fall, oxic waters flow into the basin, renewing oxygenation. The inlet is part of the Line P transect, a oceanographic time series [212]. From an OMZ research perspective, Saanich Inlet provides an opportunity to study microbial ecology and nutrient cylcing in OMZs [266, 271] particularly under oxic-anoxic shifts [279]. The Saanich Inlet time series is a valuable collection of data including both chemical and genomic measurements, and also a large collection of sequenced genomic data including 91 metagenomes, 368 SAGs [230], and 298 pyrotag SSU rRNA samples [127]. In this work, all metagenomes, all SAGs, and 112 SSU rRNA samples are used. The SAGs were sampled in August 2010 along with four other metagenomes

The environmental and SSU rRNA data sampling scheme featured a time series spanning from 2006 to 2011, but this study will only use 90 samples from 2008 onwards. This is because target variables (concentrations of oxygen, Nitrate ($NO_3^-$), and hydrogen sulfide ($H_2S$)) have missed measurements outside of this range, and model-based interpolations are eventually computationally intractable. A distribution of average chemical concentrations of shown in Figure 1.6, and it

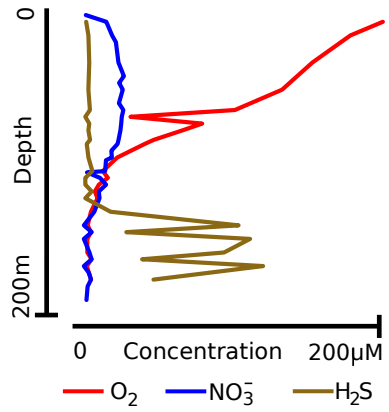



**Figure 1.6:** Saanich Inlet average chemical concentrations

can be see that the OMZ can be described in distinct sections: such as upper & lower oxycline, S/N transition zone, and sulfidic zone [125]. A common concern about this data set is for a lack of technical or biological replication, since samples were rarely taken in duplicate. However, intelligent model selection and problem phrasing overcomes these irregularities and concerns. This work will eventually study this data through multivariate regression (see chapter 3), which treats each samples' SSU rRNA counts $Y_i$ as conditionally distributed given the environmental variables $X_i$. The 90 conditional variables $(Y_i|X_i)$ are each replicates. The only constraint on the data is that it is from an observational study, which is now a common caveat for microbial ecologists due to environmental sampling. Observational studies often cannot explore a full combination nor range of variable values. However, the Saanich Inlet OMZ time series remains natural experiment describing oxic-anoxic transitions.

In this work, Saanich Inlet OMZ data will be used to demonstrate how precise network inference methods reproduce known understanding and novel perspectives on microbially mediated denitrification in the Saanich Inlet OMZ. Microbially mediated denitrification, which is closely related to oceanic nitrogen loss, is a choosen focus from many possible topics. The conceptual model posed by Hawley et al. [125] describes this process in detail (see Figure 1.7), and will often be used as the primary argumentative basis when interpreting new evidence.

### 1.4.3 Important taxa

#### SUP05 (Candidatus Thioglobus autotrophicus)

SUP05 gammaproteobacteria are a major participant in the OMZ denitrification pipeline. Following first observation [252], an early metagenomic analysis by Walsh et al. [266] used SSU rRNA data to show that SUP05 was often a dominant population in several OMZs, not just Saanich Inlet [279] but also the ESP and Namibia OMZs. The work also provides early insights into SUP05's metabolic capacity through assembly of metagenomic fosmids. This SUP05 metagenome encodes the metabolic potential for autotrophic carbon fixation of $CO_2$, dissimilatory sulfite reductase ($SO_3^{2-} \rightarrow SO^{2-}$) thereby suggesting anaerobic respiration of sulfur, and incomplete nitrate reduction genes ($NO_3^- \rightarrow N_2O$). These metabolic capabilities describe SUP05 as a sulfide-driven partial

denitrifier. These findings are strictly genomic, and therefore invite plausible scrutiny. Fortunately, the later proteomic analysis of Saanich Inlet's OMZ by Hawley et al. [125] further corroborated this perspective, while also emphasizing that sulfur reduction was also driving SUP05's carbon fixation. This narrative was confirmed with the work of Shah et al. [237], with SUP05's cultivation. Cultivation enabled on-the-bench (*in vitro*) manipulation and measurements of microbial activity, and demonstrated that sulfur was essential to SUP05's growth, that growth was increased through nitrate reduction ($NO_3^- \rightarrow NO_2^-$), and further increased in the presence of $NH_4^+$. The discovery of SUP05's reliance on sulfur reduction is important, because it is often abundance in OMZs' non-sulfidic zones. Shah et al. remarked that this suggests an ecological hypothesis: that SUP05 is somehow reliant on a sulfur oxidizer.

**Marinimicrobia (Marine Group A, MGA, or SAR406)**

The Marinimicrobia phylum is an important accomplice in OMZ denitrification. Following early observation is diversity surveys [96, 97], catalyzed reporter deposition fluorescence *in situ* hybridization and SSU rRNA analysis demonstrated [6] that Marinimicrobia is abundant in the NESAP OMZ, comprising $0.3 - 2.4\%$ of total bacterial sequences, and increases to as high as 11.0% under $O_2$ deficient conditions. Statistically significant Spearman correlation statistics show Marinimicrobia negatively correlating with $O_2$. In a later analysis [272] which combined NESAP and Saanich Inlet data sets, and provided genomic insights through the inspection of 46 fosmid libraries. Genomic information suggests a sulfur-based energy metabolism, particularly suggesting dissimilatory sulfur oxidation. Marinimicrobia has also been implicated in syntrophic reactions in methanogenic bioreactors [199].

**Planctomycetes**

The Planctomycetes phylum is known to harbour anammox, an integral process in oceanic nitrogen loss [73]. The existence of an anammox-harbouring clade of Planctomycetes was first identified in a bioreactor designed to remove ammonia from waste water [251]. Anammox abbreviates anaerobic ammonium oxidation, and converts ammonium to nitrogen ($NH_4^+ + NO_2^- \rightarrow N_2 + H_2O$), thereby making the nitrogen largely biologically inaccessible. Anammox has been implicated in the Black Sea [151] and Costa Rica [66] to account for between 28% [60] and 48% [189] of oceanic nitrogen loss.

**Thaurmarcaeota**

The Thaumarchaeota are an extremely abundant Archael phylum, making up as much as 20% of all Picoplankton (Plankton sized $0.2 - 2.0\mu$m) [141]. Thaumarchaeota conduct ammonia oxidation ($NH_4^+ \rightarrow NO_2^-$), though they likely use alternative energetic strategies under ammonia-poor conditions [214]. Ammonia oxidation is important because it makes fixed nitrogen accessible to denitrification.

**Nitrospira & Nitrospina**

Nitrate ($NO_3^-$) accounts for 88% of fixed marine nitrogen [116], and the only known biological nitrate-forming reaction is nitrite oxidation ($NO_2^- \rightarrow NO_3^-$) [170]. Nitrospina and Nitrospira are nitrite oxidizers. An enrichment culture of Nitrospira sampled from a sponge was observed to convert $NO_2^-$ to $NO_3^-$ [203]. An early Nitrospira metagenome sampled from an activated sludge enrichment culture corroborated with this result, encoding genetic capacity for nitrite oxidation, $CO_2$ fixation, and lacked classic defense mechanisms against oxidative stress [169]. While the lack of genes in a metagenome is notable, it is a poor argument for an authentic lack of genetic material. Though of different phyla, a cultivated Nitrospira genome [170] suggests participation in a similar nitrite oxidizing niche. Cultivation confirmed the lack genes encoding coping mechanisms oxidative stress, explaining the anaerobic nature of these organisms.

**SAR324 (Marine Group B)**

SAR324 are deltaproteobacteria common to the dark ocean, particularly in OMZs [41, 96, 271, 273]. Single-cell amplified genomes sampled from the Altantic ocean encoded the capacity for C1-metabolism, sulfur oxidation, and a particle associated life-stlye [253]. An ESOM binning [74] experiment concluded that SAR324 might also harbour nitrite reductase ($NO_2^- \rightarrow NO$) [241]. A later MetaBAT [140] binning experiment (CheckM statistics [207]: $> 96\%$ complete, 0% contamination) would conclude that SAR324 does not harbour nitrite reductase, but does participate in sulfur oxidation [123].

**Sulfurimonas gotlandica & Arcobacteraceae**

Sulfurimonas gotlandica is an epsilonproteobacteria with a cultured representative, strain GD1 [115], from the Baltic Sea OMZ. A sub-clade, Sulfurmonas GD17, is common to OMZs [113] and is phylogenetically similar to Arcobacteraceae. Epsilon proteobacteria are common to the Bastlic Sea, African shelf, and Black sea OMZs, respresenting up to 25% of all prokaryotic cells [38, 47, 113–115, 152, 156, 163]. *In vitro* experiments have shown Sulfurimonas gotlandica to facilitate sulfide-oxidizing ($S^{2-} \rightarrow SO_4^{2-}$) complete denitrification ($NO_3^- \rightarrow N_2$).

### 1.4.4 A conceptual model

In the work of Hawley et al. [125], a metaproteomic analysis of Saanich Inlet is conducted. Normalized spectral abundance factor (NSAF) values are used to describe how metabolic activity occurs in relation to the oxygen gradient. Whenever possible, protein expression is attributed to taxa. This information is digested into a conceptual model illustrated in Figure 1.7. The model highlights many of the taxa described in the previous section, subsection 1.4.3. A hypothesis, *SUP05 produces* $NH_4^+$, is illustrated via "?$NH_4^+$", and was later refuted with SUP05's cultivation and subsequent growth experiments [237]. It was shown that SUP05 cultures can grow by consuming $NH_4^+$. A nitrogen loss mechanism is clearly observed through attributing anammox behaviour to Planctomycetes, but the final denitrifying enyze *nosZ* ($N_2O \rightarrow N_2$) goes taxonomically unattributed despite being observed in the sulfidic zone. As this conceptual model is a description of entities and their interactions, this work considers it a network perspective of Saanich Inlet microbial ecology, and it will be revisited.

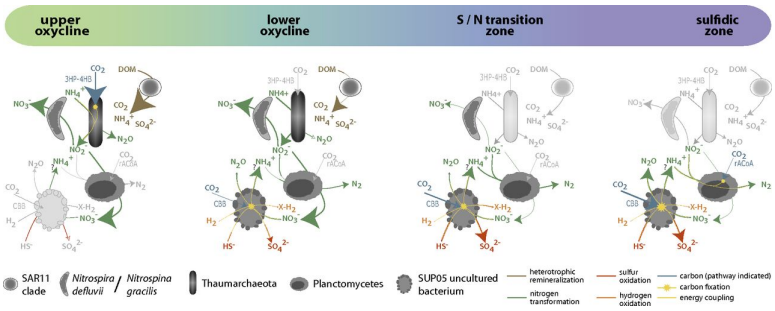



**Figure 1.7:** A conceptual model of Saanich Inlet denitrification. Image credit [125]

### 1.4.5 A differential model

In the work of Louca et al. [167], a multi-omic analysis integrates DNA, mRNA, and protein with environmental measurements of $O_2$, $NO_3^-$, and $H_2S$ concentrations to inform on Saanich Inlet's denitrifying community. The work is done with a philosophical twist, and considers genes independently of the taxa. The dynamics of gene abundance with the environment is modelled with a system of differential equations. The perspective, despite being independent of taxa, fits so well with the previous conceptual model that an illustration of the differential system Figure 1.8 includes taxa. The method is applied toward estimating variables like PDNO (partial denitrification to nitrous oxide) gene abundance through a method that is very different from statistical theory. Instead of maximum likelihood estimates, the differential system is run to steady state (it *converges*), then variables are measured. After cultivating SUP05, Shah et al. demonstrated the genomic potential for SUP05 to work with an sulfur reducer. This differential model supports that hypothesis, but also an interpretation of sulfide-driven denitrification. As this differential model is a description of entities and their interactions, this work considers it a network perspective of Saanich Inlet microbial ecology, and it will be revisited.

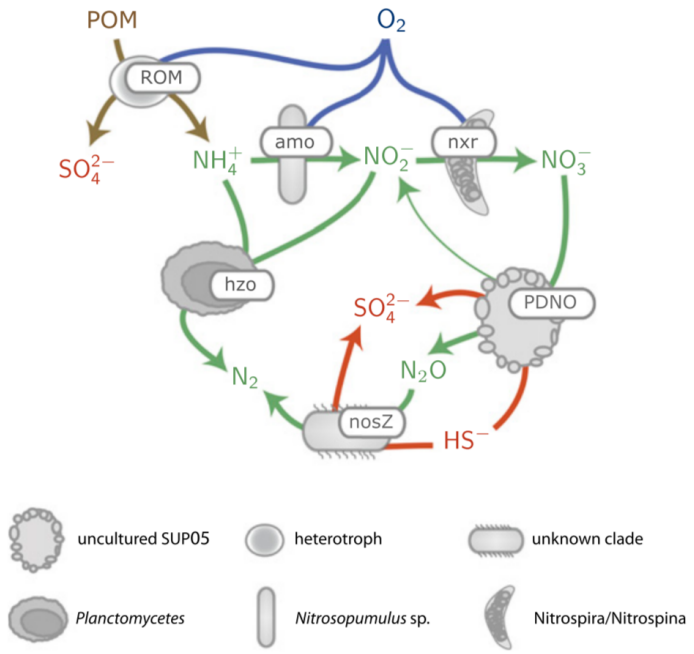



**Figure 1.8:** An illustration describing a differential model of Saanich Inlet denitrification. Image credit [167]

## 1.5 Math concepts

### 1.5.1 Set theory

A **set** $S$ is a collection of distinct **elements**. If $x$ is an element of set $S$, then it is written $x \in S$. If $y$ is not an element of $S$, then it is written $y \notin S$. For example, $S = \{1, 2, \{3\}\}$ is a set satisfying $1 \in S, 2 \in S, 3 \notin S$, and $\{3\} \in S$. If all elements in set $A$ are also elements in set $B$, then it is said that $A$ is a **subset** of $B$ and it written $A \subset B$, otherwise it is written $A \not\subset B$. An important set is the **empty set** $\varnothing = \{\}$, which is uniquely defined as the only set for which every element $x$ satisfies $x \notin \varnothing$. A set important to this work is the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ and the continuum of real numbers $\mathbb{R}$. Other useful sets are the non-negative integers (counts) $\mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \ldots\}$ and the positive reals $\mathbb{R}_{>0}$. Notice that $\mathbb{Z} \subset \mathbb{R}$, yet $\mathbb{R} \not\subset \mathbb{Z}$.

A **cartesian product** $\times$ is an operation on two sets, producing a third set. So the cartesian product of sets $A$ and $B$ is a third set $A \times B$. It is defined as the set of all pairs of elements between $A$ and $B$. Formally, for every $a \in A$ and $b \in B$, $(a, b) \in A \times B$. It is conventionally recognized that the cartesian product may be iterated so that for every $s_1 \in S_1, s_2 \in S_2, \ldots$, and $s_n \in S_n, (s_1, s_2, \ldots, s_n) \in S_1 \times S_2 \times \cdots \times S_n$. A cartesian product of a set $A$ with itself is written $A^2 = A \times A, A^3 = A \times A \times A$, and so on.

#### Dimensions

Most mathematical statements made in this work establish relationships between several variables. Hence the statements are multi-variate in nature. The cartesian product is used to construct

multi-dimensional spaces. For the purposes of this work, it is sufficient to define the **dimension** of a variable $x$ satisfying $x \in \mathbb{R}^p$ as $p$. So if $x \in \mathbb{R} = \mathbb{R}^1$, then $x$ is univariate. Also notice that since, $\mathbb{Z}_{\geq 0} \subset \mathbb{R}$, if $x \in \mathbb{Z}_{\geq 0}^{10}$, then $x$ has dimension 10. Further, if $x \in \mathbb{Z}_{\geq 0}^{a} \times \mathbb{R}_{>0}^{b}$, then the dimension of $x$ is $a + b$. For the purposes of this work, it is sufficient to define a **vector** as any list of numbers $x = (x_1, x_2, \ldots, x_p)$ satisfying $x \in \mathbb{R}^p$.

#### Hypercubes

A **hypercube** is a subset of a real space $\mathbb{R}^p$ which bounds a set of vectors per coordinate. Every hypercube can be bounded defined with two vectors $a = (a_1, a_2, \ldots, a_p)$ and $b = (b_1, b_2, \ldots, b_p)$, such that for each $j \in \{1, 2, \ldots, p\}$, $a_j \leq b_j$. A unique hypercube exists for every unique pair of $a$ and $b$, defined as $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_p, b_p]$, where each $[a_j, b_j]$ is a **bounded**, **segment** subset of $\mathbb{R}$ from $a_j$ to $b_j$. Notice that in $\mathbb{R}$ every hypercube is a segment, in $\mathbb{R}^2$ a square, and in $\mathbb{R}^3$ a cube. Visual analogues break-down in higher dimensions, only leaving analytic understanding.

### 1.5.2 Probability

A more thorough description of these concepts can be found in the work of Klenke [146].

Probability theory is founded on the observation that random trials can be imagined as an experiment with a set of possible outcomes $\Omega$, defined the **sample space**. For example, the sample space for a single dice-roll is $\{1, 2, 3, 4, 5, 6\}$. Though it is sometimes sufficient to only describe probabilities of individual outcomes, probability theory instead describes the likelihood of outcome **events** $\mathcal{F}$, which is a special set (*sigma algebra*) of subsets of $\Omega$. So for any $x \in \mathcal{F}$, it is also true that $x \subset \Omega$. For example, an outcome event for a single dice roll is $\{\text{dice} < 3\} = \{1, 2\}$. A **probability measure** $\mathbb{P}$ is a function from mapping elements of $\mathcal{F}$ to $[0, 1]$ (written as $\mathbb{P} : \mathcal{F} \to [0, 1]$). Even though $\mathbb{P}$ is a function of sets (example: $\mathbb{P}[\{\text{dice} < 3\}] = 2/6 = 1/3$), it is conventional to drop the braces ("{" and "}") (example: $\mathbb{P}[\text{dice} < 3]$). A probability measure must satisfy the probability axioms, which are not stated here.

#### Probability models

Mathematical models are essential to this work. For this work, it is sufficient to define a **mathematical model** as any set of constraints on variables. For example, two variables $(X, Y) = (O_2 \text{ concentration}, \text{Nitrospina abundance})$ could be constrained through an equation $Y = a + bX$ and real-valued constants $(a, b) \in \mathbb{R}^2$ (read as $(a, b)$ is *in* the set of two-dimensional real-valued numbers). Constraints can also be much more abstract, leveraging powerful theory to be both more realistic and holding meaningful implications. For example, variable values many be assumed to be drawn from a simple random sample. Simple random sampling holds powerful implications through statistical theory.

#### Random variables

Probability theory provides a framework for describing the likelihood of observing variable values. Variables defined with a description of likely values are **random variables**. The admission of imprecise values inherent to random variables is useful to mathematical models, because

non-random equations are so easily wrong in application. For example, no mircobial abundance will be exactly known through $O_2$ concentrations, so the earlier equation $Y = a + bX$ is wrong. However through admission of an error term $\varepsilon$, the model $Y = a + bX + \varepsilon$ cannot be wrong. Of course, in order for the $Y = a + bX + \varepsilon$ model to be useful, an understanding of $\varepsilon$ is motivated.

**Probability distributions**

For a random variable $X \in \mathbb{R}$ (read as $X$ is in the set of single-dimensional real-valued numbers) and a non-random variable $x \in \mathbb{R}$, the probability of observable values is defined through the **univariate distribution function** $F_X(x) = \mathbb{P}[X \leq x]$. Through defining the probability of $X$ falling below or equal to $x$, the probability of all other events (technically *Borel* sets) is implicitly defined. Multi-dimensional random variables $\mathbf{X} = (X_1, X_2, \ldots, X_p) \in \mathbb{R}^p$ have a **multivariate distribution function** $F_{\mathbf{X}}(\mathbf{x}) = F_{X_1,X_2,\ldots,X_p}(x_1, x_2, \ldots, x_p) = \mathbb{P}[\cap_{j=1}^p \{X_j \leq x_j\}]$ (where an **intersection** of sets $\cap_{j=1}^p A_j$ is the set of all elements $a$ satisfying $a \in A_j$ for every $j \in \{1, 2, \ldots, p\}$). It is conventional to write $\mathbb{P}[\cap_{j=1}^p \{X_j \leq x_j\}]$ as $\mathbb{P}[X_1 \leq x_1, X_2 \leq x_2, \ldots, X_p \leq x_p]$.

As the $Y = a + bX + \varepsilon$ example demonstrates, the addition of a random error term $\varepsilon$ allows a model to be valid, but potentially less useful. If it were true that $\varepsilon = 0$ constantly, the model would be perfectly predictive. Practically, $\varepsilon$ has a non-trivial distribution function which describes precisely how wrong the $Y = a + bX + 0$ model is. By understanding the distribution of $\varepsilon$, one may understand how useful a model (such as $Y = a + bX + 0$) is predictively. Two popular concepts for describing a distribution function are location and dispersion.

**Measures of location**

Define **location** as a near-typical value for a random variable. The **expected value** $\mathbb{E}X = \int_{\mathbb{R}} x dF_X(x)$ of a random variable $X$ is a common description of location, often written as $\mu$. Expected values can be understood through averaging repeated trials. The Strong Law of Large Numbers states that for a random vector $\mathbf{X} \in \mathbb{R}^n$ satisfying $F_{\mathbf{X}}(\mathbf{x}) = F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)F_{X_1}(x_2)\cdots F_{X_1}(x_n)$ (independence) and $\mathbb{E}X_1 = \mu \in \bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$ (well-defined expected value), then their average $n^{-1}\sum_{i=1}^n X_i$ converges to $\mu$ with probability one. This is written as $\mathbb{P}[\lim_{n\to\infty} n^{-1}\sum_{i=1}^n X_i = \mu] = 1$ or as $n^{-1}\sum_{i=1}^n X_i \to_{a.s.} \mu$, where *a.s.* stands for *almost surely*, equivalent to *with probability one*. Expected values are useful and theoretically well-developed, but because they are defined through an integral, not every random variable can have one. A common description of location that always exists for a random variable $X$ is the **median**, defined as any value $m$ satisfying $\mathbb{P}[X \leq m] \leq 1/2$ and $\mathbb{P}[X \geq m] \geq 1/2$. A random variable's median $m$ is *near* its mean $\mu$, as constrained by $|\mu - m| = |\mathbb{E}(X - m)| \leq \mathbb{E}|X - m| \leq \mathbb{E}|X - \mu| \leq \sqrt{\mathbb{E}(X - \mu)^2} = \sigma$ [172], where $\sigma$ is a non-random constant which is defined through integrals (and therefore $\sigma$ may not be well-defined).

**Measures of dispersion**

Define **dispersion** as a description of how near a random variable's realized values are to its location. A common description of dispersion is the **variance**, defined as $\mathrm{Var}X = \mathbb{E}(X - \mathbb{E}X)^2$ and often written as $\sigma^2$. Drawing inspiration from the law of large numbers, it is clear that $\sigma^2$ might be estimated through the average square deviation from the mean. Since variance is

defined through the expectation operator $\mathbb{E}$ which is in turn defined through an integral, not all random variables have a well-defined variance. A common description of dispersion that always exists is the **median absolute deviation**, defined as $\text{median}(X - \text{median}(X))$. This work utilizes some models which permit inexistence of an expected value or variance, so an understanding of median-based descriptors of location and dispersion is motivated.

#### Measures of dependence

A central focus of this work is the description of relationships between variables, many of which are modelled by random variables. A popular description of statistical dependence between random variables $X$ and $Y$ is the **covariance**, defined as $\text{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$. The covariance describes the strength of linear relationships between variables. To see this, let random variables $X$ and $Y$ satisfy the following linear relationship $Y = a + bX$, then $\text{Cov}(X, Y) = \text{Cov}(X, a + bX) = b\text{Var}X$, so their covariance is related to the slope of the line between them. A concept which betters describes linear relationships by building on covariance is **correlation**, defined as $\text{Cor}(X, Y) = \text{Cov}(X, Y)/\sqrt{(\text{Var}X)(\text{Var}Y)}$. For the $Y = a + bX$ example, $\text{Cor}(X, Y) = b\text{Var}X/\sqrt{(b\text{Var}X)^2} = 1$, since $Y$ is a deterministic linear function of $X$. This work frequently estimates covariance structures, motivating a description of a random vector $\mathbf{X}$'s **Covariance matrix**, defined as the symmetric matrix $\Sigma = \text{Cov}(\mathbf{X}) = \text{Cov}((X_1, X_2, \ldots, X_p)^T)$ satisfying $\Sigma_{ij} = \text{Cov}(X_i, X_j)$, where $\Sigma_{ij}$ is the element of $\Sigma$ on the $i^{th}$ row and $j^{th}$ column.

### 1.5.3 Gaussian models

In this work a recurring distribution function is the **Gaussian** or **Normal** distribution. In a single dimension, the Gaussian distribution function with mean $\mu$ and variance $\sigma^2$ is $\Phi(x; \mu, \sigma^2) = \int_{-\infty}^{x} \phi(z; \mu, \sigma^2)dz$, where $\phi$ is the Gaussian **density** function, $\phi(x; \mu, \sigma^2) = (2\pi)^{-1/2}\sigma^{-1}\exp[-(x - \mu)^2/(2\sigma^2)]$. The **standard** Gaussian is a Gaussian distribution with $(\mu, \sigma^2) = (0, 1)$. The Gaussian density function $\phi$ is shaped like a bell, with tails that drop to zero exponentially fast as $x$ is taken away from zero (see Figure 1.9). The bell-shape indicates that a Gaussian random variable $X$ likely takes values near $\mu$, and the exponentially small tails indicates that extreme values quickly become so unlikely as to effectively never occur. For example, a standard Gaussian-distributed random variable is only expected to fall below $-10$ only once in every $10^{23}$ trials.

The multivariate generalization of the Gaussian is the **multivariate Gaussian** or **multivariate Normal** distribution of a random vector $\mathbf{X} \in \mathbb{R}^p$ with mean $\mu = (\mu_1.\mu_2, \ldots, \mu_p)$ and covariance matrix $\Sigma$, defined as $\Phi(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \Phi(x_1, x_2, \ldots, x_p; \boldsymbol{\mu}, \Sigma) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_p}\phi(t_1, t_2, \ldots, t_p; \boldsymbol{\mu}, \Sigma)dt_1dt_2\cdots dt_p$, where $\phi$ is the multivaraite Guassian density function, and $\phi(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-p/2}\det(\Sigma)^{-1/2}\exp[-(\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})/2]$. Samples drawn from a multivariate Gaussian distribution are likely to fall within an elliptical region, as demonstrated in Figure 1.10. If a random vector $\mathbf{X} \in \mathbb{R}^p$ is multivariate Gaussian distributed with parameters $(\boldsymbol{\mu}, \Sigma)$, then it is written $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$. If $p = 1$, it may be written $X \sim N(\mu, \Sigma)$, in which case $X$ has a univarite Gaussian distribution.

#### Eigenvectors

Eigenvectors are analytic tools which are abstract and powerful. Their abstraction makes them widely applicable, but also difficult to understand. By adding assumptions in an example, their

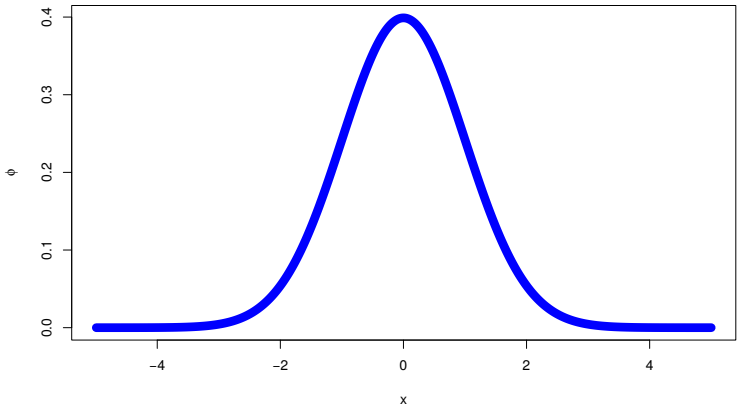



**Figure 1.9:** The standard Gaussian curve

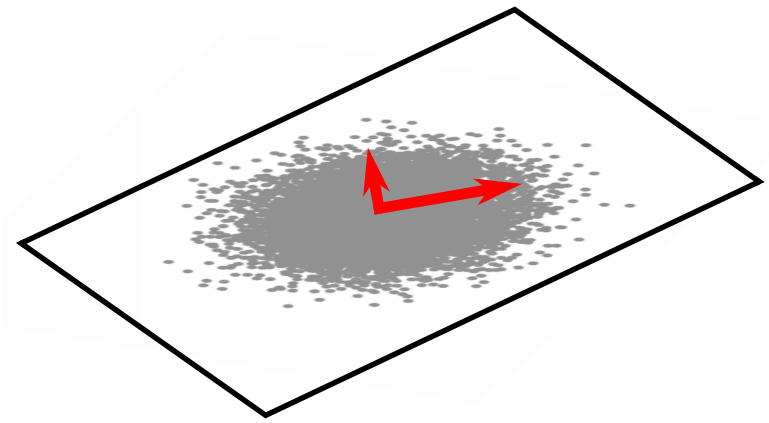


**Figure 1.10:** Bivariate Guassian simulation. Arrows are covariance matrix eigenvectors.

usefulness is made clearer. The analytic definition of an eigenvector $\mathbf{v}$ is that for some matrix $M$, there is a scalar (*eigenvalue*) $\lambda \neq 0$ such that $M\mathbf{v} = \lambda\mathbf{v}$. Eigenvalues are the magnitudes of their corresponding eigenvectors. So a matrix may have eigenvector and eigen value pairs. For example, a covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ has $p$ eigenvector-value pairs $\{(\mathbf{v}_j, \lambda_j)\}_{j=1}^{p}$ such that $\Sigma = \lambda_1\mathbf{v}_1\mathbf{v}_1^T + \lambda_2\mathbf{v}_2\mathbf{v}_2^T + \cdots \lambda_p\mathbf{v}_p\mathbf{v}_p^T$, and $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$. It is possible to approximate $\Sigma$ by truncating it's sum of eigenvectors, so $\Sigma \approx \lambda_1\mathbf{v}_1\mathbf{v}_1^T + \lambda_2\mathbf{v}_2\mathbf{v}_2^T$. This heuristic is useful for interpreting Factor analysis models and Principal Component Analyses (PCA).

For a multivariate Gaussian distribution, the eigenvectors of its covariance matrix are the axes of greatest variation. This is shown as arrows in Figure 1.10. Principal Componenet Analysis (PCA) [210] estimation uses this fact to reduce the dimensionality of data. To derive a PCA representation for some data, calculate the data's correlation matrix, then project the data into the lower-dimensional space provided by the first few (usually 2) eigenvectors.

#### Factor analysis models

**Factor Analysis models** are a constrained variant of the multivariate Gaussian, initially developed for Psychology by Spearman. The Factor Analysis model is achieved by constraining a multivariate Gaussian's covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ so the $\Sigma = LL^T + \Psi$ where $L \in \mathbb{R}^{p\times m}$ where $m \leq p$ and $\Psi \in \mathbb{R}^{p\times p}$ is a diagonal matrix. An important difference is that the Factor Analysis model has fewer free parameters than the general multivariate Gaussian. Instead of $\Sigma$ having $p^2$ parameters, $LL^T + \Psi$ has $p(m+1)$ parameters. If $m$ is small, then the model has linearly many ($O(p)$) parameters instead of quadratically many ($O(p^2)$). Parameter reductions or contraints are important from a statistical perspective, providing an avenue toward lower-variance estimates. A second important difference between the Factor Analysis model and the general Gaussian is that a random vector $X$ with mean 0 and covariance $LL^T + \Psi$ follow the Factor Analysis model is implity equivalent to $\mathbf{X} = L\mathbf{F} + \Psi^{1/2}\mathbf{E}$ where $\mathbf{L} \in \mathbb{R}^{p\times m}, \mathbf{E} \in \mathbb{R}^p$ are $m+1$ independent standard multivariate Guassian vectors (having distribution function $\Phi(\cdot; 0, I)$). The implicit decomposition of any Factor Analysis model into standard multivariate Gaussians provides an interpretation similar to PCA.

#### Gaussian mixture models

A **Gaussian Mixture model** is used to model data clusters. It is best understood through a stochastic process. If a random vector $\mathbf{X}$ follows a Gaussian Mixture model, then it is equivalent to $\mathbf{X} = \sum_{j=1}^{m} 1_{Y=j}\mathbf{Z}_j$, where $Y \in \{1, 2, \ldots, m\}$ and $\mathbf{Z}_j \sim N_p(\boldsymbol{\mu}_j, \Sigma_j)$ independently. The indicator function $1_A$ equals 1 when $A$ is true and zero otherwise. Effectively $Y$ acts like a switch, causing $\mathbf{X} \sim N_p(\boldsymbol{\mu}_j, \Sigma_j)$ with probability $\mathbb{P}[Y = j]$. Many samples drawn from the same Gaussian Mixture model will fall into $m$ multivariate Gaussians, as demonstrated in Figure 1.11.

#### Covariance decomposition

In chapter 3, a method for precise correlation network estimation is described. However the immediately resulting network has many nodes and ends, and is difficult to interpret. It is tempting to disregard large sections of the network for the sake of simplification, but that would destroy information. Viewing correlation networks more generally as covariance matrices, a

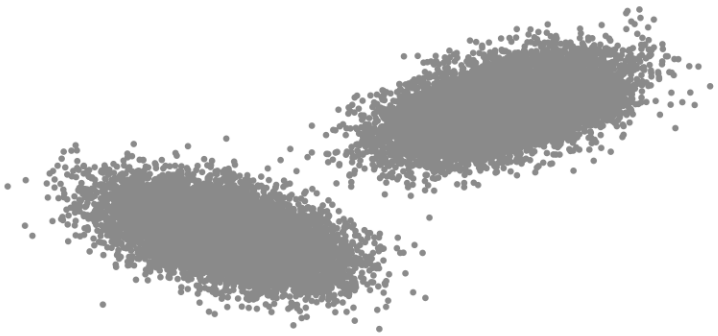


**Figure 1.11:** Data simulated from a Gaussian mixure model

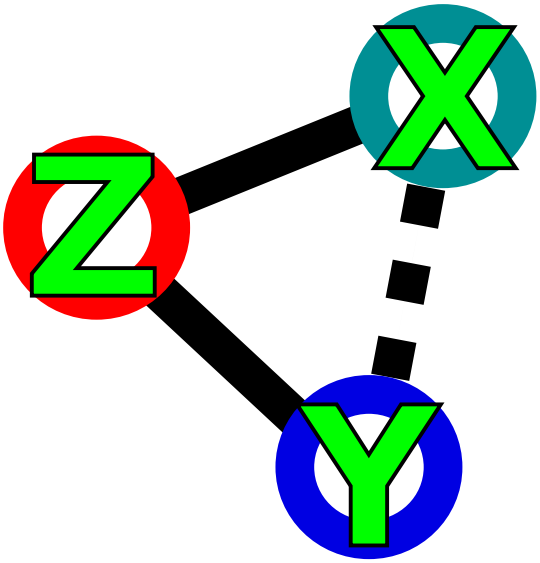



**Figure 1.12:** Illustration of two correlations between $(X, Z)$ and $(Y, Z)$, generating a partial correlation $(X, Y|Z)$

solution becomes available. Partial covariance decomposition [16–18] can break a covariance matrix into a simple and complex part without destroying information. Partial correlations can be imagined as an underlying correlation structure which may generate tertiary correlations. This work uses *tertiary correlation* to refer to a correlation for variables $(X, Y)$, such that there is a variable $Z$ and $\text{Cor}[X, Y] \neq 0$ and $\text{Cor}[X, Y|Z] =_{a.s.} 0$. The tertiary correlation is no less real, in that the correlation of $(Y, Z)$ from Figure 1.12 is non-zero. However, the tertiary correlations are less foundational, since their multivariate dependencies are entirely constructed by other variables (see Baba [16], theorem 2.1.1).

The following section will develop a $\sigma_{\mathbf{Y}} = \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \Sigma$ decomposition where $\sigma_{\mathbf{Y}}$ is the initial covariance matrix and $\Sigma$ is the simple, partial covariance matrix, and $\sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}}$ is a complicated residual structure which may not be studied at all. In this way, covariation between the dimensions of $\mathbf{Y}$ can be studied while controlling for the covariances with and between the dimensions of $\mathbf{X}$.

The following well-known theory builds the argument.

Let $\mathbf{Y} \in \mathbb{R}^p, \mathbf{X} \in \mathbb{R}^q$ be random variables with $\mathbb{E}\mathbf{X} = \mathbb{E}\mathbf{Y} = \mathbf{0}$.
Define $\sigma_{\mathbf{Y},\mathbf{X}} = [\mathrm{Cov}(Y_i, X_j)]_{p\times q}$ (asymmetric covariance matrix).
Define $\sigma_{\mathbf{Y}} = \sigma_{\mathbf{Y},\mathbf{Y}}$ (covariance matrix).
Define $\beta_{\mathbf{Y},\mathbf{X}} = \sigma_{\mathbf{Y},\mathbf{X}}\sigma_{\mathbf{X}}^{-1}$ (best linear predictor weights).
Define $\sigma_{\mathbf{Y},\mathbf{X}\cdot\mathbf{Z}} = \sigma_{(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{Z}}\mathbf{Z}),(\mathbf{X}-\beta_{\mathbf{X},\mathbf{Z}}\mathbf{Z})}$ (asymmetric partial covariance).
Define $\sigma_{\mathbf{Y}\cdot\mathbf{Z}} = \sigma_{\mathbf{Y},\mathbf{Y}\cdot\mathbf{Z}}$ (partial covariance).
Define $(\mathbf{Y}\cdot\mathbf{X}) = \mathbf{Y} - \beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}$ (partial variable).
Let $L_{\mathbf{X}} = \boldsymbol{\alpha} + B\mathbf{X}$ for some non-random $\boldsymbol{\alpha} \in \mathbb{R}^p, B \in \mathbb{R}^{p\times q}$.

**Theorem 1.** $\sigma_{(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}),L_{\mathbf{X}}} = 0$.

*Proof.* of Theorem 1.

$\sigma_{(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}),L_{\mathbf{X}}} = B\sigma_{(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}),\mathbf{X}} = B\left(\sigma_{\mathbf{Y},\mathbf{X}} + \sigma_{-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X},\mathbf{X}}\right)$
$= B\left(\sigma_{\mathbf{Y},\mathbf{X}} - \beta_{\mathbf{Y},\mathbf{X}}\sigma_{\mathbf{X},\mathbf{X}}\right) = B\left(\sigma_{\mathbf{Y},\mathbf{X}} - \sigma_{\mathbf{Y},\mathbf{X}}\sigma_{\mathbf{X},\mathbf{X}}^{-1}\sigma_{\mathbf{X},\mathbf{X}}\right) = 0$

□

**Theorem 2.** *trace* $\left(\sigma_{\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}}\right) \leq$ *trace* $\left(\sigma_{\mathbf{Y}-L_{\mathbf{X}}}\right)$, *where trace*$(\sigma) = \sum_{j=1}^{p}\sigma_{jj}$.

*Proof.* of Theorem 2.

$\sigma_{\mathbf{Y}-L_{\mathbf{X}}} = \sigma_{(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X})+(\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}-L_{\mathbf{X}})}$
$= \sigma_{\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \sigma_{(\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}-L_{\mathbf{X}}),(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X})} + \sigma_{(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}),(\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}-L_{\mathbf{X}})} + \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}-L_{\mathbf{X}}}$
$= \sigma_{\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + 0 + \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}-L_{\mathbf{X}}}$ ; ($\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X} - L_{\mathbf{X}}$ linear in $\mathbf{X}$, apply Theorem 1)
$\Rightarrow \mathrm{trace}\left(\sigma_{\mathbf{Y}-L_{\mathbf{X}}}\right) \geq \mathrm{trace}\left(\sigma_{\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + 0\right)$ ; ($\sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}-L_{\mathbf{X}}}$ positive definite)

□

In the case of multivariate Gaussian regression, define a model $\mathbf{Y} = \beta\mathbf{X} + \Sigma^{1/2}\boldsymbol{\varepsilon}$ imposed on the data $(Y, X)$, where $\beta \in \mathbb{R}^{p\times q}$ (non-random) and $\boldsymbol{\varepsilon} \sim N_p(0, I)$. Because this is multivariate Gaussian regression, $\beta$ is the best possible linear predictor of $Y$ given $X$ (*best* minimizes $\sum_{j=1}^{p}\mathbb{E}\left([\mathbf{Y} - \beta\mathbf{X}]_j^2\right)$), which is $\beta = \beta_{\mathbf{Y},\mathbf{X}}$ as given by Theorem 2. Decomposition of the covariance follows.

$\sigma_{\mathbf{Y}} = \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}+\Sigma^{1/2}\boldsymbol{\varepsilon}} = \sigma_{(\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X})+(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X})}$
$= \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \sigma_{(\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}),(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X})} + \sigma_{(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}),(\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X})} + \sigma_{\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}}$
$= \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + 0 + \sigma_{\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}}$ ; (apply Theorem 1)
$= \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \sigma_{\mathbf{Y}\cdot\mathbf{X}}$ ; (note the partial covariance)
$= \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \sigma_{\Sigma^{1/2}\boldsymbol{\varepsilon}} = \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \Sigma^{1/2}\sigma_{\boldsymbol{\varepsilon}}\Sigma^{1/2} = \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \Sigma^{1/2}I\Sigma^{1/2} = \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \Sigma$

The covariance has been decomposed into a regressor and partial covariance part $\sigma_{\mathbf{Y}} = \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \sigma_{\mathbf{Y}\cdot\mathbf{X}}$. Because this is made possible by best linear predictors, decomposition is further possible via ultimately resulting in successive partial decomposition. Define some subset $J \subsetneqq \{1, 2, \ldots, p\}$ and let $\mathbf{Z} = [\mathbf{Y} - \beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}]_J$ be sub-vector (projections) of our residuals, and decomposition may proceed as follows.

$\sigma_{\mathbf{Y}} = \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \sigma_{\mathbf{Y}\cdot\mathbf{X}} = \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \sigma_{\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} = \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \sigma_{\beta_{(\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}})\mathbf{X},\mathbf{Z}}\mathbf{Z}} + \sigma_{\mathbf{Y}-\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}\cdot\mathbf{Z}}$
$= \sigma_{\beta_{\mathbf{Y},\mathbf{X}}\mathbf{X}} + \sigma_{\beta_{(\mathbf{Y}\cdot\mathbf{X}),\mathbf{Z}}\mathbf{Z}} + \sigma_{(\mathbf{Y}\cdot\mathbf{X})\cdot\mathbf{Z}} = \sigma_{(\mathbf{Y}\cdot\mathbf{X})-\mathbf{Y}} + \sigma_{((\mathbf{Y}\cdot\mathbf{X})\cdot\mathbf{Z})-(\mathbf{Y}\cdot\mathbf{X})} + \sigma_{(\mathbf{Y}\cdot\mathbf{X})\cdot\mathbf{Z}}$

Thus the initial covariance can be decomposed into a series of covariance matrices, $\sigma_{\mathbf{Y}} = \sigma_{(\mathbf{Y}\cdot\mathbf{X})-\mathbf{Y}} + \sigma_{((\mathbf{Y}\cdot\mathbf{X})\cdot\mathbf{Z})-(\mathbf{Y}\cdot\mathbf{X})} + \sigma_{(\mathbf{Y}\cdot\mathbf{X})\cdot\mathbf{Z}}$. This elegant theoretical construction is a close approximation to what occurs in more complex regression models. Roughly, interpretation can be compartmentalized into separate covariance matrices $\sigma_{\mathbf{Y}} \approx \sigma_{[\text{Chemical concentrations}]} + \sigma_{[\text{Non-target taxa}]} + \sigma_{[\text{Target nitrogen cycling taxa}]}$.

## 1.6 Statistics concepts

Further reading on material in this section can be found in texts by Casella and Berger [52] and Murphy [193].

Statistics and Machine Learning methods are applied toward a variety of problems including modelling, descriptive summaries, and visual communication, but this work primarily applies all such techniques toward automated decision making. Surveys of large amounts of data is made possible through automated decision making. In certain situations decisions are known to be right or wrong, and protocols can be evaluated. All major contributions of this thesis are evaluated in such a way, primarily through the use of precision-recall curves (see section 1.6.5).

### 1.6.1 Estimation

#### Maximum likelihood estimates

Statistics employs probability models. Recall the example model from section 1.5.2, $Y = a + bX + \varepsilon$, where $(Y, X, \varepsilon)$ is a random vector and $(a, b)$ is a constant vector. Assume further that $Y$ represents Nitrospina abundance, and $X$ represent $O_2$ concentration. Repeated observations of $(Y, X)$ hint at the likely values of $(a, b)$ and the distribution of $\varepsilon$. Statistical theory provides methods for constructing a likelihood function $f_{Y,X}(y, x; a, b)$ which can describe how likely certain parameters ($(a, b)$ in this case), given observed vectors $(y, x)$. In more general and conventional notation, a likelihood function for a random vector $\mathbf{X}$ with parameter vector $\boldsymbol{\theta}$ is written $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta})$. Given a sample vector of observations $\mathbf{X} = \mathbf{x}$, the value $\hat{\boldsymbol{\theta}}$ which maximizes $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta})$ is the maximum likelihood estimate (MLE). MLEs are extremely popular in statistics, because they attain asymptotically minimal variance and are eventually unbiased for sufficiently large sample sizes. The MLE's asymptotically minimal variance is said to make it **efficient**.

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta})$$

#### Overfit

Small estimator variance is important, because it allows the model to realistically represent the variational structure of the data it is modelling. In Bioinformatics, a common cause of estimator variance is having too few data and too many parameters. For example if $\dim(\boldsymbol{\theta}) = p > n = \dim(\mathbf{x})$, then the model will likely **overfit** the data. An overfit model is able to describe the eccentricities of the data, but has lost the *big-picture* signal that it was meant to capture. Overfit is known to reduce models' predictive capacity.

### 1.6.2 Regression

The term **regression** is used in a variety of ways, but generally refers to any statistical process for measuring relationships among variables. *Measurement* can be interpretted as *deciding the existence of relationships* or as *modelling relationships*, and in application both perspectives are often satisfied simultaneously. Through historical precident, regression analyses tend to model relationships between pairs of variables $(\mathbf{Y}, X)$ linearly. So if $\boldsymbol{\beta}$ is a non-random parameter vector and $X$ a matrix, the model constrains $\mathbf{Y} = X\boldsymbol{\beta}$. However, $(\mathbf{Y}, X)$ are often defined in very flexible ways. For example, $\mathbf{Y}$ might be an internal model parameter, or $X$ might be a matrix of transformed variables. Effectively, a linear constraint is often applied toward modelling some very non-linear relationships.

#### Univariate regression

This work defines **univariate regression** as the regressing of a location variable (see section 1.5.2) for a univariate random variable $Y \in \mathbb{R}^1$ against some other variable $X$. For example, least squares regression can be applied to model $Y$ as a Gaussian distributed random variable with conditional expectation $\mathbb{E}[Y|X] = X\boldsymbol{\beta}$. For generalization beyond univariate Gaussian distributions, Generalized Linear Models (GLMs) [181] model the conditional expectation of $Y$ through a **link function** $g$, so $\mathbb{E}[Y|X] = g(X\boldsymbol{\beta})$. It is very common for models to have additional parameters beyond $\boldsymbol{\beta}$, such as variance. A popular GLM applied in Microbial Ecology [168, 185, 228] is the Negative Binomial, satisfying the following.

$$Y \in \mathbb{Z}_{\geq 0} \; ; \; \mathbb{P}[Y = y] = \binom{y + \mu^2/(\sigma^2 - \mu) - 1}{y} \left(\frac{\sigma^2 - \mu}{\sigma^2}\right)^y \left(\frac{\mu}{\sigma^2}\right)^{\mu^2/(\sigma^2 - \mu)} \; ; \; \sigma^2 > \mu \; ; \; \mu = e^{X\beta}$$

The Negative Binomial can be further specified as the NEGBIN P [45, 46, 110], taking P in $\mathbb{Z}_{>0}$, and $\text{Var}[Y] = \sigma^2 = \mu + \mu^P/\nu$. The NEGBIN 2 or NB2 configuration is used in this work.

#### Multivariate regression

This work defines **multivariate regression** as the regressing of location variables for a random vector $\mathbf{Y} \in \mathbb{R}^p$, while simultaneously modelling a covariance structure for the dimensions of $\mathbf{Y}$. A popular example is **multivariate Gaussian regression**, where $\mathbf{Y}$ follows a multivariate Gaussian distribution with covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ and expected values $\mathbb{E}[Y_j|X] = X\boldsymbol{\beta}_j$ for $j \in \{1, 2, \ldots, p\}$. Notice that the vectors $\boldsymbol{\beta}_j$ can be arranged into a matrix $\beta$. Multivariate regression should be contrasted with **univariate regression surveys**, where a vector of expected values $\mathbb{E}[Y_j|X]$ is estimated but a covariance structure is not. Univariate regression surveys are popular in Bioinformatics [168, 227, 228].

It is important to note that multivariate regression models tend to have more parameters than univariate regression models. For example, if $X \in \mathbb{R}^{n \times q}$ describes $n = 100$ samples for $q = 4$ variables, then a univariate Gaussian regression model has $q + 1 = 5$ parameters ($+1$ for $\sigma^2$). In a univariate regression survey of $Y \in \mathbb{R}^{n \times p}$ with $p = 50$, then there are $O(p) = p(q + 1)$ parameters. A multivariate Gaussian regression model for $\mathbf{Y}$ has every parameter in the univariate regression survey plus $p(p-1)/2$ parameters in $\Sigma$, so it has $O(p^2) = p((p-1)/2 + q + 1)$ parameters.

### 1.6.3 Model selection

#### AIC

Statistical models can be poorly selected. Sometimes data exhibit behaviour that a model necessarily describes as unlikely. Consider the Gaussian distribution for example, the likelihood of observed values drops off expoentially fast with distance from the expected value. If a Gaussian distribution is fit to data with extreme values, it will likely have an inflated variance estimate, thereby distorting later inference. In this way, poorly selected statistical models can lie. Application of Information Criterion (AIC) [2] can alleviate this issue, by perhaps comparing the Gaussian model to a Student-t distribution, which has the ability to model extreme values. AIC is a statistic which can be used to ordinate the quality of model fits. To use it, competing models are fit to the same data set, evaluated with the AIC statistic, and then the model with the lowest AIC values likely suffers the least *information loss*. AIC statistics must be constrasted with **Goodness-of-fit** tests, which are used to accept or reject the hypothesis that the data follow a particular distribution. AIC ordinates models (which might not fit at all), whereas a goodness-of-fit test simply describes binary acceptance and rejection.

$$\mathrm{AIC}_f = 2k - 2\log f_{\mathbf{X}}(\mathbf{x}; \hat{\boldsymbol{\theta}}) \; ; \; k = \dim(\boldsymbol{\theta})$$

#### Regularization

**Regularization** is a model constraint used to reduce overfit. Regularization methods are well-developed in univariate linear model selection (best exemplified by the covariance test in L1 regularization [164]), but also in many other applications [28]. Some regularization methods can work by constraining an optimization problem [131], while others work by reducing the number of variables [247]. Regularization for high-dimensional covariance matrix estimation has recently matured [218] but is actually quite narrow in its applicable scope. Most methods are only useful when applied to multivariate Gaussian distributions, which the multivariate counts of SSU rRNA data do not satisfy. This work utilizes a regularized high-dimensional covariance matrix and uses *copula* to interface the requisite multivariate Gaussian distribution with univariate count distributions.

A useful way to imagine how regularization methods work is to realize that models with too many parameters fit to too few data will overfit. The many moving parts of the model allow it to conform too well to the data, thereby exhibiting its eccentricities, and ignoring larger, more important signals in the data. Regularization methods always work by constraining the model in some way, thereby making it less flexible. A less flexible model may no-longer conform too well, and can describe more general themes in the data. In this work, the covariance matrix is constrained by requiring it to equate with a factor analysis models' covariance structure, $\Sigma = LL^T + \Psi$. Recall that for $p$ dimensions (taxa), $\Sigma$ has $O(p^2)$ correlations, but $LL^T + \Psi$ has only $O(p)$ parameters. By reducing the number of *moving parts* (parameters) in the covariance strucutre, it can better highlight the greater themes in data's covariance structure.

### 1.6.4 Copula & marginals

A statistical **copula** is a mathematical modeller's tool, which allows for great conveniences. It allows the modeller to consider the multivariate structure as a separate *back end* to the model, while also allowing largely independent selection of univariate distributions in the models' *front end*. Copula are used like theoretical glue, sticking the multivariate and univariate components together through a deterministic transform. The convenience of selecting univariate and multivariate structures independently allows for an otherwise unprecedented breadth of models to choose from. For example, this work needs special univariate distributions to allow for sufficient goodness-of-fit (see section 1.7 for why), yet also needs a special regularizing covariance (multivariate) structure. In this way, copula is a necessary solution.

Mathematically defined, a copula $C$ is a multivariate cumulative distribution function $F_C$ with uniform $U(0,1)$ marginal distributions.

Multivariate normal (*Gaussian*) distributions have already been described as convenient by allowing strategic employment of partial correlations and regularization. Unfortunately, the data studied in chapter 3 follow a multivariate count distribution (discrete) which is clearly not Gaussian (continuous), and no transform will ever map between them [49]. Fortunately, the marginal (univariate parts) and copula (multivariate parts) parts are guaranteed to be, in a way (see Theorem 3), separable through Sklar's theorem [245]. Theorem 3 is written similarly as found in Joe [138].

**Theorem 3. (Sklar's theorem):** *For a random vector* $\mathbf{Y}$ *with multivariate cumulative distribution function* $\mathbb{P}[Y_1 \leq y_1, Y_2 \leq y_n, \ldots, Y_p \leq y_p]$ *and univariate marginal distributions* $F_j(y_j) = \mathbb{P}[Y_j \leq y_j]$*, an associated copula function* $C : [0,1]^p \to [0,1]$ *satisfies the following.*

$$F_{\mathbf{Y}}(\mathbf{y}) = C(F_1(y_1), F_2(y_2), \ldots, F_p(y_p))$$

**(a)** *If* $F_{\mathbf{Y}}$ *is continuous and has quantile functions* $F_1^{-1}, F_2^{-1}, \ldots, F_p^{-1}$*, then* $C$ *is uniquely defined as follows.*

$$C(\mathbf{u}) = F_{\mathbf{Y}}(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_p^{-1}(u_p))$$

**(b)** *If* $F_{\mathbf{Y}}$ *is discrete (or partly discrete), then* $C$ *is only unique on the following set.*

$$Range(F_1) \times Range(F_2) \times \cdots \times Range(F_p)$$

From the modeller's perspective the copula function $C$ is the multivariate sturcture, and each **marginal** distribution describes the random behaviour of each univariate distribution. For example, in chapter 3, the copula function is a Gaussian copula constrained to have a factor analysis model's covariance structure, while each marginal distribution describes the random behaviour of SSU rRNA counts.

### 1.6.5 Hypothesis testing & classification

This work applies statistical methods toward automated decision making, and uses **hypothesis testing** to automate that framework. Hypothesis testing can be thought of as a comparison of two models, the null and alternative model. Imagining the alternative mathematical model as a set

of assumptions $\{A_1, A_2, \ldots, A_n\}$, the null model is merely the same model with an additional constraining assumption: the null hypothesis $H_0$. So the null model can be imagined as a slightly larger set of assumptions $\{A_1, A_2, \ldots, A_n, H_0\}$. The null hypothesis $H_0$ is not assumed like each $A_i$, it is a postulate to be tested. Statistical theory allows the models to be compared for likelihood, and should the null model be deemed sufficiently unlikely, the null model $\{A_1, A_2, \ldots, A_n, H_0\}$ must be deemed untrue and at least one of its assumptions must be untrue. Good statistical methodology requires checking each assumption $A_i$, so it is likely that the only untrue assumption is the postulated null hypothesis $H_0$. In this way, the null hypothesis may be rejected.

This abstract framework well-developed both theoretically and applicably. For example, the extremely broad category of regression software exist with a standardized interface allowing for automated hypothesis testing. Of course, different data and questions require different software. While a little bit complicated, it truly couldn't be easier. If it was easier, abstraction would be reduced so that the tools would be too narrow in scope. Null hypotheses are often formulated as $H_0 : \boldsymbol{\theta} = \mathbf{0}$, which is sufficiently general for many problems.

Plenty of theoretical tools exist for designing hypothesis testing software. Ultimately, these tests must somehow digest the comparison of models into a single test statistic $t(\mathbf{x}; \boldsymbol{\theta})$ and threshold $\tau$. Tests will be formulated as rules, such as *if* $t(\mathbf{x}; \boldsymbol{\theta}) < \tau$, *reject* $H_0$, *otherwise do not.* In order to design such tests, powerful statistical theory is used to derive the distribution of $t(\mathbf{X}; \boldsymbol{\theta})$ for random data $\mathbf{X}$ while assuming $H_0$. Thereby unlikely values of $t(\mathbf{X}; \boldsymbol{\theta})$ can be identified, leading to rejection of the null $H_0$. Because the null rejecting machinery is derived assuming the null, hypothesis tests have derivable rejection rates $\alpha$. Formally, $\alpha = \mathbb{P}[\text{reject } H_0 \mid H_0 \text{ is true}]$.

The common assumptions of independent sampling and asymptotically large samples (math is derived taking the sample size large, $n \to \infty$) allow common and powerful theory to be employed. For example, Wilks' theorem [270] says that, under the null hypothesis $H_0$, the test statistic $t(\mathbf{X}; \hat{\boldsymbol{\theta}}) = -2 \log \lambda = -2 \log(f_{\mathbf{X}|H_0}(\mathbf{X}; \hat{\boldsymbol{\theta}}_0) / f_{\mathbf{X}|H_0^c}(X; \hat{\boldsymbol{\theta}}_c))$ is asymptotically chi-square distributed with $\dim(\boldsymbol{\theta})$ degrees of freedom. So $-2 \log \lambda \sim \chi^2_{\dim(\boldsymbol{\theta})}$. Similarly, it is known that MLEs take on (multivariate) Gaussian distributions for sufficiently large sample sizes, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim N_{\dim(\boldsymbol{\theta})}(0, I_{\boldsymbol{\theta}}^{-1})$, where $I_{\boldsymbol{\theta}}$ is the Fisher information matrix of $\mathbf{X}$ following parameter $\boldsymbol{\theta}$.

In this work, hypothesis testing is used to survey for non-zero correlations. So many null hypotheses of the form $H_0 : \rho_{ij} = 0$ are tested. Because the greatest challenges in chapter 3 are due having many more parameters than data, it is unclear if asymptotic assumptions ($n \to \infty$) are valid, thus invalidating previously described statistical theory. Instead, the boostrap [83] is used, which is robust to lower sample sizes but requires more computational power to employ.

Employing powerful hypothesis testing machinery toward automated decision making leads to **classification**. In this work, it is common to classify or assert that $\theta = 0$ if a statistical test cannot reject $H_0 : \theta = 0$, and to classify $\theta \neq 0$ if a statistical test rejects $H_0 : \theta = 0$. In this way, a classifying machine $C$ is defined to take random data and produce decisions. So for some data set $X$, $C(X) = \{\theta \neq 0\}$, and for another $C(X') = \{\theta = 0\}$.

#### Precision-recall exchanges

Roughly, *precision* is the probability that one is right when making a claim, and *recall* is the probability that one makes a claim when it is right. Notice that it is often easy to claim little and be right (low recall, high precision), and easy to claim a lot and often be wrong (high recall, low

precision), but usually challenging to precisely claim enough to make an argument. Precision and recall tend to exchange with each other. If one imagines truth as the harvest of scientific endeavour, then recall is our yield, and precision is the efficiency or quality of our harvest. Under this perspective, one might imagine claims to be the unit of our harvest. Usually, an experimental or deductive logical process is required to harvest a single claim, but in modern contexts machines are participating in the harvest. The result is that machines are producing many, similarly formed claims. For example, our correlation software makes claims of the form *taxa A is correlated with taxa B*, and our binning software claims *contig A belongs to taxa B*. In bioinformatics, machine-produced claims tends to be part of a larger argument, and may be used in conjunction with many other machine-produced claims and a few human-produced claims. Inevitably, an entire argument is formed and some degree of confidence is required in the machine-produced claims. Certain arguments require high-quality machine-produced claims. If these machines must be right, they must be precise.

The primary challenge [267] in section 1.3 and an important part of the concerns described in subsection 1.2.4 [55, 175, 240] are due to poor precision-recall exchanges. Precise correlation would empower microbial ecologists to make further and more confident claims pertaining to individual taxa interactions and which taxa have which functions. Objectively, the contributions delivered by the work are measured through precision-recall exchanges.

Because precision and recall are central topics to this work, a formal interpretation of classifiers will now be developed. Classifiers make claims which may be phrased as *object or phenomena A has attribute B*. For example, one might attribute correlation to a pair of taxa. Define a set of possible attributions $\mathbb{T}$ and a set of objects or phenomena $\mathbb{S}$ which may be given an attribution. It is known that special pairs $(X, y) \in \mathbb{S} \times \mathbb{T}$ satisfy $X \in y$, and with the goal to confidently map $\mathbb{S}$ to $\mathbb{T}$ accordingly, a function $C : \mathbb{S} \to \mathbb{T}$ is needed. Define *precision* as $\mathbb{P}[X \in y | C(X) = y]$. Define *recall* as $\mathbb{P}[C(X) = y | X \in y]$.

The efficient precision-recall exchanges in this work come at a cost that is strategically paid when possible. In the two problems covered by this work, only one (binning) leverages additional data, the other (correlation) is simply a reinterpretation of existing resources. Efficient precision-recall exchanges do not come for free. It is generally true that the strategies employed in this work achieve efficient exchange by increasing the maximum attainable precision and lowering the maximum attainable recall. So by increasing our bound on precision, a bound on recall is lowered. Fortunately, despite a constrained improvement, gains in precision per fixed recall are attained.

**Precision myths**

Bioinformatic pipelines involving many machines $(C_1, C_2, \ldots)$ might suggest the opportunity to overcome imprecision through consensus. Such opportunities do exist, certainly under repeated trials. It is tempting to imagine that combining multiple strategies necessarily improves precision (consider ensemble approaches [87] in light of Weiss et al. [267]). However, it is simply not a ubiquitous truth. Effectively, the question is $\mathbb{P}[X \in y | C_1(X) = y, C_2(X) = y] \geq \mathbb{P}[X \in y | C_1(X) = y]$? The answer is contextual of course, however the following mathematical constraint applies generally.

**Result 1.** *If* $\mathbb{P}[X \in y | C_2(X) = y] \leq \mathbb{P}[X \in y | C_1(X) = y]\mathbb{P}[C_1(X) = y | C_2(X) = y]$, *then* $\mathbb{P}[X \in y | C_1(X) = y, C_2(X) = y] \leq \mathbb{P}[X \in y | C_1(X) = y]$.

*Further require* $\mathbb{P}[C_1(X) = y, C_2(X) = y] > 0$.

*Proof.* Let $A_1 = \{C_1(X) = y\}, A_2 = \{C_2(X) = y\}, B = \{X \in y\}$.
Then our hypothesis is $\mathbb{P}[B|A_2] \leq \mathbb{P}[B|A_1]\mathbb{P}[A_1|A_2]$
$\Leftrightarrow \mathbb{P}[B \cap A_2]/\mathbb{P}[A_2] \leq \mathbb{P}[B|A_1]\mathbb{P}[A_1 \cap A_2]/\mathbb{P}[A_2]$
$\Leftrightarrow \mathbb{P}[B \cap A_2] \leq \mathbb{P}[B|A_1]\mathbb{P}[A_1 \cap A_2]$
$\Leftrightarrow \mathbb{P}[B|A_1] \geq \frac{\mathbb{P}[B \cap A_2]}{\mathbb{P}[A_1 \cap A_2]} \geq \frac{\mathbb{P}[B \cap A_1 \cap A_2]}{\mathbb{P}[A_1 \cap A_2]} = \mathbb{P}[B|A_1 \cap A_2]$

□

This result can be generalized to many machines $C_i$ as follows.

**Corollary 1.** *If* $\mathbb{P}[X \in y|C_2(X) = y, C_3(X) = y, \ldots]$
$\leq \mathbb{P}[X \in y|C_1(X) = y]\mathbb{P}[C_1(X) = y|C_2(X) = y, C_3(X) = y, \ldots]$,
*then* $\mathbb{P}[X \in y|C_1(X) = y, C_2(X) = y, C_3(X) = y, \ldots] \leq \mathbb{P}[X \in y|C_1(X) = y]$.
*Also require* $\mathbb{P}[C_1(X) = y, C_2(X) = y, C_3(X) = y, \ldots] > 0$.

*Proof.* Take $A_2 = \{C_2(X) = y, C_3(X) = y, \ldots\}$ in the proof of Result 1. □

Despite its simplicity, there are meaningful interpretations of Result 1 useful to bioinformaticians. First, notice that if the precision of $C_2$ is sufficiently lower than that of $C_1$, the best precision is obtained by not employing $C_2$. Second, even if agreement is high between machines ($\mathbb{P}[C_1(X) = y|C_2(X) = y] \cong 1$), if both are sufficiently imprecise then their combined effect is no more precise. Consensus does not bestow truth. A theoretically relaxed interpretation is that no amount of agreement between low quality machines matters unless at least one of them is shown to be precise.

**Debunked myths:**

1. Imprecise methods can aid precise methods.
2. Consensus is as good as truth.

A third interpretation of Result 1 is that if both $C_1$ and $C_2$ have similar precisions but they disagree ($\mathbb{P}[C_1(X) = y|C_2(X) = y] \leq 1$), the best precision is again obtained by only employing one machine. In the sub-case where precisions are low and they disagree, the only lesson is that better methods are needed. Further, a precision must be bounded if disagreement exists, because only one machine can be right. This merely demonstrates that consensus is not sufficient for precision, but it is necessary.

In constructing data-driven arguments, the reliance on automated decision making motivates objective evaluations of methods. Without some guard against false interpretation, it is easy to make mistakes. For example, in section 1.4.3, where two binning experiments arrived at contradicting conclusions. The ESOM experiment concluded the SAR 324 might harbour nitrite reductase, whereas a MetaBAT-generated bin concluded otherwise, and was also evaluated with CheckM. Automating decision making can lead to mistakes without objective evaluations. This work advocates for the use of precision-recall curves, because the statistics evaluate the exact desired behaviour. Precision is the rate of correct attribution amongst all attributions. Recall is the rate of attributions amongst correct attributions. Losing objectivity might just let one lose touch with reality.

Further motivation for precision in data-driven argument is developed in Appendix A, where data-driven arguments are analogized to Hidden Markov Models.

## 1.7 Computation

Further reading on topics in this section can be found in work by Isaacson and Keller [136] and Boyd and Vandenberghe [34].

Bioinformatics is very much a computational science. This presents certain mathematical challenges, particularly in algorithmic theory and numerical analysis. **Numerical analysis** is the algorithmic theory of numerical approximation. There are many situations where best methods for calculating on paper and in silicon are very different. Calculation of eigenvectors has examples of this. Constrained, non-linear optimization with Lagrange multipliers does as well. Many popular statistical softwares employ the same linear algebra code libraries–such as LAPACK [11].

Many challenges arise from the use **floating point** representation, where computers use predefined quantities of memory to store a number $x$ is represented through the stored values $(s, m)$ as $x = s2^m$. The finite-memory constraint means that **roundoff** errors can occur, with $x$ being rounded to one of $\{0, \infty, -\infty\}$. **Truncation** errors are due to insufficient digits being stored in $s$. **Numerical stability** is a desired property, describing algorithms which can accurately approximate their target functions. A numerical algorithm lacks stability of it produces large errors.

It might be tempting to imagine that all computational problems can be overcome with sufficient hardware resources, but such conveniences often cannot achieve what better algorithms can. For example, truncation errors might be made slightly less frequent by using 64 bit representation of floating point values (**double precision**) rather than the typical 32 bit representation (**single precision**), but the problem is often overcome entirely by computing on log-scale (see subsection 3.2.3 for a particular example of this). In the case of GPU computing (see subsection 1.7.3), avoiding the use of double precision representation can even make software run faster.

### 1.7.1 Numerical calculus

Numerical approximation of functions is often motivated through the use of calculus. Derivatives and integrals often cause numerical approximation to become necessary. Heavily studied functions (consider hypercubic Gaussian integrals [101], beta ratios [39, 75], and the student-t distribution [129]) often have piecewise approximating solutions, broken into efficient iterations or polynomial approximations.

#### Derivatives

Derivatives are a common computational goal. While most derivatives are manually calculable in theory, many calculations are made pragmatically feasible through computer aid (consider backpropagation as an example [233]). It is also common for derivatives to be **numerically computed** (calculated through computer aid) for convenience (this is common in non-linear programming, see subsection 1.7.2). A common numerical approximation is through the stencil. Stencils are expensive to compute, because they require re-evaluation of the numerically differentiated function. Examples of two and five-point stencils follow.

$$g'(x) \approx \left[g(x+h) - g(x-h)\right] / (2h)$$

$$g'(x) \approx \left[-g(x+2h) + 8g(x+h) + 0 - 8g(x-h) + g(x-2h)\right] / (12h)$$

$$g''(x) \approx \left[-g(x+2h) + 16g(x+h) - 30g(x) - 16g(x-h) - g(x-2h)\right] / (12h^2)$$

**Integrals**

With large amounts of probability theory applied in this work, many examples of integrals have already been motivated. It is common for integrals to require numerical approximation, though sometimes models are selected because of their analytically solved integrals. For low-dimensional integrals ($\int_{\mathbb{R}^p} g(\mathbf{x}) dF(\mathbf{x})$, where $p$ is small, usually $\leq 3$), **numerical quadrature** routines can be used to approximate integrals. A software library for univariate integral approximation is QUADPACK [217]. For higher-dimensional integrals **MC-integration** (Monte Carlo-integration) is more feasible, though sufficiently many dimensions will make any computational approach infeasible. MC-integrals use random number generation to approximate integrals, and thus are ideal for computing expected values. This is shown as follows.

$$\mathbb{E}g(\mathbf{X}) = \int g(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}) =_{a.s.} \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} g(\mathbf{X}_i) \; ; \; \mathbf{X}_i \sim F_{\mathbf{X}}$$

If simulation of random variables $X_i$ from distribution function $F_X$ is fast (it often is), this method approximates $\mathbb{E}f(X)$. Note that the number of required iterates can be predicted through Chebychev's formula as follows.

$$\mathbb{P}\left[\left|n^{-1}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g(X)\right| \geq k\sqrt{n^{-1}\mathrm{Var}\left[f(X)\right]}\right] \leq k^{-2}$$

### 1.7.2 Non-linear programs

Non-linear programs are optimization problems, where a function $g(x)$ is maximized (or minimized, equivalent through maximization of $-g(x)$). Sometimes solutions are analytically known. For example, the parabola $g(x) = -3x^2 + 2x - 5$ can be analytically optimized through setting its derivative to zero $g'(x) - 6x + 2 = 0 \Rightarrow x = -3^{-1}$. In chapter 2, optimization is achieved through the EM-algorithm [69], iterating analytically solvable sub-optimizations. Optimization programs are commonly described as follows. The program is non-linear of $g$ is non-linear.

$$\text{maximize } g(x) \text{ subject to } h(x) \leq 0 \text{ and } x \in X$$

Iterative solutions are common and work by iteration some maximizing function $m(x_n) = x_{n+1}$, and the iterative series must start at some $x_0$. Choice of initial guess $x_0$ is often very important. In some situations, the iterative component is only good for tuning less-significant digits of $x$. The process by which an initial guess is produced is often at least as important as the optimizing procedure. Crafting an initialization algorithm requires some domain knowledge. There are no general solutions. This work's initialization algorithms are always a series of estimators, each

taking as input the output of the previous. The initial estimators are robust, numerically stable and inaccurate. Later estimators are accurate and more delicate.

#### Quasi-Newton methods

A well-developed category of numerical algorithms is for optimizing convex, non-linear, differentiable $g$. **Quasi-Newton methods** take advantage of how easily parabolic systems are optimized, and work by iteratively approximating $g$ with a parabolic system and optimizing it. A multivariate Taylor series approximation of $g$ expanded about $\mathbf{x}_0$ is the following.

$$g(\mathbf{x}) \approx g(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^T \nabla g(\mathbf{x}_0)^T + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla^2 g(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

$\nabla g$ is the gradient vector of $g$, so $[\nabla g(\mathbf{x})]_i = \frac{\partial}{\partial x_i} g(\mathbf{x})$. $\nabla^2 g$ is the Hessian matrix of $f$, so $[\nabla^2 g(\mathbf{x})]_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} g(\mathbf{x})$. Assuming $g$ is (at least locally) convex ensures that $\nabla^2 g$ is positive symmetric definite (for every $\mathbf{x}$ vector, $\mathbf{x}^T \nabla^2 g \mathbf{x} > 0$, and $[\nabla^2 g]^T = \nabla^2 g$). Systems satisfying these conditions have a breadth of well-developed numerical algorithms available to them [42, 92, 105, 145, 205, 206, 213, 238, 282]. These solutions take advantage of the fact that the parabolic equiations are solved by differentiating and setting to zero. The derivative of the previous Taylor series approximation is as follows.

$$\nabla g(\mathbf{x}) \approx \nabla g(\mathbf{x}_0) + \nabla^2 g(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

For MLE computation in chapter 3, having many dimensions makes $\nabla^2 g$ computation expensive. For this situation, the BFGS or L-BFGS algorithms are ideal, since $\nabla^2 g$ is either calculated implicitly [108] or not stored at all.

### 1.7.3 GPU supercomputing

Graphics processing units (GPUs) are hardware modules (see Figure 1.13) which were developed to accelerate video rendering, a process which requires calculating values for individual pixels on a screen in parallel. With so many pixels, the problem is massively parallel. With a consistently large enough consumer demand for video acceleration, specialized hardware (the GPU) has been developed precisely for this task. Conveniently, certain non-video algorithms have a similar parallelization structure and can be accelerated with the hardware. GPUs are somewhat overspecified, but capable of massive speed-ups in the right compute scenarios. For example in chapter 3, a simulation study utilized a GeForce GTX 980 Ti GPU capable of up to 5360 GFLOPS (giga floating point operations per second) and an AMD FX-8320 Eight-core CPU capable of about 40 GFLOPs, suggesting a potential increase of over 100 times. Of course, the specialization of a GPU means that this speed-up is only available for certain problems. Modern GPU technology is employed in tasks that require only the most powerful computing solutions, including the world-class Go-playing AlphaGo [243], self-driving cars [8], and high quality sequence alignment with the Smith-Waterman algorithm [173]. All GPU-accelerated software in this work is written in CUDA [200].

**Figure 1.13:** The NVidia GeForce 980 Ti GPU

### Warp divergence

GPUs' special form of immense parallelization is achieved by making trade-offs at the hardware level which exclude many numerical problems from acceleration. Accommodating for these specializations makes algorithmic design and programming more difficult for GPU software. Software is written for blocks of threads (see Figure 1.14), which need to execute very similar instruction sets. This is because GPUs achieve their acceleration by executing instructions with multiprocessors, which are capable of issuing the same instruction to several threads at a time. Threads are processed in groups called **warps**.

**Warp divergence** occurs when threads in a warp receive different instructions. Because the multiprocessor can only issue the same command, it process the different instruction sets in serial–one after another. Small amounts of warp divergence are common and still allow for accelleration, but if most threads follow different execution paths, then the entire task would be slower than run in serial on a CPU.

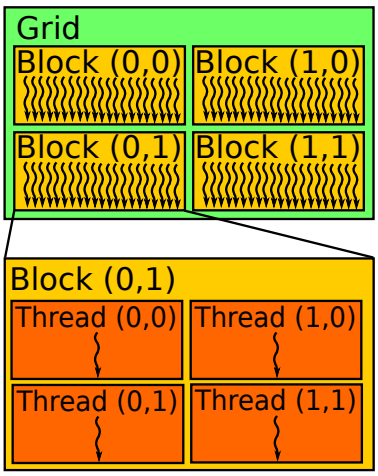



**Figure 1.14:** Blocks of GPU threads

## 1.8 Deliverables

This work makes contributions toward two bioinformatic problems relevant to Microbial ecology: (1) metagenomic binning, and (2) SSU rRNA correlation, described each in their own chapter. Objective evaluation of these contributions is made through precision-recall curve estimation. Contributed methods are applied toward better understanding the microbial ecology of denitrification in Saanich Inlet. While both chapters will describe inferences that won't be revisited, there is a unifying narrative interpreting results from a network perspective and arguing for the metabolic cooperation between two taxa, SUP05 and Marinimicrobia.

In chapter 2, a method for precise metagenomic binning is made possible by recruiting genetic material to SAGs. The work contributes to a much-desired understanding of how binners succeed and fail [5, 225, 240] by calculating precision-recall curves for different binning strategies under different conditions and from different perspectives. It turns out that metagenomic binning shares a similar error profile to that of bootstrapped phyolegenetic trees, informing on how these tools might best be applied in the future. The work also contributes to the correlation network perspective of the Saanich Inlet denitrifying community by better describing the genomic potential of different taxa in a variety of conditions.

In chapter 3 a solution for overcoming the poor precision-recall exchanges of modern SSU rRNA correlation surveys in Microbial Ecology [267] is described, objectively validated with precision-recall curves, and applied toward the Saanich Inlet denitrifying community. It is discovered that a form of covariance matrix regularization is important for good precision-recall exchanges in SSU rRNA correlation surveys. Findings in the chapter bolster metabolic syntrophy arguments with correlative support.

In chapter 4, findings from both chapters are summarized and brought together. The technical conclusions of either chapter are important for Bioinformatics, though largely independent. Findings are largely unified through the comparison of several network perspectives of the Saanich Inlet denitrifying community. A future direction is imagined for correlation networks in Microbial Ecology.

# Chapter 2

# Metagenomic binning

Metagenomic binning methods are often used to extract genomes from metagenomes and are being applied on vast scales to draw important conclusions on a variety of topics. Viewing these methods as machine learning tools, approaches include a variety of strategies and sometimes very little training data. Wide acceptance of binning products hinges on concerns including false taxonomic assignment. Precision, the probability of correct assignment, is therefore desired to be high. Using single-cell amplified genomes (SAGs) as a reference to guide metagenomic binning is shown here to make significant increases in precision without excessive loss of recall. This work introduces a binning software released as SAGEX (SAG EXtrapolator). Several binning strategies are compared and motivated, and illustrate that precision tends to increase when describing higher-ranked clades, or when more training data is used. This work suggests that evolving genomic standards require binning products to be published with their binning strategy and should encourage the use of precise techniques when possible.

## 2.1 Introduction

High-throughput sequencing technologies are rapidly uncovering the incredible genomic potential of microbial life. Despite their integral roles in mediating matter and energy transformations [85], the vast majority of microorganisms remain uncultivated [58, 89, 197]. Metagenomics, the cultivation-independent sequencing of nucleotides from an environment, is illuminating this uncultivated "dark matter" opening a taxonomic and functional window into the networks driving microbial community metabolism in natural and engineered ecosystems [225]. A combination of *metagenomic binning* and assembly methods have been popular for attributing function to taxonomy in these cultivation-independent contexts [93, 234, 261]. While pragmatic, communicating the degree of genome completion and quality is important [175], has been advocated for in genomic standards [55], and modern definitions continue to be proposed [207]. The issue that metagenomic binning might falsely assign genomic sequences to a genome is an acknowledged concern [240] that can be dealt with through a combination of understanding how binning mistakes are made and how to reduce their occurrence. Here, this work suggets that precise binning is a way to reduce such mistakes.

An alternative cultivation-independent method for obtaining genomic information from an environment is single-cell sequencing. SAGs are theoretically capable of unambiguously linking whole genomes to a taxonomy at the pinnacle level of resolution, a single cell. Unfortunately, caveats of this technology persist such as incomplete recovery of the host's genome [29]. This work demonstrates reductions in false assignment made possible through SAG-guided binning. The greatest value SAGs offer for binning is that they provide a most relevant source of training data for classifying binners, thus allowing more precise binning of novel taxa. Because SAG

sequencing tends to result in incomplete genomes, binning is further motivated by increasing genome recovery. Despite training classifying binners with SAGs, this work discovers that species-level precision is not attainable with neither the tool presented herein nor other popular binning software. Therefore, bins should be considered mixed within a narrow taxonomic range (i.e., species and genus). This taxonomic range is shown to ultimately narrow as binner precision increases.

A concern for reduced false assignment is best satisfied with precision–the probability that each assignment (or classification) is done correctly. Precise metagenomic binning provides confidence in individual assignments of genomic sequences to microbial taxa. While not necessarily essential for all applications, this precision is valuable in the later use of binning to make inferences in an oxygen minimum zone's microbial food web [125]. Because certain metabolic capabilities can be deeply meaningful toward a microbiome's ability to transform the chemical composition of its environment [154], confidence in results is key.

### 2.1.1 Definitions

Metagenomic binning has been reinterpreted over time [174, 180, 234], so generally applicable definitions will now be provided. It is generally true to define a metagenomic *bin* as a collection of related genomic sequences from a metagenome, which may or not be given a taxonomic label. *Metagenomic binning* or *binning* is a process resulting in the creation of bins. A *binner* is a tool which assists or automates binning. A binner which produces bins with a taxonomic label is a *classifying* binner, otherwise it is a *clustering* binner. Examples of binning software are described in subsection 1.2.4.

The prediction that binners are evaluated with synthetic data [180] has held true [7, 102, 182, 186, 276]. A popular and effective method for estimating classification has been to pull genomes from public databases and use them as known-label data. Examples include the Integrated Microbial Genomes (IMG) [176] system and the National Center for Biotechnology Information's (NCBI) [268] RefSeq [254]. Leveraging 368 SAGs, this work takes advantage of a novel opportunity to more rigorously scrutinize a variety of binners. This work evaluates binners with synthetic metagenomes composed of SAGs. Where known-label data derived from public databases may have been sampled from a variety of experiments and surveys, the SAGs were sampled from the same location within a radius of 200m and on the same day. The important difference between the SAGs and disparate public database entries is that the SAGs have evolved together. For example, Horizontal Gene Transfer (HGT) has had the opportunity to cause these genomes to share DNA. Because binners leverage forms of DNA dissimilarity to separate bins, the increased potential for similarity between SAGs introduces more pitfalls during binner evaluation that would have been otherwise missed.

So that binner behaviour may be exactly described, a formal description of binning and phylogeny is defined. These formalisms allow ambiguity to be avoided in descriptions of phylogeny, precision, genome-bin differentiation, and binner error behaviour. Define $\mathbb{S}$ as the set of all *sequences* to be binned, so $\mathbb{S} = \cup_{n=0}^{\infty}\{A, T, C, G\}^n$, where exponentiation stands for a Cartesian product. Define a *taxonomic label* as a set of sequences, for example $\{\text{gammaproteobacteria}\} \subset \{\text{bacteria}\}$. Define $\mathbb{T}$ as *the set of all taxonomic labels*, so $\mathbb{T} = \mathcal{P}(\mathbb{S})$. Define a *phylogenetic tree* $V$ as satisfying $V \subset \mathbb{T}$ such that it defines a rooted-tree graph $G_V = (V, E)$ where $E = \{(a, b) : (a, b) \in$

$V^2, a \subset b\}$. Any $v \in V$ can be viewed as a *clade* or *population genome*. For a given phylogenetic tree, define all leaf nodes of $G_V$ as *genomes*. This work recognizes all genomes to be the nucleotide sequences of individual organisms. Under this framework, a metagenomic bin $b$ satisfies $b \in \mathbb{T}$ and there exists $v \in V$ such that $b \subset v$. These definitions allow bins to have genetic material from several genomes and not necessarily have a recognized definition within a phylogenetic tree–a perspective shared other work [240]. Note that some works use the term *genome* in cases this work defines specifically as either a genome or a bin [133], but fortunately this is an example of ambiguity relievable by precise language.

To maintain objectivity, evaluation metrics are defined according to section 1.6.5 using the formal definitions of binning and phylogenetics. Precision-recall exchanges are the fundamental metrics of comparison. Define a classifying binner $C$ as any function from $\mathbb{S}$ to $\mathbb{T}$. Then precision is the probability that the sequence $x$ is in the taxonomic set $y$ given that the classifier $C$ has assigned it to $y$, and is formally written as $\mathbb{P}[x \in y | C(x) = y], x \in \mathbb{S}, y \in \mathbb{T}$. Precision is a desirable metric, because precise binners produce more phylogenetically homogeneous bins, and would alleviate previously acknowledged binning concerns [240]. Recall is probability that a sequence is correctly assigned to a taxonomic group given that it belongs to that group, formally written as $\mathbb{P}[C(x) = y | x \in y]$. It is important to track both precision and recall because they tend to exchange for one another, and a very precise method without recall is useless.

### 2.1.2 Software

Because a classifying binner designed to train specifically with SAGs doesn't exist yet, this work also introduces SAGEX (SAG EXtrapolator). The software is written entirely in C/C++ and the only library necessary for compilation is POSIX threads. It accepts two `.fasta` files as inputs (see an example in Figure 2.1), a SAG to train with and a metagenome to recruit from, and outputs a bin `.fasta` file containing sequences from the metagenome which should be related to the SAG (Figure 2.2). Define the concatenation of the training SAG and the SAGEX output as an *extrapolated SAG* or *extrapolation*. Notice that extrapolations are bins as well. This work uses SAGEX with assembled inputs. While tested on SAGs, note that it is possible to run SAGEX on any `.fasta` file of several contigs, thus allowing the use of Illumina Tru-Seq synthetic long reads [161], fosmids, or database entries. SAGEX is available from github.com/hallamlab/sagex. Ultimately this work describes a wide variety of precisions between binning strategies, of which SAGEX is a high-precision binning method, and demonstrates precise binning in the context of microbial ecology of an oxygen minimum zone.

Genome completion has historically been evaluated with marker genes [5, 225, 240]. A modern tool for finding and summarizing marker genes within a collection of nucleotide sequences is CheckM [207]. Both estimates of bin completion and binning precision are relevant in a series on genomic standards [90, 91].

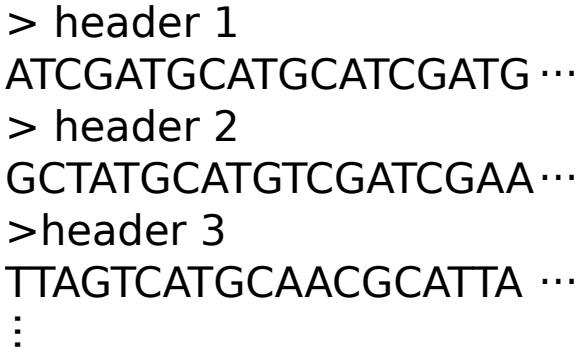



**Figure 2.1:** The first six lines of an example `.fasta` file

## 2.2 Methods

### 2.2.1 SAGs

Single cell sampling and sequencing was carried out as described in [230] and all SAGs of the SUP05 lineage in this study originate from [230]. In brief, samples for single cell sequencing were collected August 2010 at station S3 in Saanich Inlet at 100, 150 and 185m. Samples were collected directly into 10 ml glass vials and 1ml was transferred into 143 $\mu l$ of 48% beatine and frozen on dry ice. Samples were stored on dry ice in the field and transferred to -80$^o$C freezer for storage until thawing at Bigelow Laboratories Single Cell Genomics Center (SCGC; https://scgc.bigelow.org) for sorting by flow cytometry. Cells were sorted and underwent initial round of multiple displacement amplification (MDA) and PCR amplification of the small subunit rRNA (SSU rRNA) gene as described in [249, 253]. Taxonomy of single cells was determined by direct sequencing of amplicons of the SSU rRNA gene (see section D.5). Clean SSU rRNA sequences were clustered at 99% identity and representative sequences aligned with LAST [144] against GreenGenes [71] database (2010) to obtain taxonomy. Taxonomic assignments and efficiency of MDA were used to choose cells for additional MDA and subsequent sequencing. Chosen cells were sequenced as described in Roux et al. [230] at the Genome Sciences Centre, Vancouver BC, Canada and assembled using SPAdes [21].

Early SAG sequencing techniques have known contamination issues [257], also detected by CheckM (Table 2.2). Because taxonomically consistent SAGs form the argumentative foundation of this work, sequences with potential for contamination were removed. It is likely that many non-contaminant sequences were removed in the decontamination process. This policy is favourable over working with taxonomically inconsistent SAGs, because it removes ambiguity from the binning evaluation process. The rule for post-assembly contig removal was 100% identity over at least 2kbp. Because it is fair to expect such alignments to occur as non-contaminants between related SAGs, alignment between SAGs sharing a pre-defined taxonomic range (see section D.6) were not counted as potential contaminants. So perfect alignments over at least 2kbp between contigs from SAGs of sufficient taxonomic distance were removed prior to the any binner analyses. For applications which require more recall, a good software is ProDeGe [257].

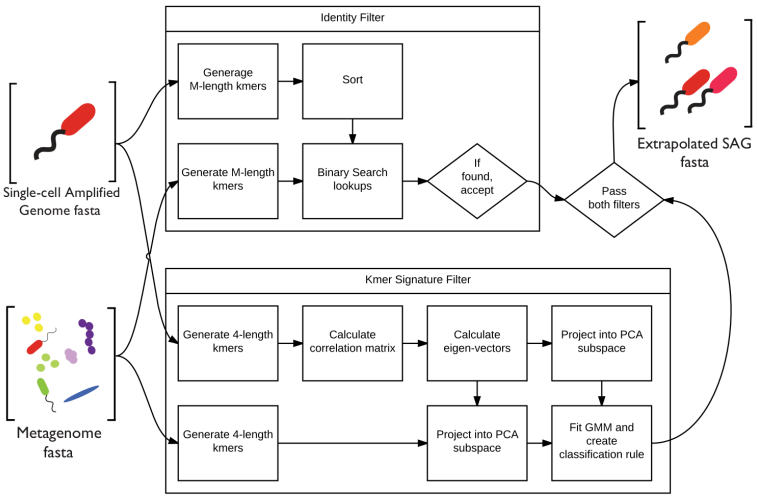



**Figure 2.2:** SAGEX pipeline

### 2.2.2 SAGEX

SAGEX is a classifying binner trained on a single SAG. Its design is inspired by a previous work [76], and automates and refined the essential binning strategy. It accepts a SAG `.fasta` and a metagenome `.fasta` file as inputs and outputs a `.fasta`. The SAG is used as a training data set, each metagenome is evaluated for recruitment, and metagenomic contigs similar to the SAG in taxonomy are recruited. Optionally tetranucleotide counts or tetranuculeotide principle component dimensional reduction data products are available. This work describes a bin from SAGEX as the output `.fasta`. The extrapolated SAG is the bin concatenated with the training SAG.

The SAGEX pipeline (Figure 2.2) carries out two tests on a metagenomic contig to determine membership to a given SAG-bin. First is the kmer (see subsection 1.2.4 for a definition) signature test, which checks that the contig and SAG have similar kmer signatures. Second is the identity filter which checks for at least one region (user defined length) of DNA with 100% identity between the contig and the SAG. A typical run with a metagenome ($\sim$ 50MB) takes 140 seconds per SAG.

The kmer signature filter utilizes a Gaussian mixture model fit to the SAG's kmer values after dimensional reduction with Principal Components (PCs). This works by first calculating the tetranuculeotide frequencies (4-mers) for a given SAG contig. This puts each contig into $256 = 4^4$ dimensions, which is too high for later statistical models thus motivating dimensional reduction. A correlation matrix is then calculated for the metagenome's kmer points. The correlation matrix eigen-decomposition is used to select three eigenvectors, the principal components. This allows for the derivation of a linear transformation which projects the metagenome's kmer values into the principal components' subspace. The linear transformation is then applied to both the SAG and metagenome's kmers. The natural behaviour of the data causes the kmers to take on multivariate Gaussian distributions (see subsection 1.5.3); the overlap of an Ecoli K12 mg1655 genome [30] and SAG [106] demonstrate this best (see Figure 2.3). This popular insight [261] then motivates the choice to employ a Gaussian Mixture Model, which is fit to the SAG's kmers utilizing the

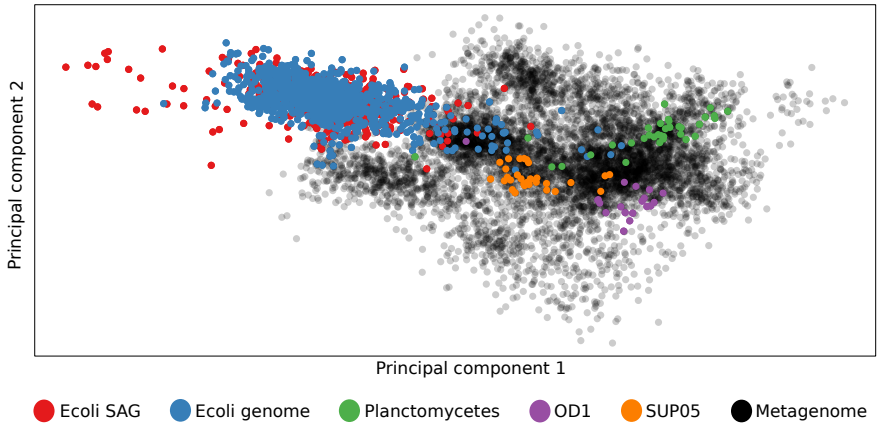



**Figure 2.3:** Tetranucleotide signatures are illustrated for various SAGs, an EColi Genome, and a 200m Saanich Inlet metagenome.

Expectation Maximization algorithm. The user may set the initial number of clusters and choose to let SAGEX decrement the number of clusters. If SAGEX is allowed to decrement the number of clusters, it does so heuristically by decrementing when round-off errors occur in estimation due to poor-fitting Gaussians. To evaluate whether or not a metagenomic contig has a similar kmer signature to the SAG, it must fall within the model's null region (see Equation 2.1). The radius of the region may be set by the user.

$$\begin{gathered}
\mathbf{X} = \sum_{k=1}^{m} \chi_{\{C=k\}} \mathbf{M}_k \\
\chi_A = \{1 \text{ if } A \text{ is true}, 0 \text{ otherwise}\} \\
C \in \{1, 2, \ldots, m\}, \text{categorical} \\
\mathbf{M}_k \sim_{iid} N_3(\boldsymbol{\mu}_k, \Sigma_k), \text{ multivariate normal} \\
\text{null hypothesis } H_0 : x =_D X
\end{gathered} \tag{2.1}$$

where $=_D$ denotes equivalence in distribution

The identity filter is simple and key to generating high-quality SAG-bins. To pass this test, the metagenomic contig must share a contiguous region of perfect identity with the SAG. The length, $S$, of this region is $S = 25$bp by default, but may be set by the user. To accelerate look-ups but not overly consume memory, all of the SAG's $S$-length contiguous regions are perfectly hashed and stored in a sorted list for log-time look-ups. This extremely simple rule is used because it is both very effective and computationally fast.

SAGEX is written in C as a modular pipeline with a user interface implemented in C++.

### 2.2.3 Precision-recall comparisons

In order to demonstrate how SAG-guided binning effects quality, a variety of binning strategies are evaluated via precision and recall statistics. Classifying and clustering binners are evaluated with two different methods because they are not directly comparable. Either paradigm is evaluated at three taxonomic *levels*, representing low (domain), medium (class), and high-level taxonomies

(strain) (see section D.7). All evaluations are done with synthetic metagenomes composed of concatenated assembled SAG genomes. While precision was motivated earlier as requisite to confident binner application, recall is also emphasized because most classifiers can be made arbitrarily precise at the expense of recall, and a classifier without recall is useless. The precision estimator is $TP/(TP+FP)$, and the recall (also called *sensitivity*) estimator is $TP/(TP+FN)$, where $TP$ is the number of true positives, $FP$ is the number of false positives, and $FN$ is the number of false negatives.

Classifying binners are evaluated on their ability to accurately assign taxonomy to contigs. Because evaluation occurs at three different resolutions of taxonomies (previously defined as low, medium, and high), attributions are counted as correct if they fall at or below a taxonomic designation. For example, if a contig is from *Bacteria* and is attributed to *Gammaproteobacteria* the attribution is considered correct.

Clustering binners have a more complex evaluation because clustering is binning without the attribution of specific taxonomic labels. Contigs with matching taxonomic labels are defined to share a cluster, otherwise they do not share a cluster. Instead of classifying contig taxonomies, relationships between contigs are classified. Specifically, define a graph of vertices and edges (Equation 2.2). Each contig has a unique vertex. If contigs are in a bin (cluster), they share an edge, otherwise no edge is shared. Thus every cluster has a clique in the graph. Notice that for every binner evaluation there are two graphs, the true graph and the attributed graph. Every attributed edge is then evaluated as a true or false positive according to the true graph. Note that a cluster of $n$ vertecies will have $n(n-1)/2 = O(n^2)$ edges, resulting in quadratically deformed counts. So clustering statistics will not be comparable with classifier evaluation statistics.

$$\begin{gathered} V \subset \mathbb{S}, E = \{\{a,b\} : a,b \in V\} \\ \{\{a,b\},\{b,c\}\} \subset E \Rightarrow \{a,c\} \in E \\ G = (V,E) \end{gathered} \tag{2.2}$$

PhylopythiaS [208] was evaluated utilizing its web interface (http://phylopythias.bifo.helmholtz-hzi.de/). While inputs had to be divided into compute jobs, the tool was evaluated as a single classifying binner. The *Generic 2013 - 500 Species* model was used. It was run once per synthetic metagenome level.

SAGEX was evaluated twice, once as a classifying binner and again as a clustering binner. In either case, because the synthetic metagenomes were made of SAGs and SAGEX is trained on single SAGs, each SAG was not allowed to recruit its own contigs thus avoiding a clear bias. To evaluate SAGEX as a classifying binner, it was run once per SAG per level, given that the SAG fell within the taxonomic range defined by the level. SAGEX was always run with arguments `-C 25 -k 6 -K`, and has the default of rejecting all contigs shorter than 2kbp. The taxonomy of the SAG used as a training data was used as the taxonomy attributed to any recruited contigs. Because there were many more SAGs than taxanomic categories, classification error statistics were averaged per taxonomic category per level.

To evaluate SAGEX as a clustering binner, each SAG's recruitments were treated as a single bin. Taxonomic attributions were disregarded. Since each SAG bin tends to be small and the number of edges grows quadratically with cluster size, recall is driven down.

MaxBin2 [276] is a clustering binner. Because it utilizes coverage estimates, metagenome reads

**Table 2.1:** Binner precision-recall statistics

| Binner | Type | Precision 1 | Precision 2 | Precision 3 |
|---|---|---|---|---|
| PhylopythiaS | classify | NA* | .36 | .78 |
| SAGEX | classify | .82 | .92 | .98 |
| SAGEX*** | cluster | .42 | .56 | .89 |
| MaxBin2 | cluster | .28 | .44 | .75 |
| ESOM + R | cluster | .33** | .14 | .53 |
| Binner | Type | Recall 1 | Recall 2 | Recall 3 |
| PhylopythiaS | classify | .0007 | .08 | .39 |
| SAGEX | classify | .04 | .03 | .02 |
| SAGEX*** | cluster | .002 | .0002 | .0002 |
| MaxBin2 | cluster | .05 | .04 | .03 |
| ESOM + R | cluster | 0.04 | .13 | 0.01 |

*9 of 10 measurements are NA, **1 of 10 measurements NA, ***Non-standard SAGEX approach

were required as an input. However, the synthetic SAG metagenomes do not have informative coverage statistics due to their creation with MDA. Instead the `.fastq` of a metagenome sampled near the SAGs at 200m was used. The three synthetic metagenomes, one per level, were then clustered.

ESOMs were used as a clustering binner. ESOMs are classically a visualization and dimensional reduction tool [147] and the U-matrix data product is used to produce clusters [1]. This work follows a modern design [74] using Databionics ESOM Tools (http://databionic-esom.sourceforge.net). One thousand 2kbp contigs were sampled randomly from each synthetic metagenome, processed through SAGEX to produce tetranucleotide frequency proportions (an optional SAGEX data product). Tetranucleotide frequencies were then loaded into the ESOM software using a 50x50 U-matrix, 100 iterations, 15% k-batch training, and all other parameters default. The U-matrix `.umx` and best match `.bm` were fed into an R script (section D.8) which clusters contigs when they share a valley in the U-matrix: any nodes sharing neighbours below a cut-off of 0.15 in the U-matrix were defined to share the same cluster (section D.9).

While statistics are available per taxonomic category per level (section D.10), statistics are averaged across categories (Table 2.1). Statistics are stated three times, enumerated by their levels. Clustering and classifying binner statistics cannot be directly compared. Note that recall statistics are not required to be large, since so many sequences are evaluated. Instead, binning strategies tend to exchange recall for increased precision.

Various bins and genomes were evaluated with CheckM with the `lineage_wf -t 32 -x fasta` command. First, the initial SAGs and their decontaminated counterparts compared. Second, each binners' output bins at level 3 were evaluated. Mean completeness, mean contamination, and a ratio of mean contamination over mean completeness are reported (Table 2.2). The ratio is reported because it makes the ESOM results easier to compare against other evaluations. This is motivated because ESOM usage requires sampling large kmers from metagenomes instead of using contigs. Thus ESOMs are evaluated with related yet different synthetic metagenomes relative to the other methods. The possibility of biases due to SAG self-recruitment do not apply here as in the precision analysis, so entire SAGEX extrapolations (SAG and recruited contigs concatenated) are used as bins. Completion average estimates are low relative to proposed standards [207],

**Table 2.2:** CheckM statistics

| Data | Mean Completeness | Mean Contamination | Ratio* |
|---|---|---|---|
| SAGs | 50.91% | 5.97% | 0.12 |
| Clean SAGs | 15.92% | 0.93% | 0.06 |
| SAGEX Extrapolations | 21.19% | 9.88% | 0.39 |
| ESOM + R** | 0.59% | 0.20% | 0.34 |
| PhylopythiaS | 9.64% | 26.19% | 3.94 |
| MaxBin2 | 23.87% | 13.22% | 0.55 |

*Mean contamination divided by mean completeness, **ESOM protocol modifies input. Only compare ratios.

suggesting a tendency toward incompletion was typical across binners. This work reports mean CheckM statistics because the same was done for precision and recall estimation. Means are reported because they approximate expected behaviour of the desired statistics.

### 2.2.4 Saanich Inlet

SAGs from Saanich Inlet (SI) were run through SAGEX against four metagenomes from 2010 along a gradient of depth and decreasing oxygen $O_2$ at 10, 100, 150 and 200m (samples SI048_S3_10, SI060_S3_100, SI060_S3_150, SI060_S3_200 accessions [127]). Metagenomes from 100, 150 and 200m were collected concomitantly with the SAGs, 10m metagenome was collected the previous year in August to match environmental conditions. SAGs from all three depths (100m, 150m and 185m) were run against all four metagenomes in order to see the variability of recruitment to SAGs from both corresponding and disparate metagenomes. SAGEX was run on the same settings as in evaluation. Extrapolated SAGs were then run through checkM [207] in order to estimate genome completeness and contamination (Table 2.2). Extrapolated SAGs were also run through Metapathways annotation and metabolic pathway finding tool [121, 149, 150] for analysis of metabolic attributes for population genomes. Using reads from the metagenomes, RPKM [191] values were calculated for all extrapolated SAG open reading frames (ORFs).

A similar and larger analysis features the use of 91 SI metagenomes [127]. All SUP05 SAGs are run through SAGEX against all 91 metagenomes. All recruited contigs' ORFs (as determined through MetaPathways) are aligned with LAST (e-value cut-off $10^{-3}$ [211]) to RefSeq-nr in search of denitrifying genes (*narG, nirS/K, norCB, nosZ*). Events of recruiting one or more denitrification gene are recorded. Logistic regression tests for statistically significant interaction between SUP05_1c recruitment rates and time while controlling for $O_2$ and metagenome size. The regression analysis surveys samples from 2009 to 2014. All statistical tests use a Type-I error rate of $\alpha = 0.05$.

## 2.3 Results

### 2.3.1 Precision-recall comparisons

A wide variety of precisions are observed (Table 2.1). The distinct difference between the classifying binners PhylopythiaS and SAGEX is the relevance of training data. For low-level taxa, the PhylopythiaS training data becomes irrelevant to this work's metagenomes, while the SAG-based training data is naturally more relevant. This is demonstrated through a drop in recall to 0.07% and majority *NA* precision estimates for PhylopythiaS at level 1, because the classifier is hardly classifying anything to such a low level. Clustering binners can be ordered by amount of

training data as well, with ESOMs only uses kmer signatures, MaxBin 2.0 uses kmer signatures, depth, and marker genes, and SAGEX uses kmer signatures, alignments, and many SAGs. As with classifiers, the clustering binners also demonstrate that more and more relevant training data result in greater precisions. Therefore the pursuit of precision motivates SAG-guided binning.

When constrained to clustering, SAGEX has the least recall, which is not surprising because it is the most precise clustering method. Because exchanges between precision and recall are trivial, it is important to note that SAGEX as a classifier (standard usage) achieves the best precisions while attaining somewhat typical recall. This favourable exchange between precision and recall is the motivating imperative for SAGEX.

CheckM statistics have some important caveats for effective interpretation. First, note that this work reports mean completeness and contamination statistics. The motivating use case for CheckM often involves discarding many nearly incomplete bins, so a CheckM-aided binning analysis may have low average completeness but report near-complete bins. This work reports means because they describe typical bin behaviour. For example, with independent or weakly correlated [171] samples, the mean will converge to the expected marker gene statistic with probability one. Second, note that CheckM's contamination statistics are biased downward when completeness is small. This bias is because CheckM's contamination statistic is a sum of over-abundant marker genes [207]. Because different applications are evaluated with CheckM, ratios of mean contamination per mean completeness is provided (Table 2.2).

The CheckM analysis must be divided into two primary cases. First, the top two rows (Table 2.2), SAGs and Clean SAGs, represents a comparison of initial SAGs against decontaminated SAGs and shows that aggressive decontamination achieved the goal of avoiding low-quality synthetic metagenomes as input to the binners. Thus ambiguity in the results of this analysis is reduced, because input data are less noisy. Second, the last four rows summarize each binner. Note that the ESOM protocol [74] requires modified inputs and thus shifted statistics, hence the ESOM's ratio should to be compared against other methods' ratios instead of raw CheckM statistics. When considering ratios, only PhylopythiaS stands out as appearing to have a higher rate of contamination per completeness. This is because PhylopythiaS bins include high level categories such as *Archaea* that are meant to be interpreted within a hierarchy of related genotypes.

### 2.3.2 Saanich Inlet

The overarching motivation for binning is the desire for a cultivation independent tool of discovery for metabolic capacity of specific organisms within an environmental context. With this in mind this work explore the metabolic pathways of the resulting population genomes produced by SAGEX using Metapathways [121, 149, 150] as the primary tool for gene annotation and pathway identification. In general, the average SAG extrapolation gained 10% more new pathways, only counting unique MetaCyc pathways [53]. Extrapolations also lost 6% of old pathways because increased genetic information allows Pathologic to specify and exclude pathways. Of particular interest is knowledge about metabolic capacity of candidate phyla and microbial dark matter [225] as well as attributing functions to specific taxa. Members of these phyla are defined primarily by SSU rRNA genes found from environmental studies and have no cultured representatives and very little if any genomic sequence information associated with them.

The OD1 (average CheckM completion: 22.10%, contamination: 4.01%) are one such candidate

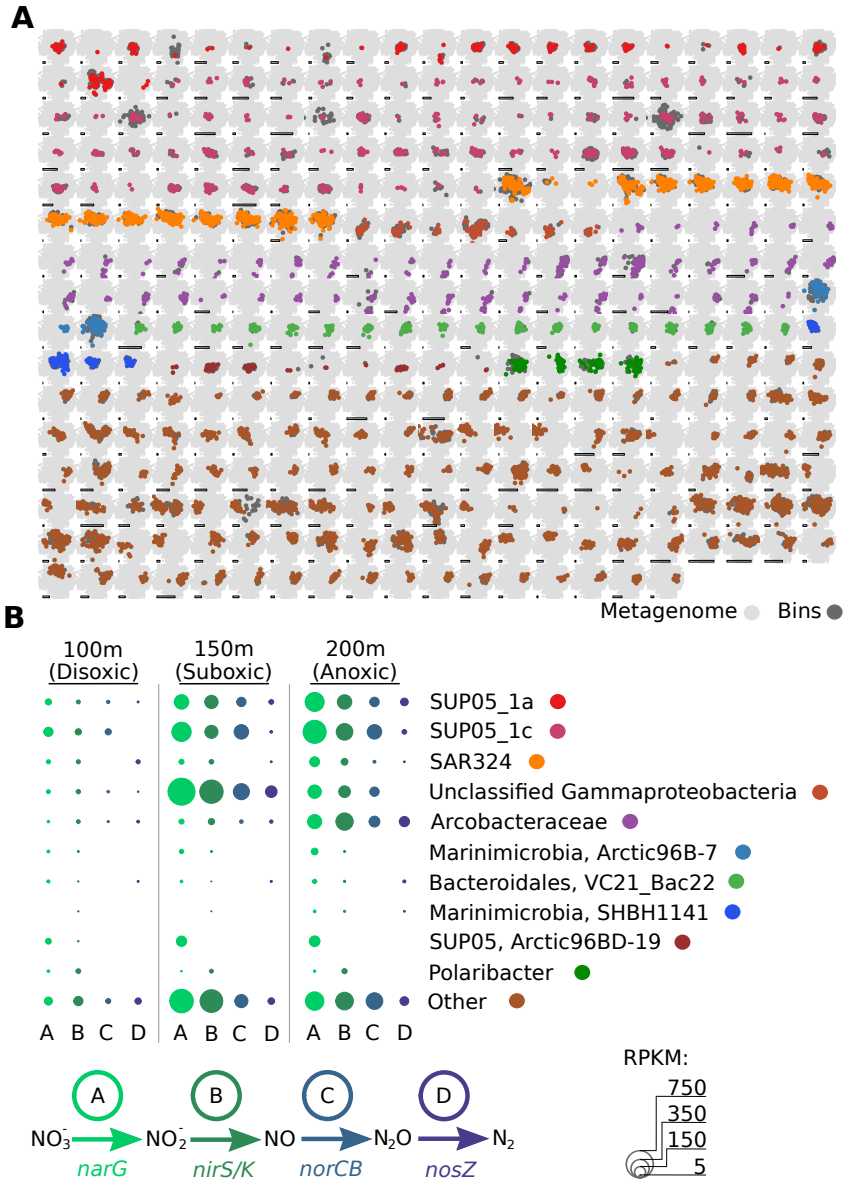
phyla which are found in aquifers [124], merometic lakes [103] and Saanich Inlet [125]. Two SAGs from Saanich inlet, collected from 185 m, were found to belong to the OD1 candidate phyla. Metablic capacity of OD1 from ground water samples following acetate addition show a fermentation based metabolism likely producing hydrogen gas ($H_2$) or hydrogen sulfide ($H_2S$) [274]. The genome bins for the two OD1 SAGs (averaging 300Kb) indeed carried genes from fermentation but genes for hydrogen production were not detected. However, while Wrighton et al. [274] described the OD1 as anaerobic, the gene superoxide dismutase for handling oxygen stress in aerobic (or microaerobic) environnments was detected in the population genome bin, which likely allows the organism to cope with fluctuations in oxygen commonly found in the Saanich Inlet environment [279].

The Planctomycetes (average CheckM completion: 41.92%, contamination: 21.81%) are found abundantly in low oxygen environments such as waste water treatment and marine oxygen minimum zones, and are responsible for carrying out anaerobic ammonium oxidation (anammox). Four Planctomycetes SAGs ideintified by small subumit rRNA gene as *Kueneniaceae scalindua* were found in Saanich Inlet (two at 150m and two at 185m) with an average genome bin size of $\sim$2Mb. Anammox genes hydroxylamine dehydrogenase and hydrazine hydrolyase were expectedly found in the population genomes. Additionally, three genes were found to be involved in sulfur reduction/oxidation including dissimilatory sulfite reductase, adenylylsulfate reductase (previously found in Kuenenia stuttgartiensis) and sulfate adenylyltransferase. The potential role of Planctomycetes in sulfur cycling has been previously unrecognised within oxygen minimum zones and these may play an important role in protecting the Planctomycetes from harmful effects of reduced sulfur species found in sulfidic basin waters of Saanich Inlet. Carbon monoxide dehydrogenase and an hydrogenase were detected in the population genome suggesting the Planctomycetes here may be involved in hydrogen production from carbon monoxide (hydrogen production VI metacyc pathway). While the Carbon monoxide dehydrogenase has been previously reported as involved in Wood-Ljungdal carbon fixation pathway [125], the detection of the hydrogenase in the population genome points to this potentially new function which feeds directly into co-metabolic pathways proposed to occur within Saanicih Inlet, namely hydrogen oxidation by the SUP05 group [125].

Canonical denitrification is one of loss of biologically available nitrogen globally. However, the denitrification pathway is often modular with different taxa carrying enzymes and few of the organisms responsible for denitrification in OMZs are known. Utilizing the SAGs and population genome bins taxa may be identified as harbouring the various steps of denitrification at three points along the oxygen gradient in Saanich Inlet (see Figure 2.4). The abundance of denitrification genes overall increased with depth, with 150m and 200m being quite similar as both are under anoxic conditions. The two SUP05 clades show very similar patterns with the exception of nitrous oxide reductase (*nosZ*), where SUP05_1a population genome (average CheckM completion: 13.66%, contamination: 6.49%) is seen to have *nosZ* at all depths but SUP05_1c population genome (average CheckM completion: 10.21%, contamination: 4.09%) is seen to have the nosZ only at the 150 m depth. This is somewhat consistent with the nosZ gene only being found in 10 out of 48 SUP05_1a SAGs and no SUP05_1c SAGs. SAR324 population genomes also have genes for denitrification, though the nitric oxide reductase (*norCB*) appears to be missing from the population genome bins. Both the unclassified Gammaproteobacteria and Arcobacteraceae have complete or nearly compete denitrification pathways and other taxa have various components of the pathway. Notably,

the Marinimicrobia Arctic96B-7, recently attributed to have nitrate reductase (*narG*) and nitrite reductase (*nir*) [126]. Marinimicrobia SHBH1141, recently attributed to have *nosZ*, population genome also contains *narG* and *nir*. SUP05, Arctic and Polaribacter population genomes carry the narG and narG and nir respectively.

The logistic regression analysis models the effect of time on the probability that a SUP05_1c SAG recruits a denitrification gene. A statistically significant negative interaction exists for between nitric oxide reductase (*norCB*). This means that SUP05_1c's *norCB* recruitment rates decrease over time. The estimated probabilities of recruitment per SAG-metagenome pair are shown in Figure 2.5.




Visualization of a metabolic analysis with SAGEX, focussing on denitrification. (A) SAGEX kmer signatures are visualized for SAGs, the 200m metagenome, and recruited metagenomic contigs. Recruitments are always attributed the same taxa as the SAG which recruited them. SAG exatrapolations (SAGs and recruited contigs) are output from SAGEX and input into Metapathways. (B) Reads from the 100m, 150m, and 200m metagenomes were aligned to extrapolation ORFs, allowing calculation of RPKM statistics. All shown dots represent cumulative RPKM values from unique ORFs within a taxonomic category.

**Figure 2.4:** A SAGEX work flow

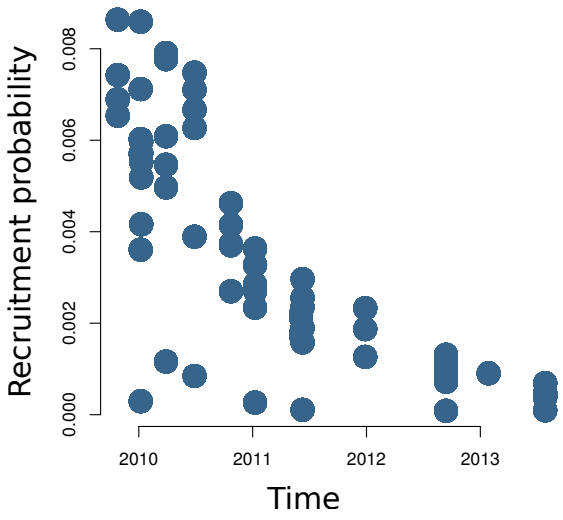



**Figure 2.5:** The probability which SUP05_1c recruits nitric oxide reductase drops off time. SAGEX was run on all pairs of metagenomes and SUP05_1c SAGs, aligned to RefSeq-nr (e-value cut-off: $10^{-3}$), then tested with logistic regression for significantly significant interactions between time and recruitment of denitrification genes. Models control for effects of $O_2$ concentrations and metagenome size.

## 2.4 Discussion

With the advent of next generation sequencing has come also a burst of binning from metagenomes on a large scale [240]. Binning has utilized marker genes for both the binning procedure (MaxBin [275, 276], Phylopythia [182]) and post-hoc testing (CheckM [207]). This work demonstrates that the advent of SAGs brings more precise binning than previous popular binning strategies. These technologies will work well together into the future of binning.

Existing binning tools were designed out of the need to elucidate the connections between taxonomy and function from the vast metagenomic space filled by next generation sequencing. Because SAG-guided binning has resulted in greater precision, these findings suggest that both fosmids and Illumina Tru-Seq synthetic long reads [161] may be used to guide binning as well.

The greatest application for SAGEX is to exploit the link between taxonomy and function which the SAG provides, expanding it to a population level, such that while a single SAG is the genome of a single organism (with varying degrees of completion) SAGEX can then bin contigs from the metagenome which are representative of that population as a whole, and thereby provide a taxonomic grounding for much more of the metagenome.

### 2.4.1 Metabolic discovery

SAGEX has shown the ability to confirm existing metabolic capacities such as fermentation in OD1 and carbon monoxide dehydrogenase in Planctomycetes showing its fidelity. Further, OD1 was shown to have an ecosystem adaptation via superoxide dismutase. Implications for assigning metabolism to taxonomy such as hydrogen production from Planctomycetes is a substantial insight into distributed metabolic coupling which has been proposed for Saanich Inlet [125, 167]. Identifying the Planctomycetes as the likely source of hydrogen for SUP05 oxidation metabolically

links these two organisms and likely serves to provide SUP05 with additional energetic substrate for growth and likely carbon fixation. With greater numbers of SUP05 nitrate reduction via partial denitrification may lead to increased nitrite production [237] and feed forward planctomycetes anammox activity ultimately increasing nitrogen loss from OMZ systems. Greater knowledge and taxonomic resolution of the energetic pathways such as hydrogen production and oxidation which fuel major players in these cycles such as SUP05 and Planctomycetes is key to understanding the future global impacts of OMZ expansion and intensification [86] in a warming planet.

Taxonomic distribution of the denitrification pathway sheds light on a recently identified but not taxonomically constrained niche for nitrous oxide reduction within the anoxic waters of Saanich Inlet [167]. Indeed, it appears that several taxa may be capable of filing this niche, specifically SUP05, SAR324, uncultured Gammaproteobacteria, Arcobacteracea, Bacteroidales VC21_Bac22 and Marinimicrobia SHBH1141. These taxonomic attributions are predominaly novel, with no other references of *nosZ* in any of these groups other then recentoy in Marinimicrobia and in Epsilonproteobacteria *Sulfurimonas gotlandica* related to Arcobacteraceae [153]. The identification of *nosZ* in SUP05 in very interesting as a recently sequenced isolate did not contain the gene [237]. It is possible that only a sub-population of SUP05 contain the *nosZ* gene, as suggested by *nosZ* only being found in the SUP05_1a and not 1c clades. The presence of *nosZ* in both the SUP05 clad population genomes at 150 m is slightly confounding and may be due one of two possibilities. One is the possible of miss-assembly in the metagenome between SUP05_1a and SUP05_1c clades, where high abundance and high similarity between the two clades created chimeric contigs which contained the SUP05_1a *nosZ* but recruited to the SUP05_1c population genome. Two would be the possibility that SAGEX, while highly precise, could not differentiate between the two SUP05 clades, indeed, binning methods may not be suitable to differentiate between such closely related groups. Attribution of *nosZ* to Marinimicrobia SHBH1141 is consistent with recent findings, though attribution of *narG* and *nir* is not and may again be the result of either miss-assembly or other cross-recruitment between closely related clades. The novelty of SAR234 involvement in the denitrification pathway is highly intriguing as this group has been implicated in other OMZs [271], the extent to which the denitrification trait exists outside of Saanich Inlet would need to be explored. As several taxonomic groups are seen to carry out various steps of the denitrification pathway the dynamics of which group is dominant under what conditions remains to be addressed and would require analysis of gene expression data such as metatranscriptomics and metaproteomics.

Application of SAGEX to all SUP05_1c SAGs over all 91 metagenomes from the Saanich Inlet time series made testing for time effects possible. Logistic regression analysis found that SUP05_1c's nitric oxide reductase recruitment probability decreased from 2009 to 2014 (p-value $< 0.05$). Precise binning makes these results more credible. Indeed, earlier observations of SUP05 [125, 266] did observe nitric oxide reductase, but the later cultivation [237] did not. Combined with the observation of other complete denitrifiers in 2010 (Figure 3), these findings support a hypothesis: SUP05_1c is evolving toward partial denitrification, opening up a niche for another complete denitrifier.

Application of SAGEX may extend beyond coupled SAGs and metagenomes. Differential metabolic pathways present in the genome bins from metageomes along ecological gradients may indicate different populations related to the SAG may differ under different environmental conditions along gradients such as depth or oxygen concentration. Additionally, the utility of

SAGEX to be used with assembled metatranscriptomes further enhances the prospect of exploring differential expression at the level of population genomes. Thus SAGEX may be useful in binning populations from different environments or over time from the same environment in efforts to explore differences in metabolic capacity or genetic drift along gradients or over time. The extent to which this variation occurs likely depends on the genetic diversity of the group of organisms. Certainly, within *Prochlorococcus* a substantial amount of diversity exists amoung sub-populations present in different relative abundances over changing seasons [142]. As long as key genomic characteristics (genomic regions of 100% similarity and kmer signatures) remain distinct, SAGEX remains precise.

### 2.4.2 Precise binning

Having evaluated both classifying and clustering metagenomic binners using a varying amounts of training data, this work has effectively surveyed some binning strategies across the spectrum of supervision. That is, the binners which use more training data are more supervised. With observed precisions ranging from 14% to 98%, it is clear that some automated binning solutions require additional curation. In general, [180] correctly inferred that more and more relevant training data results in better bins. ESOMs remain curation aids [1] and resist automation. MaxBin2 performs admirably despite reduced training data. PhylopythiaS performs well with taxa that are relevent to its models (see section D.10). SAGEX tends to get the highest precisions and requires a guiding `.fasta` to operate (this work focuses on SAGs). Of course, as phylogenetic range narrows all binners make more errors. Formally, for some phylogenetic tree $G_V$ and a bin $b$, the height of $v$ satisfying $b \subset v \in V$ would correlate positively with precision. This error pattern of confidently describing higher level clades is shared with bootstrapped phylogenetic estimation [84, 88, 132, 204], suggesting a common error profile might exist for all taxonomic estimation methods. This error pattern is a mixed blessing, because binning is being applied in both strain-level analysis [65, 261], but also in studying the Tree of Life [133] (see Figure 1.3). Binning has been applied toward understanding strains, but ultimately it is best applied in understanding the whole tree. The variable quality and pragmatism of binning strategies clearly plays a large role in these abstract analyses but has yet to be incorporated in an objective way.

The comparison of precision and contamination statistics demonstrates that binners effect different forms of bin quality. The primary difference is the scope of applicability. Binner precision describes the probability that any recruited sequence is correctly recruited. CheckM's contamination statistic describes single-copy marker genes and can be heuristically generalized to describe all sequences in a bin. In theory precision is less heuristic than the contamination statistic, but practically both provide essential information the other cannot. CheckM's contamination statistic describes a necessary condition toward individual genome recovery: that marker genes are in single multiplicity. It is possible that such a case be satisfied and have contaminant sequences in a bin, so also requiring that a binner have good precision amounts toward a sufficient argument. To claim a bin represents a single genome, it should constructed by a precise binner and also good marker gene statistics.

The advantage of precise binning is that individual attributions of genetic sequences to a taxa have a higher probability of being correct. Increased probability of correct attributions then reduces the probability of mistakes in bin construction. The result is a higher quality bin that can

be more widely trusted, thus satisfying a known issue binning [240]. Precision is the motivation and deliverable of SAGEX. However, the other binning strategies studied here have merits as well. First, precision generally comes at the cost of recall and thus reduces inferential yield. Second, increasing precision per recall often costs additional training data [180]. If methods of greater recall (consider ESOMs and MaxBin 2.0) are preferred, appropriate statistical methodology can increase precision. For example, special sequences may have a verifiably greater precision than others, but this verification will consume additional data (see Appendix C). So correct usage of imprecise binners can yield precise results, but comes at the cost of large data sets. Consider the work of [133] as a potential example of this. Notice that precision and therefore confident binning never comes for free.

It is important to note that this work has the potential for bias due to its SAG-centric lens. A way to improve on this study would be to reproduce it with curated metagenomes composed of known isolate metagenomes paired with SAGs per isolate. This would also do a better job of demonstrating the influence of chimeras, which are a non-trivial issue in metagenomic assembly [35, 240].

## 2.5 Conclusions

Single-cell amplified genome (SAG)-guided binning is shown to substantially increase precision in metagenomic binning, thereby alleviating the concern that binning may falsely attribute contigs to bins incorrectly [240]. Precision is studied because it is the probability that attributions are correct. This work's software for SAG-guided binning is released as SAGEX (SAG EXtrapolator). While evaluated with SAGs, SAGEX only requires its guiding (or training) data to be a `.fasta` file with at least 4 sequences longer than 2kbp. A thorough comparison was made possible by a collection of 368 SAGs sampled from the Saanich Inlet oxygen minimum zone. It is observed that binners which utilize a larger volume and more relevant training data obtain better precision per recall. All binners are observed to exchange precision for taxonomic specificity; genome-level bins tend to be the most error- prone, while bins at a higher phylogenetic level will have more precision. All methods are argued to have precise usages at least in theory, given sufficient training data or correct statistical manipulation combined with a sufficient sample size. Precise binning alleviates issues with false recruitment, and is attainable with genomic data of sufficient length with a defined phylogeny.

A motivating application of precise binning is explored in a microbial food web in the Saanich Inlet oxygen minimum zone [125]. Precision is valuable in this application because each interaction has the potential for immense transformative effects on oceanic chemical composition [154]. Used in combination with Metapathways [121, 149, 150], SAGEX was able to recover an additional 10% more pathways per SAG and describe a variety of metabolisms. OD1 was confirmed to support fermentation, and also superoxide dismutase in its population genome for handling oxygen stress demonstrates an adaption to the Saanich Inlet's seasonal anoxicity. Discovery of hydrogen production in Planctomycetes supports potential for coupling with SUP05. Using metagenomes from 2010, potential for complete denitrification was observed for taxa including SUP05_1a, SUP05_1c, and Marinimicrobia SHBH1141. Applying this pipeline to the Saanich Inlet time series (2009-2014) and surveying with logistic regression found SUP05_1c nitric oxide reductase recruitment drops with time, suggesting evolution toward a partial denitrification niche.

As metagenomic binning is often used to recover individual genomes with variable degrees of quality and completion, genomic reporting standards are called for. The findings of this work suggest that individual genome recovery standards should require both marker gene statistics and the binning strategy to be published with binning products. When confidence in results is desired, the most precise binning method available should be used, thus motivating guided binning when possible.

These findings contextualize agreeably with previous models of Saanich Inlet. The discovery that SUP05 might be evolving toward a partial denitrification niche agrees with the conceptual model (see subsection 1.4.4). Specialization would leave an energetic opportunity unutilized, so it really suggests niche partitioning. Further attribution of nitrous oxide reductase ($N_2O \rightarrow N_2$) provides candidate partners for metabolic syntrophy. Further contextualizing in the differential model (see subsection 1.4.5) supports this hypothesis, particularly that complete denitrification may be based on a sulfur-driven relationship. Due to known sulfur-processing metabolic capability combined with a nitrous oxide reductase attribution, Marinimicrobia is a prime candidate. In chapter 3, this argument and others will be bolstered through correlative means.

# Chapter 3

# SSU rRNA correlation

Small subunit of the ribosomal RNA gene data (SSU rRNA) correlation surveys produce networks which may be used to bolster ecological arguments with statements on microbial covariation. For example, in chapter 2 many metabolic capabilities were attributed to specific taxa, leaving many syntrophic hypotheses standing. This chapter uses correlative evidence to further constrain these hypotheses. These results agree with and extend previous network perspectives of Saanich Inlet (see subsection 1.4.4 and subsection 1.4.5). Recently, a major challenge in SSU rRNA correlations surveys was demonstrated [267], which effectively brings individual correlative edge attributions into question. Imprecise correlation networks might be able to inform on general topological characteristics, but descriptions of fine community structure have effectively lost confidence. If unmet, these concerns would make correlative evidence at most a suggestion if not a burden, in producing ecological arguments. This work meets this concern through constraining of a covariance matrix $\Sigma = LL^T + \Psi$, which reduces parameter complexity from $O(p^2)$ to $O(p)$ for $p$ taxa. The method is shown to be capable of substantial precision-recall improvements. Hence ecological arguments are stengthened through **precise correlation**.

## 3.1 Introduction

The denitrifying community of Saanich Inlet includes a variety of taxa, with no strain ever operating in isolation. The process of transforming initially fixed nitrogen in $NH_4^+$ to largely biologically inaccessible $N_2$ can span several ecological niches (as in denitrification but not anammox, see subsection 1.4.1). Due to various factors including competitive pressure, these niches might encourage metabolic specialization, as would follow from the *Black Queen Hypothesis* [190]. Indeed, an abundance of genomic data supports the perspective of a taxonomically diverse denitrification pipeline (see section 1.4). An integral component of oceanic nitrogen loss is entirely ecological.

Perspectives of such ecological machines can be conveyed through a network perspective. The microbial ecology of Saanich Inlet relating to denitrification has been described through conceptual [125] and differential [167] models (described in subsection 1.4.4 and subsection 1.4.5), both of which are abstractly communicated through network representations. As described in subsection 1.1.2, a coherent network abstraction exists to communicate these models. Particularly, these networks have microbial or chemical nodes and interactions described with edges. These networks' abstraction imposes a degree of superficiality, but the abstraction's coherence makes the networks relevant. While the underlying microbial individuals are no less complex, careful construction of networks simplifies and thereby contributes to conversations about the ecological machine.

Networks are not only created through abstraction, because they are often inferred directly.

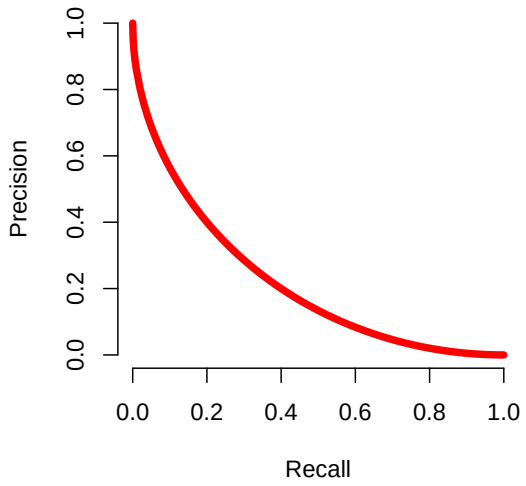



**Figure 3.1:** A simplified depiction of a poor precision-recall exchange, like those observed in Weiss et al. [267]. See Figure D.1 for actual.

SSU rRNA correlation networks (see section 1.3) can be estimated directly from specially-processed metagenomic data, producing a survey of statistical dependence (see section 1.5.2) in the microbial community. Upon estimation correlation networks remain abstract, only imposing descriptions of covariation, and invite concretization through further genomic information and prior knowledge. This work does exactly that, by first describing denitrifying taxa with metagenomic binning in chapter 2, then bolstering arguments with correlative information in this chapter.

Correlation networks are often estimated through hypothesis testing (see subsection 1.6.5). For each pair of taxa indexed $(i, j)$ with correlation $\rho_{ij}$, the null hypothesis $H_0 : \rho_{ij} = 0$ is tested. If the null is rejected in favour of $H_0^c : \rho_{ij} \neq 0$, then taxa pair $(i, j)$ is classified as *correlated*. If the null is not rejected, the pair is classified as *uncorrelated*. Thus correlation network estimation can be treated as a classification problem. In the language of section 1.6.5), the task is to classify pairs of taxa $\mathbb{S} = \{\{\text{Cyanobacteria}, \text{Planctomycetes}\}, \{\text{Planctomycetes}, \text{SUP05}\}, \ldots\}$ as correlated or not (so $\mathbb{T} = \{\text{correlated}, \text{uncorrelated}\}$) with correlation classifier (network estimation protocol) $C$. Interpreting correlation network estimation as a classification problem implicitly describes each $C$ with precision-recall exchanges. Thus, correlation network estimation methods can be objectively compared and measured.

Unfortunately, it was recently observed that popular correlation survey techniques in Microbial Ecology suffer from poor precision-recall exchanges [267]. This means that precision-recall curves tend be shaped according to Figure 3.1, implying that networks can be estimated with precision or recall, but rarely both. This result means that individual edges in estimated correlation networks are in doubt. General graph structure may be preserved in some approximate sense, but individual edges attributions are in doubt. So if one points to an edge and asks *"does this correlation really exist?"*, the answer is merely *"maybe"*. The discovery of poor precision-recall exchanges for correlation network estimation in Microbial Ecology threatens their viable application, because it

introduces incoherence into the abstraction.

### 3.1.1 The overfit hypothesis

SSU rRNA can be counted in almost arbitrarily-many clusters (see subsection 1.2.3), because the resolution is adjustable. Studying few low-resolution clusters from a phylogenetically diverse community will inevitably lead to oversimplifications. So studying many fine-resolution clusters is motivated. From a statistical modelling perspective, this means that for a sample's vector of SSU rRNA counts $\mathbf{Y} \in \mathbb{Z}_{\geq 0}^{p}$, $p$ becomes large–often in the hundreds, thousands, or tens-of-thousands. All correlation network estimation methods, not just those surveyed by Weiss et al. [267], must evaluate every pair of these many dimensions for a statistically significant correlation. Indeed, despite applying some of the best existing theoretical solutions, this work must also reduce $p$ for a feasible solution.

Correlation estimation can be viewed generally as a form of multivariate regression. Through the modelling of dependencies, multivariate regression models tend to have many more parameters than their univariate counter-parts (see subsection 1.6.2). Simple multivariate models tend to have quadratically-many ($O(p^2)$) parameters, as in the following equation. So if $p$ is one thousand, then there are about one million parameters in need of estimation. In subsection 1.6.1, overfit was described as an estimation failure due to having too many parameters. In studying the Saanich Inlet community 112 SSU rRNA samples are modelled. Any large SSU rRNA data set will never be large enough to adequately describe a one-million-dimensional space. Without more creative modelling, overfit is inevitable. The correlation tests evaluated by Weiss et al. [267] employ no mechanisms to significantly reduce the number of parameters estimated, and are thus prone the overfit. Overfit is known to reduce predictive capacity and is thus a possible cause of poor precision-recall exchanges in correlation network estimation in Microbial Ecology.

$$\mathrm{Cor}(\mathbf{Y}) = \left[\rho_{ij}\right]_{p \times p} = \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,p} \\ \rho_{1,2} & 1 & \rho_{2,3} & \cdots & \rho_{2,p} \\ \rho_{1,3} & \rho_{2,3} & 1 & \cdots & \rho_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1,p} & \rho_{2,p} & \rho_{3,p} & \cdots & 1 \end{bmatrix}$$

This line of reasoning supports the hypothesis that overfit has caused the high error rates observed by Weiss et al. [267]. Regularization is a broad category of methods for reducing overfit through some form of constraint (see subsection 1.6.3). While a statistical frontier, methods for dealing with high-dimensional dependence structures are available [37, 44, 218]. This work effectively pulls known solutions from the frontiers of statistical theory to implement a practically feasible correlation network estimation software. To deal directly with the quadractic explosion ($O(p^2)$) of parameters, a factor model's covariance structure (see subsection 1.5.3) is used to model quadratically-many correlations with only linearly-many ($O(p)$) parameters. So the utilized regularization constraint is $\Sigma = LL^T + \Psi$ (expanded in the following equation), where $L \in \mathbb{R}^{p \times m}$, $\Psi \in \mathrm{diagonal}(\mathbb{R}_{>0}^{p \times p})$, $m$ is small. This work uses $m = 3$.

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} & \cdots \\ \sigma_{1,2} & \sigma_{2,2} & \sigma_{2,3} & \cdots \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_{2,2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = LL^T + \Psi = \begin{bmatrix} l_{1,1} & \dots & l_{1,m} \\ l_{2,1} & \dots & l_{2,m} \\ l_{3,1} & \dots & l_{3,m} \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} l_{1,1} & \dots & l_{1,m} \\ l_{2,1} & \dots & l_{2,m} \\ l_{3,1} & \dots & l_{3,m} \\ \vdots & \vdots & \vdots \end{bmatrix}^T + \begin{bmatrix} \psi_{1,1} & 0 & 0 & \cdots \\ 0 & \psi_{2,2} & 0 & \cdots \\ 0 & 0 & \psi_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Of available regularization approaches, the factor structure ($LL^T + \Psi$) has its own strengths and weaknesses relative to other solutions. Constrained optimization or *shrinkage* approaches are popular regularization methods [44], extending well beyond multivariate dependence modelling [164]. Shrinkage methods are not used in this work because they tend to the symptoms of the parameters' quadratic explosion ($O(p^2)$), and with $p$ large, the problem should be dealt with directly. This leaves a single primary alternative to consider, *truncated vine copula* [37], which construct multivariate dependencies through a hierarchy of latent bivariate variables. Both truncated vine copula and factor models can represent $O(p^2)$ correlations through $O(p)$ parameters, but there are some important differences. An advantage of vines over factor models is that they are capable of representing correlations as exactly zero, $\rho_{ij} = 0$ is possible. Factor models cannot achieve this without adverse effects, so for a sufficiently large sample size (well out of reach for most any SSU rRNA study) every test of $H_0 : \rho_{ij} = 0$ will be rejected. An advantage of using a factor model is that it is capable of detecting and modelling unobserved regressors, effectively adding a missed column to the $X$ matrix (see subsection 1.6.2). The factor model is used in this work, because an essential and unobserved variable is long-hypothesized to exist in SSU rRNA and RNA seq studies: sequencing depth. This is advocated through the *mixed effect* perspective of SSU rRNA correlation analysis (see subsection 1.3.1). For this work, the validity of this perspective over the *compositional* perspective is advocated in subsection 3.3.2.

It is important to note that high-dimensional covariance estimation has previously been interpretted as a multiple comparison error problem [283]. *False discovery Rate* (FDR) [24, 25] has been successfully applied toward univariate regression surveys [168, 227, 228]. However, multiple comparison corrections have not been designed for high-dimensional correlation surveys. This work does not evaluate them as a potential solution.

### 3.1.2 Precision-recall comparison

To maintain objectivity, this work's regularized model is compared against existing methods. Having phrased correlation network estimation as a classification problem, precision and recall are the metrics of comparison. The regularied model is compared against **Pearson correlation coefficient** or **PCC** and **SparCC** (Sparse Correlations for Compositional data). The Pearson correlation statistic converges to $\text{Cor}(X, Y)$ for sufficiently large sample sizes when expectation integrals exist. SparCC is a popular method in Microbial Ecology, designed to accommodate for compositional effects [95], and was survey by Weiss et al. [267]. To allow for confident precision-recall calculations, the correlation methods are compared in a simulation study.

This simulation study demonstrates that regularized estimation is necessary for increasing precision without loss of recall, thereby alleviating the concerns highlighted by Weiss et al. [267]. This bolters the applicability of correlation networks in Microbial Ecology. Surveying a community for correlations is only useful if attributed correlations can be done with confidence. Without precision, there is no confidence.

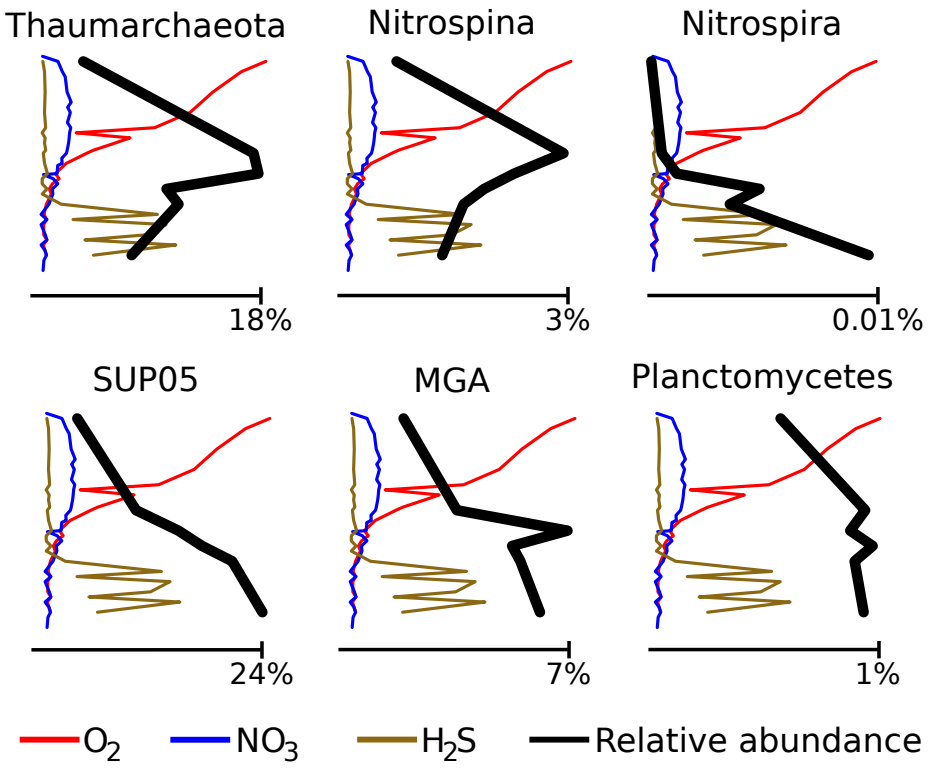



**Figure 3.2:** Average abundances of select chemical concentrations and taxa

### 3.1.3 Saanich Inlet

Beyond precision-recall comparisons, objectivity is further maintained by applying the regularized correlation method to Saanich Inlet SSU rRNA data and findings are contextualized in literature. Any discrepancies between existing understandings of Saanich Inlet community structure and the correlation network are explored. First, the findings will be contrasted against the conceptual (see subsection 1.4.4) and differential (see subsection 1.4.5) models. While all three models use environmental data, a significant difference between is that the correlation network only utilizes SSU rRNA data, whereas the other use proteomic or multi-omic data. An advantage the correlative approach has is that its statistical survey over few data types enables high-throughput analysis, thus providing a description of years of data (2008-2011) instead of days. Second, known community structure (see Figure 3.2) should also be recovered. Third, of all correlations surveyed, the subset relating to SUP05 will be inspected, because observations in chapter 2 suggest the potential for SUP05 to currently be evolving toward a metabolically syntrophic relationship built on sulfur-driven denitrification.

Application of the regularized model to real data also allows an opportunity to test several modelling hypotheses. First, the negative binomial distribution is popular in SSU rRNA regression analyses [168, 185, 228], but there may be better models. Through model selection via AIC testing (see subsection 1.6.3), this work finds such an alternative. Further, the question of whether SSU rRNA data follow the *mixed effect* or *compositional* perspectives will be commented on.

## 3.2 Methods

### 3.2.1 Multivariate construction

As stated in subsection 3.1.1, this work tests the ability of a regularization strategy to improve precision-recall exchanges in correlation network estimation. Interpreting regularization as a model constraint meant to reduce overfit, this work constrains the covariance matrix $\Sigma$ to a lower-dimensional structure, $\Sigma = LL^T + \Psi$. Model dimension (the number of free parameters) is long-suspect in causing errors, reaching over modern statistical theory [2]. For $p$ taxa, the constraint achieves a reduction in parameters from $O(p^2)$ to $O(p^2)$. In the applied problem, SSU rRNA correlation, $p$ is consistently large, so the reduction is motivated from the perspective of overfit.

However, choosing to constraining an abstract parameter is purely hypothetical without a particularly specified model. Factor models specify covariance matrices for multivariate Guassian distributions. However SSU rRNA data are never Gaussian-distributed (in $\mathbb{R}$), because they are count data (in $\mathbb{Z}_{\geq 0}$). So the regularizing solution complicates model selection. To overcome this issue, this work employs statistical copula (see subsection 1.6.4), a tool which eases multivariate modelling by allowing the multivariate and univariate components to be selected separately. Via the application of copula, the multivariate structure is now defined without specification of which marginal distributions are used. So the model is abstractly defined as follows for arbitrary marginal distributions $F_{Y_{ij}}$.

**Data:**
$Y \in \mathbb{Z}_{\geq 0}^{n \times p}$ observed counts,
$X \in \mathbb{R}^{n \times q}$ regressors,
**Parameters:**
$L \in \mathbb{R}^{p \times m}$ factor weights, $m$ small,
$\Psi \in \text{diagonal}(\mathbb{R}_{>0}^{p \times p})$ factor model errors,
**Implicit parameters:**
$\Sigma \in \mathbb{R}^{p \times p}, \Sigma = LL^T + \Psi$ latent covariance matrix,
**Random variables:**
$Z \in \mathbb{R}^{n \times p}, \mathbf{Z}_i \sim N_p(0, \Sigma)$ latent Gaussians,
$Y_{ij} \sim F_{Y_{ij}}$ observed marginal counts,
**Functions:**
$F_{Z_{ij}}(z_{ij}) = \Phi(z_{ij}; 0, [LL^T + \Psi]_{jj})$ marginal Gaussian distributions (see subsection 1.5.3),
$F_{Y_{ij}}^{-1}(y_{ij})$ marginal counts' generalized inverse distribution function,
**Constraints:**
$Y_{ij} = F_{Y_{ij}}^{-1}(F_{Z_{ij}}(Z_{ij}))$ copula mechanism

### 3.2.2 Marginal model selection

Modelling with copula has conveniently allowed marginal distributions to be dealt with abstractly, but they must be specified prior to model implementation. While copula allows individual dimensions' (taxa's) marginal distributions to be different, having arbitrarily many dimensions

(see subsection 1.2.3) motivates selection of a single model which is flexible enough to model every dimension. This work surveys a selection of models with AIC testing (see subsection 1.6.3).

There are many count models to choose from [46, 63], but an effective way of reducing candidates is to require a certain kind of flexibility: the ability to satisfy both complete under-dispersion and over-dispersion. An **over-dispersed** model $Y$ satisfies $\text{Var}[Y] > \mathbb{E}Y$. An **under-dispersed** model satisfied $Y < \mathbb{E}Y$. Not all count models satisfy both of these properties. For example the Poisson distribution satisfies $Y \sim \text{Poisson}(\lambda) \Rightarrow \text{Var}[Y] = \mathbb{E}Y$. By the law of total variance, the large class of Poisson mixtures is never under-dispersed, and thereby the negative binomial and lognormal Poisson [100] are both strictly over-dispersed. Similarly, the binomial and thereby all multinomial marginals are strictly under-dispersed. Requiring no mean-variance constraints not only reduces candidate models, but also requires that models be able to reflect the realities the data are expressing. Constraints should come from data, not from models.

Requiring complete mean-variance flexibility implicity and ironically requires that candidate count models have a single mean-variance constraint. By 2, if $Y$ is a count model ($Y \in \mathbb{Z}_{\geq 0}$), then $\text{Var}[Y] \geq (\mathbb{E}Y - \lfloor \mathbb{E}Y \rfloor)(\lceil \mathbb{E}Y \rceil - \mathbb{E}Y)$. This is the minimum variance bound satisfied by every count model.

**Result 2.** *If $N$ is a count variable (random variable in $\mathbb{Z}_{\geq 0}$ with probability one) with mean $\mu$ and variance $\sigma^2$, then $\sigma^2 \geq (\mu - \lfloor \mu \rfloor)(\lceil \mu \rceil - \mu)$.*

*Proof sketch.* of Result 2

Condition the variance of $N$ on $\mathbb{E}[N|N < \mu]1_{N<\mu} + \mathbb{E}[N|N \geq \mu]1_{N \geq \mu}$, where $1_X$ is an indicator function.

□

The negative binomial distribution has been popular in SSU rRNA analysis [168, 185, 228]. It is a good choice because SSU rRNA data tend to be over-dispersed and it is mathematically simple, making computation faster and easier to implement. It is an imperfect choice, because it cannot model under-dispersed data and it cannot model data from distributions with very heavy tails. Not being able to model under-dispersed data denies modellers the ability to accurately predict taxa values when it might actually be possible. Infinite variance occurs when expectation is finite ($\mathbb{E}Y < \infty$), but variance is not ($\mathbb{E}Y^2 = \infty$). This work explores the possibility of SSU rRNA models with non-finite and undefined variance. For these reasons, other models are considered in this work. Univariate models are objectively compared with AIC statistics subsection 3.3.1. The following models are considered.

1. **Negative Binomial distribution**

   Despite not satisfying the mean-variance freedom requirements postulated by this work, this popular model is a candidate for AIC testing. This work uses the MASS [264] (chapter 7.4) library's implementation of the Negative Binomial, with location parameter $\mu$, dispersion parameter $\nu$, and with the following probability mass function.

$$f_Y(y; \mu, \nu) = \frac{\Gamma(\mu + \nu)}{\Gamma(\nu) y!} \frac{\mu^y \nu^\nu}{(\mu + \nu)^{\mu+\nu}}, \mathbb{E}[Y] = \mu, \text{Var}[Y] = \mu + \mu^2/\nu$$

2. **Conway-Maxwell-Poisson Distribution**

   The Conway-Maxwell Poisson distribution (CMP) [62] was initially derived from a queuing model and has seen recent interest [117, 165, 166, 235, 242] because of its ability to model both under-dispersion ($\mathbb{E}[Y] \geq \text{Var}[Y]$) and over-dispersion ($\mathbb{E}[Y] \leq \text{Var}[Y]$) in count data. The CMP has the following mass function.

$$\mathbb{P}[X = x] = f_{CMP}(x; \lambda, \nu) = \frac{\lambda^x}{(x!)^\nu} \left( \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} \right)^{-1}$$

   Notice that the CMP generalizes the Poisson distribution from $\nu = 1$. While the capacity for both over and under-dispersion suggested the CMP may potentially satisfy our requirement for separate and unbounded location and dispersion parameters, it was not meant to be. In Appendix B, this work proves that a CMP with mean $\mu$ and variance $\sigma^2$ satisfies $\sigma^2 < \mu(\mu + 1)$. Because SSU rRNA data sometimes exhibits extreme variability, a variance upper-bound is inadmissible.

3. **Hagmark class**

   In Hagmark [118], the following transform is proven to construct count distributions from absolutely continuous, positive random variables' distributions with unconstrained means and variances. Most importantly, the count variable $N$ and continuous, positive variable $Z$ satisfy $\mathbb{E}[N] = \mathbb{E}[Z]$, and $N$ will have a variance *near* the variance of $Z$ (see Hagmark [118] for exact details). The transform defining a count model $N$ from a continuous model $Z$ is the following.

$$\mathbb{P}[N \leq n] = \int_n^{n+1} \mathbb{P}[Z \leq z] dz$$

   In Hagmark [119], a special Poisson distribution generalization capable of all means and variances is derived.

   This work applies the Hagmark transform to the Gamma and Log Normal distirbutions, because their expected values are guaranteed to exist. This allows link functions to always link to the expected value.

4. **Floor class**

   This class of models is motivated through computational pragmatism. To transform and continuous, positive random variable $Z$ into a count variable $N$ with the floor transform, simply take the floor $N = \lfloor Z \rfloor$. Instead of linking to the expected value, $e^{X\beta}$ links to the median $\text{med}(N) = \text{med}(\lfloor Z \rfloor) = \lfloor \text{med}(Z) \rfloor$ by linking directly to $\text{med}(Z)$. Further, $N$ is capable of arbitrarily large and small variances. Continuous, positive random variables are selected for their known medians. This work transforms the Log Normal, Log Cauchy, and Log Student t distributions. Despite the Log Student t distribution generalizing both the Log Normal and Log Cauchy, it does so at the cost of an additional parameter, and so may perform worse in AIC testing.

### 3.2.3 Full model definition

In subsection 3.3.1, it is shown that the log Student-t distribution reliably achieves the lowest AIC values most often. With both the multivariate structure defined abstractly, contigent only on specification of marginal distributions, a specific model is defined as follows.

**Data:**
$Y \in \mathbb{Z}_{\geq 0}^{n \times p}$ taxa count data,
$X \in \mathbb{R}^{n \times q}$ environmental data, $q < p$,
**Parameters:**
$\beta \in \mathbb{R}^{q \times p}$ regressor weights,
$L \in \mathbb{R}^{p \times m}, \Psi \in diag(\mathbb{R}^{p \times p})$ factor model parameters,
$\boldsymbol{\sigma^2} \in \mathbb{R}_{>0}^{p}$ marginal scale parameters,
$\boldsymbol{\nu} \in \mathbb{R}_{>0}^{p}$ marginal tail parameters,
**Implicit parameters:**
$\mu = \exp(X\beta) \in \mathbb{R}_{>0}^{n \times p}$ marginal location parameters,
$\Sigma = LL^T + \Psi \in \mathbb{R}_{>0}^{p \times p}$ covariance matrix,
**Random variables:**
$T_\nu \in \mathbb{R}_{>0}^{n \times p}, [T_\nu]_{ij} \sim Student - t(\nu)$ marginal variables,
$Z \in \mathbb{R}^{n \times p}, [Z]_i \sim_{iid} Normal(0, \Sigma)$ latent Gaussians,
**Functions:**
$F_{Z_{ij}}(z) = \mathbb{P}[Z_{ij} \leq z]$ a distribution function,
$F_{Y_{ij}}(y) = \mathbb{P}[Y_{ij} \leq y]$ a distribution function,
$F_{Y_{ij}}^{-1}(p), F_{Y_{ij}}(y)$'s generalized inverse,
**Constraints:**
$Y_{ij} = \lfloor \left( \mu_{ij} e^{[T_{\nu_j}]_{ij}} \right)^{\sigma_j} \rfloor$ marginal model,
$Y_{ij} = F_{Y_{ij}}^{-1}(F_{Z_{ij}}(Z_{ij}))$ copula mechanism

Calculation of probabilities $\mathbb{P}[\mathbf{Y}_i = \mathbf{y}_i]$ are necessary during estimation protocols, but require integration. This is because for any observed $\mathbf{Y}_i$, $\mathbf{Z}_i$ is only constrained to a potentially unbounded hypercubic region in $\mathbb{R}^p$. This is a high-dimensional numerical integral which would be computationally intractable (see subsection 1.7.1) if not for the application of the factor model. Through conditional expecations (see subsection 3.2.4), the $p$-dimensional integral can be broken into $p$ $m$-dimensional integrals. This work uses $m = 3$, so the integrals are computationally tractable.

### 3.2.4 Estimation

The probability model is parameterized by $\boldsymbol{\theta} = (\beta, \log \boldsymbol{\sigma}, \log \boldsymbol{\nu}, L, \log \Psi) \in \mathbb{R}^{p(3+q+m)}$. It is flexible. Within $\theta$ there are $p(3 + q + m)$ knobs which adjust our model when turned. The likelihood function $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}[\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}]$ is the probability that the entire experimental sample $(\mathbf{y}, \mathbf{x})$ is observed under a particular parameterization $boldsymbol\theta$. For most choices of $\boldsymbol{\theta}$ the model is absurd, and for one $\hat{\boldsymbol{\theta}}$ the model is a most likely representation of the data. $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (MLE) (see subsection 1.6.1). MLEs are solutions to non-linear programs (see subsection 1.7.2). Practical solving of non-linear programs recognizes that

optimization protocols often require a good *first guess* to work reliably. Therefore MLE calculation is broken into two major steps: (1) a heuristic initialization, followed by (2) running a non-linear program solving protocol.

**Initialization**

Calculation of the initial guess for the non-linear program $\hat{\boldsymbol{\theta}}_0$ is done in series of layered heuristics. Early layers are numerically stable and inaccurate, while later layers are more numerically delicate and much more accurate. Except for the first layer, every layer needs an initial guess. This layering of methods from robust to accurate is important, because the final stages are delicate. For example, if a probability model is initialized at to an absurd initial guess $\boldsymbol{\theta}_0$, then $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\theta}_0) \approx 0$ and the computer's floating point representation system will underflow (see section 1.7), deciding the value is exactly zero. Optimization cannot occur if all values near the initial value are so absurd that they decide the data set only occurs with probability zero. The layers of estimates are the following.

1. $\hat{\beta}_0 = \text{OLS}(\log(Y+1) \sim X)$

   Calculate $\hat{\beta}_0$ with an ordinary least squares (OLS) estimation of $\log(Y+1)$ regressed on $X$. Counts are converted to log scale because the exponential link function is used. One is added to $Y$ to avoid non-finite values after the logarithm.

2. $(\hat{\boldsymbol{\sigma}}_0, \hat{\boldsymbol{\nu}}_0) = \text{MM}(Y, X; \hat{\mu}_0 = \exp[X\hat{\beta}_0])$

   With a first guess for each marginal's regressor weights $\hat{\beta}_0$, calculate the remaining marginal distribution parameters $(\hat{\boldsymbol{\sigma}}_0, \hat{\boldsymbol{\nu}}_0)$ via method of moment (MM) estimators. However, since the Student t$(\nu)$ doesn't always have a finite expected value, this work uses median-based estimates. With other models, MM estimates are sufficient and this step.

3. $(\hat{\beta}_1, \hat{\boldsymbol{\sigma}}_1, \hat{\boldsymbol{\nu}}_1) = \text{Univariate MLE}(Y, X; \hat{\beta}_0, \hat{\boldsymbol{\sigma}}_0, \hat{\boldsymbol{\nu}}_0)$

   With robust estimates computed, non-linear optimization protocols are employed. However these solvers at not applied toward the entire system, but to each marginal distribution separately. This produce a high-quality initial maximum likelihood estimates for each marginal distribution.

4. $\hat{z} = [\hat{\mathbf{z}}_j] = [\Phi^{-1}(F_{\mathbf{Y}_j}(\mathbf{Y}_j, \mathbf{X}_j; [\hat{\beta}_1]_j, [\hat{\boldsymbol{\sigma}}_1]_j, [\hat{\boldsymbol{\nu}}_1]_j))]$

   This step does not actually calculate an estimate, but prepares data for the final layer of estimation. Here, approximate samples are generate for the latent multivariate Gaussian $Z$. The approximate sample $\hat{z}$ will be used to calculate initial estimates for the copula structure, because it's easier to estimate parameters for an observed variable than an unobserved variable. This is done by calculating the distribution functions, $F_Y(Y, X; \hat{\beta}_1, \hat{\boldsymbol{\sigma}}_1, \hat{\boldsymbol{\nu}}_1)$, for each count in $Y$. The result is a matrix in $[0,1]^{n \times p}$ of probabilities which are then inverted into approximate observations with the multivariate normal distribution function $\Phi(\cdot) = F_{N_p}^{-1}(\cdot; 0, I)$.

5. $(\hat{L}_1, \hat{\Psi}_1) = \text{Eigen-decomposition Factor Estimate}(\text{Cor}(\hat{z}))$

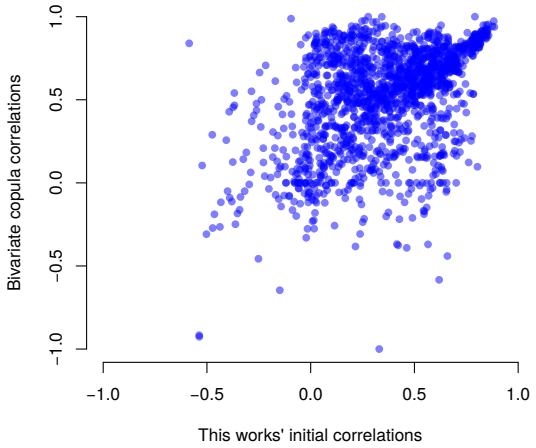



**Figure 3.3:** This work's initial correlation estimates compared to a more robust method

To initialize the $(L, \Psi)$ estimate, this work relies on a method described in Johnson and Wichern [139], section 9.3. First, the correlation matrix $R = \text{Cor}(\hat{z})$ is calculated. Then, because the model has $m$ factors, calculate the first $m$ eigen vectors $\mathbf{v}_k$ and values $\lambda_k$. Each column of $\hat{L}_1$ will be $[\hat{L}_1]_k = \sqrt{\lambda_k}\mathbf{v}_k$, and each diagonal element of $\hat{\Psi}_1$ will be $[\hat{\Psi}_1]_{jj} = I - LL^T$.

6. $\hat{\boldsymbol{\theta}}_0 = (\hat{\beta}_1, \hat{\sigma}_1, \hat{\nu}_1, \hat{L}_1, \hat{\Psi}_1)$

   The initial value for the entire model's non-linear optimization routine $\hat{\boldsymbol{\theta}}_0$ is just a concatenation of all best estimates thus far.

It is important to note that a more robust method exists for calculating $(\hat{L}_1, \hat{\Psi}_1)$. Instead, the correlation matrix which generated them could have been generated by calculating a matrix of latent correlations. Each correlation is an MLE for a bi-variate Gaussian copula model. The matrix of correlations can then be projected to a nearest covariance matrix, then converted to a correlation matrix. This robust method is compared with the initialization proceedure employed by this work in Figure 3.3. The estimates have correlation of 0.419, and so have an operable similarity in this case. These initial estimates need only be close enough, since their differences would be eliminated after applying a non-linear optimizing protocol. The bi-variate copula method is more robust, because it works for small $p$, where the implemented method may not.

### Non-linear optimization

With an initial estimate $\hat{\theta}_0$, non-linear optimization may begin. Ideally, this work would employ the L-BFGS routine (subsection 1.7.2) with high-quality gradient calculations, but for for the sake of project brevity, a series of univariate optimizations are performed with simple Newton-Raphson optimizations, and five-point stencils are used to estimate first and second derivatives (see subsection 1.7.1). While this software is interfaced through R [221], GPU acceleration makes the pre-implemented L-BFGS routine cannot be used due to inefficient job batching. The protocol requests too few simultaneous jobs to overcome warp divergence (see subsection 1.7.3).

GPU acceleration is used for this project, because of substantial compute requirements. There are three major causes of substantial compute requirements. First, the hypercubic integrals (see subsection 3.2.3) are calculated with MC-integration (see subsection 1.7.1). Second, statistical significance for correlation tests $H_0 : \rho_{ij} = 0$ is computed through bootstrapping. Third, many networks are generated for a full-factorial simulation study (see subsection 3.2.5) is used to evaluate the method. GPU calculation is motivated over grid compute methods such as Map Reduce [68] or MPI [112], because the overall floating point operations per second were expected to be higher.

The primary caveat in programming for GPUs is that warp divergence must be minimized. Warp divergence, the issuance of different instructions to the same batch of threads, can cause a GPU compute job to run slower than if run on a CPU. To avoid warp divergence, code execution patterns must be predictable. Divergences must be organized into separate thread blocks. This work achieves predictable log likelihood $\log f_{Y|X}(y|x;\theta)$ calculation primarily through two strategies: (1) $\log f_{Y|X}(y|x;\theta)$ is algebraically broken into summable components, allowing separate calculations to be organized accordingly; and (2) integration over the $m$-dimensinal Gaussian hypercubes is calculated with an MC-integral, likening it to the work of Genz [101], rather than an unpredictable quadrature routine [217]. The algebraic decomposition of $\log f_{Y|X}(y|x;\theta)$ into summed components is done as follows.

$$
\begin{aligned}
&\log f_{Y|X}(y|x;\boldsymbol{\theta}) = \log \mathbb{P}[Y = y|X = x] \\
&= \log \mathbb{P}[Y = y] \text{ (suppress conditional notation for brevity)} \\
&= \log \textstyle\prod_{i=1}^{n} \mathbb{P}[\mathbf{Y}_i = \mathbf{y}_i] = \sum_{i=1}^{n} \log \mathbb{P}[\mathbf{Y}_i = \mathbf{y}_i] \\
&= \textstyle\sum_{i=1}^{n} \log \left( \mathbb{P}(\mathbf{Z}_i \in [\mathbf{a}, \mathbf{b}]_i) \prod_{j=1}^{p} \mathbb{P}[Y_{ij} = y_{ij}] \right) \\
&= \textstyle\sum_{i=1}^{n} \left( \log \mathbb{P}(\mathbf{Z}_i \in [\mathbf{a}, \mathbf{b}]_i) + \sum_{j=1}^{p} \log \mathbb{P}[Y_{ij} = y_{ij}] \right) \\
&= \textstyle\sum_{i=1}^{n} \left( \log \mathbb{P}(L\mathbf{F}_i + \Psi^{1/2}\mathbf{E}_i \in [\mathbf{a}, \mathbf{b}]_i) + \sum_{j=1}^{p} \log \mathbb{P}[Y_{ij} = y_{ij}] \right) \\
&= \textstyle\sum_{i=1}^{n} \left( \log \mathbb{E}[\mathbb{P}(\Psi^{1/2}\mathbf{E}_i \in [\mathbf{a}, \mathbf{b}]_i - L\mathbf{F}_i | \mathbf{F}_i)] + \sum_{j=1}^{p} \log \mathbb{P}[Y_{ij} = y_{ij}] \right) \\
&=_{a.s.} \textstyle\sum_{i=1}^{n} \left( \log \lim_{K\to\infty} K^{-1} \sum_{k=1}^{K} \mathbb{P}(\Psi^{1/2}\mathbf{E}_i \in [\mathbf{a}, \mathbf{b}]_i - L\mathbf{F}_i | \mathbf{F}_i = \mathbf{f}_{ik}) + \sum_{j=1}^{p} \log \mathbb{P}[Y_{ij} = y_{ij}] \right) \\
&= \lim_{K\to\infty} \textstyle\sum_{i=1}^{n} \left( -\log K + \log \sum_{k=1}^{K} \mathbb{P}(\Psi^{1/2}\mathbf{E}_i \in [\mathbf{a}, \mathbf{b}]_i - L\mathbf{F}_i | \mathbf{F}_i = \mathbf{f}_{ik}) + \sum_{j=1}^{p} \log \mathbb{P}[Y_{ij} = y_{ij}] \right) \\
&= \lim_{K\to\infty} \textstyle\sum_{i=1}^{n} \left( -\log K + \log \sum_{k=1}^{K} \prod_{j=1}^{p} \mathbb{P}(\Psi_{jj}^{1/2} E_{ij} \in [a, b]_{ij} - [L\mathbf{F}_i]_j | \mathbf{F}_i = \mathbf{f}_{ik}) + \sum_{j=1}^{p} \log \mathbb{P}[Y_{ij} = y_{ij}] \right) \\
&= \lim_{K\to\infty} \textstyle\sum_{i=1}^{n} \left( -\log K + [LS]_{k=1}^{K} \log \prod_{j=1}^{p} \mathbb{P}(\Psi_{jj}^{1/2} E_{ij} \in [a, b]_{ij} - [L\mathbf{F}_i]_j | \mathbf{F}_i = \mathbf{f}_{ik}) + \sum_{j=1}^{p} \log \mathbb{P}[Y_{ij} = y_{ij}] \right) \\
&= \lim_{K\to\infty} \textstyle\sum_{i=1}^{n} \left( -\log K + [LS]_{k=1}^{K} \sum_{j=1}^{p} \log \mathbb{P}(\Psi_{jj}^{1/2} E_{ij} \in [a, b]_{ij} - [L\mathbf{F}_i]_j | \mathbf{F}_i = \mathbf{f}_{ik}) + \sum_{j=1}^{p} \log \mathbb{P}[Y_{ij} = y_{ij}] \right)
\end{aligned}
$$

where $[\mathbf{a}, \mathbf{b}]_{ij} = F_{Z_{ij}}^{-1} \left( G_{Y_{ij}}(y_{ij}; \mu_{ij}, \sigma_j, \nu_j); [LL^T + \Psi]_{jj} \right)$,
and $G_{Y_{ij}}(y_{ij}; \mu_{ij}, \sigma_j, \nu_j) = \{u \in [0, 1] : y_{ij} = F_{Y_{ij}}^{-1}(u; \mu_{ij}, \sigma_j, \nu_j)\}$,
and $[\mathbf{a}, \mathbf{b}]_i = \times_{j=1}^{p} [a, b]_{ij}$ is a hypercube,
and $\mathbf{f}_{ik} \sim N_m(0, I_m)$ is an MC-simulant,
and "$=_{a.s.}$" is equivalence with probability one,
and the $\lim_{K\to\infty}$ is applied via the strong law of large numbers,
and $K$ is the number of MC iterates,
and $[LS]_{j=1}^{p}$ is a log sum.

A log sum possible by applying the identity $\log(a+b) = \log a + \log\left(1 + \exp[\log b - \log a]\right)$. Log sums are used to maintain logarithmic scale during integration. Logarithmic scale is important because the high-dimensional integral is over small-measured (potentially unbounded) hypercubes in $\mathbb{R}^p$. The hypercubic probabilities are so small that they tend to cause underflows if not kept on log scale.

An additional detour from popular methods is required to avoid warp divergence. This algorithm must compute many Student $\mathrm{t}(\nu)$ distribution functions, which is usually calculated via the following identity.

$$\int_{-\infty}^{t} f_T(u;\nu)du = 1 - I_{\frac{\nu}{t^2+\nu}}\left(\frac{\nu}{2}, \frac{1}{2}\right), \text{ where } I_x(a,b) = \frac{B(x;a,b)}{B(1;a,b)}, B(x;a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt$$

The $I_x(a,b)$ function is the incomplete beta ratio function. On CPUs processing in double precision, $I_x(a,b)$ is calculated with the TOMS708 library [39, 75]. This work's software computes in single precision (which is faster for many GPUs), and cannot afford to break the computation into the many cases used by the TOMS708 algorithm, because unpredictable code execution results in warp divergence. To remedy this problem, $I_x(a,b)$ is broken into the following three easy-to-predict cases, and computed to single-precision accuracy.

1. When $\nu > 10^5$, a normal-distribution approximation is used.
2. When $t^2 < \nu$, the identity $I_x(a+1,b) = I_x(a,b) + \frac{x^a(1-x)^b}{aB(1;a,b)}$ is iterated.
3. Otherwise apply an older t-distribution algorithm. Particularly, ACM Algorithm 395 is applied [129].

**Statistical significance via bootstrapping**

All estimation exists to serve statistical testing, which effectively decides which taxa-pairs are correlated or not. With one test for significance per correlation $H_0 : \rho_{ij} = 0$, likelihood ratio tests are infeasible, because it would require calculating a different MLE per $\rho_{ij}$. A good alternative might be employing a normal approximation with observed inverse Fisher Information estimating covariances. However with so many parameters, it is not obvious that asymptotic assumptions ($n \to \infty$) are truly satisfied. Further, testing for $L$ or $\Psi$ significance is insufficient, because the implicit $\Sigma = LL^T + \Psi$ must be tested per-entry instead. Fisher Information of transformed variables is calculable but assumes the existence of many derivatives, which may not actually be well-defined. With so many uncertainties, a bootstrapping [83] approach is motivated (see subsection 1.6.5). Employing the bootstrap overcomes delicate theory with expensive computational work. The bootstrap works by randomly resampling from the data set and calculating an estimate $\hat{\boldsymbol{\theta}}$ per resample, thereby empirically reconstructing the sampling distribution.

In subsection 1.3.1, the compositional and mixed effect perspectives of SSU rRNA multivariate structure were described. Both are effects which obfuscate the community's authentic correlation structure. The primary strategy of this chapter is to increase correlation-attribution precision through the factor model's parameter reduction, but the factor model also implicitly agrees with the mixed effect perspective. The factor model $\mathbf{Z}_i \sim_{iid} N_p(0, LL^T + \Psi)$ is implicitly equivalent to

the sum of two Gaussian random variables $(\mathbf{F}_i, \mathbf{E}_i) \in \mathbb{R}^m \times \mathbb{R}^p$, via $\mathbf{Z}_i = \mathbf{F}_i L + \mathbf{E}_i \Psi^{1/2}$. Each of the $m$ dimensions of $\mathbf{F}_i$ is a called a *factor*. This allows any of the $m$ dimensions of each $\mathbf{F}_i$ to act as a mixed effect. In subsection 3.3.2, this work also builds evidence toward the hypothesis that Saanich Inlet SSU rRNA data does indeed follow the mixed effect perspective, and also argues that the first dimension of $\mathbf{F}_i$ models the mixed effect of sequencing depth.

Embracing the mixed effect perspective, this work modifies bootstrapped estimates of $\Sigma$ by removing the mixed effect. This is equivalent to separating $L$ into two components $L = [L_1, L_{-1}]$, where $L_1$ is the first column of $L$ and $L_{-1}$ is the matrix of the remaining $m - 1$ columns of $L$. So while $\Sigma = LL^T + \Psi$, this work actually estimates bootstrapped values of $\Sigma_{-1} = L_{-1}L_{-1}^T + \Psi$.

### 3.2.5 Precision-recall comparison

As described in subsection 3.1.2, a simulation study is used to objectively compare correlation network estimation methods via the metric of precision-recall exchanges. However, many variables important exist beyond the choice of correlation statistic. Further, all three correlation statistics compared are designed for different use cases. To understand the inter-dependencies between these variables, the simulation study will be conducted as a *full-factorial* experiment. Full-factorial experiments follow a special experimental design which enables for the full testing of all pair-wise dependencies between variables [188]. Recall each correlation statistic is only a classifier for a specific Type-I error rate $\alpha$ (p-value cutoff). Note that the full-factorial experiment is replicated once per $\alpha \in \{0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4\}$, and precision-recall curves are generated over the $\alpha$ values. The variables of the full-factorial experiment are the following.

1. Statistic: Regularized, SparCC, or Pearson
2. Number of samples, one of 200 or 1000.
3. Regressor effects: yes or no. If yes, simulants' marginal location parameters are regressed against environmental parameters. Environmental measurements and $\beta$ weights are randomly selected from actual model fits.
4. Sparsity: one of 0.01 or 0.5. This is the proportion of simulants' correlation structure elements which are non-zero.
5. Simulant model: compositional or mixed effect. Specific probability models are used to generate the random data with known correlations. One model follows the compositional perspective, while the other follows the mixed effect perspective. Model definitions are provided below. Note that neither model agrees perfectly with the specifications of either this work's regularized method, nor SparCC.

The *compositional* model is a logistic normal multinomial (LNM) [277]. It is specified as follows.

**Data:**
$Y \in \mathbb{Z}_{\geq 0}^{n \times p}$ observed counts,
$X \in \mathbb{R}^{n \times q}$ regressors,
**Parameters:**

$\beta \in \mathbb{R}^{q \times p}$ regressor weights,
$\Sigma \in \mathbb{R}^{p \times p}$ covariance matrix,
**Implicit parameters:**
$\mu \in \mathbb{R}^{n \times p}, \mu_{ij} = e^{\mathbf{X}_i \boldsymbol{\beta}_j}$ regressed means,
$P \in (0,1)^{n \times p}, P_{ij} = e^{Z_{ij}} / (1 + \sum_{j=1}^{p} e^{Z_{ij}})$ probabilities,
**Random variables:**
$Z \in \mathbb{R}^{n \times p}, Z_i \sim_{iid} N_p(\mu_i, \Sigma)$ latent Gaussians,
$N \in \mathbb{Z}_{>0}^{n}, N \sim_{iid}$ Empirically sampled,
$\mathbf{Y}_i \sim_{iid} \text{Multinomial}(\mathbf{P}_i, N_i)$

The *mixed effect* model uses a copula mechanism (see subsection 1.6.4) to join negative binomial distributions to a multivarite structure with a mixed effect. This work refers to it as the negative binomial with Gaussian copula (NBGC) (see section 1.6.2 for the Negative Binomial definition used). It is defined as follows. Notice that the mixed effect componenet obfuscates the underlying correlation structure.

**Data:**
$Y \in \mathbb{Z}_{\geq 0}^{n \times p}$ observed counts,
$X \in \mathbb{R}^{n \times q}$ regressors,
**Parameters:**
$\beta \in \mathbb{R}^{q \times p}$ regressor weights,
$\Sigma \in \mathbb{R}^{p \times p}$ covariance matrix,
$\sigma^2 \in \mathbb{R}_{>0}^{p}$ marginal variances,
$\mathbf{L} \in \mathbb{R}^{p}$ mixed effect vector,
**Implicit parameters:**
$\mu \in \mathbb{R}_{>0}^{n \times p}, \mu_{ij} = e^{X_i \beta_j}$ marginal expected values,
**Random variables:**
$Y_{ij} \sim \text{NegativeBinomial}(\mu_{ij}, \sigma_j^2)$ marginal distributions,
$Z \in \mathbb{R}^{n \times p}, \mathbf{Z}_i \sim_{iid} N_p(0, \mathbf{L}\mathbf{L}^T + \Sigma)$ latent Gaussians,
**Functions:**
$F_{Y_{ij}}(y_{ij}; \mu_{ij}, \sigma_j^2)$ marginal distribution function,
$F_{Y_{ij}}^{-1}(p; \mu_{ij}, \sigma_j^2)$ generalized inverse of a marginal distribution function,
$F_{Z_j}(z_{ij}; 0, [\mathbf{L}\mathbf{L}^T + \Sigma]_{jj})$ latent Gaussian's marginal distribution function,
**Contstraints:**
$Y_{ij} = F_{Y_{ij}}^{-1}(F_{Z_j}(Z_{ij}; 0, [\mathbf{L}\mathbf{L}^T + \Sigma]_{jj}); \mu_{ij}, \sigma_j^2)$

### 3.2.6 Saanich Inlet

As described in subsection 3.1.3, objective evaluation of the regularized correlation method is extended beyond a simulation study and into real-world SSU rRNA data. Using data from Saanich Inlet (see section 1.4), finding can be contextualized amongst current models of the denitrifying community's structure. Further, the data provides meaningful evidence in evaluating which univariate count models best-describe SSU rRNA data (see subsection 3.3.1), and also in

evaluating whether a compositional or mixed effect perspective better-describes the data (see subsection 3.3.2).

A total of 91 SSU rRNA samples are used, sampled from 2007 to 2011, from the depths of 10m to 215m. Concentrations of $O_2$, $NO_3^-$, and $H_2S$ are paired with each sample. The SSU rRNA data are converted to multivariate count data with QIIME (see subsection 1.2.3). From a regression perspective the SSU rRNA data are represented in the dependent (count) variable $Y \in \mathbb{Z}_{\geq 0}^{n \times p}$, and the environmental concentrations are the regressors (independent variables) $X \in \mathbb{R}^{n \times q}$, where $n = 91$, $q = 3$, and $p = 57$. Regessors used are concentrations of $O_2$, $NO_3^-$, and $H_2S$, because they are known to be important variables. Depth is omitted as a regressor, because it is strongly correlated with $O_2$, and thus might decrease statistical power in later testing. Choosing $q = 3$ is justified through compute time restrictions. These data are processed through the estimation pipeline described in subsection 3.2.4, and statistical significance is decided with 1000 bootstrap iterates (see subsection 1.6.5).

**Reducing phylogenetic resolution**

Choosing $p = 57$ is motivated by the primary contribution of chapter, precise correlations through parameter reduction (illustrated in subsection 3.1.1). Pragmatically, the $O(p^2) \to O(p)$ reduction still requires $p$ to be at least comparable to $n$. So while estimating a model of the whole community is desirable, it is not yet feasibly estimable. Keeping the number of taxa $p$ small can be achieved through two methods: (1) throwing out clades, and (2) summing up SSU rRNA counts within clades. The advantage of throwing out clades is that high-resolution descriptions of remaining clades will be made possible, but it runs the risk of falsely attributing correlations due to omitted taxa (see section 1.5.3). So discarding information runs the risk of supporting false interpretations which are otherwise avoidable. Instead this work takes the second option, summing counts within clades, because it reduces the chance of false inference only at the cost of a coarser-grain description of the system. Fortunately clades needn't always be selected at the same taxonomic level, allowing greater description of a few target taxa. All clades are summed up to the phylum level, except for SUP05, Nitrospina, and Nitrospira. In total this produces $p = 57$ counted clades. Just as in precision-recall exchanges, the desire for descriptive breadth will always be at odds with confident inference. In section 4.3, an $O(\log p)$ reduction is explored, but it comes at the cost of even more complex modelling, and thereby increased abstraction in results.

## 3.3 Results

### 3.3.1 Marginal model selection

In determination of which univariate marginal model best describes observed SSU rRNA, several univariate count distributions were surveyed with AIC statistics (see subsection 3.2.2). Since there are many dimensions of SSU rRNA counts, each model has several AIC statistics. Descriptive statistics are provided in Table 3.1. AIC values behave like a measurement of error, the best model will have consistently small AIC values. Because all marginal models are joined together in a single multivariate model, a marginal model with only a few very large AIC values (high-variance dimensions) is inadmissible, because it threatens the entire model's ability to fit. Extremely

**Table 3.1:** AIC score statistics per marginal model

| | Negative Bin. | Hagmark Gam. | Hagmark Log N. | Floor Log N. | Floor Log Cau. | Floor Log t |
|---|---|---|---|---|---|---|
| Min | 13.64 | 13.81 | 14.49 | 13.56 | 13.36 | 14.00 |
| 1st Quant. | 14.23 | 14.32 | 14.87 | 14.12 | 14.18 | 14.23 |
| Median | 15.09 | 14.97 | 15.25 | 14.71 | 14.83 | 15.11 |
| Mean | 41.56 | 18.53 | 15.99 | 15.50 | 15.71 | 15.60 |
| 3rd Quant. | 19.85 | 16.87 | 16.18 | 15.93 | 16.15 | 16.29 |
| Max | 4043.07 | 640.69 | 33.82 | 33.83 | 42.95 | 26.47 |

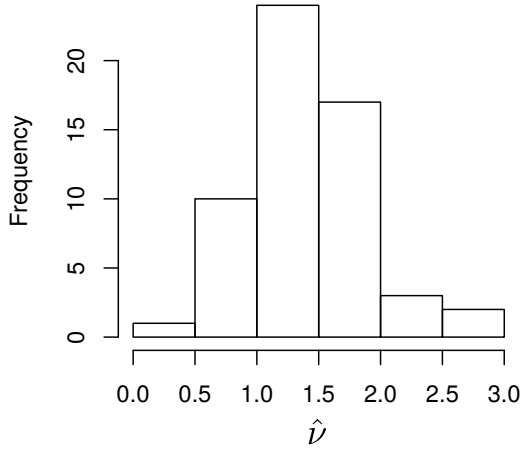



**Figure 3.4:** Histogram of final estimated values $\hat{v}$

unlikely values can cause a few data points to overpower the model or simply fail estimation through underflow errors (see section 1.7).

The selected marginal model is the Floor Log t distribution, because it consistently achieved the lowest AIC values. This model has one more parameter than the others, which permits it more flexibility. Fortunately, the AIC statistic accounts for the number of parameters, so the low AIC values indicate the additional parameter is motivated. The additional parameter is the Student-t's shape parameter $\nu > 0$, which controls the thickness of the model's tail. A distribution with thick tails has more extreme values. A Student-t model has more extreme values when $\nu$ is small. This can be seen analytically, because its expected value is undefined for $\nu \leq 1$, and its variance is undefined for $\nu \leq 2$. With these special values in mind, the histogram of all estimates $\hat{v}$ (see Figure 3.4) meaningfully demonstrates that all distributions exhibit heavy tails, and many don't even have defined variances.

### Descriptive statistics

Interpretation of the AIC statistics is easiest when contextualized in descriptive statistics of the actual data, because the different models are best-suited to different distributions of data. The data is characterized by strongly skewed count distributions as illustrated by histograms A, B, & C of Figure 3.5. This skew shape is supported by all candidate models, but not necessarily all are capable of capturing the extreme skew. Another important quality of the data is that observed variances can be much larger than their corresponding means, as illustrated by $\hat{\sigma}/\hat{\mu}$

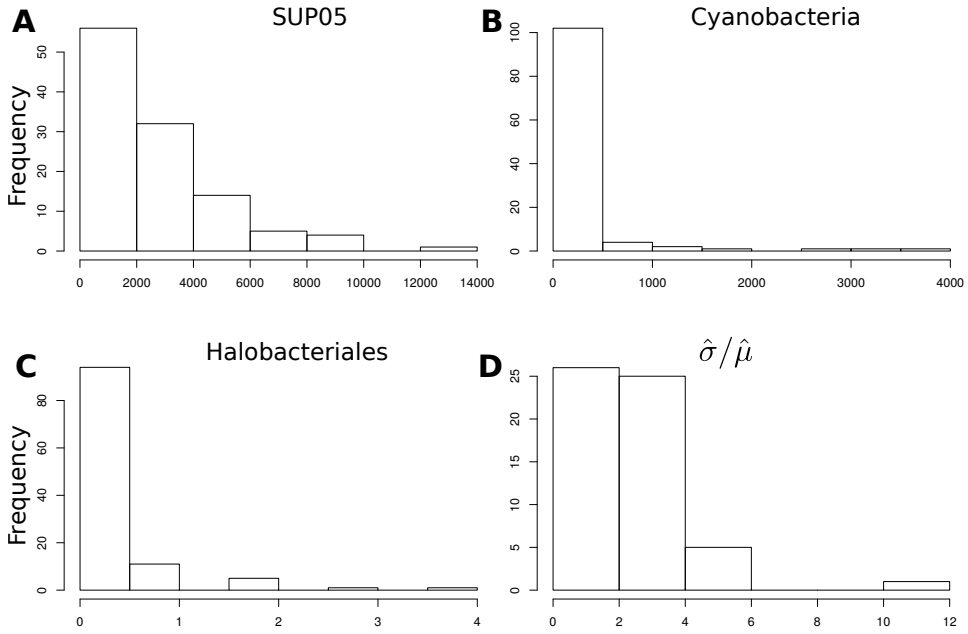



**Figure 3.5:** Descriptive statistics for the SSU rRNA data set

ratios in histogram D of Figure 3.5. Having standard deviations 6 or 12 times larger than the mean highlight the importance of considering models capable of infinite variances, particularly Cauchy or Student-t distributions. Another important statistic is that of all counts, 40% are zero. Thus selected model must be both capable of very large values amongst many zeros. Again, generating count models from infinite-variance-capable distributions makes this quality possible. Finally, histograms A shows that the SUP05 distribution's skew is not so extreme, thereby supporting a model with is capable of finite variances. Of the models considered, such flexibility is only possible through the Floor Log t distribution, which the AIC statistic favoured.

### 3.3.2 Compositional vs. mixed effect perspective

In subsection 1.3.1, two perspectives on SSU rRNA multivariate structure are proposed: the compositional and mixed effect perspectives. Under the compositional perspective, taxa are expected to compete for sequencing depth, and thus have a negative correlative effect included upon their counts. Under the mixed effect perspective, taxa counts rise and fall together with sequencing depth, and thus have a positive correlative effect induced upon their counts. Both are obfuscating effects that need to be controlled for, if present. Of the $p = 57$ taxa observed, 75% of their observed correlations are positive, suggesting the mixed effect perspective is correct.

Observing a majority of positive correlations does not conclusively support a mixed effect perspective, because it is merely an unobserved and linear latent effect, and a compositional mechanism may be overpowered by a highly inter-dependent community. The strongest evidence for the data following a mixed effect perspective is in Figure 3.6. The $\mathbf{L}_1$ vector has statistically significant, positive values for nearly every taxa. Because $\mathbf{L}_1$ is a factor model weight, the data support the existence of a single, linear effect shared amongst the taxa. Such is the exact design of the mixed effect perspective. These findings describe Saanich Inlet SSU rRNA multivariate

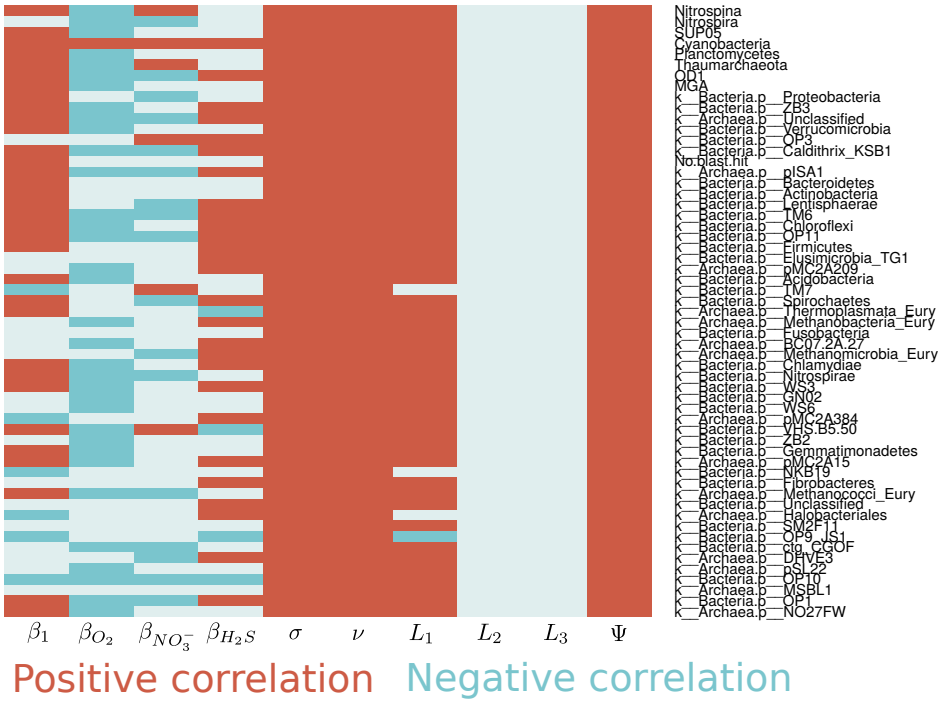



**Figure 3.6:** Statistically significant parameter values per taxa, testing equality with zero. Each $\boldsymbol{\beta}_x$ describes regressor weight against variable $x$. For example, $\boldsymbol{\beta}_1$ is the intercept, $\boldsymbol{\beta}_{O_2}$ is the weight against $O_2$ concentration, and so on. The parameters $\boldsymbol{\sigma}, \boldsymbol{\nu}$ and $\Psi$ must always be positive. Majority-positive values for $\mathbf{L}_1$ demonstrate the observation of a mixed effect. The lack of significant values for $\mathbf{L}_2$ and $\mathbf{L}_3$ does not stop their associated covariance matrix $\Sigma = L_{-1}L_{-1}^T + \Psi$ from attaining significant values.

structure as following the mixed effect perspective.

### 3.3.3 Exploring GPU necessity

Setting up a CUDA-enabled GPU is not always easy, so it is worth exploring the necessity of GPU-acceleration in estimating for the regularized model. The estimation scheme is a two-stage process, with an initialization step and non-linear optimization step (see subsection 3.2.4). The initialization step is fast and does not require a GPU, whereas the non-linear optimization step is slow and does require a GPU. Avoiding the non-linear optimization process would not just simplify hardware requirements, but also drastically shorten compute times. For example, the two-week-long computational experiment described in subsection 3.2.5 would have taken under one day without non-linear optimization. That said, the initialization produces a merely approximate estimate, whereas applying the GPU allows for optimal estimates. Therefore the GPU is only valuable if MLE optimality is.

Having fully applied the GPU and its associated CUDA software, this study now has bootstraps which describe the sampling distribution produced from Saanich Inlet SSU rRNA data (see subsection 3.2.6). This data is valuable in deciding the necessity of GPU-driven non-linear optimization. By comparing an initialized estimate $\hat{\boldsymbol{\theta}}_0$ to the sampling distribution of $\hat{\boldsymbol{\theta}}$, it can be decided if $\hat{\boldsymbol{\theta}}_0$ falls within the natural variation fully optimized estimates. If it does not, then $\hat{\boldsymbol{\theta}}_0$

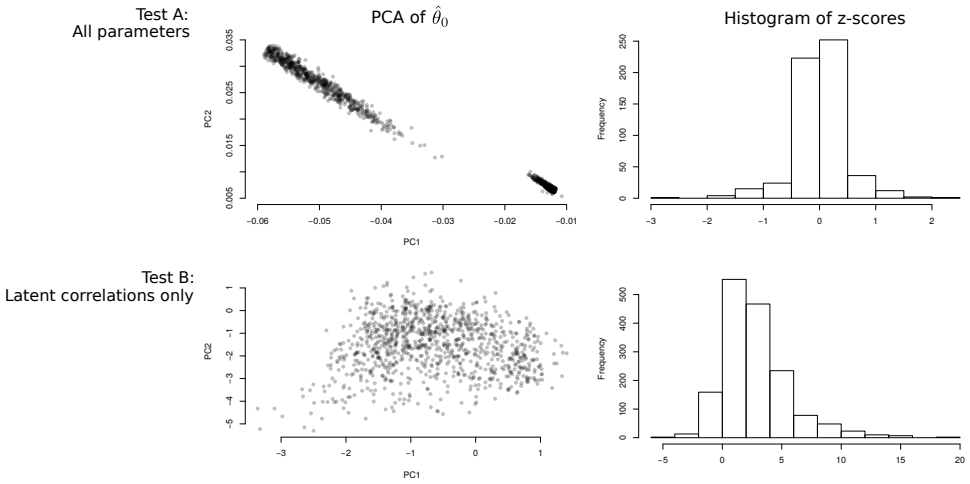



**Figure 3.7:** Testing the necessity of GPU acceleration in estimation. Test A shows GPU acceleration is not necessary for general model parameters. Test B shows GPU acceleration is necessary for correctly estimating correlations.

certainly has sufficient bias to obfuscate real results.

The problem is scrutinized through a combination of hypothesis testing and descriptive statistics. Asymptotic statistical theory provides the following approximation, which roughly applies to this problem. The hypothesis test assumes that the bootstraps $\hat{\boldsymbol{\theta}}$ follow a multivariate Gaussian distribution with empirical mean $\bar{\boldsymbol{\theta}}$. As shown in PCA A of Figure 3.7 this assumption imperfectly applied, though it is justifiable in PCA B.

$$\hat{t} = (\hat{\boldsymbol{\theta}}_0 - \bar{\boldsymbol{\theta}})^T \hat{\Sigma}^{-1} (\hat{\boldsymbol{\theta}}_0 - \bar{\boldsymbol{\theta}}) \sim \chi^2_p$$

Two tests are conducted to demonstrate the nuanced effect the GPU provides. In test **A**, $\boldsymbol{\theta}$ is tested in its $\boldsymbol{\theta}_A = (\beta, \log \boldsymbol{\sigma}, \log \boldsymbol{\nu}, L, \log \Psi)$ form, thereby describing estimation quality for the entire model's structure. In test **B**, only a subset of the parameters are used $\boldsymbol{\theta}_B = (L, \log \Psi)$, and are transformed into latent correlation estimates $\{\hat{\rho}_{ij}\}_{i \neq j}$, thereby describing estimation quality only for the correlative structure. In both cases, the statistics $\hat{\Sigma}$ and $\bar{\boldsymbol{\theta}}$ are estimated from the bootstraps $\hat{\boldsymbol{\theta}}$, the approximate Gaussian distribution of $\hat{\boldsymbol{\theta}}$ is scrutinized through PCA, and a robust argument is provided through z-scores $(\hat{\boldsymbol{\theta}}_0 - \bar{\boldsymbol{\theta}})/(\text{diag}(\Sigma))^{1/2}$. The null hypothesis is always $H_0 : \hat{\boldsymbol{\theta}}_0 =_D \hat{\boldsymbol{\theta}}$.

The combined approach of hypothesis testing and z-score histograms provides different argumentative qualities, and ultimately agree in their results. Hypothesis testing is theoretically powerful but logically delicate and prone to assumption failures, whereas the z-score histograms provide logically robust but ambiguous results. The hypothesis tests are imperfectly applied. In test A, $\hat{\boldsymbol{\theta}}$ is not Gaussian distributed as indicated in Figure 3.7 PCA A, where a bimodal distribution is shown and existing outliers are hidden. In test B, the mere 1000 bootstraps is insufficient to estimate $\hat{\Sigma}$ due to the $p(p-1)/2 = 1596$ correlations it describes. Instead, $\hat{\Sigma}^{-1}$ is constructed through diagonalization omitting any eigenvalues less than $2.22 \times 10^{-14}$ (which happens to be the first 999 vectors). The hypothesis test A fails to reject with a p-value rounding off to one (supporting $H_0 : \hat{\boldsymbol{\theta}}_{A0} =_D \hat{\boldsymbol{\theta}}_A$), and hypothesis test B rejects with a p-value rounding

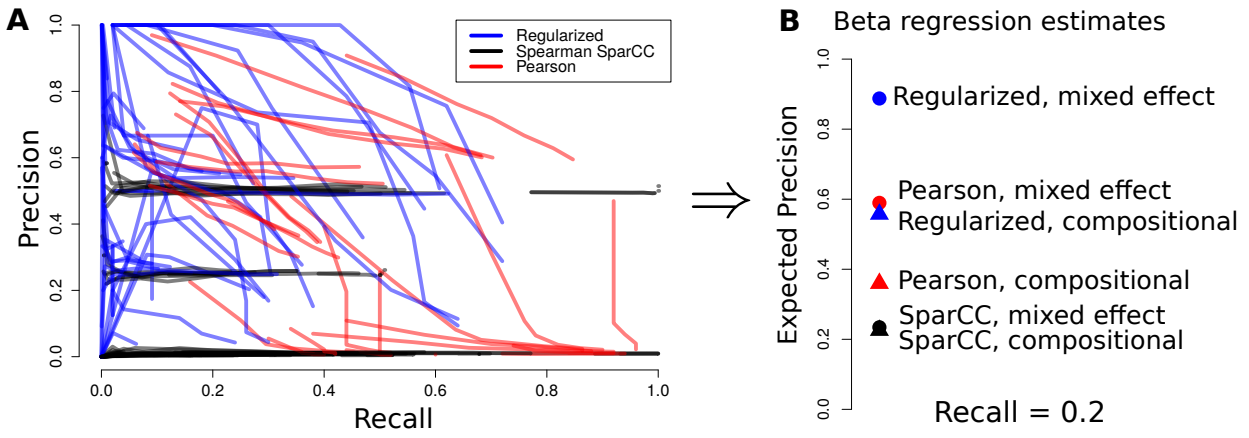



**Figure 3.8:** (A) Precision-recall curves, (B) Expected precisions after beta regression

off to zero (rejecting $H_0 : \hat{\theta}_{0B} =_D \hat{\theta}_B$). Similarly, 0.3% of test A's z-scores fall outside of the univariate Gaussian's 95% confidence interval (supporting $H_0 : \hat{\theta}_{A0} =_D \hat{\theta}_A$), while 56.2% of test B's z-scores fall outside of the confidence interval (rejecting $H_0 : \hat{\theta}_{0B} =_D \hat{\theta}_B$). These results support the conclusion that GPU acceleration is necessary for extracting fine correlative structure, while it is unecessary for describing general model structure.

### 3.3.4 Precision-recall comparison

The simulation study described in subsection 3.2.5 yields precision-recall curves illustrated in Figure 3.8 (A). The full-factorial experimental design is great for building strong inferential arguments, but they produce enough complicated data that a regression analysis is usually required to correctly interpret findings. In this case, the simulation study produces so many precision-recall curves, that while a few suggested themes can be seen, over-plotting obfuscates clear conclusions, and beta regression is used to clean up the findings in Figure 3.8 (B).

The experiment reveals the following observations.

1. Under the mixed effect model perspective (which is supported by data, see subsection 3.3.2), the regularized model greatly improves precisions. In Figure 3.8 (B), for a recall of 20%, the expected precision is about 90%.

2. Under the compositional perspective, the regularized model performs comparably to the best of other methods on average. In Figure 3.8 (B), for a recall of 20%, the expected precision drops to about 55% under the compositional perspective. It is worth noting that the Pearson method's precision drops as well. It is possible that drop may be partially due to simulant model selection. Particularly, the mixed effect simulates from the LNM, which hides correlation structure behind a mixture model, and thereby should add variance, making inference less efficient. However, the regularized model inherently follows the mixed effect perspective, and likely had built-in biases that could cause it to under-perform under a compositional condition.

3. Computing the regularized portion of the experiment took two weeks, whereas the other methods took no longer than two days. This demonstrates that the precision increases are

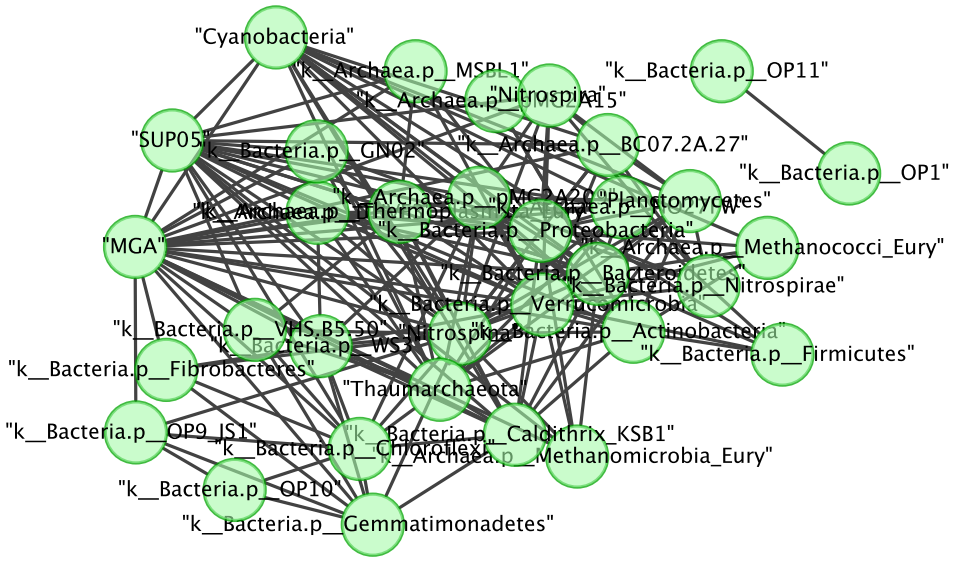



**Figure 3.9:** All statistically significant correlations

not free, and come at the cost of increased compute requirements. Despite the substantially increased compute time, algorithmic complexity only grows linearly $O(p)$ with $p$ taxa, so larger jobs should scale well if a GPU grid may be used. The experiment was run over $2^4$ conditions $\times 2$ replicates per statistic. The software was run on a single terminal with an AMD FX(tm)-8320 Eight-Core Processor, 32GB RAM, GeForce GTX 980 Ti GPU, running Ubuntu 14.04.

4. Pearson outperformed SparCC. This is likely due to a combined effect of two factors. First, sample sizes tested were large (200 or 1000), and because Pearson correlations are consistent, they should be fairly accurate. Second, SparCC precisions tended to cluster, which is likely a symptom of its known tendency to over-attribute correlations (see figure 1b of Friedman and Alm [95]). The precisions of any method which attributes too many correlations will by dictated more by the natural abundance of correlations, than the method itself.

### 3.3.5 Saanich Inlet

As described in subsection 3.1.3, the Saanich Inlet SSU rRNA data set is run through this work's regularized method so that it may (1) be contextualized against previous descriptions of the Saanich Inlet denitrifying community, (2) be evaluated for its ability to recover observed community structure, and (3) further constrain understanding of how denitrification is occurring via SUP05's partial denitrification. The estimated correlation network is not simple (see Figure 3.9), and requires further processing so that it may be digested. It might be tempting to simply discard nodes from the network that seem disinteresting, but such an approach destroys information. For example, if edges in the remaining network are merely tertiary correlations (see section 1.5.3), then the discarded nodes are actually deciding the observed correlations. So simply discarding taxa risks ignoring the primary drivers of the community's covariational structure.

Network subsetting can incorporate all relevant covariational information by subsetting with

partial covariance decompositions (see section 1.5.3). To ensure that only statistically significant edges are observed in the sub-network, bootstraps are calculated for partially decomposed $\Sigma_{-1}$ matrices (see section 3.2.4). The resulting sub-network only includes decided taxa as nodes, but the whole system's information is represented via the exclusion of edges strictly due to discarded taxa. An illustration of the Saanich Inlet denitrifying community's sub-network is shown in Figure 3.10. The representation of community-wide information in the sub-network can be seen through including special nodes. For example, including taxa without know participation in denitrification does not add edges, as was the case for *OD1, Chlamydiae, Methanococci Eury*, and *Acidobacteria*. Further, edges can be added by including groups which superset taxa which ought to be correlated for phylogenetic reasons, as was the car for *No.blast.hit, Unclassified Bacteria*, and *Proteobacteria*.

An nuanced caveat for interpreting Figure 3.10 is that the entire analysis is conditioned on the environmental observations ($O_2$, $NO_3^-$, $H_2S$). This is due to the model's origination from a multivariate regression paradigm (see section 1.6.2). While tha taxa's environmental and partial correlations are estimated simultaneously, the environmental correlations take precedence, because the partial correlations are only able to describe system variation that is not due to environmental variables. The mechanism is very similar to a partial correlation decomposition (see section 1.5.3), however inequivalent due to the non-Gaussian nature of the count data. This nuanced perspective is important, because it provides a deeper perspective into postulated ecological dependences which might be attributed to environmental accumulations of public goods. For example, the hypothesis of sulfur-driven denitrification via SUP05 and Marinimicrobia is strengthened, because the organisms share a partial correlation while not correlating with environmental $H_2S$.

Several observations are relevant to the motivations described in subsection 3.1.3. First, leveraging the method's ability to more-precisely survey for correlations, it is notable that SUP05 shares a partial-correlative edge with Marinimicrobia, but not with Planctomycetes. It is notable that all three are negatively correlated with $O_2$ concentrations, but have no significant interactions with $NO_3^-$. Second, the environmental correlations agree with known environmental abundances illustrated in Figure 3.2. Note that all taxa in the network are negatively correlated with $O_2$, despite existing at different oxygenation levels within the OMZ. This is OMZ $O_2$ levels are only subtly different relative to $O_2$ levels near the surface. This $O_2$ phenomena is so strong that the only Cyanobacteria are positively correlated with $O_2$ (see Figure 3.6). A particularly interesting univariate differentiation is that Nitrospira are negatively correlated with $NO_3^-$ concentrations, while Nitrospina are positively correlated.

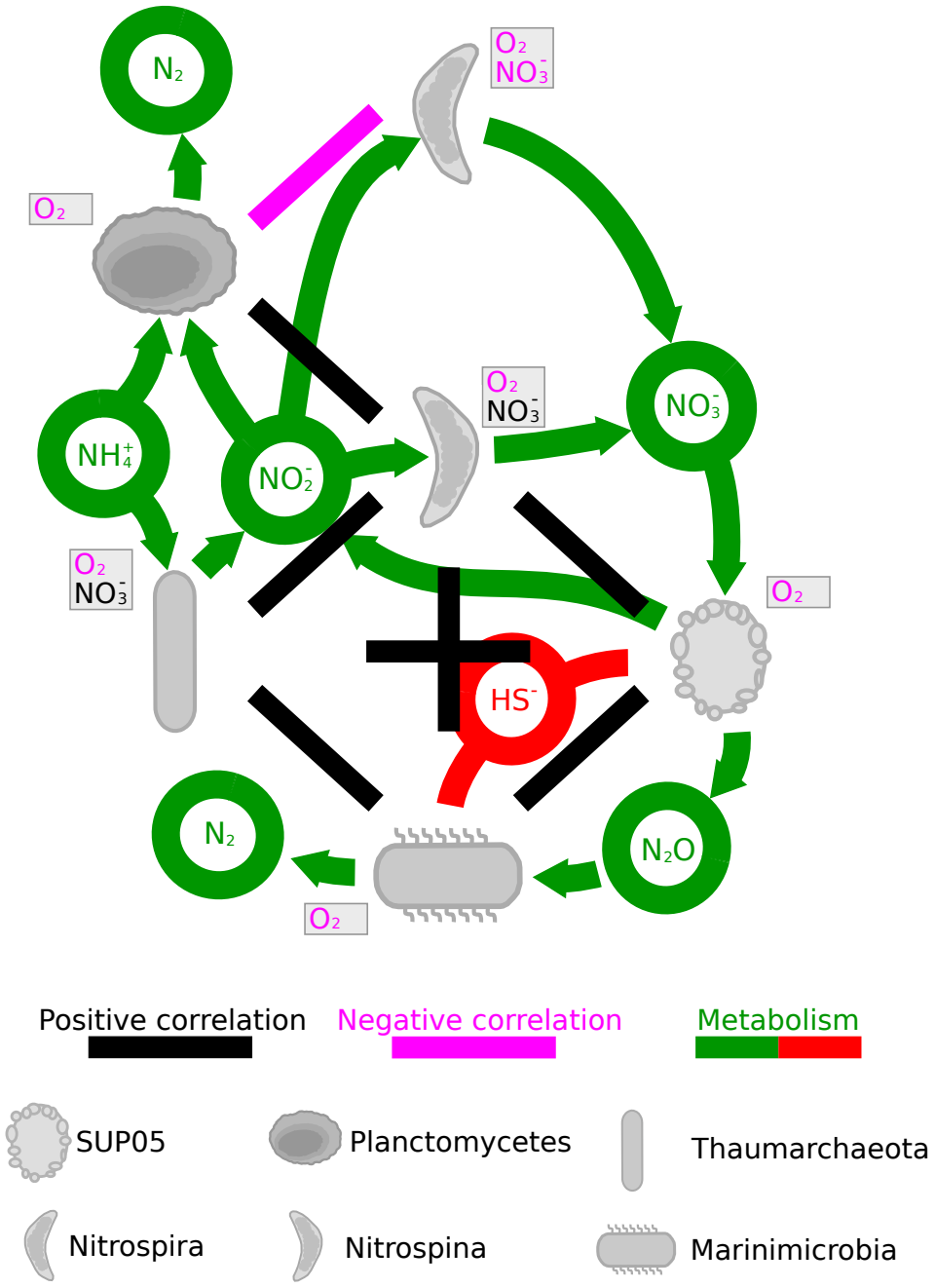



**Figure 3.10:** Statistically significant partial correlations and regressors superimposed over metabolisms. Metabolic relationships reflect both previous interpretations described in subsection 1.4.3 and observations from chapter 2.

## 3.4 Discussion

### 3.4.1 Precision-recall comparison

In subsection 3.3.4, a full-factorial experiment demonstrates the ability of model regularization to make precision attribution possible for SSU rRNA correlation surveys, directly meeting the concerns of Weiss et al. [267]. Taking the perspective that regularization is a model constraint used to reduce overfit, the regularizer is a covariance matrix constraint $\Sigma = LL^T + \Psi$ which is effective through reducing the parameter's dimensional complexity from $O(p^2)$ to $O(p)$ for $p$ taxa. Without precise attribution, correlation networks may only be a guide for approximate topologies, perhaps effectively describable through graph statistics [14, 107, 109, 201, 281] such as centrality [32], betweenness centrality [94], connectivity, or power law distributions [3]. Of course, some statistics might not actually converge meaningfully on certain random graphs. With precise attribution, correlation networks become meaningful to the finest level of resolution: the edge. These findings support the conclusion that the future of SSU rRNA correlation surveys almost entirely requires some form of high-dimensional accommodation [44, 138, 218] such as regularization.

### 3.4.2 Univariate SSU rRNA models

In subsection 3.3.1, univariate count regression models are surveyed with AIC statistics (see Table 3.1). The conclusion is not that the popular [168, 185, 226, 228] negative binomial is a reliable good fit, but instead the floor log t distribution is. While the Student t distribution requires an additional parameter thereby inviting overfit, the AIC statistic accounts for such effects. The Student t shape parameter estimates (see Figure 3.4) indicate that extreme values are commonplace, and therefore make less-robust models including the negative binomial prone to poor fits. This observation is meaningful for authors of univariate regression survey software, and suggests that such software might be more robust through application of different models.

### 3.4.3 Multivariate SSU rRNA models

In subsection 3.3.2, a data-driven argument is built to support the mixed effect perspective for the Saanich Inlet data set. Particularly, 75% of all observed correlations were positive, perhaps cultivated by the common mixed effect variable of sequencing depth. However, it is the mostly-positive $L_1$ factor weight (see Figure 3.6) which provides the strongest support for the perspective, because it observes a mixed effect mechanism. This data support the mixed effect mechanism over the compositional perspective, but do not conclusively decide the question of which is *right*. The compositional perspective has been a foundational insight for modellers in Microbial Ecology [95, 277], but any such negative correlations may only be due to the use of relative abundances.

Further efforts to model SSU rRNA data would benefit from deciding which perspective is *more* correct. However, the idea that SSU rRNA's observed correlations tend to either postivity or negativity could easily be a falsely-imposed dichotomy. The idea that most counts tend to rise and fall with sequencing depth (mixed effect) is no more absurd than the idea that some will compete for depth (compositional). The goal of the modeller should be to allow the data communicate its underlying message, not impose idealizations upon it. Instead of attempting to decide which way

correlational biases tend toward, it may be better to acknowledge the correlational continuum these perspectives share. If possible, it would be best to allow models to express both compositional and mixed effects when possible. The challenge is to do this while avoiding over-parameterization. The primary contribution of this chapter is a more succinct representation of a high-dimensional space, but even it only has enough space for just the mixed effect perspective. A more elegant solution is motivated.

### 3.4.4 Saanich Inlet

In Figure 3.10, a partial correlation network is presented. It is the result of applying the regularized correlation framework to Saanich Inlet SSU rRNA data. The environmental correlations ($O_2$, $NO_3^-$, $H_2S$) illustrate known community structure as observed in Figure 3.2. First, SUP05, Marinimicrobia, and Planctomycetes' only environmental correlations are negative correlations with $O_2$. Their negative correlation with $O_2$ is expected, but their metabolisms (see subsection 1.4.3) suggest that a positive correlation $NO_3^-$ might be expected. However, inspecting the sparklines of Figure 3.2 shows none of their abundances drop offs below the nitrate-sulfidic transition. In interpreting this result, it is important to remember vast genomic variety in each of these clades (see subsection 3.2.6), and that this work's binning experiment did observe the potential for some complete denitrification for both SUP05 and Marinimicrobia (see Figure 2.4). Second, Thaumarchaeota are know to be highly abundant, and harbour ammonia-oxidizing capabilities. The regression analysis finds Thaumarchaeota negatively correlated with $O_2$ and positively correlated with $NO_3^-$. Knowing that ammonia oxidation requires oxygen might make the negative correlation with $O_2$ surprising, but observed Tharmarchaeota abundances (see Figure 3.2) support this fact, and demonstrate its known ecological niche as nitrifier under ammonium-poor conditions [177]. The regression software is merely representing Thaumarchaeota's narrow niche as a linear construction of $O_2$ and $NO_3^-$ concentrations. Third, while Nitrospina and Nitrospira are known anaerobic $NO_2^-$ oxidizers (see subsection 1.4.3), they are observed in Saanich Inlet to sit in different niches (see Figure 3.2). The regression software models this fact through positively correlating Nitrospina with $NO_3^-$, and negatively correlating Nitrospira with $NO_3^-$. Nitrite oxidizing opportunities should be rarest in the sulfidic portion of the OMZ, providing fewer energetic opportunities, and this is reflected in Nitrospira's significantly smaller population than Nitropsina. These observations and known facts agree with the model's representation of the denitrifying community.

In contextualizing the correlation network amongst previous network representations of Saanich Inlet (see subsection 1.4.4 and subsection 1.4.5), the partial correlation components become relevant. The most important differences in comparing these networks pertain to breadth and depth of description. Both the conceptual and differential models represent a *deep dive* into a few samples (2 depth profiles separated by 5 months and 5 depth profiles over 6 months, respectively), providing an extremely detailed description of ecological mechanisms, and they are informed by far more genomic information than just SSU rRNA data, including proteomics. In contrast, the correlation network is informed by the 20 depth profiles from 2006 to 2011, but it only leverages SSU rRNA data and environmental measurements. The conceptual and differential networks are focused descriptions of target denitrifiers or ecological roles, whereas the correlation network is a description of all taxa described by the SSU rRNA data. An advantage of surveying

the entire community is that sub-networks generated by partial correlation decompositions account for effects due to taxa excluded from the network, protecting later inferences against false attribution. A disadvantage is that correlation networks are inherently abstract and thereby provide superficial descriptions of ecology, and therefore require a combined-methods approach to be made useful. This work not only uses previous literature, but also enhances its correlation network with genomic descriptions produced via metagenomic binning (see chapter 2).

The conceptual, differential, and correlation networks agree with each other, and work together to further describe denitrifier community structure in Saanich Inlet. First, the metabolic interactions of Figure 1.7 do not repeat the correlative structure of Figure 3.10 exactly, but instead represent a more modern understanding. Cultivation refuted [237] the hypothesis posed by Hawley et al. [125] that SUP05 might provide Planctomycetes with $NH_4^+$ for anammox. This is reflected in the correlation network as the inexistence of a partial correlative edge between SUP05 and Planctomycetes. Second, the correlational model is able to differentiate roles for Nitrospina and Nitrospira, not just describing them as exemplar nitrite oxidizers, but also describing Nitrospira as Nitrospina's low-abundance, sulfidic zone counter-part. Third, the correlational topology agrees strongly with the metabolic topology of the differential network (see Figure 1.8), reflecting strong support for SUP05 to play the role of partial denitrifier to a sulfur-driven complete denitrifying counter-part. Fourth, when further contextualized in the findings of chapter 2, SUP05's correlative edge with Marinimicrobia is becomes meaningful, because not only is SUP05 described as tending toward partial denitrification (see Figure 2.5), but Marinimicrobia recruits complete denitrifying genes (see Figure 2.4) while also supporting known sulfur metabolism (see subsection 1.4.3). These findings suggest Marinimicrobia is SUP05's sulfur-driven denitrifying partner. Given the lack of $H_2S$ correlation, it is likely that this process is not contributing to environmental sulfur accumulations, and is thus a cryptic sulfur cycle [48]. Fifth, leveraging the correlation survey's precision, it is important to note that SUP05 sharing an edge with Marinimicrobia instead of Planctomycetes supports the argument that nitrogen loss in Saanich Inlet continues to exist through denitrification instead of SUP05 shunting $NO_2^-$ to Planctomycetes for anammox. Limited recall invites the possibility of SUP05 maintaining both relationships. Without precision, this argument becomes substantially weaker, effectively losing correlative support.

### 3.4.5 Partial correlations and succinct representation

This correlation analysis has two convenient problems. First, in subsection 3.2.6 the summation of SSU rRNA counts within large clades (often up to phylum) was motivated as a necessary dimensional reduction, despite working with a massive parameter reduction from $O(p^2)$ to $O(p)$ for $p$ taxa. The problem is not only that the SSU rRNA counting pipeline (see subsection 1.2.3) allows for arbitrary-high resolution and thus arbitrarily-large $p$, but also that there is an immense level of genomic complexity throughout the Saanich Inlet water column. Unfortunately, $O(p)$ parameters is simply too many when $p$ is effectively infinite. This motivates the usage of fewer parameters.

Second, in subsection 3.3.5 partial correlation decomposition (see section 1.5.3) is used to turn a complex correlation network (see Figure 3.9) into a simple one (see Figure 3.10). This process has the advantage of providing digestible simplicity while still using all information from the complex network, but it also highlights an applied fact: not all information in the network needs

to be modelled to a perfect quality, while certain parts do. From a statistical perspective, this means that many expensive parameters are being discared in application. Using the techniques available in subsection 3.3.5, it would be possible for users to specify which taxa to provide high quality modelling for, but such a method risks only reproducing the modeller's assumptions, instead of allowing the data to guide analyses. This shows that the usage of more parameters is not even desired.

With the recognition that correlation analyses need fewer parameters and that many parameters are not even desired, it is clearly time to search for a more elegant solution. The bioinformatic realities of problems in Microbial Ecology show that brute-force application of large statistical models to SSU rRNA data misdirect precious resources. There will never be enough data to describe the correlation structure infinitely-many taxa. Better models are needed. Ideally parameter complexity should not grow with $p$, however explorations of multivariate $O(\log p)$ representations would likely yield effective results. In chapter 4 a succinct model is described, which achieves parameter reduction through assuming evolution follows a stochastic process. The problem of SSU rRNA correlation is certainly not yet solved, but does have feasible avenues into the future.

## 3.5 Conclusions

This work has responded the concerns of Weiss et al. [267] by presenting a correlation network estimation paradigm which meets the statistical needs of Microbial Ecology. The essential mechanic is the constraint of the a covariance matrix $\Sigma = LL^T + \Psi$, achieving a reduction in parameters from $O(p^2)$ to $O(p)$ for $p$ taxa, while still representing all correlations. The methodology is objectively measured through precision-recall curves and application to a Saanich Inlet SSU rRNA data set. The methodology is shown to be capable of substantial increases in precision, largely only at the cost of increased computational time. In application it allows further understanding of the Saanich Inlet denitrifier community by bolstering claims of a SUP05-Marinimicrobia sulfur-driven denitrification pathway with correlative evidene, but also supports the perspective that nitrogen loss through denitrification continues alongside anammox despite a having observed SUP05's shift toward partial denitrification in chapter 2. It is clear that if SSU rRNA correlation surveys are to contribute to any further confident descriptions of fine community structure, some form of regularization must be considered. It would be fruitful to develop models which are able to represent community structure in fewer than $O(p)$ parameters.

# Chapter 4

# Future directions

This work contributes toward two bioinformatic tasks in Microbial Ecology: metagenomic binning and SSU rRNA correlation. Major contributions are evaluated in an objective manner, primarily through precision-recall curves. Evaluations extend into application by applying contributions toward improving understanding of the Saanich Inlet denitrifying community. Application provided both an opportunity to catch discrepancies in inferences, but also to further understanding. The primary conclusions of chapter 2 are that metagenomic binning can be made more precise through use of a good reference (as in assembly), and that taxonomic attributions are more precise nearer to the phylogenetic root (as in bootstrapped phylogenetic estimation). The primary conclusion of chapter 3 is that regularization is necessary for precise estimation of fine correlative structure. In application toward the Saanich Inlet denitrifying community these methods have allowed various inferences, but a single argument has been built on from previous network perspectives (see subsection 1.4.4 and subsection 1.4.5): SUP05 is observed to be taking on a partial denitrifying role, and is likely working with Marinimicrobia through cryptic sulfur cycling to sustain complete denitrification. Despite these contributions, important questions still remain.

## 4.1 Denitrification in Saanich Inlet

Building on previous network interpretations of the Saanich Inlet denitrifying community (see subsection 1.4.4 and subsection 1.4.5), this work has attributed metabolic capabilities to certain taxa in chapter 2, then constrained interpretation with correlative evidence in chapter 3. At this point it appears that SUP05, a major denitrifier, is moving toward partial denitrification. Metabolic and correlative evidence suggests that SUP05 may continue to play a role in complete denitrification by driving it via a cryptic sulfur-cycling relationship with Marinimicrobia. Despite the evidence behind this hypothesis, other must also be considered. First, Arcobacteraceae is known to be an active complete denitrifier in the sulfidic zone (see subsection 1.4.3), and is observed to be operating in Saanich Inlet (see Figure 2.4). So an alternative hypothesis is that Arcobacteraceae or another complete denitrifier may be taking over the niche. Second, nitrogen loss may be occurring through anammox instead of denitrification. Since SUP05 is taking on a nitrite producing role, and Planctomycetes is known to harbour anammox capabilities, one would expect a positive correlation to develop between them, though one was not observed (see Figure 3.10). The lack of a correlation is not conclusive, because the result is not unlikely to be a false negative, or the relationship could exist, but in a non-obligate manner (which is also likely). Third, denitrification may actually be slowing. The ecological consequences energetic opportunities are not fully understood, and just because energy exists to be taken, does not mean it is best to do so.

These hypotheses can be further narrowed bioinformatically with existing data. For example,

the third alternative hypothesis can be ruled out with existing chemical concentration data. If nitrous oxide concentrations are non-increasing over time, the hypothesis is unlikely. This could be argued informatically with a univariate regression analysis or simple plotting. The second alternative hypotheses is contingent on Planctomycetes increasing in anammox activity while SUP05 decreases in nitric oxide reduction. This could be argued via a binning experiment similar to the one used to construct Figure 2.5. Similarly, the first alternative hypothesis describes Arcobacteraceae (or another organism) increasing in complete denitrifying behaviour, and could also be refuted through a binning experiment. If taxonomic attributions are not desired, binning isn't even necessary.

Ruling out alternative hypotheses does not necessarily support the initial hypothesis either, because the SUP05-Marinimicrobia metabolic syntrophy argument is built on imperfect statistical inferences. Such methods are important for building hypotheses, but ultimately the verification should be performed with an isolation and rate measurement experiment. If the hypothesis is true, then hydrogen sulfide should be consumed, not build up in the environment, and should accelerate denitrification. This should not discredit efforts to rule out alternative hypotheses however, because such bioinformatic analyses are ultimately much less expensive than isolation and rate measurement experiments. In this way, these binning and correlation-based arguments are a typical bioinformatic precursor work which helps place a few, expensive, high-quality verifications.

## 4.2 Regularization as reduced parameter complexity

In subsection 1.6.3 regularization is defined as constraining a model to reduce overfit. Theory for regularization via constrained optimization is well-developed [28, 44, 131, 164], but in chapter 3 constraint is implemented through equating a higher dimensional parameter with a lower dimensional representation [218, 247]. It is unclear that such a method should somehow increase precision-recall exchanges. Indeed, it would be worth directly comparing the higher and lower dimensional models. However, the whole process also highlights an alternative perspective of regularization in general, where the essential mechanic is the reduction of parameter complexity. Model constraints inevitably invite reduced dimensionality, implicit or not. Even theoretically developed L1-regularization is applied toward explicit parameter reduction [164]. Embracing this perspective, it becomes important to ask how smaller models generally improve statistical power.

This perspective can be developed theoretically, thereby providing broadly applicable results. Interpreting classification as a hypothesis testing problem, the Neyman-Pearson Lemma [196] can be leveraged to establish a test statistic: under the right conditions a likelihood ratio test (LRT) statistic $\lambda$ is most-powerful, and so only it will be considered. To gain perspective on a breadth of models this work will invoke a large sample size assumption, thereby making the work of Wilks [270] and Wald [265] applicable. Under the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, a large sample size, and regularity assumptions, the distribution of the LRT statistic $\lambda$ is known [270] via $-2\lambda \sim_{H_0} \chi^2(p, 0)$, where $p$ is the dimension of the parameter tested ($i\boldsymbol{\theta} \in \mathbb{R}^p$), and $\chi^2(p, \delta)$ is the non-central chi-square distribution with degrees of freedom $p$, non-centrality parameter $\delta$, and distribution function $F_{\chi^2(p,\delta)}$. Under the alternative hypothesis $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, the LRT distribution is similarly known [67, 265] with $-2\lambda \sim_{H_1} \chi^2(p, \delta)$, where $\delta = \boldsymbol{\partial\theta}^T I_{\boldsymbol{\theta}_1}^{-1} \boldsymbol{\partial\theta}$, and $\boldsymbol{\partial\theta} = \boldsymbol{\theta}_0 - \boldsymbol{\theta}_1$. Then the rejection region for the test statistic $-2\lambda$ with false-rejection rate $\alpha$ is any value less than

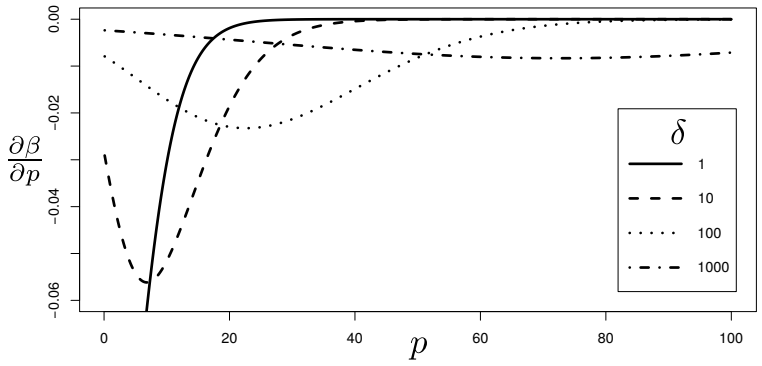



**Figure 4.1:** Statistical power decreases as dimension increases for $\alpha = 0.05$.

$F^{-1}_{\chi^2(p,0)}(1-\alpha)$. Then power is calculable under the null as $1-\beta = 1 - F_{\chi^2(p,\delta)}\left(F^{-1}_{\chi^2(p,0)}(1-\alpha)\right)$.

Regularization by parameter constraint effectively defines a function $\boldsymbol{\theta}(\boldsymbol{\eta}) \in \mathbb{R}^p$ and $\boldsymbol{\eta} \in \mathbb{R}^q$ where $q < p$. It is convenient to impose an unbiased representation assumption of $\boldsymbol{\theta}_0 = \boldsymbol{\theta}(\boldsymbol{\eta}_0)$ and $\boldsymbol{\theta}_1 = \boldsymbol{\theta}(\boldsymbol{\eta}_1)$ for some $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ in $\mathbb{R}^q$, so that representations are more directly comparable. Where only biased representations exist, the approximation may still be valuable through variance-bias trade-off considerations. Under the $\boldsymbol{\theta}$-representation in $p$ dimensions, the non-centrality parameter is $\delta_{\boldsymbol{\theta}} = \boldsymbol{\partial\theta}^T I_{\boldsymbol{\theta}_1}^{-1} \boldsymbol{\partial\theta}$. Under the $\boldsymbol{\eta}$-representation, the non-centrality parameter is $\delta_{\boldsymbol{\eta}} = \boldsymbol{\partial\eta}^T (J^T I_{\boldsymbol{\theta}_1} J)^{-1} \boldsymbol{\partial\eta}$, where $\boldsymbol{\partial\eta} = \boldsymbol{\eta}_0 - \boldsymbol{\eta}_1$, $J \in \mathbb{R}^{p\times q}$, and $J_{ij} = (\partial\theta_i/\partial\eta_j)(\boldsymbol{\eta}_1)$. Therefore the power of the $\boldsymbol{\eta}$-representation is $1-\beta_{\boldsymbol{\eta}} = 1 - F_{\chi^2(q,\delta_{\boldsymbol{\eta}})}\left(F^{-1}_{\chi^2(q,0)}(1-\alpha)\right)$.

Having developed this perspective of regularization theoretically, it is now clear that a tersely defined set of functions has broadly applicable implications. Regularization constraints can be generalized to any functions $\boldsymbol{\theta}(\boldsymbol{\eta})$ which increase statistical power by satisfying the following equation and previously stated assumptions. It's important to note that this is not yet a perfectly posed mathematical problem, because $\boldsymbol{\theta}(\boldsymbol{\eta})$ likely needs to satisfy certain pragmatic qualities. For example, in chapter 3 the $LL^T + \Psi$ constraint is continuously differentiable everywhere, and is capable of describing a useful breadth of covariance matrices. Even so, developing theory for $\boldsymbol{\theta}(\boldsymbol{\eta})$ has the potential to guide modelling choices in high-dimensional spaces such as for SSU rRNA data.

$$1-\beta_{\boldsymbol{\theta}} \leq 1-\beta_{\boldsymbol{\eta}} \Leftrightarrow F_{\chi^2(q,\delta_{\boldsymbol{\eta}})}\left(F^{-1}_{\chi^2(q,0)}(1-\alpha)\right) \leq F_{\chi^2(p,\delta_{\boldsymbol{\theta}})}\left(F^{-1}_{\chi^2(p,0)}(1-\alpha)\right)$$

A heuristic understanding of these conditions can be produced through numerical approximation, thereby bolstering the claim that reducing parameter complexity (dimension) can improve statistical power. Since $F_{\chi^2(p,\delta)}$ is defined continuously over $p$, a sense of continuity between parameter dimensions can be established. In chapter 3, it is reasonable to imagine that regularization causes far greater changes in model dimension $p$ than in $\delta$. So behaviour of statistical power over model dimension with fixed $\delta$ can be examined through numerical approximation of $\frac{\partial\beta}{\partial p} = \frac{\partial}{\partial p} F_{\chi^2(p,\delta)}\left(F^{-1}_{\chi^2(p,0)}(1-\alpha)\right)$ with a five-point stencil (see subsection 1.7.1). It can be seen in Figure 4.1 that for each test point, $\beta$ decreases with model dimension, thereby causing statistical power $(1-\beta)$ to increase.

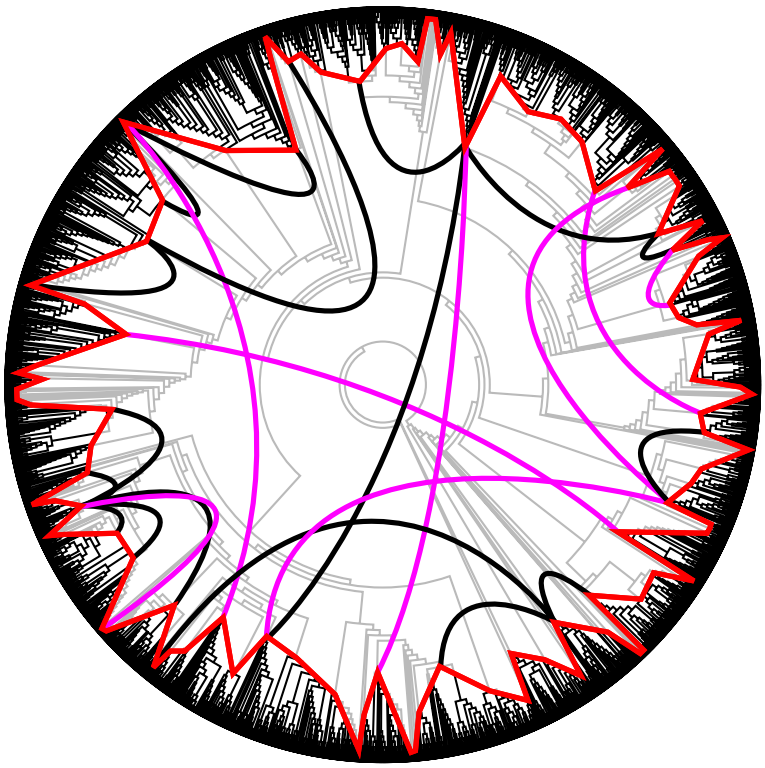


**Figure 4.2:** Simplified Tree of Life superimposed with a succinct correlation structure. The red line is a cutting line, which separates the entire tree into clades. Each clade's correlation structure is dictated entirely by its own tree and clade parameters. Clade parameters are latent random variables with a complete correlation structure. Correlations are illustrated with black and magenta lines. Tree image credit: [130]

## 4.3 A more succinct representation

In chapter 3, precise correlation inference was made possible via a model constraint $\Sigma = LL^T + \Psi$ which reduces the parameter complexity from $O(p^2)$ to $O(p)$ for $p$ taxa. Despite the effort, in subsection 3.2.6 dimensional reduction is still employed: many clades' SSU rRNA counts are summed up to the phylum level, and all remaining taxa are sub-optimally descriptive. Dimensional reduction is employed, because for very large $p$, $O(p)$ is still far too many parameters. A more succint representation is motivated, meaning that a parameter complexity of $O(\log p)$ or $O(1)$ needs to be described (see subsection 3.4.5). This work now proposes one such representation.

Instead of trying to correlate taxa, it might be more pragmatic to only correlate clades. Drawing inspiration from phylogenetic regression (see subsection 1.2.2), evolution is modelled as an unobserved Brownian motion processes [224]. Every phylogenetic tree can be described this way, including clades, because they are also trees. If the tree is cut correctly, it is broken into separate clades. For example, cutting along the red line in Figure 4.2 produces many clades. If one was to specify the cut along a bifurcating tree according to node height, there would be $O(\log p)$ clades. The essential mechanic here is that pair-wise correlations are only estimated between clades, while the brownian motion process defines all correlations within clades.

Formally, for each of $p$ taxa dimensions (not clades) define a marginal distribution function

$F_{\mathbf{Y}_j}(y; \mu_j, \sigma_j)$, where $\mu_j$ is a location parameter and $\sigma_j$ is a dispersion parameter. Recognizing that $O(p)$ is too many parameters, each $(\mu_j, \sigma_j)$ will not be estimated. Instead $(\mu_j, \sigma_j)$ is random, following a bivariate lognormal distribution. Each $\mu_j$ in clade $C_k$ is dependent according to the brownian motion process of tree $C_k$, and originates from an initial Gaussian-distributed random value $Z_k$ at the root of the its clade, so $\mathbb{E}[\mu_j | Z_k] = e^{Z_k}$. Similarly constraint each $\sigma_j$, independently of every $\mu_j$. The random vector $\mathbf{Z} = [Z_k]$ controls the location parameters of each clade. Since $\mathbf{Z} \in \mathbb{R}^{O(\log p)}$ it can describe any regression mechanisms and be given a full pair-wise correlation structure with fewer than $O(p)$ parameters, so $\mathbf{Z} \sim N(X\boldsymbol{\beta}, \Sigma)$. So despite being capable of modelling all taxa, this model only suffers a parameter complexity of $O((\log p)^2)$.

This proposed model is succinct in that it has a parameter complexity less than $O(p)$ for $p$ taxa. It is a demonstration of how a high-dimensional model can have a succinct parameterization. It is not designed to have numerically feasible properties however. It is not even uniquely defined, because the marginal distribution functions and $\sigma_j$ distributions are defined abstractly. Much more work is required before such models can be made useful to Microbial Ecologists. For example, pragmatic tree cutting requires exploration. The work in chapter 3 demonstrates that precise correlation networks are possible, but still requires destructive dimensional reduction (such as the summation of many clades up to the phylum level). These succinct models offer the opportunity to have precise correlation networks without the need to resort to dimensional reduction.

# Bibliography


[1] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura. Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Research*, 12(5):281–290, 2006.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[4] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406: 378–382, 2000.

[5] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 31:533–538, 2013.

[6] E. Allers, J. J. Wright, K. M. Konwar, C. G. Howes, E. Beneze, S. J. Hallam, and M. B. Sullivan. Diversity and population structure of marine group a bacteria in the northeast subarctic pacific ocean. *The ISME Journal*, 7:256–268, 2013.

[7] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11:1144–1146, 2014.

[8] D. Altavilla. Nvidia doubles down on self-driving cars with xavier ai chip and a hat tip to next gen volta gpu, 2016. http://www.forbes.com/sites/davealtavilla/2016/09/28/nvidia-doubles-down-on-self-driving-cars-with-xavier-ai-chip-and-a-hat-tip-to-next-gen-volta-gpu; accessed online 30-September-2016.

[9] S. F. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[10] K. Anantharaman, C. T. Brown, L. A. Hug, I. Sharon, C. J. Castelle, A. J. Probst, B. C. Thomas, A. Singh, M. J. Wilkins, U. Karaoz, E. L. Brodie, K. H. Williams, S. S. Hubbard, and J. F. Banfield. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, 7:13219, 2016.

[11] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. Lapack users' guide. 1999.

[12] J. J. Anderson and A. H. Devol. Deep water renewal in saanich inlet, an intermittently anoxic basin. *Estuarine and Coastal Marine Science*, 1(1):1–10, 1973.

[13] M. J. Anderson, T. O. Crist, J. M. Chase, M. Vellend, B. D. Inouye, A. L. Freestone, N. J. Sanders, H. V. Cornell, L. S. Comita, K. F. Davies, S. P. Harrison, N. J. B. Kraft, J. C. Stegen, and N. G. Swenson. Navigating the multiple meanings of b diversity: a roadmap for the practicing ecologist. *Ecology Letters*, 14:19–28, 2011.

[14] M. Arumugam, J. Raes, E. Pelletier, D. L. Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, MetaHIT Consortium, J. Weissenbach, S. D. Ehrlich, and P. Bork. Enterotypes of the human gut microbiome. *Nature*, 473: 174–180, 2011.

[15] J. R. Ashford and R. R. Snowden. Multivariate probit analysis. *Biometrics*, 26:535–546, 1970.

[16] K. Baba. *Partial, Conditional and Multiplicative Correlation Coefficients*. Keio University, 2004.

[17] K. Baba and M. Sibuya. Equivalence of partial and conditional correlation coefficients. *Journal of the Japan Statistical Society*, 35(1):1–19, 2005.

[18] K. Baba and R. S. M. Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

[19] A. R. Babbin, R. G. Keil, A. H. Devol, and B. B. Ward. Organic matter stoichiometry, flux, and oxygen control nitrogen loss in the ocean. *Science*, 344(6182):406–408, 2014.

[20] D. Baetens. *Enhanced biological phophorus removal: modelling and experimental design*. Ghent University, 2001.

[21] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner. Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

[22] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113, 2004.

[23] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[24] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 51(1):289–300, 1995.

[25] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[26] S. Bennett. Solexa ltd. *Pharmacogenomics*, 5(4):433–438, 2004.

[27] A. A. Berryman. The orgins and evolution of predator-prey theory. *Ecology*, 73(5):1530–1535, 1992.

[28] P. Bickel, B. Li, A. Tsybakov, S. Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. Vaart. Regularization in statistics. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 15(2): 271–344, 2006.

[29] P. C. Blainey. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiology Reviews*, 73(3):407–427, 2013.

[30] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. The complete genome sequence of escherichia coli k-12. *Science*, 277(5331): 1453–1462, 1997.

[31] N. Bokulich and C. Bamforth. The microbiology of malting and brewing. *Microbiology and Molecular Biology Reviews*, 77(2):157–172, 2013.

[32] S. P. Borgatti. Centrality and network flow. *Social Networks*, 27:55–71, 2005.

[33] A. Bowe, T. Onodera, K. Sadakane, and T. Shibuya. *Succinct de Bruijn Graphs*. Springer, 2012.

[34] S. Boyd and L. Vandenberghe. Convex optimization. 2004.

[35] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W. Chou, J. Corbeil, C. D. Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, Élénie Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. MacCallum, M. D. MacManes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2:10, 2013.

[36] J. R. Bray and J. T. Curtis. An ordination of upland forest communities of southern wisconsin. *Ecological Monographs*, 27:325–349, 1957.

[37] E. C. Brechmann and H. Joe. Parsimonious parameterization of correlation matrices using truncated vines and factor analysis. *Computational Statistics & Data Analysis*, 77:233–251, 2014.

[38] I. Brettar, M. Labrenz, S. Flavier, J. Bötel, H. Kuosa, R. Christen, and M. G. Höfle. Identification of a thiomicrospira denitrificans-like epsilonproteobacterium as a catalyst for autotrophic denitrification in the central baltic sea. *Applied and Environmental Microbiology*, 72(2):1364–1372, 2006.

[39] B. Brown and L. Levy. Certification of algorithm 708: Significant digit computation of the incomplete beta function ratios. *ACM Transactions on Mathematical Software*, 20(3):393–397, 1994.

[40] C. T. Brown, L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, and J. F. Banfield. Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, 52:208–211, 2015.

[41] M. V. Brown and S. P. Donachie. Evidence for tropical endemicity in the deltaproteobacteria marine group b/sar324 bacterioplankton clade. *Aquatic Microbial Ecology*, 46:107–115, 2007.

[42] C. G. Broyden. The convergence of a class of double-rank minimization algorithms: Part 2. *Journal of the Institute of Mathematics and its Applications*, 6:222–231, 1970.

[43] M. Burrows and D. Wheeler. A block-sorting lossless data compression algorithm. *Digital Equipment Corporation*, Technical Report 124, 1994.

[44] T. Cai, W. Liu, and X. Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *American Statistical Association*, 106(494):594–607, 2011.

[45] A. C. Cameron and P. K. Trivedi. Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1:29–54, 1986.

[46] A. C. Cameron and P. K. Trivedi. Regression analysis of count data. 2013.

[47] B. J. Campbell, A. S. Engel, M. L. Porter, and K. Takai. The versatile epsilon-proteobacteria: key players in sulphidic habitats. *Nature Reviews Microbiology*, 4:458–468, 2006.

[48] D. E. Canfield, F. J. Stewart, B. Thamdrup, L. De Brabandere, T. Dalsgaard, E. F. Delong, N. P. Revsbech, and O. Ulloa. A cryptic sulfur cycle in oxygen-minimum-zone waters off the chilean coast. *Science*, 330(6009):1375–1378, 2010.

[49] G. Cantor. Ueber eine elementare frage der mannigfaltigkeitslehre. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 1:75–78, 1892.

[50] D. G. Capone and D. A. Hutchins. Microbial biogeochemistry of coastal upwelling regimes in a changing ocean. *Nature Geoscience*, 6:711–717, 2013.

[51] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, 7:335–336, 2010.

[52] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Thomson Learning, 2002.

[53] R. Caspi, R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, and P. D. Karp. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 44:D471–D480, 2016.

[54] S. Chaffron, H. Rehrauer, J. Pernthaler, and C. von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20:947–959, 2010.

[55] P. S. G. Chain, D. V. Grafham, R. S. Fulton, M. G. FitzGerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. C. Bruce, C. Buhay, J. R. Cole, Y. Ding, S. Dugan, D. Field, G. M. Garrity, R. Gibbs, T. Graves, C. S. Han, S. H. Harrison, S. Highlander, P. Hugenholtz, H. M. Khouri, C. D. Kodira, E. Kolker, N. C. Kyrpides, D. Lang, A. Lapidus, S. A. Malfatti, V. Markowitz, T. Metha, K. E. Nelson, J. Parkhill, S. Pitluck, X. Qin, T. D. Read, J. Schmutz, S. Sozhamannan, P. Sterk, R. L. Strausberg, G. Sutton, N. R. Thomson, J. M. Tiedje, G. Weinstock, A. Wollam, Genomic Standards Consortium, Human Microbiome Project, Jumpstart Consortium, and J. C. Detter. Genome project standards in a new era of sequencing. *Science*, 326(5950):236–237, 2009.

[56] A. Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11(4):265–270, 1984.

[57] S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998.

[58] J.-C. Cho and S. J. Giovannoni. Cultivation and growth characteristics of a diverse group of oligotrophic marine gammaproteobacteria. *Applied and Environmental Microbiology*, 70(1):432–440, 2004.

[59] J. E. Clarridge. Impact of 16s rrna gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, 17(4):840–862, 2004.

[60] L. A. Codispoti. The limits to growth. *Nature*, 387:237, 1997.

[61] L. A. Codispoti, J. A. Brandes, J. P. Christensen, A. H. Devol, S. A. Naqvi, H. W. Paerl, and T. Yoshinari. The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene? *Scientia Marina*, 65(2):85–105, 2001.

[62] R. Conway and W. Maxwell. A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132–136, 1962.

[63] D. R. Cox. Renewal theory. 1970.

[64] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[65] F. D. F. da Silva, A. R. J. Lima, P. H. G. Moraes, A. S. Siqueira, L. T. Dall'Agnol, A. R. F. Baraúna, L. C. Martins, K. G. Oliveira, C. P. S. de Lima, M. R. T. Nunes, J. L. S. G. Vianez-Júnior, and E. C. Gonçalves. Draft genome sequence of limnobacter sp. strain caciam 66h1, a heterotrophic bacterium associated with cyanobacteria. *Genome Announcements*, 4(3):e00399–16, 2016.

[66] T. Dalsgaard, D. E. Canfield, J. Petersen, B. Thamdrup, and J. Acuña-González. N2 production by the anammox reaction in the anoxic water column of golfo dulce, costa rica. *Nature*, 422:606–608, 2003.

[67] R. R. Davidson and W. E. Lever. The limiting distribution of the likelihood ratio statistic under a class of local alternatives. *Sankhyā: The Indian Journal of Statistics*, 32(2):209–224, 1970.

[68] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[69] Dempster, N. M. Arthur P.; Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1976.

[70] Y. Deng, Y.-H. Jiang, Y. Yang, Z. He, F. Luo, and J. Zhou. Molecular ecological network analyses. *BMC Bioinformatics*, 13:113, 2012.

[71] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied Environmental Microbiology*, 72:5069–5072, 2006.

[72] P. M. S. Desmond G. Higgins. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244, 1988.

[73] A. H. Devol. Nitrogen cycle: Solution to a marine mystery. *Nature*, 422:575–576, 2003.

[74] G. J. Dick, A. F. Andersson, B. J. Baker, S. L. Simmons, B. C. Thomas, A. P. Yelton, and J. F. Banfield. Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, 10:R85, 2009.

[75] A. R. Didonato and A. H. Morris. Algorithm 708: Significant digit computation of the incomplete beta function ratios. *ACM Transactions on Mathematical Software*, 18(3):360–373, 1992.

[76] J. A. Dodsworth, P. C. Blainey, S. K. Murugapiran, W. D. Swingley, C. A. Ross, S. G. Tringe, P. S. G. Chain, M. B. Scholz, C. Lo, J. Raymond, S. R. Quake, and B. P. Hedlund. Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the op9 lineage. *Nature Communications*, 4:1854, 2013.

[77] S. C. Doney. The growing human footprint on coastal and open-ocean biogeochemistry. *Science*, 328(5985):1512–1516, 2010.

[78] J. C. Doyle, D. L. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, , and W. Willinger. The "robust yet fragile" nature of the internet. *Proceedings of the National Academy of Sciences*, 102(41): 14497–14502, 2005.

[79] W. E. Durno, N. W. Hanson, K. M. Konwar, and S. J. Hallam. Expanding the boundaries of local similarity analysis. *BMC Genomics*, 14(Suppl 1):S13, 2013.

[80] R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[81] R. C. Edgar. Local homology recognition and distance measures in linear time using compressed amino acid alphabets. *Nucleic Acids Research*, 32(1):380–385, 2004.

[82] R. C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19): 2460–2461, 2010.

[83] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7:1–26, 1979.

[84] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetictrees. *Proceedings of the National Academy of Sciences*, 93(23):13429–13434, 1996.

[85] P. G. Falkowski, T. Fenchel, and E. F. Delong. The microbial engines that drive earth's biogeochemical cycles. *Science*, 320(5879):1034–1039, 2008.

[86] P. G. Falkowski, T. Algeo, L. Codispoti, C. Deutsch, S. Emerson, B. Hales, R. B. Huey, W. J. Jenkins, L. R. Kump, L. A. Levin, T. W. Lyons, N. B. Nelson, O. S. Schofield, R. Summons, L. D. Talley, E. Thomas, F. Whitney, and C. B. Pilcher. Ocean deoxygenation: Past, present, and future. *Earth & Space Science News*, 92(46):409–410, 2011.

[87] K. Faust and J. Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10:538–550, 2012.

[88] J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4): 783–791, 1985.

[89] R. L. Ferguson, E. N. Buckley, and A. V. Palumbo. Response of marine bacterioplankton to differential filtration and confinement. *Applied and Environmental Microbiology*, 47:49–55, 1984.

[90] D. Field, G. Garrity, T. Gray, N. Morrison, J. Selengut, P. Sterk, T. Tatusova, N. Thomson, M. J. Allen, S. V. Angiuoli, M. Ashburner, N. Axelrod, S. Baldauf, S. Ballard, J. Boore, G. Cochrane, J. Cole, P. Dawyndt, P. D. Vos, C. dePamphilis, R. Edwards, N. Faruque, R. Feldman, J. Gilbert, P. Gilna, F. O. Glöckner, P. Goldstein, R. Guralnick, D. Haft, D. Hancock, H. Hermjakob, C. Hertz-Fowler, P. Hugenholtz, I. Joint, L. Kagan, M. Kane, J. Kennedy, G. Kowalchuk, R. Kottmann, E. Kolker, S. Kravitz, N. Kyrpides, J. Leebens-Mack, S. E. Lewis, K. Li, A. L. Lister, P. Lord, N. Maltsev, V. Markowitz, J. Martiny, B. Methe, I. Mizrachi, R. Moxon, K. Nelson, J. Parkhill, L. Proctor, O. White, S.-A. Sansone, A. Spiers, R. Stevens, P. Swift, C. Taylor, Y. Tateno, A. Tett, S. Turner, D. Ussery, B. Vaughan, N. Ward, T. Whetzel, I. S. Gil, G. Wilson, and A. Wipat. The minimum information about a genome sequence (migs) specification. *Nature Biotechnology*, 26:541–547, 2008.

[91] D. Field, L. Amaral-Zettler, G. Cochrane, J. R. Cole, P. Dawyndt, G. M. Garrity, J. Gilbert, F. O. Glöckner, L. Hirschman, I. Karsch-Mizrachi, H. Klenk, R. Knight, R. Kottmann, N. Kyrpides, F. Meyer, I. S. Gil, S.-A. Sansone, L. M. Schriml, P. Sterk, T. Tatusova, D. W. Ussery, O. White, and J. Wooley. The genomic standards consortium. *PLoS Biology*, 9(6):e1001088, 2011.

[92] R. Fletcher. A new approach to variable metric algorithm. *The Computer Journal*, 13:317–322, 1970.

[93] J. A. Frank, Y. Pan, A. TommingKlunderud, V. G. H. Eijink, A. C. McHardy, A. J. Nederbragt, and P. B. Pope. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Scientific Reports*, 6:25373, 2016.

[94] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[95] J. Friedman and E. J. Alm. Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, 8(9):e1002687, 2012.

[96] J. A. Fuhrman and A. A. Davis. Widespread archaea and novel bacteria from the deep sea as shown by 16s rrna gene sequences. *Marine Ecology Press Series*, 150:275–285, 1997.

[97] J. A. Fuhrman, K. McCallum, and A. A. Davis. Phylogenetic diversity of subsurface marine microbial communities from the atlantic and pacific oceans. *Applied and Environmental Microbiology*, 59(5): 1294–1302, 1993.

[98] J. A. Fuhrman, J. A. Cram, and D. M. Needham. Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, 13:133–146, 2015.

[99] B. Gallone, J. Steensels, T. Prahl, L. Soriaga, V. Saels, B. Herrera-Malaver, A. Merlevede, M. Roncoroni, K. Voordeckers, L. Miraglia, C. Teiling, B. Steffy, M. Taylor, A. Schwartz, T. Richardson, C. White, G. Baele, S. Maere, and K. J. Verstrepen. Domestication and divergence of saccharomyces cerevisiae beer yeasts. *Cell*, 166(6):1397–1410, 2016.

[100] M. Gallopin, A. Rau, and F. Jaffrézic. A hierarchical poisson log-normal model for network inference from rna sequencing data. *PLoS ONE*, 8(10):e77503, 2013.

[101] A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–150, 1992.

[102] T. S. Ghosh, M. H. M, and S. S. Mande. Discribinate: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics*, 11(Suppl 7):S14, 2010.

[103] E. A. Gies, K. M. Konwar, J. T. Beatty, and S. J. Hallam. Illuminating microbial dark matter in meromictic sakinaw lake. *Applied and Environmental Microbiology*, 80(21):6807–6818, 2014.

[104] Gigabyte. Gv-n98twf3oc-6gd, 2017. http://www.gigabyte.com ; accessed online 10 March 2017.

[105] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[106] J. Gole, A. Gore, A. Richards, Y. Chiu, H. Fung, D. Bushman, H. Chiang, J. C. andYu-Hwa Lo, and K. Zhang. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nature Biotechnology*, 31:1126–1132, 2013.

[107] J. K. Goodrich, J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. V. Treuren, R. Knight, J. T. Bell, T. D. Spector, A. G. Clark, and R. E. Leycorrespondence. Human genetics shape the gut microbiome. *Cell*, 159(4):789799, 2014.

[108] R. M. Gower and P. Richtárik. Randomized quasi-newton updates are linearly convergent matrix inversion algorithms. *ArXiv*, 1602:01768v3, 2016.

[109] S. Greenblum, P. J. Turnbaugh, and E. Borenstein. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*, 109(2):594–599, 2011.

[110] W. Greene. Functional forms for the negative binomial model for count data. *Economics Letters*, 99: 585–590, 2008.

[111] I. Gregor, J. Dröge, M. Schirmer, C. Quince, and A. C. McHardy. Phylopythias+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ*, 4: e1603, 2016.

[112] W. Gropp, E. Lusk, N. Doss, and A. Skjellum. A high-performance, portable implementation of the mpi message passing interface standard. *Parallel Computing*, 22(6):789–828, 1996.

[113] J. Grote, M. Labrenz, B. Pfeiffer, G. Jost, and K. Jürgens. Quantitative distributions of epsilonproteobacteria and a sulfurimonas subgroup in pelagic redoxclines of the central baltic sea. *Applied and Environmental Microbiology*, 73(22):7155–7161, 2007.

[114] J. Grote, G. Jost, M. Labrenz, G. J. Herndl, and K. Jürgens. Epsilonproteobacteria represent the major portion of chemoautotrophic bacteria in sulfidic waters of pelagic redoxclines of the baltic and black seas. *Applied and Environmental Microbiology*, 74(24):7546–7551, 2008.

[115] J. Grote, T. Schott, C. G. Bruckner, F. O. Glöckner, G. Jost, H. Teeling, M. Labrenz, and K. Jürgens. Genome and physiology of a model epsilonproteobacterium responsible for sulfide detoxification in marine oxygen depletion zones. *Proceedings of the National Academy of Sciences*, 109(2):506–510, 2012.

[116] N. Gruber. The marine nitrogen cycle: Overview and challenges. 2008.

[117] S. Guikema and J. Coffelt. A flexible count data regression model for risk analysis. *Risk Analysis*, 28(1):213–223, 2008.

[118] P.-E. Hagmark. On construction and simulation of count data models. *Mathematics and Computers in Simulation*, 77:72–80, 2008.

[119] P.-E. Hagmark. An exceptional generalization of the poisson distribution. *Open Journal of Statistics*, 2: 313–318, 2012.

[120] A. S. Hahn, K. M. Konwar, S. Louca, N. W. Hanson, and S. J. Hallam. The information science of microbial ecology. *Current Opinion in Microbiology*, 31:209–216, 2016.

[121] N. W. Hanson, K. M. Konwar, A. K. Hawley, T. Altman, P. D. Karp, and S. J. Hallam. Metabolic pathways for the whole community. *BMC Genomics*, 15:619, 2014.

[122] G. H. Hardy, J. E. Littlewood, and G. Plya. *Inequalities*. Cambridge University Press, 1934.

[123] M. F. Haroon, L. R. Thompson, and U. Stingl. Draft genome sequence of uncultured sar324 bacterium lautmerah10, binned from a red sea metagenome. *Genome Announcements*, 4(1):e01711–15, 2016.

[124] J. K. Harris, S. T. Kelley, and N. R. Pace. New perspective on uncultured bacterial phylogenetic division op11. *Applied and Environmental Microbiology*, 70(2):845–849, 2004.

[125] A. K. Hawley, H. M. Brewer, A. D. Norbeck, L. Paša-Tolić, and S. J. Hallam. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proceedings of the National Academy of Sciences*, 111(31):11395–11400, 2013.

[126] A. K. Hawley, M. K. Nobu, J. J. Wright, W. E. Durno, C. Morgan-Lang, B. Sage, P. Schwientek, B. K. Swan, C. Rinke, M. Torres-Beltrán, K. Mewis, W. Liu, R. Stepanauskas, T. Woyke, and S. J. Hallam. Co-metabolic innovation along eco-thermodynamic gradients. *Submitted*, 2016.

[127] A. K. Hawley, M. Torres-Beltrán, M. Bhatia, E. Zaikova, D. A. Walsh, A. Mueller, M. Scofield, S. Kheirandish, C. Payne, L. Pakhomova, O. Shevchuk, E. A. Gies, D. Fairle, S. A. Malfatti, A. D. Norbek, H. M. Brewer, L. Pasa-Tolic, T. Glavina del Rio, C. A. Suttle, S. Tringe, and S. J. Hallam. A compendium of water column multi-omic sequence information from a seasonally anoxic fjord saanich inlet. *Scientific Data*, (Submitted), 2017.

[128] G. J. Herndl and T. Reinthaler. Microbial control of the dark end of the biological pump. *Nature Geoscience*, 6:718–724, 2013.

[129] G. W. Hill. Algorithm 395: Students tdistribution. *Communications of the ACM*, 13(10):617–619, 1970.

[130] D. M. Hillis. Hillis laboratory, 2017. http://www.zo.utexas.edu/faculty/antisense/Download.html;accessed online 29 March 2017.

[131] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[132] S. Holmes. Bootstrapping phylogenetic trees: Theory and methods. *Statistical Science*, 18(2):241–255, 2003.

[133] L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield. A new view of the tree of life. *Nature microbiology*, 1:16048, 2016.

[134] J. B. Hughes, J. J. Hellmann, T. H. Ricketts, and B. J. M. Bohannan. Counting the uncountable: Statistical approaches to estimating microbial diversity. *Applied Environmental Microbiology*, 67(10): 4399–4406, 2001.

[135] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. Megan analysis of metagenomic data. *Genome Research*, 17:377–386, 2007.

[136] E. Isaacson and H. B. Keller. Analysis of numerical methods. 1994.

[137] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[138] H. Joe. *Dependence Modeling with Copulas*. CRC Press, 2014.

[139] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education Inc., 2007.

[140] D. D. Kang, J. Froula, R. Egan, and Z. Wang. Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 2015.

[141] M. B. Karner, E. F. DeLong, and D. M. Karl. Archaeal dominance in the mesopelagic zone of the pacific ocean. *Nature*, 409:507–510, 2001.

[142] N. Kashtan1, S. E. Roggensack, S. Rodrigue, J. W. Thompson, S. J. Biller, A. Coe, H. Ding, P. Marttinen, R. R. Malmstrom, R. Stocker, M. J. Follows, R. Stepanauskas, and S. W. Chisholm. Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus. *Science*, 344(6182):416–420, 2014.

[143] R. F. Keeling, A. Kortzinger, and N. Gruber. Ocean deoxygenation in a warming world. *Annual Review of Marine Science*, 2:199–229, 2010.

[144] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith. Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493, 2011.

[145] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. *Institute for Catastrophic Loss Reduction*, 2015.

[146] A. Klenke. Probability theory: A comprehensive course. 2013.

[147] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[148] K. O. Konhauser, E. Pecoits, S. V. Lalonde, D. Papineau, E. G. Nisbet, M. E. Barley, N. T. Arndt, K. Zahnle, and B. S. Kamber. Oceanic nickel depletion and a methanogen famine before the great oxidation event. *Nature*, 458:750–753, 2009.

[149] K. M. Konwar, N. W. Hanson, A. P. Pagé, and S. J. Hallam. Metapathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC bioinformatics*, 14(1), 2013.

[150] K. M. Konwar, N. W. Hanson, M. P. Bhatia, D. Kim, S. Wu, A. S. Hahn, C. Morgan-Lang, H. K. Cheung, and S. J. Hallam. Metapathways v2.5: quantitative functional, taxonomic and usability improvements. *Proceedings of the 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, 31(20):3345–3347, 2015.

[151] M. M. M. Kuypers, A. O. Sliekers, G. Lavik, M. Schmid, B. B. Jorgensen, J. G. Kuenen, J. S. S. Damsté, M. Strous, and M. S. M. Jetten. Anaerobic ammonium oxidation by anammox bacteria in the black sea. *Nature*, 422:608–611, 2003.

[152] M. Labrenz, I. Brettar, R. Christen, S. Flavier, J. Bötel, and M. G. Höfle. Development and application of a real-time pcr approach for quantification of uncultured bacteria in the central baltic sea. *Applied and Environmental Microbiology*, 70(8):4971–4979, 2004.

[153] M. Labrenz, J. Grote, K. Mammitzsch, H. T. S. Boschker, M. Laue, G. Jost, S. Glaubitz, and K. Jürgens. Sulfurimonas gotlandica sp. nov., a chemoautotrophic and psychrotolerant epsilonproteobacterium isolated from a pelagic redoxcline, and an emended description of the genus sulfurimonas. *International Journal of Systematic and Evolutionary Microbiology*, 63:4141–4148, 2013.

[154] P. Lam and M. M. Kuypers. Microbial nitrogen cycling processes in oxygen minimum zones. *Annual Review of Marine Science*, 3:317–345, 2011.

[155] M. G. I. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight, R. G. Beiko, and C. Huttenhower. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature Biotechnology*, 31: 814–821, 2013.

[156] G. Lavik, T. Stührmann, V. Brüchert, A. V. der Plas, V. Mohrholz, P. Lam, M. M. zligmann, B. M. Fuchs, R. Amann, U. Lass, and M. M. M. Kuypers. Detoxification of sulphidic african shelf waters by blooming chemolithotrophs. *Nature*, 457:581–584, 2009.

[157] P. Legendre and L. Legendre. *Numerical Ecology*. Elsevier, 3 edition, 2012.

[158] T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber. Tipping elements in the earth's climate system. *Proceedings of the National Academy of Sciences*, 105(6):1786–1793, 2007.

[159] D. Li, C. M. Liu, R. Luo, K. Sadakane, and T. W. Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10): 1674–1676, 2015.

[160] H. Li and R. Durbin. Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[161] R. Li, C.-L. Hsieh, A. Young, Z. Zhang, X. Ren, and Z. Zhao. Illumina synthetic long read sequencing allows recovery of missing sequences even in the finished c. elegans genome. *Scientific Reports*, 5: 10814, 2015.

[162] G. Lima-Mendez, K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux, F. Vincent, L. Bittner, Y. Darzi, J. Wang, S. Audic, L. Berline, G. Bontempi, A. M. Cabello, L. Coppola, F. M. Cornejo-Castillo, F. d'Ovidio, L. D. Meester, I. Ferrera, M.-J. Garet-Delmas, L. Guidi, E. Lara, S. Pesant, M. Royo-Llonch, G. Salazar, P. Snchez, M. Sebastian, C. Souffreau, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, G. Gorsky, F. Not, H. Ogata, S. Speich, L. Stemmann, J. Weissenbach, P. Wincker, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, E. Karsenti, C. Bowler, C. de Vargas, and J. Raes. Determinants of community structure in the global plankton interactome. *Science*, 348(6237), 2015.

[163] X. Lin, S. G. Wakeham, I. F. Putnam, Y. M. Astor, M. I. Scranton, A. Y. Chistoserdov, and G. T. Taylor. Comparison of vertical distributions of prokaryotic assemblages in the anoxic cariaco basin and black sea by use of fluorescence in situ hybridization. *Applied and Environmental Microbiology*, 72(4): 2679–2690, 2006.

[164] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[165] D. Lord, S. Guikema, and S. Geedipally. Application of the conway-maxwell-poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention*, 40(3):1123–1134, 2008.

[166] D. Lord, S. Geedipally, and S. Guikem. Extension of the application of conway-maxwell-poisson models: Analyzing traffic crash data exhibiting under-dispersion. *Risk Analysis*, 30(8):1268–1276, 2010.

[167] S. Louca, A. K. Hawley, S. Katsev, M. Torres-Beltran, M. P. Bhatia, S. Kheirandish, C. C. Michiels, D. Capelle, G. Lavik, M. Doebeli, S. A. Crowe, and S. J. Hallam. Integrating biogeochemistry with multiomic sequence information in a model oxygen minimum zone. *Proceedings of the National Academy of Sciences*, 113(40):E5925–E5933, 2016.

[168] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014.

[169] S. Lücker, M. Wagner, F. Maixner, E. Pelletier, H. Kocha, B. Vacherieb, T. Ratteie, J. S. S. Damsté, E. Spieck, D. L. Paslier, and H. Daims. A nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proceedings of the National Academy of Sciences*, 107(30):13479–13484, 2010.

[170] S. Lücker, B. Nowka, T. Rattei, E. Spieck, and H. Daims. The genome of nitrospina gracilis illuminates the metabolism and evolution of the major marine nitrite oxidizer. *Frontiers in Microbiology*, 4:27, 2013.

[171] R. Lyons. Strong laws of large numbers for weakly correlated random variables. *The Michigan Mathematical Journal*, 35(3):353–359, 1988.

[172] C. Mallows. Another comment on o'cinneide. *The American Statistician*, 45(3):257, 1991.

[173] S. A. Manavski and G. Valle. Cuda compatible gpu cards as efficient hardware accelerators for smith-waterman sequence alignment. *BMC Bioinformatics*, 9(Suppl 2):S10, 2008.

[174] S. S. Mande, M. H. Mohammed, and T. S. Ghosh. Classification of metagenomic sequences: methods and challenges. *Briefings in Bioinformatics*, 13(6):669–681, 2012.

[175] E. Mardis, J. McPherson, R. Martienssen, R. K. Wilson, and W. R. McCombie. What is finished, and why does it matter. *Genome Research*, 12(5):669–671, 2002.

[176] V. M. Markowitz, I. A. Chen, K. Palaniappan, K. Chu, E. Szeto, M. Pillay, A. Ratner, J. Huang, T. Woyke, M. Huntemann, I. Anderson, K. Billis, N. Varghese, K. Mavromatis, A. Pati, N. N. Ivanova, and N. C. Kyrpides. Img 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research*, 42:D560–D567, 2014.

[177] W. Martens-Habbena, P. M. Berube, H. Urakawa, J. R. de la Torre, and D. A. Stahl. Ammonia oxidation kinetics determine niche separation of nitrifying archaea and bacteria. *Nature*, 461:976–981, 2009.

[178] E. P. Martins and T. F. Hansen. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667, 1997.

[179] M. Maurer and M. Boller. Modelling of phosphorus precipitation in wastewater treatment plants with enhanced biological phosphorus removal. *Water science and technology*, 39(1):147–163, 1999.

[180] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, 4:495–500, 2007.

[181] P. McCullagh and J. Nelder. *Generalized Linear Models*. Springer, 1983.

[182] A. C. McHardy, H. G. Martn, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length dna fragments. *Nature Methods*, 4:63–72, 2007.

[183] R. A. McLean, W. L. Sanders, and W. W. Stroup. A unified approach to mixed linear models. *The American Statistician*, 45(1):54–64, 1991.

[184] K. D. McMahon and E. K. Read. Microbial contributions to phosphorus cycling in eutrophic lakes and wastewater. *Annual Review of Microbiology*, 67:199–219, 2013.

[185] P. J. McMurdie and S. Holmes. Waste not, want not: Why rarefying microbiome data is inadmissible. *PLOS Computational Biology*, 10(4):e1003531, 2014.

[186] M. H. Mohammed, T. S. Ghosh, N. K. Singh, and S. S. Mande. Sphinx–an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics*, 27(1):22–30, 2011.

[187] F. Monteiro. *Mechanistic models of oceanic nitrogen fixation*. Massachusetts Institute of Technology, 2009.

[188] D. C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, Inc., 7 edition, 2009. ISBN 9780471661597.

[189] B. A. V. Mooy, R. G. Keila, and A. H. Devol. Impact of suboxia on sinking particulate organic carbon: Enhanced carbon flux and preferential degradation of amino acids via denitrification. *Geochimica et Cosmochimica Acta*, 66(3):457–465, 2002.

[190] J. J. Morris, R. E. Lenski, and E. R. Zinser. The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *mBio*, 3(2):e00036–12, 2012.

[191] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, 5:621–628, 2008.

[192] P. J. Mumby. Statistical power of non-parametric tests: A quick guide for designing sampling strategies. *Marine Pollution Bulletin*, 44:85–87, 2002.

[193] K. P. Murphy. Machine learning: A probabalistic perspective. 2012.

[194] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter. A whole-genome assembly of drosophila. *Science*, 287(5461):2196–2204, 2000.

[195] N. Nagarajan and M. Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14:157–167, 2013.

[196] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 231:289–337, 1933.

[197] D. Nichols, N. Cahoon, E. M. Trakhtenberg, L. Pham, A. Mehta, A. Belanger, T. Kanigan, K. Lewis, and S. S. Epstein. Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species. *Applied Environmental Microbiology*, 76(8):2445–2450, 2010.

[198] P. H. Nielsen, A. T. Mielczarek, C. Kragelund, J. L. Nielsen, A. M. Saunders, Y. Kong, A. A. Hansen, and J. Vollertsen. A conceptual ecosystem model of microbial communities in enhanced biological phosphorus removal plants. *Water Research*, 44:5070–5088, 2010.

[199] M. K. Nobu, T. Narihiro, C. Rinke, Y. Kamagata, S. G. Tringe, T. Woyke, and W.-T. Liu. Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. *The ISME Journal*, 9:1710–1722, 2015.

[200] Nvidia. Cuda c programming guide, 2016. docs.nvidia.com/cuda/cuda-c-programming-guide; accessed online 30-September-2016.

[201] B. B. Oakley, C. A. Morales, J. Line, M. E. Berrang, R. J. Meinersmann, G. E. Tillman, M. G. Wise, G. R. Siragusa, K. L. Hiett, and B. S. Seal. The poultry-associated microbiome: Network analysis and farm-to-fork characterizations. *PLoS ONE*, 8(2):e57190, 2013.

[202] A. Oehmen, A. M. Saunders, M. T. Vives, Z. Yuan, and J. Keller. Competition between polyphosphate and glycogen accumulating organisms in enhanced biological phosphorus removal systems with acetate and propionate as carbon sources. *Journal of Biotechnology*, 123:22–32, 2006.

[203] S. Off, M. Alawi, and E. Spieck. Enrichment and physiological characterization of a novel nitrospira-like bacterium obtained from a marine sponge. *Applied and Environmental Microbiology*, 76(14): 4640–4646, 2010.

[204] N. R. Pace. Mapping the tree of life: Progress and prospects. *Microbiology and Molecular Biology Reviews*, 73(4):565–576, 2009.

[205] J. M. Papakonstantinou. *Historical Development of the BFGS Secant Method and Its Characterization Properties*. Rice University, 2009.

[206] J. M. Papakonstantinou and R. A. Tapia. Origin and evolution of the secant method in one dimension. *The American Mathematical Monthly*, 120(6):500–518, 2013.

[207] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25:1043–1055, 2015.

[208] K. R. Patil, L. Roune, and A. C. McHardy. The phylopythias web server for taxonomic assignment of metagenome sequences. *PLoS One*, 7(6):e38581, 2012.

[209] A. Paulmier and D. Ruiz-Pino. Oxygen minimum zones (omzs) in the modern ocean. *Progress in Oceanography*, 80(3-4):113–128, 2009.

[210] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2: 559–572, 1901.

[211] W. R. Pearson. An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, Chapter 3:Unit3.1, 2013.

[212] M. A. Pena and S. J. Bograd. Time series of the northeast pacific. *Progress in Oceanography*, 75(2): 115–119, 2007.

[213] A. Perry. A class of conjugate gradient algorithms with a two-step variable metric memory. *Northwestern University, Center for Mathematical Studies in Economics and Management Science, Evanston, IL*, Discussion Paper 269, 1977.

[214] M. Pester, C. Schleper, and M. Wagner. The thaumarchaeota: an emerging view of their phylogeny and ecophysiology. *Current Opinion in Microbiology*, 14(3):300–306, 2011.

[215] P. A. Pevzner, H. Tang, and M. S. Waterman. An eulerian path approach to dna fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, 2001.

[216] G. L. Phyllis Lam and, M. M. Jensen, J. van de Vossenberg, M. Schmid, D. Woebken, D. Gutiérrez, R. Amanna, M. S. M. Jetten, and M. M. M. Kuypersa. Revising the nitrogen cycle in the peruvian oxygen minimum zone. *Proceedings of the National Academy of Sciences*, 106(12):4752–4757, 2009.

[217] R. Piessens, E. deDoncker Kapenga, C. Uberhuber, and D. Kahaner. *Quadpack: a Subroutine Package for Automatic Integration*. Springer, 1983.

[218] M. Pourahmadi. *High-Dimensional Covariance Estimation*. Wiley, 2013.

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

[220] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 2013.

[221] R Core Team. Writing r extensions, 2016. cran.r-project.org/doc/manuals/r-release/R-exts.html;accessed online 4 Oct 2016.

[222] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[223] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062): 1518–1524, 2011.

[224] L. J. Revell. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, 1(4):319–329, 2010.

[225] C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, and T. Woyke. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499:431–437, 2013.

[226] D. Risso, J. Ngai, T. P. Speed, and S. Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32:896–902, 2014.

[227] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[228] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[229] R. Rohde, R. A. Muller, R. Jacobsen, E. Muller, S. Perlmutter, A. Rosenfeld, J. Wurtele, D. Groom3, and C. Wickham. A new estimate of the average earth surface land temperature spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview*, 1, 2012.

[230] S. Roux, A. K. Hawley, M. T. Beltran, M. Scofield, P. Schwientek, R. Stepanauskas, T. Woyke, S. J. Hallam, and M. B. Sullivan. Ecology and evolution of viruses infecting uncultivated sup05 bacteria as revealed by single-cell- and meta-genomics. *eLife*, 3:e03125, 2014.

[231] Q. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics*, 22(20):2532–2538, 2006.

[232] W. Rudin. *Principles of Mathematical analysis*. McGraw-Hill, 1976.

[233] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[234] N. Sangwan, F. Xia, and J. A. Gilbert. Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, 4:8, 2016.

[235] K. Sellers and G. Shmueli. A flexible regression model for count data. *Annals of Applied Statistics*, 4(2): 943–961, 2010.

[236] A. L. Sessions, D. M. Doughty, P. V. Welander, R. E. Summons, and D. K. Newman. The continuing puzzle of the great oxidation event. *Current Biology*, 19(14):R567–R574, 2009.

[237] V. Shah, B. X. Chang, and R. M. Morris. Cultivation of a chemoautotroph from the sup05 clade of marine bacteria that produces nitrite and consumes ammonium. *The ISME Journal*, 2016.

[238] D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

[239] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[240] I. Sharon and J. F. Banfield. Genomes from metagenomics. *Science*, 342(6162):1057–1058, 2013.

[241] C. S. Sheik, S. Jain, and G. J. Dick. Metabolic flexibility of enigmatic sar324 revealed through metagenomics and metatranscriptomics. *Environmental Microbiology*, 16(1):304–317, 2013.

[242] G. Shmueli, T. Minka, J. Kadane, S. Borle, and P. Boatwright. A useful distribution for fitting discrete data: revival of the conway-maxwell-poisson distribution. *Journal of the Royal Statistical Society: Series C*, 54(1):127–142, 2005.

[243] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[244] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. Abyss: A parallel assembler for short read sequence data. *Genome Research*, 19:1117–1123, 2009.

[245] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris*, 8:229–231, 1959.

[246] C. Spearman. 'general intelligence', objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

[247] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[248] A. Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[249] R. Stepanauskas and M. E. Sieracki. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proceedings of the National Academy of Sciences*, 104(21):9052–9057, 2007.

[250] L. Stramma, G. C. Johnson, J. Sprintall, and V. Mohrholz. Expanding oxygen-minimum zones in the tropical oceans. *Science*, 320(5876):655–658, 2008.

[251] M. Strous, J. A. Fuerst, E. H. M. Kramer, S. Logemann, G. Muyzer, K. T. van de Pas-Schoonen, R. Webb, J. G. Kuenen, and M. S. M. Jetten. Missing lithotroph identified as new planctomycete. *Nature*, 400:446–449, 1999.

[252] M. Sunamura, Y. Higashi, C. Miyako, J. ichiro Ishibashi, and A. Maruyama. Two bacteria phylotypes are predominant in the suiyo seamount hydrothermal plume. *Applied and Environmental Microbiology*, 70(2):1190–1198, 2004.

[253] B. K. Swan, M. Martinez-Garcia, C. M. Preston, A. Sczyrba, T. Woyke, D. Lamy, T. Reinthaler, N. J. Poulton, D. P. Maslandm, M. L. Gomez, M. E. Sieracki, E. F. DeLong, G. J. Herndl, and R. Stepanauskas. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science*, 333(6047):1296–1300, 2011.

[254] T. Tatusova, S. Ciufo, B. Fedorov, K. O'Neill, and I. Tolstoy. Refseq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Research*, 42(1):D553–9, 2014.

[255] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, 6(9):938–947, 2004.

[256] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, 5:163, 2004.

[257] K. Tennessen, E. Andersen, S. Clingenpeel, C. Rinke, D. S. Lundberg, J. Han, J. L. Dangl, N. Ivanova, T. Woyke, N. Kyrpides, and A. Pati. Prodege: a computational protocol for fully automated decontamination of genomes. *The ISME Journal*, 10:269–272, 2016.

[258] C. J. F. Ter Braak. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5):1167–1179, 1986.

[259] The NIH HMP Working Group, J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, C. C. Baker, V. D. Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A. R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer. The nih human microbiome project. *Genome Research*, 19:2317–2323, 2009.

[260] T. Thomas, J. Gilbert, and F. Meyer. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2:3, 2012.

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

[262] O. Ulloa, D. E. Canfield, E. F. DeLong, R. M. Letelier, and F. J. Stewart. Microbial oceanography of anoxic oxygen minimum zones. *Proceedings of the National Academy of Sciences*, 109(40):15996–16003, 2012.

[263] A. Ultsch and F. Mörchen. Esom-maps: tools for clustering, visualization, and classification with emergent som. *Germany: Data Bionics Research Group, University of Marburg*, 2005.

[264] W. N. Venables and B. D. Ripley. Modern applied statistics with s. 2002.

[265] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):426–482, 1943.

[266] D. A. Walsh, E. Zaikova, C. G. Howes, Y. C. Song, J. J. Wright, S. G. Tringe, P. D. Tortell, and S. J. Hallam. Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science*, 326(5952):578–582, 2009.

[267] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, , and R. Knight. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, 10:1669–1681, 2016.

[268] D. L. Wheeler, D. M. Church, A. E. Lash, D. D. Leipe, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, T. A. Tatusova, L. Wagner, and B. A. Rapp. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 29(1):11–16, 2001.

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

[270] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62, 1938.

[271] J. J. Wright, K. M. Konwar, and S. J. Hallam. Microbial ecology of expanding oxygen minimum zones. *Nature Reviews Microbiology*, 10:381–394, 2012.

[272] J. J. Wright, K. Mewis, N. W. Hanson, K. M. Konwar, K. R. Maas, and S. J. Hallam. Genomic properties of marine group a bacteria indicate a role in the marine sulfur cycle. *The ISME Journal*, 8:455–468, 2014.

[273] T. D. Wright, K. L. Vergin, P. W. Boyd, and S. J. Giovannoni. A novel delta-subdivision proteobacterial lineage from the lower ocean surface layer. *Applied and Environmental Microbiology*, 63(4):1441–1448, 1997.

[274] K. C. Wrighton, B. C. Thomas, I. Sharon, C. S. Miller, C. J. Castelle, N. C. VerBerkmoes, M. J. Wilkins, R. L. Hettich, M. S. Lipton, K. H. Williams, P. E. Long, and J. F. Banfield. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, 337(6102):1661–1665, 2012.

[275] Y. Wu, Y. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer. Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2:26, 2014.

[276] Y.-W. Wu, B. A. Simmons, and S. W. Singer. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, 2015.

[277] F. Xia, J. Chen, W. K. Fung, and H. Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69:1053–1063, 2013.

[278] L. C. Xia, J. A. Steele, J. A. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman, and F. Sun. Extended local similarity analysis (elsa) of microbial community and other time series data with replicates. *BMC Systems Biology*, 5(Suppl 2):S15, 2011.

[279] E. Zaikova, D. A. Walsh, C. P. Stilwell, W. W. Mohn, P. D. Tortell, and S. J. Hallam. Microbial community dynamics in a seasonally anoxic fjord: Saanich inlet, british columbia. *Environmental Microbiology*, 12(1):172–191, 2009.

[280] D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18:821–829, 2008.

[281] J. Zhou, Y. Deng, F. Luoe, Z. He, and Y. Yang. Phylogenetic molecular ecological network of soil microbial communities in response to elevated co2. *mBio*, 2(4):e00122–11, 2011.

[282] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[283] Y. Zuo, G. Yu, M. G. Tadesse, and H. W. Ressom. Biological network inference using low order partial correlation. *Methods*, 69(3):266–273, 2014.

# Appendix A

# Data-driven argument as a Hidden Markov Model

Recognizing microbial ecology as an information science [120] emphasizes the data-driven nature of our arguments. Modern microbial ecology layers data products, feeding the output of one machine (or scientist) into another. For example, to achieve the goal of metabolic pathway prediction from a genome, (1) DNA must be extracted from a sample, then prepared and sent for sequencing; (2) a DNA sequencing machine [26] then translates the DNA into digital representation as many fragments of As,Ts,Cs, and Gs; (3) fragments of DNA must then be assembled into larger, more useful contiguous sequences (contigs) by an assembling software [159]; (4) then assembly may be translated into metabolic pathway predictions by a software [149] which is also a pipeline of further sub-modules. Another important example for ecology is the counting of microbes (more accurately, their 16S genes), where (1) after DNA extraction, polymerase chain reaction (PCR) is used to amplify target DNA; (2) amplified DNA is sequenced; (3) reads are counted with a software, QIIME [51]; (4) then 16S genes must be aligned to a reference database such as GreenGenes [71] or SILVA [220] with local-alignment software like BLAST [9] or LAST [144]. At the end of these two examples human interpretation relies heavily on the successful layering of data products. In either example, information is processed in a factory-style assembly line, with information being passed from machine to machine. The desired end product results in each assembly line occasionally sharing steps, and inevitably diverging as unique products are desired. Each machine may be viewed as a singular unit, or decomposed into its own assembly lines, and relies on a degree of precision from previous machines to reliably build its data product.

Elaborating with the formalisms defined in section 1.6.5, we may imagine our assembly line of machines $C_i$ attributing $x_i \in x_{i+1}$ whenever $C_i(x_i) = x_{i+1}$. This is where ecological narratives meet predictive power. The reality of Microbial ecology is that we often cannot observe true attributions $\{x_i \in x_{i+1}\}$, and we may not know if the narrative or model they constitute is correct, because we only every observe our own attributions $\{C_i(x_i) = x_{i+1}\}$. We indirectly observe. Fortunately we are able to construct precision estimators, which tie our narratives to reality. Without precision, these narratives are irrelevant to reality.

## A.1 Theoretical argument

In this section, we emphasize the importance of precision in data-driven argument by demonstrating how it may effect the precision of our entire argument. The conclusion is that precision is required at key points in our arguments and pipelines in order to be confident in our eventual conclusions. This is done by assuming indirect observation is routine, and likening our data analysis pipelines to a Hidden Markov Model (HMM).

To necessarily sophisticate our argument, allow a capital letter $X_i$ to denote the random variable representation of data product $x_i$. When we require that an $X_i$ equate with a particular value $x_i$ via $X_i = x_i$, we constrain our argument. For example, $X_i$ may stand in for a random taxa (which we may attribute via $C_{i-1}(x_{i-1}) = X_i$), but if we require that taxa to be $x_i = \{\text{Thaumarcheota}\}$, then have the constraint $X_i = x_i$. Constraining data products allows us to construct data-driven arguments. A thorough example of this process is in section A.2.

Define a *narrative* to have the form $\{X_1 \in X_2, X_2 \in X_3, \ldots, X_{n-1} \in X_n, X_n \in x_{n+1}\}$, where the final conclusion is non-random, and we may constrain certain data products to be non-random, $X_i = x_i$.

$$\begin{array}{ccccccc} [x_1 \in x_2] & \longrightarrow & [x_2 \in x_3] & \longrightarrow & \ldots & \longrightarrow & [x_n \in x_{n+1}] \\ \downarrow & & \downarrow & & & & \downarrow \\ [C_1(x_1) = x_2] & & [C_2(x_2) = x_3] & & & & [C_n(x_n) = x_{n+1}] \end{array}$$

**Figure A.1:** A linearly dependent Hidden Markov Model analogizing an inferential pipeline.

We actually make a data-driven argument for our narrative via our *pipeline*, $\{C_1(X_1) = X_2, C_2(X_2) = X_3, \ldots, C_{n-1}(X_{n-1}) = X_n, C_n(X_n) = x_{n+1}\}$, where again the final conclusion is non-random and certain data products may be constrained to non-random. We observe all pipeline events $\{C_i(X_i) = X_{i+1}\}$, and we observe none of the narrative events $\{X_i \in X_{i+1}\}$. A *data-driven argument* is of the form $\{X_n \in x_{n+1} | C_1(X_1) = X_2, C_2(X_2) = X_3, \ldots, C_{n-1}(X_{n-1}) = X_n, C_n(X_n) = x_{n+1}\}$. So we constrain our pipeline to produce a data-driven arugment for a narrative.

To leverage pre-existing theory for HMMs [23, 222], we conveniently assume that $\{C_i(X_i) = X_{i+1}\}$ only conditionally depends on $\{X_i \in X_{i+1}\}$, and that $\{X_i \in X_{i+1}\}$ only conditionally depends on $\{X_{i-1} \in X_i\}$, where conditional dependence is a probability concept. This provides the dependency structure illustrated in Figure A.1. Define the data-driven argument's *final precision* as $\mathbb{P}[X_n \in x_{n+1} | \cap_{i=1}^{n} \{C_i(X_i) = X_{i+1}\}]$, and $i^{th}$ *local precision* as $\mathbb{P}[X_i \in x_{i+1} | C_i(X_i) = x_{i+1}]$. Our goal is to understand the final precision through a series of known local precisions. Because we sometimes must constrain our arguments, the following conditional decomposition is useful. Let $1 < m < n$ and constrain $X_m = x_m$.

$$\begin{aligned} &\mathbb{P}[X_n \in x_{n+1}, X_{m-1} \in x_m | \cap_{i=1}^{n} \{C_i(X_i) = X_{i+1}\}] \\ &= \mathbb{P}[X_n \in x_{n+1} | X_{m-1} \in x_m, \cap_{i=1}^{n} \{C_i(X_i) = X_{i+1}\}] \\ &\times \mathbb{P}[X_{m-1} \in x_m | \cap_{i=1}^{n} \{C_i(X_i) = X_{i+1}\}] \\ &= \mathbb{P}[X_n \in x_{n+1} | X_{m-1} \in x_m, \cap_{i=m}^{n} \{C_i(X_i) = X_{i+1}\}] \\ &\times \mathbb{P}[X_{m-1} \in x_m | \cap_{i=1}^{m} \{C_i(X_i) = X_{i+1}\}] \end{aligned}$$

We've decomposed our entire final precision, breaking at the constrained point, into two final precisions for sub-pipelines, the later of which is conditioned on the other. Unfortunately, $\{X_{m-1} \in x_m\}$ is never observed, so instead we require that the initial precision is near one, $\mathbb{P}[X_{m-1} \in x_m | \cap_{i=1}^{m} \{C_i(X_i) = X_{i+1}\}] \approx 1$. If this is true, then we may assume that $\{X_{m-1} \in x_m\}$ and thus $\mathbb{P}[X_n \in x_{n+1} | X_{m-1} \in x_m, \cap_{i=m}^{n} \{C_i(X_i) = X_{i+1}\}] \approx \mathbb{P}[X_n \in x_{n+1} | \cap_{i=m}^{n} \{C_i(X_i) = X_{i+1}\}]$. So we arrive at the following approximation.

$$\begin{aligned} &\mathbb{P}[X_n \in x_{n+1}, X_{m-1} \in x_m | \cap_{i=1}^{n} \{C_i(X_i) = X_{i+1}\}] \\ &\approx \mathbb{P}[X_n \in x_{n+1} | \cap_{i=m}^{n} \{C_i(X_i) = X_{i+1}\}] \times \mathbb{P}[X_{m-1} \in x_m | \cap_{i=1}^{m} \{C_i(X_i) = X_{i+1}\}] \\ &\text{when } \mathbb{P}[X_{m-1} \in x_m | \cap_{i=1}^{m} \{C_i(X_i) = X_{i+1}\}] \approx 1 \end{aligned}$$

This approximation gives us a way to understand our final precision in terms of other final precisions of our decomposed pipeline. This decomposition is necessary for developing precise and constrained data-driven arguments.

Next we leverage the forward algorithm from HMM theory to realize how important local precision at constrained points is for achieving high final precisions. The forward algorithm is an iterative (and dynamic programming) method for calculating $\mathbb{P}[X_n \in x_{n+1}, \cap_{i=1}^{n} \{C_i(X_i) = X_{i+1}\}]$ from $\mathbb{P}[X_{n-1} \in x_n, \cap_{i=1}^{n-1} \{C_i(X_i) = X_{i+1}\}]$. We will equivalently rephrase it in terms of precisions as follows.

$$\begin{aligned} &\mathbb{P}[X_n \in x_{n+1} | \cap_{i=1}^{n} \{C_i(X_i) = X_{i+1}\}] \\ &= \mathbb{P}[X_n \in x_{n+1} | C_n(X_n) = x_{n+1}] \\ &\times \mathbb{P}[C_n(X_n) = x_{n+1}] \left(\mathbb{P}[X_n \in x_{n+1}] \mathbb{P}[C_n(X_n) = x_{n+1} | \cap_{i=1}^{n} \{C_i(X_i) = X_{i+1}\}]\right)^{-1} \end{aligned}$$

$\times \sum_{x_{n-1}} \mathbb{P}[X_n \in x_{n+1} | x_{n-1} \in X_n] \mathbb{P}[x_{n-1} \in X_n | \cap_{i=1}^{n-1} \{C_i(X_i) = X_{i+1}\}]$

This puts our final precision into a form equivalent to a product of a local precision and sub-pipeline's final precisions as follows.

$$[\text{final precision } X_n] = [\text{local precision } X_n] a_n \sum_{x_{n-1}} b_n(x_{n-1}) [\text{final precision } x_{n-1}]$$

This shows that a low local precision can lower the final precision. For example, if $C_n(\cdot) = x_{n+1}$ constantly, then our final precision is bounded above by $\mathbb{P}[X_n \in x_{n+1}]$. We have shown that pipeline component precision can bound the final precision of our data driven arguments. Therefore an imprecise pipeline component may reduce confidence in our narrative.

## A.2 An example

The HMM representation necessarily admits more complexity than in Figure A.1, which is best communicated through an example. Consider 16S correlation problem mentioned earlier. Imagine a scientist is using that pipeline among others to infer microbial syntrophy, perhaps cyclical redox of sulfur between two taxa. The pipelines likely share steps (genetic sequencing is popular) and meet at the end where interpretation occurs, but we focus on the correlation pipeline. Consider the following final precision.

$$\mathbb{P}[X_6 \in x_7 | C_1(X_1) = X_2, C_2(X_2) = x_3, C_3(x_3) = X_4, C_4(X_4) = X_5, C_5(X_5) = X_6, C_6(X_6) = x_7]$$

Where we might have the following event values.
$\{C_6(x_6) = x_7\} = \{$Scientist asserts syntrophy, perhaps utilizing other pipelines$\}$
$\{x_6 \in x_7\} = \{$Syntrophy is genuine, correlation is not illusory$\}$
$\{C_5(x_5) = x_6\} = \{$taxa correlations are statistically significant$\}$
$\{x_5 \in x_6\} = \{$taxa abundances do covary, inference is not an illusion of natural variation$\}$
$\{C_4(x_4) = x_5\} = \{$16S genes align to database entries$\}$
$\{x_4 \in x_5\} = \{$alignment corresponds to actual source$\}$
$\{C_3(x_3) = x_4\} = \{$QIIME counts clustered 16S reads$\}$
$\{x_3 \in x_4\} = \{$QIIME counts resemble authentic phylogenetic structure$\}$
$\{C_2(x_2) = x_3\} = \{$DNA is sequenced$\}$
$\{x_2 \in x_3\} = \{$sequenced DNA resembles true DNA$\}$
$\{C_1(x_1) = x_2\} = \{$16S genes are amplified$\}$
$\{x_1 \in x_2\} = \{$16S amplification primers are not biased against final taxa$\}$

In this example, imagine a highly precise short-read Illumina platform was used for sequencing. So if we discover target taxa reads via $\{C_2(X_2) = x_3\}$ we basically observe $\{X_2 \in x_3\}$, and therefore we may consider bias against our target taxa irrelevant. Read counts may be low, but their covariation is still detectable. Under this interpretation (and disregarding others), we have Markovian behaviour as follows.

$\mathbb{P}[X_6 \in x_7 | \cap_{i=1}^{n} \{C_i(X_i) = X_{i+1}\}] \approx \mathbb{P}[X_6 \in x_7 | \cap_{i=3}^{n} \{C_i(X_i) = X_{i+1}\}]$

Now, further imagine that our scientific narrative requires that we constrain $X_6 = x_6$, a particular correlation does truly exist. According to our exlorations of the forward algorithm, a low precision $\mathbb{P}[X_5 \in x_6 | C_5(X_5) = x_6]$ could jeopardize the precision of our narrative.

# Appendix B

# CMP variance bound proof

In this appendix we prove that a CMP with mean $\mu$ and variance $\sigma^2$ satisfies $\sigma^2 < \mu(\mu+1)$. Let $\mathbb{R}_{>0} = (0, \infty)$. Our proof strategy will start with strategic preliminary proofs. Then because the CMP is parameterized in terms of parameters $(\lambda, \nu) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}$, we reparameterize the CMP to have known mean $\mu$ by setting $\lambda = \lambda_{\mu,\nu}$, where $\lambda_{\mu,\nu}$ is a function of both $\mu$ and $\nu$. Notice that the reparameterized CMP thus has a reparameterized variance $\sigma^2_{\mu,\nu}$, which is also a function of $\mu$ and $\nu$. We then study the reparameterized CMP to eventually discover the following two results.

**Definition 1.** *A random variable $X$ is CMP$(\lambda, \nu)$ distributed if $\mathbb{P}[X = x] = (\lambda^x / x!^\nu) / (\sum_{j=0}^{\infty} \lambda^j / j!^\nu)$ for $x \in \mathbb{Z}_{\geq 0}$. This is written as $X \sim$ CMP$(\lambda, \nu)$.*

**Result 3.** *If $X$ follows a CMP distribution and has mean $\mu$ and variance $\sigma^2$, then $\sigma^2 < \mu(\mu+1)$.*

**Result 4.** *If $X$ follows a CMP distribution with parameters $\lambda$ and $\nu$, then $X$ is over-dispersed ($\sigma^2_X > \mu_X$) when $\nu < 1$ and under-dispersed ($\sigma^2_X < \mu_X$) when $\nu > 1$.*

## B.1 Preliminaries

**Lemma 1.** *Let functions $f, g, h$ be from $\mathbb{R}$ to $\mathbb{R}$, $h$ strictly monotone, and a random variable $X$ such that at least one of $\mathbb{E}[f(X)|h(X)]$ or $\mathbb{E}[g(X)|h(X)]$ is non-linear in $h(X)$.*

*For brevity, let $f = f(X), g = g(X), h = h(X), \sigma_{A,B} = Cov(A, B), \rho_{A,B} = Cor(A, B)$,*
*$\beta_{A,B} = \sigma_{A,B}/\sigma_{B,B}$, $\sigma_{A,B\cdot C}$ be the covariance of $A$ and $B$ partial $C$,*
*and $\sigma_{A,B|C}$ be the conditional covariance of $A$ and $B$ conditioned on $C$.*
*Then we have the following.*

$$\rho_{f,g} - \rho_{f,h}\rho_{g,h} \propto \sigma_{f,g\cdot h} \neq 0$$

*Where proportionality $\propto$ indicates scaling by a positive real value.*

*Proof.* of Lemma 1

$\rho_{f,g} - \rho_{f,h}\rho_{g,h} \propto \sigma_{f,g} - \beta_{g,h}\sigma_{f,h} = \sigma_{f,g} - \beta_{g,h}\sigma_{f,h} - \beta_{f,h}\sigma_{g,h} + \beta_{f,h}\beta_{g,h}\sigma_{h,h}$
$= \sigma_{f-\beta_{f,h}h, g-\beta_{g,h}h} = \sigma_{f,g\cdot h}$
$= \mathbb{E}\left[\sigma_{f-\beta_{f,h}h, g-\beta_{g,h}h|h}\right] + \sigma_{\mathbb{E}[f-\beta_{f,h}h|h], \mathbb{E}[g-\beta_{g,h}h|h]}$ ; (total covariance)
$= \mathbb{E}\left[\sigma_{f,g|h}\right] + \sigma_{\mathbb{E}[f-\beta_{f,h}h|h], \mathbb{E}[g-\beta_{g,h}h|h]}$ ; ($h$ constant under $\mathrm{Cov}(\cdot|h)$)
$= 0 + \sigma_{\mathbb{E}[f-\beta_{f,h}h|h], \mathbb{E}[g-\beta_{g,h}h|h]}$ ; ($h$ strict monotone $\Rightarrow f, g$ are known given $h$)

By Theorem 6 section B.4, $\sigma_{f,g\cdot h} = \mathbb{E}\left[\sigma_{f,g|h}\right] = 0$
$\Leftrightarrow \mathbb{E}[f|h]$ and $\mathbb{E}[g|h]$ are linear in $h$.
By assumption this is false, so $\sigma_{f,g\cdot h} = \sigma_{\mathbb{E}[f-\beta_{f,h}h|h], \mathbb{E}[g-\beta_{g,h}h|h]} \neq 0$.
Further, sign $\left(\rho_{f,g} - \rho_{f,h}\rho_{g,h}\right) =$ sign $\left(\sigma_{f,g\cdot h}\right)$.

□

**Lemma 2.** *(a) For each $\nu > 0, \sum_{j=0}^{N} \frac{\lambda^j}{j!^\nu} \to Z_{\lambda,\nu} \in \mathbb{R}$ uniformly in $\lambda$ as $N \to \infty$. (b) $\frac{\partial}{\partial\nu} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!^\nu} = \sum_{j=0}^{\infty} \frac{\partial}{\partial\nu} \frac{\lambda^j}{j!^\nu}$*

*Proof.* of Lemma 2 (a)

$\sum_{j=0}^{\infty} \frac{\lambda^j}{j!^\nu}$ is a power series in $\lambda$ and $\lim_{j\to\infty} \left| \frac{(j+1)!^{-\nu}}{j!^{-\nu}} \right| = 0 \Rightarrow \sum_{j=0}^{\infty} \frac{\lambda^j}{j!^\nu}$ is uniformly convergent for all $\lambda$.

□

**Lemma 3.** *For all $c > 1$, and $\lambda > 0, j \in \mathbb{Z}_{\geq 1}$, there exists an $L \in \mathbb{R}_{>0}$ such that $j^c \lambda^j \leq L^j$ and $L > \lambda$*

*Proof.* of Lemma 3

For all $(c, \lambda, j) \in \mathbb{R}^2_{>0} \times \mathbb{Z}_{\geq 0}$, there exists an $L = e^{c+\log\lambda} > \lambda$ *implies* $L^j = \exp[j(c + \log\lambda)] \geq \exp[c \log j + j \log\lambda] = j^c \lambda^j$

□

The following lemma is implicit in result 2.2 of Shmueli et al. [242]. Let $\overline{\mathbb{R}}_{\geq 0}$ denote $\mathbb{R}_{\geq 0} \cup \{\infty\}$.

**Lemma 4.** *If $X \sim CMP(\lambda, \nu)$, then $X$ has finite positive integer moments.*

*Proof.* of Lemma 4

$h_{n,i} = (1 - \frac{1}{n}) i^m \frac{\lambda^i}{i!^\nu}$ *for* $n \in \mathbb{Z}_{\geq 2} \Rightarrow h_{n,i} < h_{n+1,i}$

*and* $\lim_{n\to\infty} h_{n,i} = i^m \frac{\lambda^i}{i!^\nu}$

*then* $\mathbb{E}[X^m] = \lim_{n\to\infty} \left(\sum_{i=1}^{\infty} h_{n,i}\right) Z_{\lambda,\nu}^{-1} = \left(\sum_{i=1}^{\infty} i^m \frac{\lambda^i}{i!^\nu}\right) Z_{\lambda,\nu}^{-1} \in \overline{\mathbb{R}}_{>0}$

(by Monotone Convergence Theorem (MCT)).

For all $m \in \mathbb{Z}_{\geq 1}$, there exists $L_m$ such that

$\mathbb{E}[X^m] = \sum_{i=1}^{\infty} i^m \frac{\lambda^i}{i!^\nu} Z_{\lambda,\nu}^{-1} \leq \sum_{i=0}^{\infty} \frac{L_m^i}{i!^\nu} Z_{\lambda,\nu}^{-1}$ ; (Lemma 3)

$= Z_{L_m,\nu} Z_{\lambda,\nu}^{-1} < \infty$ ; (Lemma 2 (a))

□

Let $a \downarrow b$ mean that $a$ decreases as $b$ increases. Let $a \uparrow b$ mean that $a$ increases as $b$ increases.

*Proof.* of Lemma 2 (b)

$f_n(\nu) = \sum_{i=0}^{n} \frac{\lambda^i}{i!^\nu}$ then

$\frac{\partial}{\partial\nu} f_n(\nu) = -\sum_{i=0}^{n} \log(i!) \frac{\lambda^i}{i!^\nu} \uparrow \nu \; \forall \nu \in \mathbb{R}_{>0}$

and $\log(i!) \leq i^2$ implies $\frac{\partial}{\partial\nu} f_n(\nu) \to \cdot$ (Lemma 4).

Thus $\frac{\partial}{\partial\nu} f_n(\nu)$ is locally uniformly convergent

(by the mean value theorem)

and $f_n(\nu) \to \cdot$ (Lemma 2 (a))

implies $\forall \nu \in \mathbb{R}_{>0}, \lim_{n\to\infty} \frac{\partial}{\partial\nu} f_n(\nu) = \frac{\partial}{\partial\nu} \lim_{n\to\infty} f_n(\nu)$

Apply Rudin's theorem Theorem 5.

□

**Lemma 5.** $\frac{\partial}{\partial\lambda} \log\left(1 + \lambda + \sum_{j=2}^{\infty} \frac{\lambda^j}{j!^\nu}\right) > \frac{d}{d\lambda} \log(1 + \lambda)$

*Proof.* of Lemma 5

$\frac{\partial}{\partial\lambda} \log\left(1 + \lambda + \sum_{j=2}^{\infty} \frac{\lambda^j}{j!^\nu}\right) > \frac{d}{d\lambda} \log(1 + \lambda)$

if and only if $\frac{1 + \sum_{j=2}^{\infty} j \frac{\lambda^{j-1}}{j!^\nu}}{1 + \lambda + \sum_{j=2}^{\infty} \frac{\lambda^j}{j!^\nu}} > \frac{1}{1+\lambda}$

if and only if $\left(1 + \sum_{j=2}^{\infty} j \frac{\lambda^{j-1}}{j!^\nu}\right)(1 + \lambda) = 1 + \sum_{j=2}^{\infty} j \frac{\lambda^{j-1}}{j!^\nu} + \lambda + \sum_{j=2}^{\infty} j \frac{\lambda^j}{j!^\nu} > 1 + \lambda + \sum_{j=2}^{\infty} \frac{\lambda^j}{j!^\nu}$

□

## B.2 Properties of $\lambda_{\mu,\nu}$

In this section we develop, $\lambda_{\mu,\nu}$, a tool for fixing the expected value of the CMP distribution while manipulating the variance. It is shown that, given a $\nu$, $\mu$ and $\lambda$ are in one-to-one correspondence through the function $\lambda_{\mu,\nu}$. It should be noted that $\lambda_{\mu,\nu}$ is calculated algorithmically in practice.

**Definition 2.**

$$\textit{For each } (\mu,\nu) \in \mathbb{R}^2_{>0},\ \lambda_{\mu,\nu} := \lambda \textit{ such that } \mathbb{E}[X] = \mu \textit{ and } X \sim CMP(\lambda,\nu)$$

Defining such a $\lambda$ as above does not guarantee its existence nor uniqueness. This section its proves uniqueness and existence.

Let $\mu_{\lambda_\nu} = \mathbb{E}[X]$ for $X \sim CMP(\lambda,\nu)$.
Let $\sigma^2_{\lambda,\nu} = \mathrm{Var}[X]$ for $X \sim CMP(\lambda,\nu)$.

**Lemma 6.** *For a fixed $\nu \in \mathbb{R}_{>0}$, and for all $\mu \in \mathbb{R}_{>0}$, there exists a unique $\lambda \in \mathbb{R}_{>0}$ such that $\mu_{\lambda,\nu} = \mu$.*

**Lemma 7.** *For fixed $\nu \in \mathbb{R}_{>0}$, $\mu_{\lambda,\nu} \to 0$ as $\lambda \to 0^+$*

*Proof.* of Lemma 7
$\lim_{\lambda\to 0^+} \mu_{\lambda,\nu} = \left(\lim_{\lambda\to 0^+} \sum_{i=1}^{\infty} i\frac{\lambda^i}{i!^\nu}\right)\left(\lim_{\lambda\to 0^+} Z_{\lambda,\nu}\right)^{-1}$
$= \left(\sum_{i=1}^{\infty} \lim_{\lambda\to 0^+} i\frac{\lambda^i}{i!^\nu}\right)\left(\sum_{j=0}^{\infty} \lim_{\lambda\to 0^+} \frac{\lambda^j}{j!^\nu}\right)^{-1}$ ; (MCT)
$= \left(\sum_{i=1}^{\infty} 0\right)(1+0)^{-1} = \frac{0}{1+0} = 0$

□

**Lemma 8.** *For each $\nu \in \mathbb{R}_{>0}$, $\mu_{\lambda,\nu}$ is strictly increasing in $\lambda$, and $\frac{\partial}{\partial\lambda}\mu_{\lambda,\nu} = \sigma^2_{\lambda,\nu}/\lambda$*

*Proof.* of Lemma 8
$\mu_{\lambda,\nu}$ is a ratio of convergent power series
(apply Lemmas 2 (a), 3)
Convergent power series are continuous.
$Z_{\lambda,\nu} > 0$ for each $(\lambda,\nu) \in \mathbb{R}^2_{>0}$.
$\frac{\partial}{\partial\lambda}\mu_{\lambda,\nu} = \frac{\partial}{\partial\nu}\left(\sum_{j=1}^{\infty} j\frac{\lambda^j}{j!^\nu}\right) Z^{-1}_{\lambda,\nu} =$
$\left[\left(\frac{\partial}{\partial\lambda} j\frac{\lambda^j}{j!^\nu}\right) Z_{\lambda,\nu} - \left(\sum_{j=1}^{\infty} j\frac{\lambda^j}{j!^\nu}\right)\left(\sum_{j=0}^{\infty} \frac{\partial}{\partial\lambda}\frac{\lambda^j}{j!^\nu}\right)\right] Z^{-2}_{\lambda,\nu}$ ; (Lemma 2 (b))
$= \lambda^{-1}\sigma^2_{\lambda,\nu} > 0$

□

**Lemma 9.** *For fixed $\nu \in \mathbb{R}_{>0}$, $\mu_{\lambda,\nu} \to \infty$ as $\lambda \to \infty$*

Let $a \wedge b = \min(\{a,b\})$ and $a \vee b = \max(\{a,b\})$.

*Proof.* of Lemma 9
Assume to the contrary that $\mu_{\lambda,\nu} \not\to \infty$,
then by Lemma 8 $\lim_{\lambda\to\infty} \mu_{\lambda,\nu}$ exists and
$\lim_{\lambda\to\infty} \mu_{\lambda,\nu} < \infty \Rightarrow$ there is a $c = \lim_{\lambda\to\infty} \mu_{\lambda,\nu} \in \mathbb{R}_{>0}$.
$\Rightarrow 0 \le c - \mu_{\lambda,\nu} = c - \sum_{i=1}^{\infty} i\frac{\lambda^i}{i!^\nu} Z^{-1}_{\lambda,\nu}$
$\Leftrightarrow 0 \le cZ_{\lambda,\nu} - \sum_{i=1}^{\infty} i\frac{\lambda^i}{i!^\nu} = c + c\sum_{i=1}^{\infty}\frac{\lambda^i}{i!^\nu} - \sum_{i=1}^{\infty} i\frac{\lambda^i}{i!^\nu}$
$= c + \sum_{i=1}^{\infty}(c-i)\frac{\lambda^i}{i!^\nu} = c + \sum_{i=1}^{\lfloor c\rfloor}(c-i)\frac{\lambda^i}{i!^\nu} + \sum_{i=\lfloor c\rfloor+1}^{\infty}(c-i)\frac{\lambda^i}{i!^\nu}$
$< c + c\sum_{i=1}^{\lfloor c\rfloor}(c-i)\frac{\lambda^i}{i!^\nu} + 0 < c\sum_{i=0}^{\lfloor c\rfloor}\lambda^i = c\frac{1-\lambda^{\lfloor c\rfloor+1}}{1-\lambda} \Leftrightarrow 0 < \frac{c}{1-\lambda}(1-\lambda^{\lfloor c\rfloor+1})$
$\Leftrightarrow 0 < 1-\lambda^{\lfloor c\rfloor+1}$ ; (for $\lambda > 1$) $\Leftrightarrow 1 > \lambda^{\lfloor c\rfloor+1} \Leftrightarrow 0 > \log\lambda$
Contradiction for $\lambda > 1$, which is given with the limit.

□

**Lemma 10.** *$\mu_{\lambda,\nu}$ is continuously differentiable in $\lambda$.*

*Proof.* of Lemma 10

For all $\lambda \in \mathbb{R}_{>0}, \sum_{j=1}^{\infty} j\frac{\lambda^j}{j!^\nu} \in \mathbb{R}_{>0}$ ; (Lemma 4)

Both $Z_{\lambda,\nu}$ & $\sum_{j=1}^{\infty} j\frac{\lambda^j}{j!^\nu}$ are power series in $\lambda$ and are thus continuously differentiable.

$\Rightarrow \mu_{\lambda,\nu} = \left(\sum_{j=1}^{\infty} j\frac{\lambda^j}{j!^\nu}\right) / Z_{\lambda,\nu}$ is continuously differentiable when $Z_{\lambda,\nu} \neq 0$.

$Z_{\lambda,\nu} > 0$ for all $(\lambda,\nu) \in \mathbb{R}^2_{>0}$ since it is an infinite sum of positive terms.

□

*Proof.* of Lemma 6

Fix $\nu \in \mathbb{R}_{>0}$. Let $g(\lambda) = \mu_{\lambda,\nu}$. Then, because $g(\lambda)$ is continuous (Lemma 10) and strictly increasing (Lemma 8), it is bijective. Also, it's domain and range are $\mathbb{R}_{>0}$ (Lemma 7 & 9). This implies there exists unique function $g^{-1}(\mu) = \lambda_{\mu,\nu}$. Thus, given $\nu$ and through $g$, $\lambda$ & $\mu$ are in one-to-one correspondence.

□

We now have that $\lambda_{\mu,\nu}$ has a valid definition, in that existence and uniqueness is proven.

## B.3 Properties of $\sigma^2_{\mu,\nu}$

Let $\sigma^2_{\mu,\nu} = \text{Var}[X_{\mu,\nu}]$ for $X_{\mu,\nu} \sim CMP(\lambda_{\mu,\nu}, \nu)$. We will now study $\sigma^2_{\mu,\nu}$ so that we may prove $\sigma^2_{\mu,\nu} < \mu(\mu+1)$.

**Lemma 11.** *$\frac{\partial}{\partial\mu}\lambda_{\mu,\nu} = \lambda_{\mu,\nu}/\sigma^2_{\mu,\nu}$ and $\lambda_{\mu,\nu}$ is continuously differentiable in $\mu$.*

*Proof.* of Lemma 11

Apply the inverse function theorem.

$\frac{\partial}{\partial\mu}\lambda_{\mu,\nu} = \left(\frac{\partial}{\partial\lambda}\mu_{\lambda,\nu}\right)^{-1}\big|_{\lambda=\lambda_{\mu,\nu}} = \left(\sigma^2_{\lambda,\nu}/\lambda\right)^{-1}\big|_{\lambda=\lambda_{\mu,\nu}}$ (by Lemma 8)

$= \lambda_{\mu,\nu}/\sigma^2_{\mu,\nu}$.

Since $\mu_{\lambda,\nu}$ is continuously differentiable (by Lemma 10) and

$\frac{\partial}{\partial\mu}\lambda_{\mu,\nu}$ always exists, $\lambda_{\mu,\nu}$ is continuously

differentiable in $\mu$.

□

**Lemma 12.** *If $X \sim CMP(\lambda,\nu)$, then for each $\lambda \in \mathbb{R}_{>0}$, $\frac{\partial}{\partial\nu}\mu_{\lambda,\nu} = \frac{\partial}{\partial\nu}\mathbb{E}[X] = -Cov[X, \log(X!)] \in \mathbb{R}_{\leq 0}$*

*Proof.* of Lemma 12

For $p \in \{0,1\}$, $\frac{\partial}{\partial\nu}\sum_{j=0}^{\infty} j^p\frac{\lambda^j}{j!^\nu} = \sum_{j=0}^{\infty}\frac{\partial}{\partial\nu}j^p\frac{\lambda^j}{j!^\nu}$ ; (Lemma 2 (b))

$\Rightarrow \frac{\partial}{\partial\nu}\mathbb{E}[X] = \mathbb{E}[X]\mathbb{E}[\log(X!)] - \mathbb{E}[X\log(X!)]$

$X \in \mathbb{Z}_{\geq 0} \Rightarrow X$ & $\log(X!)$ are co-increasing,

Apply Theorem 4.

All expectations are bounded by finite moments (Lemma 3).

□

**Lemma 13.** *$\lambda_{\mu,\nu}$ is continuously differentiable in $\nu$, and for each $\nu$ and $X_{\mu,\nu} \sim CMP(\lambda_{\mu,\nu},\nu)$, $\frac{\partial}{\partial\nu}\lambda_{\mu,\nu} = Cov[X_{\mu,\nu}, \log(X_{\mu,\nu}!)]\lambda_{\mu,\nu}/\sigma^2_{\mu,\nu}$.*

*Proof.* of Lemma 13

For every $\nu \in \mathbb{R}_{>0}$, $\frac{\partial}{\partial\nu}\mu_{\lambda,\nu}$ exists by Lemma 12, and $\frac{\partial}{\partial\lambda}\mu_{\lambda,\nu} > 0$ exists by Lemma 8. So by the implicit function theorem, $\frac{\partial}{\partial\nu}\lambda_{\mu,\nu}$ exists in an open set of $\mathbb{R}_{>0}$ containing $\nu$. Since this is true for all $\nu \in \mathbb{R}_{>0}$, $\lambda_{\mu,\nu}$ is continuously differentiable in $\nu$. Further let $g(\lambda,\nu) = \mu_{\lambda,\nu}$, then $g(\lambda_{\mu,\nu},\nu) = \mu$ by Lemma 6 and the following holds.

$$\frac{\partial}{\partial \nu} g(\lambda_{\mu,\nu}, \nu) = \frac{\partial g}{\partial \lambda}(\lambda_{\mu,\nu}, \nu) \left( \frac{\partial}{\partial \nu} \lambda_{\mu,\nu} \right) + \frac{\partial g}{\partial \nu}(\lambda_{\mu,\nu}, \nu) = 0$$

$$\Leftrightarrow \frac{\partial}{\partial \nu} \lambda_{\mu,\nu} = - \left( \frac{\partial g}{\partial \lambda}(\lambda_{\mu,\nu}, \nu) \right) \left( \frac{\partial g}{\partial \nu}(\lambda_{\mu,\nu}, \nu) \right)^{-1} = - \left( -\text{Cov}[X_{\mu,\nu}, \log(X_{\mu,\nu}!)] \right) \left( \sigma^2_{\mu,\nu} / \lambda_{\mu,\nu} \right)^{-1}$$

□

The following application of the DCT seems odd when MCT may (eventually) apply, but this lemma exists to avoid a circular argument.

**Lemma 14.**

*(a) If for each $\mu \in \mathbb{R}_{>0}$, there is an $M_\mu \in \mathbb{R}_{>0}$ such that for each $\nu \in \mathbb{R}_{>0}$ and $M_\mu > \lambda_{\mu,\nu}$, then $\lim_{\nu\to\infty} \sum_{j=0}^{\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu} = \sum_{j=0}^{\infty} \lim_{\nu\to\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu}$*
*(b) If for each $\mu \in \mathbb{R}_{>0}$, there is a $\lim_{\nu\to 0^+} \lambda_{\mu,\nu} \in \mathbb{R}$, then $\lim_{\nu\to 0^+} \sum_{j=0}^{\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu} = \sum_{j=0}^{\infty} \lim_{\nu\to 0^+} \frac{\lambda_{\mu,\nu}^j}{j!^\nu}$*

*Proof.* of 14 (a)
$\sum_{j=0}^{\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu} \leq \sum_{j=0}^{\infty} \frac{M_\mu^j}{j!^\nu} \in \mathbb{R}_{>0}$ ; (Lemma 4)
Apply dominated convergence theorem.

□

*Proof.* of Lemma 14 (b)
$\lambda_{\mu,\nu} \uparrow \nu$ (Lemma 13) $\Rightarrow \frac{\lambda_{\mu,\nu}^j}{j!^\nu} \uparrow \nu$
Apply monotone convergence theorem.

□

**Lemma 15.** *If $\mu \in (0,1)$ , then $\lambda_{\mu,\infty} := \lim_{\nu\to\infty} \lambda_{\mu,\nu} \in \mathbb{R}_{>0}$*

*Proof.* of Lemma 15
Assume to the contrary that $\lambda_{\mu,\infty} = \infty$.
Then for $X_{\mu,\nu} \sim CMP(\lambda_{\mu,\nu}, \nu)$, $\mu = \lim_{\nu\to\infty} \mathbb{E}[X_{\mu,\nu}]$
$= \lim_{\nu\to\infty} \left[ \lambda \frac{d}{d\lambda} \log \left( \sum_{j=0}^{\infty} \frac{\lambda^j}{j!^\nu} \right) \right] \Big|_{\lambda=\lambda_{\mu,\nu}}$
$\geq \lim_{\nu\to\infty} \left[ \lambda \frac{d}{d\lambda} \log(1+\lambda) \right] \Big|_{\lambda=\lambda_{\mu,\nu}}$ ; (Lemma 5)
$= \lim_{\nu\to\infty} \left[ \frac{\lambda}{1+\lambda} \right] \Big|_{\lambda=\lambda_{\mu,\nu}} = \lim_{\nu\to\infty} \frac{\lambda_{\mu,\nu}}{1+\lambda_{\mu,\nu}} = 1$
implying $\Rightarrow \mu \geq 1$ and $\mu \in (0,1)$, a contradiction.
So $\lambda_{\mu,\infty} \in \mathbb{R}_{>0}$.

□

**Lemma 16.** *If $\mu \in (0,1)$, then $\lambda_{\mu,\infty} = \frac{\mu}{1-\mu} \in \mathbb{R}_{>0}$*

*Proof.* of Lemma 16
$X \sim CMP(\lambda_{\mu,\nu}, \nu)$. $\mu = \lim_{\nu\to\infty} \mathbb{E}[X_{\mu,\nu}]$
$= \lim_{\nu\to\infty} \left( \sum_{j=1}^{\infty} j \frac{\lambda_{\mu,\nu}^j}{j!^\nu} \right) \left( \sum_{j=0}^{\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu} \right)^{-1}$
$\left( \sum_{j=1}^{\infty} \lim_{\nu\to\infty} j \frac{\lambda_{\mu,\nu}^j}{j!^\nu} \right) \left( \sum_{j=0}^{\infty} \lim_{\nu\to\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu} \right)^{-1}$ ; (Lemma 14 (a))

$= \frac{\lambda_{\mu,\infty}}{1+\lambda_{\mu,\infty}}$ (Exists by Lemma 15) $\Rightarrow \lambda_{\mu,\infty} = \frac{\mu}{1-\mu}$

□

Corollaries 2 and 3 are analogous to results in (Sellers and Shmueli [235]), but are specific to the reparameterized CMP. They describe how the CMP generalizes the Bernoulli and Geometric distributions.

**Corollary 2.** *If $X_{\mu,\nu} \sim CMP(\lambda_{\mu,\nu}, \nu), \mu \in (0,1), Y \sim Bernoulli(\mu)$, then $X_{\mu,\nu} \to Y$ in distribution, as $\nu \to \infty$.*

*Proof.* of Corollary 2

$\lim_{\nu\to\infty} f_{CMP}(x;\lambda_{\mu,\nu},\nu) = \lim_{\nu\to\infty} \frac{\lambda_{\mu,\nu}^x}{x!^\nu} \left(\sum_{j=0}^{\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu}\right)^{-1}$

$= \left(\lim_{\nu\to\infty} \frac{\lambda_{\mu,\nu}^x}{x!^\nu}\right) \left(\sum_{j=0}^{\infty} \lim_{\nu\to\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu}\right)^{-1}$ ; (Lemma 14 (a))

$= \frac{\lambda_{\mu,\infty}^x}{1+\lambda_{\mu,\infty}} 1_{x<2}$ (Lemma 16) $= \mu 1_{x=1} + (1-\mu)1_{x=0}$

□

**Lemma 17.** *If $\mu \in \mathbb{R}_{>0}, \lambda = \frac{\mu}{1+\mu}$, then $\mu_{\lambda,\nu} \to \mu$ as $\nu \to 0^+$*

*Proof.* of Lemma 17

If $\lambda \in (0,1)$, then $\lim_{\nu\to 0^+} \mu_{\lambda,\nu}$

$= \left(\sum_{i=1}^{\infty} \lim_{\nu\to 0^+} i\frac{\lambda^i}{i!^\nu}\right) \left(\sum_{j=0}^{\infty} \lim_{\nu\to 0^+} \frac{\lambda^j}{j!^\nu}\right)$ ; (MCT)

$= \left(\sum_{i=1}^{\infty} i\lambda^i\right) \left(\sum_{j=0}^{\infty} \lambda^j\right)^{-1} = \frac{\lambda}{1-\lambda} = \frac{\mu/(1+\mu)}{1-\mu/(1+\mu)} = \mu$

□

**Lemma 18.** *If $\mu \in \mathbb{R}_{>0}$, then $\lambda_{\mu,0} := \lim_{\nu\to 0^+} \lambda_{\mu,\nu} = \frac{\mu}{1+\mu}$*

*Proof.* of Lemma 18

$X \sim CMP(\frac{\mu}{1+\mu}, \nu)$. $Y \sim CMP(\lambda_{\mu,0}, \nu)$.

$\lambda_{\mu,\nu} \uparrow \mu$ (by Lemma 12) & $\mu = \lim_{\nu\to 0^+} \mathbb{E}[X]$ ; (by Lemma 17)

$\Rightarrow \lambda_{\mu,0} \geq \frac{\mu}{1+\mu}$.

Assume to the contrary that $\lambda_{\mu,0} > \frac{\mu}{1+\mu}$.

$\lambda_{\mu,0} > \frac{\mu}{1+\mu} \Rightarrow \mu = \lim_{\nu\to 0^+} \mathbb{E}[Y]$

$> \mathbb{E}[X]$ (by Lemma 8) $= \mu \Rightarrow \mu > \mu$. Contradiction.

So $\lambda_{\mu,0} = \frac{\mu}{1+\mu}$.

□

**Corollary 3.** *If $X_{\mu,\nu} \sim CMP(\lambda_{\mu,\nu}, \nu)$, & $Y \sim$ Geometric $\left([1+\mu]^{-1}\right)$, then $X_{\mu,\nu} \to Y$ in distribution, as $\nu \to 0^+$.*

*Proof.* of Corollary 3

$\lim_{\nu\to 0^+} f_{CMP}(x;\lambda_{\mu,\nu},\nu) = \lim_{\nu\to 0^+} \frac{\lambda_{\mu,\nu}^x}{x!^\nu} \left(\sum_{j=0}^{\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu}\right)^{-1}$

$= \left(\frac{\mu}{1+\mu}\right)^x \left(\sum_{j=0}^{\infty} \left(\frac{\mu}{1+\mu}\right)^j\right)^{-1}$ ; (by Lemmas 14 (b) & 18)

$= \left(\frac{\mu}{1+\mu}\right)^x \left(1 - \frac{\mu}{1+\mu}\right) = \left(\frac{\mu}{1+\mu}\right)^x \left(\frac{1}{1+\mu}\right)$

□

**Corollary 4.** *If $\mu \in (0,1)$, then $\sigma_{\mu,\nu}^2 \to \mu(1-\mu)$ as $\nu \to \infty$*

*Proof.* of Corollary 4

$X_{\mu,\nu} \sim CMP(\lambda_{\mu,\nu}, \nu)$. $p \in \{1, 2\}$. $\lim_{\nu\to\infty} \mathbb{E}[X_{\mu,\nu}^p]$

$= \left(\sum_{i=1}^{\infty} \lim_{\nu\to\infty} i^p \frac{\lambda_{\mu,\nu}^i}{i!^\nu}\right) \left(\sum_{j=0}^{\infty} \lim_{\nu\to\infty} \frac{\lambda_{\mu,\nu}^j}{j!^\nu}\right)$ ; (by Lemma 14 (a))

$= \lambda_{\mu,\infty}(1 + \lambda_{\mu,\infty})^{-1}$ (by Lemma 15) $= \frac{\mu}{1-\mu}\left(1 - \frac{\mu}{1-\mu}\right)$ ; (by Lemma 16)

$= \mu \Rightarrow \lim_{\nu\to\infty} \sigma_{\mu,\nu}^2$

$= \lim_{\nu\to\infty} \left(\mathbb{E}[X_{\mu,\nu}^2] - (\mathbb{E}[X_{\mu,\nu}])^2\right) = \mu(1 - \mu)$

□

**Lemma 19.** *For fixed $\mu \in \mathbb{R}_{>0}$, $\sigma_{\mu,\nu}^2 \to \mu(1 + \mu)$ as $\nu \to 0^+$*

*Proof.* of Lemma 20

$X_{\mu,\nu} \sim CMP(\lambda_{\mu,\nu}, \nu)$. $\lim_{\nu\to 0^+} \mathbb{E}[X_{\mu,\nu}^2]$

$= \left(\sum_{i=1}^{\infty} i^2 \left(\frac{\mu}{1+\mu}\right)^i\right) \left(\sum_{j=0}^{\infty} \left(\frac{\mu}{1+\mu}\right)^j\right)^{-1}$ ; (by Lemmas 14 (b) & 18)

$= \sum_{i=1}^{\infty} i^2 \left(\frac{\mu}{1+\mu}\right)^i (1 + \mu)^{-1} = \sum_{i=1}^{\infty} i^2 \left(1 - \frac{1}{1+\mu}\right)^i \left(\frac{1}{1+\mu}\right)$

$= \mu(1 + \mu) + \mu^2$ ; (Geometric distribution)

□

We will now put $\frac{\partial}{\partial\nu}\sigma_{\mu,\nu}^2$ into a form useful to Lemma 1.

**Lemma 20.** *$\frac{\partial}{\partial\nu}\sigma_{\mu,\nu}^2 \leq 0$ if and only if $X_{\mu,\nu} \sim CMP(\lambda_{\mu,\nu}, \nu)$ &*
*$Cor[X_{\mu,\nu}^2, \log(X_{\mu,\nu}!)] \geq Cor[X_{\mu,\nu}^2, X_{\mu,\nu}]Cor[X_{\mu,\nu}, \log(X_{\mu,\nu}!)]$*

*Proof.* of Lemma 20

$\frac{\partial}{\partial\nu}\sigma_{\mu,\nu}^2 = \frac{\partial}{\partial\nu} Var[X] = \frac{\partial}{\partial\nu}\left(\mathbb{E}[X_{\mu,\nu}^2] - \mu^2\right)$

$= \frac{\partial\lambda_{\mu,\nu}}{\partial\nu}\frac{\partial\mathbb{E}[X_{\mu,\nu}^2]}{\partial\lambda} + \frac{\partial\mathbb{E}[X_{\mu,\nu}^2]}{\partial\nu}$ ; (Multivariate chain rule)

$\leq 0 \Leftrightarrow \mathrm{Cov}[X_{\mu,\nu}^2, X_{\mu,\nu}]\mathrm{Cov}[X_{\mu,\nu}, \log(X_{\mu,\nu}!)] \leq \mathrm{Cov}[X_{\mu,\nu}^2, \log(X_{\mu,\nu}!)]\mathrm{Var}[X_{\mu,\nu}]$

□

**Lemma 21.** *$\sigma_{\mu,\nu}^2$ decreases as $\nu$ increases.*

*Proof.* of Lemma 21

For each $\mu > 0$, $\sigma_{\mu,\nu}^2|_{\nu=1} = \mu$ by Poisson generalization,
and $\sigma_{\mu,\nu}^2|_{\nu=0} = \mu(\mu + 1) > \mu$,
so $\frac{\partial}{\partial\nu}\sigma_{\mu,\nu}^2 < 0$ for some $\nu \leq 1$.
For every $(\mu, \nu)$, $\mathrm{Cor}[X_{\mu,\nu}^2, \log(X_{\mu,\nu}!)] \neq \mathrm{Cor}[X_{\mu,\nu}^2, X_{\mu,\nu}]\mathrm{Cor}[X_{\mu,\nu}, \log(X_{\mu,\nu}!)]$ by Lemma 1.
$\frac{\partial}{\partial\nu}\sigma_{\mu,\nu}^2 \leq 0 \Leftrightarrow \mathrm{Cor}[X_{\mu,\nu}^2, \log(X_{\mu,\nu}!)] \geq \mathrm{Cor}[X_{\mu,\nu}^2, X_{\mu,\nu}]\mathrm{Cor}[X_{\mu,\nu}, \log(X_{\mu,\nu}!)]$ by Lemma 20,
and since $\sigma_{\mu,\nu}^2$ is continuously differentiable, $\frac{\partial}{\partial\nu}\sigma_{\mu,\nu}^2$ cannot ever be zero.
But there is some $\nu \leq 1$ such that $\frac{\partial}{\partial\nu}\sigma_{\mu,\nu}^2 < 0$, so $\frac{\partial}{\partial\nu}\sigma_{\mu,\nu}^2$ is always negative.

□

Recall that Result 3 states that if $X_{\mu,\nu} \sim \mathrm{CMP}(\lambda, \nu)$ and $\mathbb{E}[X_{\mu,\nu}] = \mu$ and $\mathrm{Var}[X_{\mu,\nu}] = \sigma^2$, then $\sigma^2 < \mu(\mu + 1)$.

*Proof.* of Result 3

Since $\sigma_{\mu,\nu}^2 \to \mu(\mu + 1)$ as $\nu \to 0^+$ (by Lemma 20),
and because $\sigma_{\mu,\nu}^2$ decreases in $\nu$ (by Lemmma 21),
$\sigma_{\mu,\nu}^2 < \mu(\mu + 1)$ for each $(\mu, \nu) \in \mathbb{R}_{>0}^2$.

□

Recall that Result 4 states that if $X_{\mu,\nu} \sim \text{CMP}(\lambda, \nu)$, then $X_{\mu,\nu}$ is over-dispersed when $\nu < 1$ and under-dispersed when $\nu > 1$.

*Proof.* of Result 4
$\sigma^2_{\mu,\nu}$ decreases in $\nu$ (by Lemma 21),
and $\sigma^2_{\mu,\nu}|_{\nu=1} = \mu$ by Poisson generalization.

□

## B.4 Borrowed material

**Theorem 4.** *If functions $f, g$ are co-monotone (couter-monotone), then*
$\mathbb{E}[f(X)g(X)] - \mathbb{E}[f(X)]\mathbb{E}[g(X)] \geq 0$ $(\leq 0)$.

*Proof.* Derive from $\mathbb{E}(f(X) - f(Y))(g(X) - g(Y)) \geq 0, X =_D Y$ independently,
or see Hardy et al. [122].

□

**Theorem 5.** *For each $n, f_n : \mathbb{R} \to \mathbb{R}$ differentiable on $[a, b]$, there exists an $x_0 \in [a, b] : f_n(x_0) \to \cdot, f_n' \to \cdot$ uniformly on $[a, b] \Rightarrow f_n \to f$ uniformly on $[a, b]$ and*

$$f'(x) = \lim_{n\to\infty} f_n'(x) \text{ for each } x \in [a, b]$$

*Proof.* See Rudin [232], theorem 7.17.

□

**Theorem 6.** *For any random vectors $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$, the following are equivalent.*

1. $\mathbb{E}[\mathbf{X}|\mathbf{Y}] = \alpha + B\mathbf{Y}$ *for a vector $\alpha$ and matrix $B$.*
2. $\Sigma_{\mathbf{X}\cdot\mathbf{Y}} = \mathbb{E}\left(\Sigma_{\mathbf{X}|\mathbf{Y}}\right)$.
   *Where $\Sigma_{\mathbf{X}\cdot\mathbf{Y}}$ is the covariance matrix of $\mathbf{X}$ partial $\mathbf{Y}$, and $\Sigma_{\mathbf{X}|\mathbf{Y}}$ is the covariance matrix of $\mathbf{X}$ conditioned on $\mathbf{Y}$.*

*Proof.* See Baba [16] theorem 2.1.1, or Baba and Sibuya [18] theorem 1.

□

# Appendix C

# Precision with imprecise binners

Here we develop strategies for improving the precision of our claims with imprecise classifiers, thus providing insight on strategies which may allow us to make confident inferences while constrained by error-prone tools. These strategies are theoretical and untested, but are inspired by current approaches. Our earlier results in section 1.6.5 which describe constraints on imprecise classifiers hold true and are mathematical facts that cannot be avoided. Instead, we produce two strategies which overcome these caveats by either changing our classifier or claim. In both strategies it is necessary to perform assumption checking with known-label data. In either strategy we follow the paradigm of section 1.6.5, classifier $C$ attempts attribution of object or phenomenon $X$ to some label $y$. In binning, $X$ is a metagenomic sequence, $C$ is a binner, and $y$ is a taxonomic label. Both of these strategies reduce the applicable scope of their methods, but only one actually consumes additional metagenomic sequences to improve its precisions, the other does have a greatly increased information requirement.

## C.1 Marker gene strategy

It has been common to evaluate binning attempts with marker genes [5, 207, 225, 240]. As a post-hoc analysis, marker gene evaluation can be viewed as a modification of the initial binning proceedure without loss of generality. So this approach modifies our classifier $C$ to leverage additional information. It is no longer enough to attribute $C(X) = y$, we now also require $X$ to satisfy additional requirements $X \in z$. For marker gene analyses specifically, $z$ would be a marker gene requirement. So while our initial precision $p = \mathbb{P}[X \in y | C(X) = y]$ may be miserable, the modified process' precision $q = \mathbb{P}[X \in y | C(X) = y, X \in z]$ can be much better. Given a known-label data set, these precisions may be estimated as $\hat{p} = \sum_{i=1}^{n} 1_{X_i \in y, C(X_i)=y} / \sum_{C(X_i)=y}$ and $\hat{q} = \sum_{i=1}^{n} 1_{X_i \in y, C(X_i)=y, X_i \in z} / \sum_{C(X_i)=y, X_i \in z}$ ($1_A$ is an indicator variable, $1_A = \{1 \text{ if } A; 0 \text{ if } A^c\}$). Of course, it is possible that these estimates are only different due to sampling variation, and is not actually meaningfuly different. To test for statistically significant difference Fisher's Exact Test may be applied for small samples whereas result 5 is applicable when samples are large and a most statistically powerful test is desired.

Evaluating the added value of further constraining $X$ by $X \in z$ can only be achieved with known-label data, and thus is only a concern during classifier evaluation. For binning with marker genes, this means that this assumption should be evaluated when the binner itself is evaluated with a synthetic data set [180]. However in application of the binner, this is assumption-checking is no longer a concern (requiring the assumption was confidently observed as true in the evaluation). The $X \in z$ requirement will likely reduce the number of sequences $C$ may be applied to, but within its scope of application, the modified binner simply has an increased precision per individual metagenomic sequence.

## C.2 Common trait strategy

Taking inspiration from modern applications of binning [10], binning may not actually used make claims attributing relationships between taxonomy $y$ and metagenomic sequences $X$, but instead is used to attribute relationships between taxonomy $y$ and common traits $z$. For example, this trait could be the encoding of particular reactions in a biogeochemical pathway. Our focus has been sufficiently fixed on the search for statistical dependence in $(1_{C(X)=y}, 1_{X \in y})$, but it might be more pragmatic to search for dependence in $(1_{X \in y}, 1_{X \in z})$. One possible way to test for such dependence is to search for a correlation between $1_{X \in y}$

and $1_{X\in z}$. Unfortunately we cannot observe $1_{X\in y}$, but can observe $1_{C(X)=y}$. So we might use an imprecise binner $C$ to inform on taxonomy $y$. It might be possible to test for dependence within $(1_{X\in y}, 1_{X\in z})$ by testing for a statistically significant correlation between observations $1_{C(X_i)=y}$ and $1_{X_i\in z}$.

Unfortunately it is entirely possible that our classifiers are not only wrong but also biased, and their classifications are overwhelmed by artifactual constructs. However, if our classifiers are wrong in a right and unbiased way, we may aggregate their behaviour conclusions in a more precise way. We formulate one such concept of bias, by assuming that our data $X_i = (1_{X_i\in y}, 1_{X_i\in z}, 1_{C(X_i)=y})^T$ follow a multivariate probit model [15, 57]. During the classifier evaluation, when $1_{X_i\in y}$ is observable, it might be decided that the latent partial correlation $\sigma_{C,Z\cdot Y}$ (formally defined in result 6) is bounded within a range $|\sigma_{C,Z\cdot Y}| < b$. If such an assumption is demonstrated to be viable, then there are correlations of the observable $(1_{C(X_i)=y}, 1_{X_i\in z})$ which imply $(1_{X\in y}, 1_{X\in z})$ is probably correlated. Of course a smaller bound $b$ results in more confidently inferrable correlations between taxonomy and biogeochemical pathways.

As in the previous strategy, a test must be conducted during classifier evaluation to conclude that an essential assumption is satisfied. However, this strategy requires further statistical testing, and we have yet to argue that it may more precisely evaluate a target claim. Given this strategy's assumption is satisfied, application requires discovering a common trait $z$ amongst some metagenomic sequences $X_i \in z$. For example, they might encode reactions participating in the denitrification pathway. Then a Fisher Exact Test could conclude that observations $(1_{C(X_i)=y}, 1_{X_i\in z})$ are significantly correlated (suitable for $b = 0$), or a likelihood ratio test could be used (suitable for $b > 0$). In this way, we may evaluate the a claim that particular taxa are correlated with a particular biogeochemical pathway. Notice tha t this test consumes several metagenomic sequences to evaluate a single claim.

To argue an increase in precision, we observe that the we are classifying our claim through hypothesis testing. We make our claim through rejection of a null hypothesis, $H_0 : \mathrm{Cor}[C(X) = y, X \in z] = 0$. We reject the null correctly with probability $1 - \beta = \mathbb{P}[\text{reject } H_0 | H_1]$, this is the statistical power of our test. We reject our null incorrectly with probability $\alpha = \mathbb{P}[\text{reject } H_0 | H_0]$, this is the type-1 error rate. Assuming that our hypothesis alternatives make up a true dichotomy of our sample space $H_0 \cup H_1 = \Omega$ on our probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we may conclude that the probability of a correct claim is $(1 - \beta)/(1 - \beta + \alpha)$. So to increase our precision (probability of a true claim), we must achieve a statistical power $(1 - \beta)$ which is large relative to our type-1 error rate ($\alpha$). Methods for increasing statistical power can be sophisticated [192], but it is heuristically true that power increases with sample size. For binning, this means consuming more metagenomic sequences with an unbiased binner can produce more precise claims.

## C.3 Formal arguments

**Result 5.** *Let $A_i := 1_{\{X_i\in y\}}$ for each of the $X_i$ such that $X_i \in z$.*

*Then we have an independent sample of n Bernoulli random variables $A_i$ such that if $\sum_{i=1}^n A_i = k$, we successfully reject the null hypothesis $H_0 : \mathbb{P}[A_i = 1] = p$ for each i with probability $\alpha$ if the following statement holds.*

$$2k \log \frac{k}{np} + 2(n-k) \log \frac{n-k}{n(1-p)} > \chi^2_{n-1}(\alpha)$$

*Proof.*

Let $\mathbb{P}[A_i = 1] = q$. The likelihood function of sample $A_{i\in\{1,2,\ldots,n\}}$ is $f(n,k,q) = q^k(1-q)^{n-k}$ with maximum likelihood estimate $\hat{q} = k/n$. Then Wilks' $-2\log\Lambda$ statistic is the following.

$-2\log\Lambda = -2\log\frac{f(n,k,p)}{f(n,k,\hat{q})} = 2\log\frac{f(n,k,p)}{f(n,k,k/n)}$

$= 2k\log\frac{k}{np} + 2(n-k)\log\frac{n-k}{n(1-p)}$

And $-2\log\Lambda$ follows a $\chi^2_{n-1}$ distribution under $H_0$ when $n$ is large, according to Wilks' theorem [270]. □

**Result 6.** *Let $X \in \{0,1\}^3$ be multivarite probit-distributed [15, 57] as follows.*

$$X = \begin{bmatrix} 1_{X\in y} \\ 1_{X\in z} \\ 1_{C(X)=y} \end{bmatrix} = \begin{bmatrix} 1_{Y>0} \\ 1_{Z>0} \\ 1_{C>0} \end{bmatrix} \; ; \; \begin{bmatrix} Y \\ Z \\ C \end{bmatrix} \sim N_3 \left( \begin{bmatrix} \mu_Y \\ \mu_Z \\ \mu_C \end{bmatrix}, \begin{bmatrix} \sigma_{Y,Y} & \sigma_{Y,Z} & \sigma_{Y,C} \\ \sigma_{Y,Z} & \sigma_{Z,Z} & \sigma_{Z,C} \\ \sigma_{Y,C} & \sigma_{Z,C} & \sigma_{C,C} \end{bmatrix} \right)$$

*where the covariance notation from section B.1 or section 1.5.3 is used.*

*If $|\sigma_{C,Z\cdot Y}| \leq b$, then any test which accepts $H_0 : |\sigma_{C,Z}| > b$ implies $\sigma_{C,Y} \neq 0$ and $\sigma_{Z,Y} \neq 0$, but also a test which accepts $H_1 : |\sigma_{C,Z}| \leq b$ admits the possibility that $\sigma_{C,Y} = 0$ or $\sigma_{Z,Y} = 0$ (without further realized constraints).*

*Proof.*

$\sigma_{C,Z} = \sigma_{C,Z\cdot Y} + \sigma_{C,Y}\sigma_{Z,Y}\sigma_{Y,Y}^{-1}$
$|\sigma_{C,Z}| > b \Rightarrow |\sigma_{C,Z\cdot Y} + \sigma_{C,Y}\sigma_{Z,Y}\sigma_{Y,Y}^{-1}| \leq |\sigma_{C,Z\cdot Y}| + |\sigma_{C,Y}\sigma_{Z,Y}\sigma_{Y,Y}^{-1}| \leq b + |\sigma_{C,Y}\sigma_{Z,Y}\sigma_{Y,Y}^{-1}|$
$\Rightarrow |\sigma_{C,Y}\sigma_{Z,Y}\sigma_{Y,Y}^{-1}| \geq |\sigma_{C,Z}| - b > 0 \Rightarrow \sigma_{Z,Y} \neq 0$ and $\sigma_{Z,Y} \neq 0$,
giving us our necessary implication.
$|\sigma_{C,Z}| < b \Leftarrow \Big(\sigma_{C,Z\cdot Y} = \sigma_{C,Z} < b \Rightarrow 0 = \sigma_{C,Y}\sigma_{Z,Y}\sigma_{Y,Y}^{-1} \Rightarrow \sigma_{C,Y} = 0 \text{ or } \sigma_{Z,Y} = 0\Big)$,
giving us our admission.

□

# Appendix D

# Miscellaneous

## D.1 Factorial experiment regression summaries

```
> summary( lm( ll$precision[,"gpu_0.05"] ~ m$mdl + m$n + m$q + m$s ) )

Call:
lm(formula = ll$precision[, "gpu_0.05"] ~ m$mdl + m$n + m$q +
    m$s)

Residuals:
     Min       1Q   Median       3Q      Max
-0.81194 -0.20880  0.01995  0.21959  0.65136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9518     0.2210   4.307 0.000262 ***
m$mdlmln     -0.4323     0.1877  -2.303 0.030691 *
m$n1000      -0.1398     0.1521  -0.920 0.367355
m$q4         -0.1709     0.1521  -1.124 0.272809
m$s0.5        0.1233     0.1670   0.738 0.467718
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3735 on 23 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared:  0.2039,Adjusted R-squared:  0.06545
F-statistic: 1.473 on 4 and 23 DF,  p-value: 0.2429

> summary( lm( ll$precision[,"pearson_0.05"] ~ m$mdl + m$n + m$q + m$s ) )

Call:
lm(formula = ll$precision[, "pearson_0.05"] ~ m$mdl + m$n + m$q +
    m$s)

Residuals:
      Min        1Q    Median        3Q       Max
-0.239552 -0.026302 -0.005363  0.026910  0.206462

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.064861   0.062489   1.038    0.311
m$mdlmln    -0.043095   0.047347  -0.910    0.373
m$n1000      0.018219   0.039277   0.464    0.648
m$q4        -0.009096   0.049585  -0.183    0.856
m$s0.5       0.549366   0.039277  13.987 4.11e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.09917 on 21 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.904,Adjusted R-squared:  0.8857
F-statistic: 49.42 on 4 and 21 DF,  p-value: 2.169e-10

> summary( lm( ll$precision[,"sparcc_0.05"] ~ m$mdl + m$n + m$q + m$s ) )
```

```
Call:
lm(formula = ll$precision[, "sparcc_0.05"] ~ m$mdl + m$n + m$q +
    m$s)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17990 -0.06180  0.01208  0.06135  0.15689

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02234    0.02620  -0.853   0.3972
m$mdlmln    -0.03278    0.02343  -1.399   0.1671
m$n1000      0.04249    0.02343   1.813   0.0749 .
m$q4         0.04816    0.02343   2.055   0.0443 *
m$s0.5       0.37918    0.02343  16.182   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.09373 on 59 degrees of freedom
Multiple R-squared:  0.8214,Adjusted R-squared:  0.8093
F-statistic: 67.83 on 4 and 59 DF,  p-value: < 2.2e-16
```

## D.2 Taxa regressed

The following taxa were regressed in our 16S multivariate regression analysis.

```
 [1] "Nitrospina"                          "Nitrospira"
 [3] "SUP05"                               "Cyanobacteria"
 [5] "Planctomycetes"                      "Thaumarchaeota"
 [7] "OD1"                                 "MGA"
 [9] "k__Bacteria.p__Proteobacteria"       "k__Bacteria.p__ZB3"
[11] "k__Archaea.p__Unclassified"          "k__Bacteria.p__Verrucomicrobia"
[13] "k__Bacteria.p__OP3"                  "k__Bacteria.p__Calditrix_KSB1"
[15] "No.blast.hit"                        "k__Archaea.p__pISA1"
[17] "k__Bacteria.p__Bacteroidetes"        "k__Bacteria.p__Actinobacteria"
[19] "k__Bacteria.p__Lentisphaerae"        "k__Bacteria.p__TM6"
[21] "k__Bacteria.p__Chloroflexi"          "k__Bacteria.p__OP11"
[23] "k__Bacteria.p__Firmicutes"           "k__Bacteria.p__Elusimicrobia_TG1"
[25] "k__Archaea.p__pMC2A209"              "k__Bacteria.p__Acidobacteria"
[27] "k__Bacteria.p__TM7"                  "k__Bacteria.p__Spirochaetes"
[29] "k__Archaea.p__Thermoplasmata_Eury"   "k__Archaea.p__Methanobacteria_Eury"
[31] "k__Bacteria.p__Fusobacteria"         "k__Archaea.p__BC07.2A.27"
[33] "k__Archaea.p__Methanomicrobia_Eury" "k__Bacteria.p__Chlamydiae"
[35] "k__Bacteria.p__Nitrospirae"          "k__Bacteria.p__WS3"
[37] "k__Bacteria.p__GN02"                 "k__Bacteria.p__WS6"
[39] "k__Archaea.p__pMC2A384"              "k__Bacteria.p__VHS.B5.50"
[41] "k__Bacteria.p__ZB2"                  "k__Bacteria.p__Gemmatimonadetes"
[43] "k__Archaea.p__pMC2A15"               "k__Bacteria.p__NKB19"
[45] "k__Bacteria.p__Fibrobacteres"        "k__Archaea.p__Methanococci_Eury"
[47] "k__Bacteria.p__Unclassified"         "k__Archaea.p__Halobacteriales"
[49] "k__Bacteria.p__SM2F11"               "k__Bacteria.p__OP9_JS1"
[51] "k__Bacteria.p__ctg_CGOF"             "k__Archaea.p__DHVE3"
[53] "k__Archaea.p__pSL22"                 "k__Bacteria.p__OP10"
[55] "k__Archaea.p__MSBL1"                 "k__Bacteria.p__OP1"
[57] "k__Archaea.p__NO27FW"
```

## D.3 Marginal regression survey results

In the following R output, +1 indicates a positive and statistically significant correlation, -1 indicates a negative and statistically significant correlation, and 0 indicates a statistically insignificant relationship at the $\alpha = 0.05$ level. Note that some taxa will correlate with a variable even if its metabolic relationship is with a correlated but unmeasured environmental variable.

| | o2 | no3 | h2s |
|---|---|---|---|
| Nitrospina | -1 | 1 | 0 |
| Nitrospira | -1 | -1 | 0 |
| SUP05 | -1 | 0 | 0 |
| Cyanobacteria | 1 | 1 | 1 |
| Planctomycetes | -1 | 0 | 0 |
| Thaumarchaeota | -1 | 1 | 0 |
| OD1 | -1 | -1 | 1 |
| MGA | -1 | 0 | 0 |
| k__Bacteria.p__Proteobacteria | 0 | -1 | 0 |
| k__Bacteria.p__ZB3 | -1 | 0 | 1 |
| k__Archaea.p__Unclassified | -1 | -1 | 1 |
| k__Bacteria.p__Verrucomicrobia | -1 | 0 | 0 |
| k__Bacteria.p__OP3 | 0 | 1 | 1 |
| k__Bacteria.p__Caldithrix_KSB1 | -1 | -1 | 1 |
| No.blast.hit | 0 | 0 | 0 |
| k__Archaea.p__pISA1 | -1 | -1 | 1 |
| k__Bacteria.p__Bacteroidetes | 0 | 0 | 0 |
| k__Bacteria.p__Actinobacteria | 0 | 0 | 0 |
| k__Bacteria.p__Lentisphaerae | 0 | -1 | 1 |
| k__Bacteria.p__TM6 | -1 | -1 | 1 |
| k__Bacteria.p__Chloroflexi | -1 | 0 | 1 |
| k__Bacteria.p__OP11 | -1 | -1 | 1 |
| k__Bacteria.p__Firmicutes | 0 | 0 | 1 |
| k__Bacteria.p__Elusimicrobia_TG1 | 0 | 0 | 1 |
| k__Archaea.p__pMC2A209 | -1 | 0 | 1 |
| k__Bacteria.p__Acidobacteria | -1 | 0 | 0 |
| k__Bacteria.p__TM7 | 0 | 1 | 0 |
| k__Bacteria.p__Spirochaetes | 0 | -1 | 1 |
| k__Archaea.p__Thermoplasmata_Eury | 0 | 0 | -1 |
| k__Archaea.p__Methanobacteria_Eury | -1 | 0 | 1 |
| k__Bacteria.p__Fusobacteria | 0 | 0 | 0 |
| k__Archaea.p__BC07.2A.27 | -1 | 0 | 1 |
| k__Archaea.p__Methanomicrobia_Eury | 0 | -1 | 1 |
| k__Bacteria.p__Chlamydiae | -1 | 0 | 1 |
| k__Bacteria.p__Nitrospirae | -1 | -1 | 0 |
| k__Bacteria.p__WS3 | -1 | 0 | 1 |
| k__Bacteria.p__GN02 | -1 | 0 | 0 |
| k__Bacteria.p__WS6 | -1 | 0 | 0 |
| k__Archaea.p__pMC2A384 | 0 | 0 | 1 |
| k__Bacteria.p__VHS.B5.50 | -1 | 1 | -1 |
| k__Bacteria.p__ZB2 | -1 | 0 | 0 |
| k__Bacteria.p__Gemmatimonadetes | -1 | 0 | 0 |
| k__Archaea.p__pMC2A15 | -1 | 0 | 1 |
| k__Bacteria.p__NKB19 | 0 | 0 | 0 |
| k__Bacteria.p__Fibrobacteres | 0 | 0 | 1 |
| k__Archaea.p__Methanococci_Eury | -1 | -1 | 0 |
| k__Bacteria.p__Unclassified | 0 | 0 | 1 |
| k__Archaea.p__Halobacteriales | 0 | 0 | 1 |
| k__Bacteria.p__SM2F11 | 0 | 0 | 0 |
| k__Bacteria.p__OP9_JS1 | 0 | 0 | -1 |
| k__Bacteria.p__ctg_CGOF | -1 | -1 | 0 |
| k__Archaea.p__DHVE3 | 0 | -1 | 1 |
| k__Archaea.p__pSL22 | -1 | 0 | 0 |
| k__Bacteria.p__OP10 | -1 | -1 | -1 |
| k__Archaea.p__MSBL1 | 0 | 0 | 0 |
| k__Bacteria.p__OP1 | -1 | -1 | 1 |
| k__Archaea.p__NO27FW | -1 | 0 | 0 |

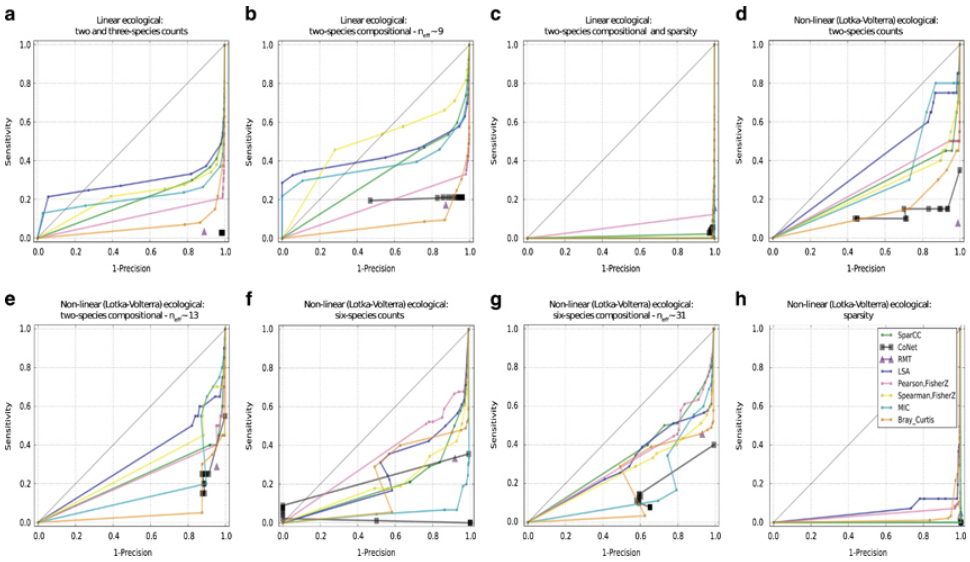



**Figure D.1:** Precision recall curves for popular 16S correlation techniques (lines) on several models (plots). Image credit: [267]

## D.4 Poor precision-recall exchanges

See Figure D.1

## D.5 SAGs sequenced

See Figure D.2.

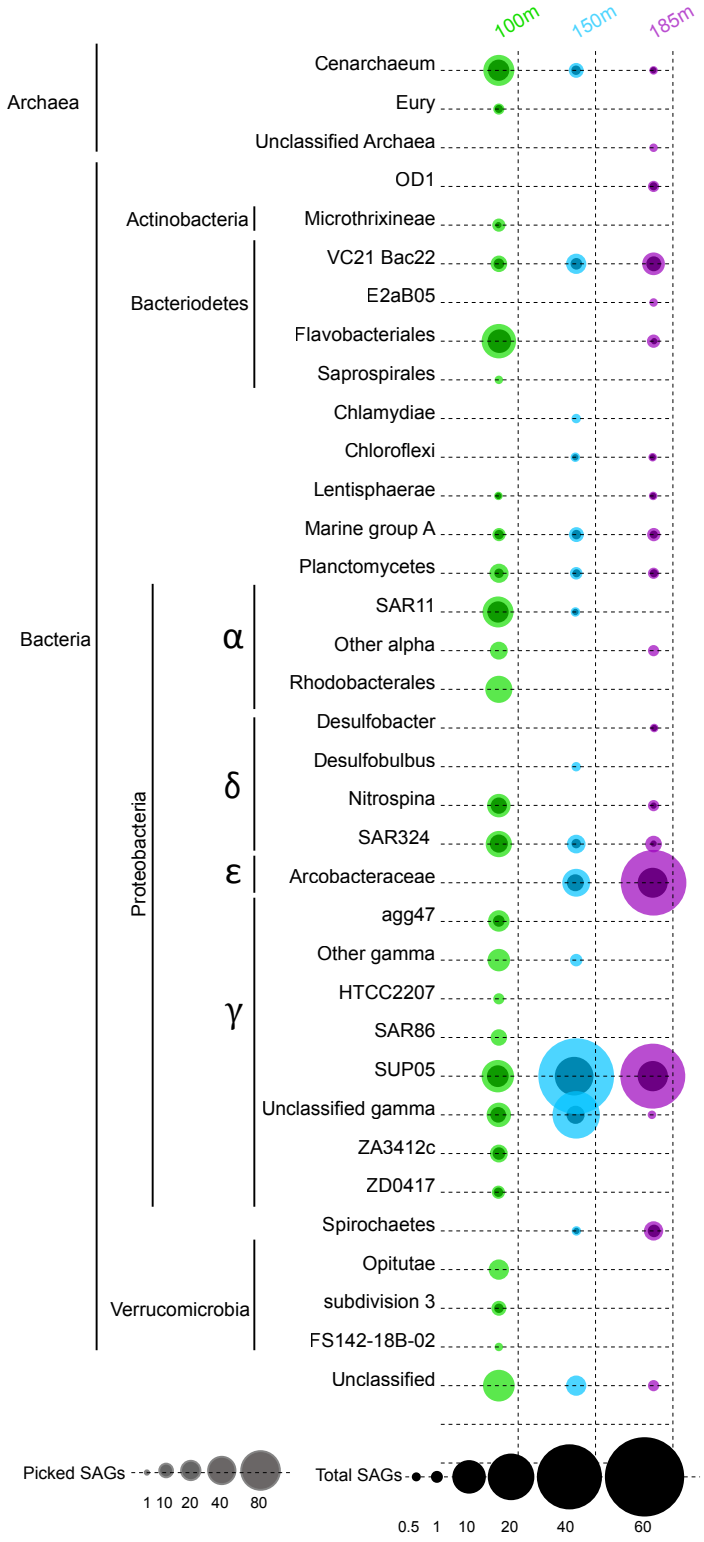



**Figure D.2:** SAGs sampled and sequenced (picked). Image credit: Alyse Hawley

## D.6 SAG decontamination taxa ranges

Bacteria;Proteobacteria;Gammaproteobacteria;SUP05
Bacteria;Proteobacteria;Gammaproteobacteria;SUP05 (Arctic)
Archaea;Thaumarchaeota;Cenarchaeales;Cenarchaeum
Bacteria;Bacteroidetes;Flavobacteriales;Cytophaga
Bacteria;Verrucomicrobia;Verrucomicrobia_subdivision_3

## D.7 Evaluation levels

**Level 1**, the low level
Archaea;Thaumarchaeota;Cenarchaeales;Cenarchaeum;Unclassified;OTU
Bacteria;Bacteroidetes;Flavobacteriales;Cytophaga;Unclassified;OTU
Bacteria;Bacteroidetes;Bacteroidales;VC21_Bac22;OTU
Bacteria;Proteobacteria;Alphaproteobacteria;Consistiales;Pelagibacter;SAR11;Candidatus_Pelagibacter_ubique;OTU
Bacteria;Proteobacteria;Deltaproteobacteria;Nitrospina;OTU
Bacteria;Proteobacteria;Deltaproteobacteria;Sva0853;SAR324;OTU
Bacteria;Proteobacteria;Epsilonproteobacteria;Arcobacteraceae;Unclassified;OTU
Bacteria;Proteobacteria;Gammaproteobacteria;SUP05;Unclassified;OTU_Arctic
Bacteria;Proteobacteria;Gammaproteobacteria;SUP05;Unclassified;OTU_SUP05_1a
Bacteria;Proteobacteria;Gammaproteobacteria;SUP05;mussel_thioautotrophic_gill_symbiont_MAR1;OTU_SUP05_1c
**Level 2**, the mid level
Archaea Bacteria;Bacteroidetes;Flavobacteriales
Bacteria;Bacteroidetes;Bacteroidales
Bacteria;Proteobacteria;Alphaproteobacteria
Bacteria;Proteobacteria;Deltaproteobacteria
Bacteria;Proteobacteria;Epsilonproteobacteria
Bacteria;Proteobacteria;Gammaproteobacteria
**Level 3**, the high level
Archaea Bacteria

## D.8 ESOM R script

```
# esom_binner.R

library("igraph")
library("RColorBrewer")

umx = as.matrix( read.table("work/fa53_sub.fa.kmer_50x50e100.umx",skip=1) )
bm = as.matrix( read.table("work/fa53_sub.fa.kmer_50x50e100.bm",skip=2) )

# build the node name
nn = function(a,b) paste0( "v" , a , "_" , b )

# clusters are built of sufficiently near nodes
# Distances can be no greater than 'cut' to share a group
build_graph = function(cut=0.2,u=umx)
{
        nr = nrow(u)
        nc = ncol(u)
        grid = expand.grid( 1:nr , 1:nc )
        check_neighbours = function(i)
        {
                if( u[ grid[i,1] , grid[i,2] ] <= cut )
                {
                        out = NULL
```

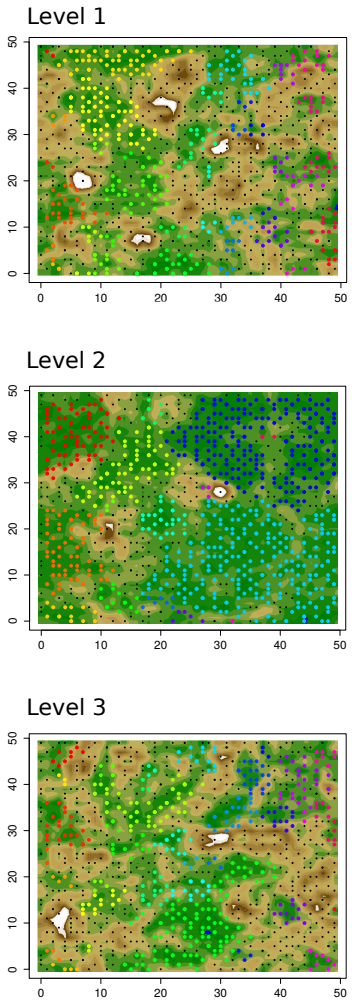
```
                        if( grid[i,1] > 1 ){ out = c( out ,
                                nn(grid[i,1],grid[i,2]) , nn(grid[i,1]-1 , grid[i,2]) ) }
                        if( grid[i,1] < nr ){ out = c( out ,
                                nn(grid[i,1],grid[i,2]) , nn(grid[i,1]+1 , grid[i,2]) ) }
                        if( grid[i,2] > 1 ){ out = c( out ,
                                nn(grid[i,1] , grid[i,2]) , nn(grid[i,1],grid[i,2]-1) ) }
                        if( grid[i,1] < nc ){ out = c( out ,
                                nn(grid[i,1] , grid[i,2]) , nn(grid[i,1] , grid[i,2]+1) ) }
                        if( grid[i,1] > 1 & grid[i,2] > 1 ){ out = c( out ,
                                nn(grid[i,1] , grid[i,2]) , nn(grid[i,1]-1 , grid[i,2]-1) ) }
                        if( grid[i,1] > 1 & grid[i,2] < nc ){ out = c( out ,
                                nn(grid[i,1] , grid[i,2]) , nn(grid[i,1]-1 , grid[i,2]+1) ) }
                        if( grid[i,1] < nr & grid[i,2] > 1 ){ out = c( out ,
                                nn(grid[i,1] , grid[i,2]) , nn(grid[i,1]+1 , grid[i,2]-1) ) }
                        if( grid[i,1] < nr & grid[i,2] < nc ){ out = c( out ,
                                nn(grid[i,1] , grid[i,2]) , nn(grid[i,1]+1 , grid[i,2]+1) ) }
                        return(out)
                }
                NULL
        }
        edges = lapply( 1:(nr*nc) , check_neighbours )
        edges = t( matrix( unlist(edges) , nrow=2 ) )
        edges = data.frame( from=edges[,1] , to=edges[,2] )
        graph.data.frame(edges)
}

construct_bins = function(cut=0.2,u=umx,b=bm)
{
        g = build_graph(cut,u)
        cmps = components(g)
        bin = function(i)
        {
                if( nn(b[i,2],b[i,3]) %in% names( cmps$membership ) )
                {
                        return( cmps$membership[ nn(b[i,2],b[i,3]) ] )
                }
                0 # zero indicates no membership
        }
        bins = sapply( 1:nrow(b) , bin )
        cbind( b , bins )
}

library("png")

plot_bins = function(bins,coef=1,bg_path=NA,...)
{
        plot( bins[,3] , bins[,2] , pch=16 , cex=0.5*coef , ... )
        clr = rainbow( max(bins[,4]) )
#       clr = brewer.pal(max(bins[,4]),"Set1")( max(bins[,4]) ) # too many colours for this palette
        idx = (1:nrow(bins))[ bins[,4] > 0 ]
        points( bins[idx,3] , bins[idx,2] , pch=16 , col=clr[ bins[,4] ] , cex=coef )
        if( ! is.na(bg_path) )
        {
                bg = readPNG( bg_path )
                lim = par()
                rasterImage(bg, lim$usr[1], lim$usr[3], lim$usr[2], lim$usr[4])
                points( bins[,3] , bins[,2] , pch=16 , cex=0.5*coef )
                points( bins[idx,3] , bins[idx,2] , pch=16 , col=clr[ bins[,4] ] , cex=coef )
        }
}
```

## D.9 ESOM U-matrices and bins




ESOM U-matrices are shown as height maps. Unique colours within levels are separate bins. Black dots are not assigned bins.

**Figure D.3:** ESOM U-matrices and bins

## D.10 All binner precision recall statistics

| Taxa | Level | Precision | Sensitivity |
|---|---|---|---|
| PhylopythiaS | | | |
| Thaumarchaeota;Cenarchaeum | 1 | NA | .00 |
| Flavobacteriales;Cytophaga | 1 | NA | .00 |
| Bacteroidetes;Bacteroidales | 1 | 1.00 | .007 |
| Alphaproteobacteria;SAR11 | 1 | NA | .00 |
| Deltaproteobacteria;Nitrospina | 1 | NA | .00 |
| Deltaproteobacteria;Sva0853;SAR324 | 1 | NA | .00 |
| Epsilonproteobacteria;Arcobacteraceae | 1 | NA | .00 |
| Gammaproteobacteria;SUP05_Arctic | 1 | NA | .00 |
| Gammaproteobacteria;SUP05_1a | 1 | NA | .00 |
| Gammaproteobacteria;SUP05_1c | 1 | NA | .00 |
| Archaea | 2 | .66 | .45 |
| Bacteroidetes;Flavobacteriales | 2 | .97 | .09 |
| Bacteroidetes;Bacteroidales | 2 | .92 | .01 |
| Alphaproteobacteria | 2 | .00 | .00 |
| Deltaproteobacteria | 2 | .00 | .00 |
| Epsilonproteobacteria | 2 | .00 | .00 |
| Gammaproteobacteria | 2 | .00 | .00 |
| Bacteria | 3 | .91 | .34 |
| Archaea | 3 | .65 | .45 |
| SAGEX (classify) | | | |
| Thaumarchaeota;Cenarchaeum | 1 | .99 | .04 |
| Flavobacteriales;Cytophaga | 1 | .77 | .008 |
| Bacteroidetes;Bacteroidales | 1 | .99 | .10 |
| Alphaproteobacteria;SAR11 | 1 | .97 | .04 |
| Deltaproteobacteria;Nitrospina | 1 | .93 | .04 |
| Deltaproteobacteria;Sva0853;SAR324 | 1 | .95 | .10 |
| Epsilonproteobacteria;Arcobacteraceae | 1 | .90 | .012 |
| Gammaproteobacteria;SUP05_Arctic | 1 | .36 | .003 |
| Gammaproteobacteria;SUP05_1a | 1 | .62 | .06 |
| Gammaproteobacteria;SUP05_1c | 1 | .77 | .04 |
| Archaea | 2 | .97 | .03 |
| Bacteroidetes;Flavobacteriales | 2 | .76 | .01 |
| Bacteroidetes;Bacteroidales | 2 | .98 | .05 |
| Proteobacteria;Alphaproteobacteria | 2 | .95 | .05 |
| Proteobacteria;Deltaproteobacteria | 2 | .93 | .05 |
| Proteobacteria;Epsilonproteobacteria | 2 | .90 | .01 |
| Proteobacteria;Gammaproteobacteria | 2 | .93 | .01 |
| Archaea | 3 | .97 | .03 |
| Bacteria | 3 | .997 | .004 |

**Table D.1:** All PhylopythiaS and SAGEX (classify) precision-recall statistics

| Taxa | Level | Precision | Sensitivity |
|---|---|---|---|
| SAGEX (cluster) | | | |
| Thaumarchaeota;Cenarchaeum | 1 | .85 | 2.8e-04 |
| Flavobacteriales;Cytophaga | 1 | .20 | 2.0e-05 |
| Bacteroidetes;Bacteroidales | 1 | .77 | 4.2e-04 |
| Alphaproteobacteria;SAR11 | 1 | .37 | 1.1e-04 |
| Deltaproteobacteria;Nitrospina | 1 | .32 | 4.3e-05 |
| Deltaproteobacteria;Sva0853;SAR324 | 1 | .85 | 1.2e-05 |
| Epsilonproteobacteria;Arcobacteraceae | 1 | .28 | 6.7e-04 |
| Gammaproteobacteria;SUP05_Arctic | 1 | .02 | 2.5e-05 |
| Gammaproteobacteria;SUP05_1a | 1 | .15 | 9.8e-03 |
| Gammaproteobacteria;SUP05_1c | 1 | .41 | 7.5e-03 |
| Archaea | 2 | .78 | 2.4e-04 |
| Bacteroidetes;Flavobacteriales | 2 | .32 | 9.4e-05 |
| Bacteroidetes;Bacteroidales | 2 | .72 | 9.5e-05 |
| Proteobacteria;Alphaproteobacteria | 2 | .34 | 1.2e-04 |
| Proteobacteria;Deltaproteobacteria | 2 | .85 | 1.7e-05 |
| Proteobacteria;Epsilonproteobacteria | 2 | .27 | 6.7e-04 |
| Proteobacteria;Gammaproteobacteria | 2 | .66 | 3.7e-04 |
| Archaea | 3 | .78 | 2.4e-04 |
| Bacteria | 3 | .999 | 7.2e-05 |
| MaxBin2 | | | |
| Thaumarchaeota;Cenarchaeum | 1 | .49 | .05 |
| Flavobacteriales;Cytophaga | 1 | .11 | .04 |
| Bacteroidetes;Bacteroidales | 1 | .36 | .04 |
| Alphaproteobacteria;SAR11 | 1 | .34 | .03 |
| Deltaproteobacteria;Nitrospina | 1 | .30 | .03 |
| Deltaproteobacteria;Sva0853;SAR324 | 1 | .37 | .01 |
| Epsilonproteobacteria;Arcobacteraceae | 1 | .54 | .10 |
| Gammaproteobacteria;SUP05_Arctic | 1 | .17 | .14 |
| Gammaproteobacteria;SUP05_1a | 1 | .04 | .008 |
| Gammaproteobacteria;SUP05_1c | 1 | .11 | .005 |
| Archaea | 2 | .43 | .05 |
| Bacteroidetes;Flavobacteriales | 2 | .21 | .02 |
| Bacteroidetes;Bacteroidales | 2 | .55 | .08 |
| Alphaproteobacteria | 2 | .38 | .01 |
| Deltaproteobacteria | 2 | .40 | .01 |
| Epsilonproteobacteria | 2 | .61 | .09 |
| Gammaproteobacteria | 2 | .48 | .01 |
| Bacteria | 3 | .95 | .004 |
| Archaea | 3 | .54 | .06 |

**Table D.2:** All MaxBin2.0 and SAGEX (cluster) precision-recall statistics

| Taxa | Level | Precision | Sensitivity |
|---|---|---|---|
| ESOM + R | | | |
| Thaumarchaeota;Cenarchaeum | 1 | NA | .00 |
| Flavobacteriales;Cytophaga | 1 | .03 | .01 |
| Bacteroidetes;Bacteroidales | 1 | .22 | .06 |
| Alphaproteobacteria;SAR11 | 1 | .71 | .01 |
| Deltaproteobacteria;Nitrospina | 1 | .41 | .07 |
| Deltaproteobacteria;Sva0853;SAR324 | 1 | .61 | .06 |
| Epsilonproteobacteria;Arcobacteraceae | 1 | .40 | .05 |
| Gammaproteobacteria;SUP05_Arctic | 1 | .08 | .03 |
| Gammaproteobacteria;SUP05_1a | 1 | .23 | .04 |
| Gammaproteobacteria;SUP05_1c | 1 | .29 | .04 |
| Archaea | 2 | .03 | .05 |
| Bacteroidetes;Flavobacteriales | 2 | .09 | .12 |
| Bacteroidetes;Bacteroidales | 2 | .05 | .13 |
| Alphaproteobacteria | 2 | .10 | .10 |
| Deltaproteobacteria | 2 | .26 | .19 |
| Epsilonproteobacteria | 2 | .04 | .03 |
| Gammaproteobacteria | 2 | .41 | .28 |
| Bacteria | 3 | .998 | .02 |
| Archaea | 3 | .06 | .002 |

ESOM protocol modifies input. Results are indirectly comparable.

**Table D.3:** All ESOM+R precision-recall statistics