IMPROVED METHODS FOR UNDERSTANDING SPARSE, MULTI-DIMENSIONAL, HIGH THROUGHPUT SEQUENCING DATA

by

Sophie J. Weiss

B.S. University of Pennsylvania, 2009

M.S. University of Colorado, 2013

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Chemical and Biological Engineering

2015

This thesis entitled:
Improved Methods for Understanding Sparse, Multi-Dimensional, High Throughput Sequencing Data
written by Sophie J. Weiss
has been approved for the Department of Chemical and Biological Engineering

_____
Ryan Gill

_____
Rob Knight

Date_____

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation standards
of scholarly work in the above mentioned discipline.

Abstract

Weiss, Sophie J. (Ph.D., Chemical and Biological Engineering)

Improved Methods for Understanding Sparse, Multi-Dimensional, High Throughput Sequencing Data

Thesis directed by Professor Ryan Gill and Professor Rob Knight

Tremendous advances in genetic and sequencing technology are enabling unprecedented insight into human disease, forensics, and cellular mechanisms, to name a few. Conclusions drawn from these studies are strongly influenced by the interpretation of their associated massive data sets. The goal of this thesis is to understand, develop, and apply algorithms to help overcome common ecological and biological sequencing study challenges: contamination, differences in sampling efforts, a very large amount of zeroes, and compositionality.

We use simulations and experimental data to understand how different matrix normalization strategies mitigate the effects of the aforementioned challenges on downstream analyses, particularly principal coordinate analysis (PCoA). PCoA is very useful to researchers as a summary of overall differences in the studied populations, e.g. case vs. control. For determining specifically which taxa in the studied populations differ, we focus on methods for differential expression/abundance testing. Our benchmarking of nonparametric and parametric models, designed to increase rare taxa detection power, leads to recommendations for which strategy to use depending on a specific data set's properties. Using these normalization and differential abundance detection guidelines, we apply them in a forensics study of how carcass mass influences the resulting microbial community, which has implications for post-mortem interval calculation. Then we move from studying changes in abundance of individual taxa to changes in abundance of multiple taxa; ultimately deriving how taxa inter-relate in pairwise and even higher-order interactions. Correlation analysis is critical because all microbial communities

and biological systems are highly interconnected, however correlations are especially adversely affected by sparsity and compositionality.

Finally, while this thesis has focused on improved analysis in the context of microbial communities, the same methodologies apply to any extremely multi-dimensional and sparse high-throughput sequencing data set. In particular, we turn to the data arising from individual microbes, such as the Trackable Multiplex Recombineering (TRMR) approach used in strain engineering. We adapt the TRMR approach from outdated microarray technology to high-throughput sequencing, and integrate it with streamlined bioinformatics software. This approach can be used to study the bacterial response to any inhibitory chemical. We focus here on the alleles contributing to antimicrobial resistance and susceptibility, and identify a unique allelic and proteomic fingerprint for each antibiotic. Collectively, we present advances towards addressing the major sequencing data set challenges of contamination, uneven library sizes, a plethora of zeroes, and compositionality, and apply them to a wide range of topics in microbial ecology and biological engineering.

Acknowledgements

First, I would like to thank my advisors, Dr. Ryan Gill and Dr. Rob Knight for their support and guidance, as well as for letting me choose my own research topics. I also thank Dr. Joel Kaar for helpful advice over the years. I'd also like to thank my thesis committee members: Dr. Manuel Lladser and Dr. Anushree Chatterjee.

I thank Gill and Knight lab members past and present for their friendship and scientific guidance. Specifically, thanks to Tirzah Glebes, Nich Sandoval, Lauren Woodruff, and Tom Mansell for teaching me about proper experimental design/techniques. I also give a big thanks to Will Van Treuren, Antonio Gonzales, Amnon Amir, Zech Xu, and Jessica Metcalf.

Pooneh Mortazavi and Tom Mansell were co-authors on Chapter 6. Amnon Amir, Embriette Hyde, Jessica Metcalf, and Se Jin Song contributed to Chapter 2. Zech Xu, Amnon Amir, Shyamal Peddada, Kyle Bittinger, Antonio Gonzales, Catherine Lozupone, Jesse Zaneveld, Yoshiki Vázquez-Baeza, and Amanda Birmingham contributed to Chapter 3. Jessica Metcalf and David O. Carter contributed to Chapter 4. Will Van Teuren, Catherine Lozupone, Zech Xu, Luke Ursell, and Amanda Birmingham contributed to Chapter 5. Additionally, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, as well as their advisors contributed to Chapter 5. I greatly appreciate your help. I also thank my funding sources: the NIH/CU Biophysics Training Grant Program, as well as separate NIH and National Human Genome Research Institute grants.

Finally, I thank my family and friends for their love and support.

Table of Contents

Tables

Figures

Chapter 1

Introduction

The development of synthetic, molecular, and genome biology methods, combined with the advent of next-generation sequencing, has rapidly advanced understanding of biological systems both at the ecosystem and cellular levels. For example, we can now better identify the types and interactions of microbial communities, and the genes within a microbe contributing to a certain trait. Previously, scientists could only identify a small fraction of the microbial community through laborious culture-based methods, and the entire genome sequence of an organism was largely unknown. The interactions of microbial communities with each other and their hosts have recently been implicated, through human correlative studies and experimental mouse models, in numerous conditions including obesity and metabolic syndrome [1-4], cardiovascular disease [5], *C. difficile* colitis [6], inflammatory bowel diseases [7], HIV [8], multiple sclerosis [9], autism [10, 11], and others. These communities are influenced by diet, culture, geography, age, and antibiotic use among other factors [12]. Microbes also hold great forensic potential to be a reliable estimation of post-mortem interval [13, 14]. At the cellular level, chemical genomics has made strides towards mapping genotype to phenotype [15-19]; however, throughput and analysis strategies are slow, and combinatorial optimization of genes to better engineer a certain trait remains a barrier [20].

The conclusions of these sequencing studies are frequently derived from vast count matrices. A typical example is an n x m count matrix M composed of n features (either genes, or different types of microbes) and m samples, with the count of feature i in sample j represented as M_{ij}. The task of correctly analyzing these matrices, which with current technology have the potential to reach sizes of 25,000 rows and 200,000 columns, remains challenging [21]. The goal of

my thesis project is to determine the best algorithms to infer microbial dynamics from sequencing data. Then we apply these methods to genotype to phenotype high-dimensional matrices, and speed up matrix construction, with a focus on antibiotic resistance.

Challenges that are a barrier to all high-throughput sequencing data are differences in sampling efforts, and compositionality. Differences in sampling efforts due to sequencing technology manifest as samples with different numbers of sequences, resulting in feature x sample matrices with uneven column sums. A natural instinct when faced with such a count matrix is to normalize by dividing each entry by its total column sum. However, this can be problematic to interpretation since samples with more reads exhibit higher variability and detect more rare species. For example, if a sample from a healthy patient ('Control') has one million sequences and a sample from a sick patient ('Case') has ten thousand, 'Control' will have some counts of rare species, whereas 'Case' will likely have zero counts. If a researcher divides the matrix by total column sum, without further mathematical modeling, this could result in the mistaken conclusion that 'Case' lacks the rare species seen in 'Control', and that missing rare species may be causing disease.

The constrained sum (compositionality) of the sample counts is especially problematic for inferring correlations, as first recognized by Karl Pearson on other compositional data types [22]. Since then, there have been repeated demonstrations that inferring correlations on compositional data could be misleading, and even lead to opposite conclusions [23-27]. Because the sampling process fixes the total number of sequences to a smaller amount than actually present in the environment, the counts of these sequences are only relative to each other, and the absolute abundances are unknown. Compositionality also affects other analysis types, like whether a specific species is significantly differentially abundant between two sample types, and

contributing to e.g. disease in the 'Case' samples vs. healthy 'Control'. For example, we consider the case when both the 'Case' and 'Control' patients have exactly the same abundance of all species, except species 'A' vastly increases in the 'Case' patients. If both the 'Case' and 'Control' samples have e.g. 100 sequences, the 'Case' sample will have fewer sequences left for species other than 'A', artificially depressing their counts in comparison to the 'Control'. This may lead to the mistaken conclusion that species other than 'A' have decreased in abundance in 'Case' samples, whereas no such even happened in the environment. Compositionality, as well as the effect of contaminating sequences, is further aggravated by low species diversity and few sequences per sample, which we explore in Chapter 2. We also make recommendations, both at the experimental and analysis stages, on how to avoid spurious results due to contamination.

A challenge specific to metagenomic sequencing data, and some strain engineering data, such as that resulting from Trackable Multiplex Recombineering (TRMR, pronounced 'tremor') is sparsity, or an extremely large amount of zeroes. This sparsity is due to both biological and technical reasons: there are many different possible types of microbes, some microbes are found in only a small fraction of samples, and there are samples with low sequencing depth [28]. Zeroes commonly reach 97% of the possible matrix counts [28]. Even without the issue of contaminating sequences, library size differences, compositionality, and sparsity represent a large barrier to proper interpretation of research results for all analysis types. Also, all three challenges are inter-related. For example, samples with low sequencing depth will have more zeroes than samples with high sequencing depth, and all the samples are compositional.

Before any analysis is done, all major analysis pipelines [29-31] implement matrix normalization to account for some of the three challenges; however, no one normalization method can account for all three, which we explore in Chapter 3. Some normalization methods

also deal with contamination better than others. We visualize the effect of normalization on the data by principal coordinates analysis (PCoA), which is a very popular analysis for showing how similar/different the populations in sample types are. PCoA attempts to collapse the most important components of the multi-dimensional variation into three dimensions. Samples sharing (not sharing) populations will cluster closer (farther) together in PCoA space. We then zoom in to the problem of determining which specific taxa differ significantly in abundance between the sample type clusters. We assess these differential abundance testing statistical tests for their robustness to the three challenges. Particularly, recently developed parametric methods attempt to increase detection power compared to non-parametric techniques for rare taxa. In Chapter 4, we turn to a practical application on forensics research, and utilize the previously recommended normalization and differential abundance testing techniques to evaluate the effect of carcass mass on gravesoil microbial communities. While differential abundance testing focuses on individual taxa, we next move to correlation analysis in Chapter 5 to provide recommendations on the best method for inferring how microbes interact as a community, as well as again ability to address uneven column sums, compositionality, and sparsity.

While the previous discussions on contamination, normalization, differential abundance testing, and correlation analysis were focused on metagenomic data, the same guidelines apply to TRMR data. In Chapter 6, we therefore apply these techniques to a novel problem: developing a rapid workflow, from experiment to analysis, for mapping genotype to phenotype. The method can be applied to tolerance for any chemical from toxic metabolites to next-generation biofuels and, our current focus, antibiotics. Microbial resistance to antibiotics is a growing crisis in clinical, agricultural, and industrial settings, and novel mechanisms of resistance challenge even the most powerful antibiotics [32, 33]. Sub-lethal concentrations of antibiotics are an important

contributor to this problem because the mutations that enable survival at lethal concentrations often have a higher fitness cost, there are fewer of them, and there is less room for enrichment [34]. Sub-lethal antibiotic concentrations are commonly found in wastewater and agricultural runoff [35]. Previous attempts to characterize genome-scale responses to antibiotic challenges [15, 17, 36-38] relied on either (1) the low-throughput construction of large libraries or (2) many generations of adaptive evolution, where characterization was limited by sequencing only the surviving colonies. However, the increasing throughput and decreasing cost of multiplex oligonucleotide synthesis [39] and high-throughput sequencing [40] has enabled unprecedented advances in throughput of genome engineering and analysis technologies [18, 19, 41-44]. In Chapter 6, we adapt the methodology of Warner et al. [19] from microarray hybridization to rapid high-throughput sequencing and multivariate analysis, and examine the genomic modifications contributing to sub-lethal concentrations of antibiotic resistance.

For the multivariate analysis, we adapt the Quantitative Insights into Microbial Ecology (QIIME) software package [29]. In addition to the aforementioned recommendations for normalization, differential abundance analysis, and correlations, QIIME also contains other very useful analyses. Procrustes analysis enables comparison of the similarity of two distance matrices in PCoA space by stretching, rotating, and scaling the two datasets to see if similar conclusions can be drawn [45]. Machine learning is another important analysis for large datasets, with the random forest classifier generally the most useful [46-48]. By training the classifier on a portion of the data, the predictive accuracy can be tested on the other portion. This is useful, for example, to see if the disease a person has can be predicted based on their microbial signature [49]. Also helpful are a variety of group significance tests for distance matrices, like non-parametric multivariate analysis of variance (PERMANOVA) [50] and analysis of similarities (ANOSIM) [51].

We use these techniques to identify antibiotic resistance and susceptibility alleles, observe that the allelic response mimics the proteomic response, and find a unique fingerprint for all antibiotics regardless of mechanism of action. Finally, in Chapter 7, we make concluding remarks and discuss future directions.

Chapter 2

Tracking Down the Sources of Experimental Contamination in Microbiome Studies

As published in *Genome Biology*, 2014 15:564

2.1 Introduction

High-throughput sequencing has revolutionized our understanding of the microbial world, providing a means by which we can characterize microbial communities in considerable detail without being affected by biases introduced by culture-based protocols that might reveal only a small fraction of the community. We have learned that, although humans share over 99.9% of their genomic DNA sequence with one another, they might share as little as 10% of their microbes at a given body site. Therefore, an intriguing hypothesis is that some aspects of the human phenotype might be determined more by microbial DNA than human DNA. Over the past five years, an enormous push in microbiome research has elucidated many of the factors that can affect this microbial individuality – the human microbiome is affected by diet, culture, geography, age and antibiotic use, among other factors [12]. Importantly, the microbiome has been implicated in numerous health conditions through correlative studies in humans and experimental research in mouse models. These conditions range from obesity [52] to multiple sclerosis [9]. However, if samples are not collected, processed, and analyzed properly, this may lead to erroneous conclusions.

2.2 Microbes are Increasingly Studied in Low-Biomass Environments

Microbes play crucial roles not just in human-associated ecosystems – they are ubiquitous in every environment, from deep ocean vents to the arctic. However, this ubiquity

also poses major challenges in controlling for background contamination present in the air, laboratory surfaces, the skin and clothing of researchers, and in laboratory reagents. In the November issue of BMC Biology, Salter and colleagues [53] present a comprehensive study of contaminant sources in microbiome experiments and demonstrate the great influence that contamination can have on readouts of microbial communities based on DNA. These effects are especially important in studies focusing on samples of low biomass.

Much of recent high-impact microbiome research has focused on the gut, which is characterized using fecal samples as a proxy for the distal large intestine. Fecal samples have such high biomass that the DNA of fecal microbes almost certainly overwhelms contaminating background microbial DNA from reagents and other sources. However, as microbiome research expands in scope to include samples of lower biomass, such as the airways, placenta or even blood plasma, the standard high-throughput approaches often used for fecal samples will probably not be sufficient to generate reliable readouts of the microbial communities or assemblages associated with such samples. This problem arises because, as the 'true' biomass becomes smaller, the potential for contaminants occupying a larger fraction of the sequences will become greater. For example, a recent study by Kennedy and colleagues [54] showed that PCR template concentration, which is associated with sample biomass (especially when extracted DNA concentrations are not normalized before downstream processing, which is common in high-throughput settings), significantly affects the resulting microbial community profile.

2.3 Sample Contamination Can Come From Many Sources

Several sources can contribute to sample contamination and can occur at several steps, occurring between collection and sequencing. The use of non-sterile equipment, or accidental exposure to the environment or researcher, can contaminate the sample. However, it should be

noted that microbial DNA could be present even in sterile equipment. Therefore, strict protocols, such as the use of cleansuits, gloves, facemasks, and bleach and UV for cleaning equipment, could be needed to prevent contamination during sample collection. Microbial DNA can also be introduced during sample processing, either during initial microbial DNA extraction or during PCR amplification, in the case of marker gene amplification and sequencing (multiple displacement amplification (MDA) and related techniques can also amplify reagent contaminants during library preparation for shotgun metagenomic sequencing). In reality, microbial DNA that is not endogenous to the samples being studied probably contaminates every microbiome dataset to some extent. The work by Salter et al. [53] takes important steps in helping us to determine what these contaminants are, where they come from and how large an effect they can have on research results.

To investigate the diversity of microbial contaminants, the researchers used an elegant combination of positive and negative (blank) controls. They used a pure culture of *Salmonella bongori*, which has not been observed as a common contaminant, in a series of five 10-fold dilutions to assess the effect of background contamination on samples with varying biomass [53]. Using 16S ribosomal RNA (rRNA) gene amplification and high-throughput sequencing, along with typical PCR-amplified 'blank' controls comprising ultrapure water, they distinguished contaminants arising from DNA extraction kits and other sources, including PCR kit reagents, laboratory consumables and personnel. Salter and colleagues [53] show very clearly that contaminating organisms became increasingly dominant as the biomass of *S. bongori* decreased, with contaminants representing the majority of the microbial biomass by the fifth dilution.

Sixty-three taxa were unique to the diluted samples compared with the PCR 'blank' control, implicating the DNA extraction kit as a likely contaminant source. Salter and colleagues

also analyzed metagenomes produced through shotgun sequencing of non-amplified bacterial DNA, which, unlike the 16S rRNA gene-sequencing protocol, does not include a targeted PCR step and thus eliminates the introduction of contamination through PCR. Nonetheless, the authors observed similar results, with contaminants dominating in low-biomass samples, and again implicating the DNA extraction kit as the source of contaminants [53] Interestingly, of the four DNA extraction kits that Salter *et al*. tested, the lowest levels of contamination appeared to result from the use of the MoBio kit, which is the kit used by most of the major microbiome studies, such as the Human Microbiome Project [55] and Earth Microbiome Project [21].

2.4 Contamination Can Affect Biological Conclusions, Especially When Confounded With Other Variables

Salter and colleagues [53] then demonstrated how contamination could affect interpretation of biological studies by analyzing low-biomass samples from a recent study of nasopharyngeal microbes during infant development [56]. The authors found that, in the original dataset, contaminant operational taxonomic units (OTUs) associated with different batches of the same extraction kit drove the clustering patterns found in principal coordinate analysis (PCoA) space, which led to the misleading conclusion that the composition of the nasopharyngeal microbiome changed with age. Once contaminant OTUs were removed from the dataset and the primary samples were reprocessed using a different extraction kit, samples no longer clustered by age, thereby significantly altering the research results and interpretation [53].

Such batch effects have already been observed in genomic data [57]. As suggested by Leek and colleagues, a good way to check that an experimental, rather than biological, variable is driving the PCoA clustering is to test whether the experimental variable correlates strongly with the major principal components. This procedure assumes that the samples have been randomly

assigned to DNA extraction batches, PCR batches and DNA sequencing-instrument runs: a common mistake, which should clearly be avoided, is to confound experimental variables (such as time-point) or clinical variables (such as case versus control status) with one or more of these variables, making resolution of the biological effect against the background of these technical effects in principle impossible. OTU-based analyses, such as correlation networks or differential-abundance testing, are even more sensitive to any type of contaminant. This sensitivity arises because each sample has a constrained total number of sequences; therefore, any change in one OTU affects all others in that sample. Furthermore, any taxa that are present in the blanks should be monitored carefully during the rest of the analysis, as recommended by Salter *et al*. [53].

The implications of this study are that microbiome researchers might need to take additional precautions in the laboratory and develop both laboratory and bioinformatics workflows for monitoring contamination. As part of their conclusion, the authors recommend a reasonable set of steps for minimizing the effects of contaminants before, during and following sequencing, including the use of negative controls, technical replicates, sample randomization and keeping records of kits and other reagents [53]. However, this study also highlights the need for additional studies that benchmark methods and protocols in microbiome research. For example, researchers might want to consider using different concentrations of a single bacterial culture as a control, which could produce better estimates of the degree and nature of contamination than reagent blanks.

2.5 Conclusion

Owing to the high sensitivity of high-throughput sequencingbased microbiome analysis, reproducibility (how well the results repeat themselves) and bias (how well the results reflect the reality) can be a major concern. The work of Salter and colleagues [53] is a springboard from

which microbiome researchers, who have been controlling for contamination primarily within individual labs, can begin to build a consensus for laboratory and bioinformatics approaches, thus helping researchers avoid spurious results and saving valuable money, time and effort. This work builds on previous studies [58-60], and recently the Microbiome Quality Control project, that rigorously tested variability introduced by differences in methodology, such as storage, preservation, extraction and analysis, and, especially, highlights taxa that might systematically point to reagent contamination [58]. However, contamination from other biological sources, and especially the mouth and skin of the investigators conducting the studies, should also be considered as a possibility when reviewing results that are surprising in the light of prior knowledge of the biological niches of the organisms involved. Together, all these efforts are beginning to close important gaps of knowledge in microbiome research and provide essential resources that inform better study design and practices for all microbiome researchers.

Chapter 3

Effects of Library Size Variance, Sparsity, and Compositionality on the Analysis of Microbiome Data

PeerJ, submitted, 2015

3.1 INTRODUCTION

Although data produced by high-throughput sequencing has proven extremely useful for understanding microbial communities, the interpretation of these data is complicated by several statistical challenges. To ease data interpretation, data are often normalized to account for the sampling process and differences in sequencing efforts. Ordination analysis, such as principal coordinates analysis (PCoA) [61], is subsequently applied to these normalized data to visualize broad trends of how similar or different bacteria are in certain sample types, such as healthy *vs.* sick patients). Samples containing similar bacteria will group, or cluster, close together, while differences in bacterial composition will cause separation in PCoA space. Next, researchers may wish to determine, through statistical testing, which specific bacteria are significantly differentially abundant between two sample type clusters.

For example, patients with *Clostridium difficile* infection cluster separately from healthy patients in PCoA plots, and these overall differences in community composition are driven by differences in microbial relative abundances [62-64]. Restoration of each intestinal bacteria type to healthy levels leads to patient recovery, and causes samples from treated patients to overlap with healthy individuals in PCoA plots. Significant changes in certain bacterial species abundances has also been linked to inflammatory bowel diseases [7], diarrhea [65], obesity [1, 2, 4], HIV [8], diet [66],

culture, age, and antibiotic use [67], among many other factors. However, the veracity of these discoveries depends upon how well the chosen normalization and differential abundance testing techniques address the statistical challenges posed by the underlying community sequence data.

Following initial quality control steps to account for errors in the sequencing process, microbial community sequencing data is typically organized into large matrices where the columns represent samples, and rows contain observed counts of clustered sequences commonly known as Operational Taxonomic Units, or OTUs, that represent bacteria types. These tables are often referred to as OTU tables. Several features of OTU tables can cause erroneous results in downstream analyses if unaddressed. First, the microbial community in each biological sample may be represented by very different numbers of sequences, reflecting differential efficiency of the sequencing process rather than true biological variation. This problem is exacerbated by the observation that the full range of species is rarely saturated, such that more bacterial species are observed with more sequencing. (Similar trends by sequencing depth hold for discovery of genes in shotgun metagenomic samples [68, 69]). Thus, samples with relatively few sequences can have inflated beta (β, or between sample) diversity, because authentically shared OTUs are erroneously scored as unique to samples with more sequences [70]. Second, most OTU tables are sparse, meaning that they contain a high proportion of zero counts [28]. This sparsity means that the counts of rare OTUs are uncertain, since they are at the limit of sequencing detection ability in large library size samples, and are undetectable in small library size samples. Third, each sample is only a small percentage of its original environment, constraining the total number of rRNA sequences to a constant sum; in such "compositional" data, researchers do not know the absolute counts of each type of OTU but only their relative abundances in relation to each other [24, 71, 72]. Uneven sampling depth, sparsity, and compositionality represent serious challenges for

interpreting these data. No normalization method or differential abundance testing method simultaneously addresses all of these challenges. Thus, investigators must choose methods based on relevant features of the dataset under consideration.

3.1.1 Normalization

Normalization is critical to address variability in sampling depths and number of zeros. Microbial ecologists in the era of high-throughput sequencing have commonly normalized their OTU matrices by rarefying, or drawing without replacement from each sample such that all samples have the same number of total counts. Samples with total counts below the defined threshold are excluded, sometimes leading researchers to face difficult trade-offs between sampling depth and the number of samples evaluated. To ensure the proper total sum is chosen, rarefaction curves can be constructed [73]. These curves plot the number of counts sampled (rarefaction depth) vs. the expected value of species diversity. Rarefaction curves provide guidance that allows users to avoid negatively impacting the species diversity found in samples by choosing too low a rarefaction depth. The origins of rarefying sample counts are mainly in sample species diversity measures, or alpha diversity [73, 74]. However, more recently rarefying has been used in the context of β-diversity [75, 76]. Rarefying samples for normalization is now the standard in microbial ecology, and is present in all major data analysis toolkits for this field [29, 31, 77, 78]. While rarefying is not an ideal normalization method, as it reduces statistical power by removing some data, and was not designed to address compositionality, alternatives to rarefying have not been sufficiently developed until recently.

Normalization alternatives to rarefying all involve some type of transformation, the most common of which are scaling or log-ratio transformations. Effects of scaling methods depend on the scaling factor chosen; often, a particular quantile of the data is used for normalization, but

choosing the correct quantile is difficult [28, 79-81], and scaling can overestimate or underestimate the prevalence of zero fractions, depending on whether zeroes are left in or thrown out of the scaling [24, 82]. This is because putting all samples of varying sampling depth on the same scale ignores the differences in sequencing depth, and therefore resolution of species, between the samples. For example, a rare species having zero counts in a small rRNA sample can have a small fractional abundance in a large rRNA sample (unless further mathematical modeling beyond simple proportions is applied to correct for this). Scaling can also distort OTU correlations across samples, again due to zeroes, differences in sequencing depth, and sum constraints [24, 25, 71, 72, 83].

While rarefying and some scaling techniques, such as total sum scaling (proportions), treat OTU sequence counts as absolute environmental abundances, the counts are compositional and only a fraction from the original environment, making only their relative ratios known [24, 71]. In contrast, log ratio transformations correct for compositionality by exploiting this relative ratio information, and can also alleviate some noise in the data [24, 25, 71, 72]. However, because the log transformation cannot be applied to zeros (which are often well over half of microbial data counts [28]), sparsity is extremely problematic for methods that rely on this transformation. One approach to this issue is to replace zeros with a small value, known as a pseudocount. Despite active research on selection of pseudocount values for scaling methods [84, 85], the choice of pseudocount values can dramatically change the results [86, 87].

3.1.2 Differential Abundance Testing

For OTU differential abundance testing between conditions (e.g. case vs. control), a common approach is to first rarify the count matrix to a fixed depth and then apply a non-parametric test (e.g. Mann-Whitney test for tests of two classes; Kruskal-Wallis test for tests of

multiple groups). Non-parametric tests are often preferred because most OTU counts are not normally distributed [88]. However, this approach does not account for the fact that the OTU counts are compositional. Also, nonparametric tests such as the Kruskal-Wallis test do not fare well in terms of power when the data are sparse, but perform well when the data are not sparse [28]. Recently, promising parametric models that make stronger assumptions about the data have been developed in the subfields of transcriptomics ('RNA-Seq') and metagenomic sequencing. These may additionally be useful for microbial marker gene data [28, 30, 89-94]. Such models have greater detection power if their assumptions about the data are correct; however, studies of these models on RNA-Seq data have shown that they can yield poor results [95] if relevant assumptions are not valid.

These parametric models are composed of a generalized linear model (GLM) that assumes a distribution [96], and there is considerable debate about which distribution to use [28, 30, 95, 97-103]. In the genomics field, the negative binomial (NB) GLM has replaced the Poisson GLM to allow for estimating overdispersion [89, 90, 92]. This model type was also one of the first in the RNA-Seq field, and developed for use with a low number of replicates. NB models accommodate low replication by assuming that OTUs of similar mean expression strength have similar variance in their sample count distributions, estimating model parameters using this assumption, and then leveraging the GLM to perform exact statistical tests. These NB models, like rarefying with a non-parametric test, do not address compositionality. Also, while allowing for some overdispersion, the NB often yields a poor fit in the case of a large number of zeroes, which is very typical in microbiome data [28, 99]. Zero-inflated GLMs, the most promising of which is the zero-inflated Gaussian (ZIG), attempts to overcome this limitation [28]. The ZIG tries to address compositionality, sparsity, and unequal sampling depth by separately modeling

'structural' zero counts generated by e.g. undersequencing and zeros generated by the biological distribution of taxa. Log transformation of the non-zero counts yields the Gaussian. However, this mixture model distribution is designed for continuous data rather than discrete microbiome data. Hence, it is expected to do best in study designs that have large sample sizes and high sequencing depths, and thus best approximate continuous distributions.

Here, we evaluate some of the most widely used or promising techniques for analyzing sequencing data in the context of microbial ecology, with a focus on normalization and OTU differential abundance testing. In addition to these widely used or promising methods, we also test the naïve approaches of no normalization, and proportions (i.e. total sum scaling) for comparison purposes. Such comparisons are important, because while potential issues with many methodologies are known, the balance of sensitivity and specificity for these methods in situations commonly facing microbial ecologists is currently largely unknown. Recent work in this area [30], provides insights into the performance of parametric normalization and differential abundance testing approaches for microbial ecology studies. However, the work is primarily focused on estimating proportions from discrete data. We update and expand these recent findings using both real and simulated datasets exemplifying the additional combined challenges of uneven library sizes, sparsity, and compositionality.

3.2 MATERIALS AND METHODS

3.2.1 Normalization

The basic test of how well broad differences in microbial sample composition are detected, as assessed by clustering analysis, was conducted as in 'Simulation A' from McMurdie and Holmes [30]. Briefly, the 'Ocean' and 'Feces' microbiomes (the microbial data from ocean and human feces samples, respectively) from the 'Global Patterns' dataset [104] were used as

templates, modeled with a multinomial, and taken to represent distinct classes of microbial community because they have few OTUs in common. These two classes were mixed in many defined proportions (the 'effect size') in independent simulations in order to generate simulated samples of varying clustering difficulty. Samples were generated in sets of 40, as in McMurdie and Holmes [30]. We also tested smaller and larger sample sizes but saw little difference in downstream results. Additional sets of 40 samples were simulated for varying library sizes (1000, 2000, 5000, and 10000 sequences per sample). These simulated samples were then used to assess normalization methods by the proportion of samples correctly classified into the two clusters by the partitioning around medioids (PAM) algorithm [105, 106].

McMurdie and Holmes [30] evaluated clustering accuracy with five normalization methods (none, proportion, rarefying with replacement as in the multinomial model [107], DESeqVS [89], and UQ-logFC (in the edgeR package) [92]) and six beta diversity metrics (Euclidean, Bray-Curtis [108], PoissonDist [109], top-MSD [92], unweighed UniFrac [110], and weighted UniFrac [111]). We modified the normalization methods to those in Table S1 (none, proportion, rarefying without replacement as in the hypergeometric model [107], CSS [28], logUQ [80], DESeqVS [89], and edgeR-TMM [79]) and the beta diversity metrics to those in Fig2 and Fig. S1 (binary Jaccard, Bray-Curtis [108], Euclidean, unweighed UniFrac [110], and weighted UniFrac [111]), thus including more recent normalization methods [28, 80], and only those beta diversity metrics that are most common in the literature. We amended the rarefying method to the hypergeometric model [107], which is much more common in microbiome studies [29, 31]. Negatives in the DESeq normalized values [89] were set to zero as in McMurdie and Holmes [30], and a pseudocount of one was added to the count tables [30]. McMurdie and Holmes [30] penalized the rarefying technique for dropping the lowest fifteenth percentile of sample library sizes in their simulations by counting the dropped samples as 'incorrectly

clustered'. Because the 15th percentile was used to set rarefaction depth, this capped clustering accuracy at 85%. We instead quantified cluster accuracy among samples that were clustered following normalization to exclude this rarefying penalty (Fig. S1). Conversely, it has since been confirmed that low-depth samples contain a higher proportion of contaminants (rRNA not from the intended sample) [54, 112]. Because the higher depth samples that rarefying keeps may be higher quality and therefore give rarefying an unfair advantage, Fig. 2 compares clustering accuracy for all the techniques based on the same set of samples remaining in the rarefied dataset.

On the real datasets, non-parametric multivariate ANOVA (PERMANOVA) [50] was calculated by fitting a Type I sequential sums of squares model (y ~ Library_Size + Biological_Effect). Thus, we control for library size differences before assessing the effects on the studied biological effect. All data was retrieved from QIITA (https://qiita.microbio.me).

3.1.2 Differential Abundance Testing

The simulation test for how well truly differentially abundant OTUs are recognized by various parametric and non-parametric tests was conducted as in 'Simulation B' in McMurdie and Holmes [30], with a few changes. The basic data generation model remained the same, but the creation of 'true positive' OTUs was either made symmetrical through duplication or moved to a different step, to avoid introducing compositionality artifacts (see below) depending on the simulation. The 'Global Patterns' [104] dataset was again used, because it was one of the first studies to apply high-thoughput sequencing to a broad range of environments, which includes 9 environment types from 'Ocean', to 'Soil'; all simulations were evaluated for all environments. Additionally, we verified the results on the 'Lean' and 'Obese' microbiomes from a different study [113]. As in McMurdie and Holmes, significant changes were controlled for multiple comparisons using the Benjamini & Hochberg [48] False Discovery Rate (FDR) threshold of 0.05.

A simple overview of the two methods used for simulating differential abundance is presented in Figure 3-8A. In McMurdie and Holmes' [30] 'Original' simulation (second row), the distribution of counts from one environment (e.g. 'Ocean') was modeled off of a multinomial template (first row) for two similar groups ('Ocean_1' and 'Ocean_2'), ensuring a baseline of all 'true negative' OTUs. Following the artificial inflation of specific OTUs in the 'Ocean_1' samples to create 'true positives', fold-change estimates for every other OTU are affected. Thus, 'true negatives' are possible 'true positives.' This is because the counts in an OTU table are compositional, or relative abundances constrained to a sum. To control for this we inflate OTUs by pairs of differentially abundant OTUs in both the 'Ocean_1' and 'Ocean_2' samples (third row), creating a new 'Balanced' simulation.

We also tested the effect of differentially abundant organisms dominating one type of community by drawing from a multinomial distribution where solely that organism's template value is increased. This 'Compositional' approach is explained in Figure 3-8B, and the results are shown in Figure 3-14. In Figure 3-14, the environmental abundances of 25% of the OTUs in one group are increased.

Besides the above procedural changes to the McMurdie and Holmes [30] simulation, we also modified the rarefying technique from sampling with replacement (multinomial) to sampling without replacement (hypergeometric - as in the previous Normalization simulations) [107]. The testing technique was modified from a two-sided Welch t-test to non-parametric Mann-Whitney test, which is widely used and more appropriate because the OTU distributions in microbiome data usually deviate from normality. The techniques used (Table 3-2) differ only by the addition of another RNA-Seq method, Voom [93]. Finally, we corrected the FPR definition [30]

from FP/(TP + FP) to FP/(TN + FP), where FP = number of false positive OTUs, TP = number of true positive OTUs, and TN = number of true negative OTUs.

3.2.3 Power Curve Calculations

Similar to Table S1 in McMurdie and Holmes [27], we considered a very simplistic set-up to evaluate the effect of rarefying on power when comparing two groups, labeled A and B. As in McMurdie and Holmes [27], we considered the extreme case of a microbial population consisting of only 2 species (or 2 OTUs), with OTU1 + OTU2 = library size. For power calculations, we assumed that the amount of OTU1 in group B is 85% of the amount of OTU1 in group A. Thus, it is enough to quantify the proportion of OTU1 in group A and library sizes of groups A and B to specify the whole system.

We considered varied patterns of proportions of OTU1 in group A ranging from very rare to common (0.5% to 50%). The library size of group A was fixed at either 500, 1000 or 10,000 sequences per sample. Meanwhile, the library size of group B was always taken to be at least as large as that of group A and was either 10,000 or 100,000 sequences per sample. Various rarefied percentages of the group B library size were considered. The percent-rarefied calculation for the first set of power curves is exemplified below using a library size of 500 for library A and an unrarefied library size of 10,000 for B:

Library size for A	Library size for B
500	10,0000 (unrarefied case)
500	5,000 (50% rarefied)

500	1,000	(90% rarefied)
500	500	(95% rarefied)

For each scenario of proportion of OTU1 and library sizes, power was computed using Fisher's exact test. Power calculations were done using the statistical software SAS. Power calculation results are provided in Figure 3-10.

3.2.4 Software Package Versions

R version 3.1.0 [114] was used with Bioconductor [115] packages phyloseq version 1.10.0, DESeq version 1.16.0, DESeq2 version 1.4.5, edgeR version 3.6.8, metagenomeSeq version 1.7.31, and Limma version 3.20.9. Also, we used python-based QIIME version 1.9.0, with Emperor [116].

3.3 RESULTS AND DISCUSSION

3.3.1 Normalization

When there is a strong biological signal, and normalization is done properly, PCoA can yield clear clustering and insight into microbial community differences (Figure 3-1A). However, low-depth samples can lead to poor cluster resolution (Figure 3-1B), both by reducing information on community structure, and by being more readily influenced by contamination [54, 112]. Furthermore, if no data normalization is applied, or the normalization method fails to properly correct for differences in sequencing efficacy, the original library size of the samples can confound biological differences (Figure 3-1C). This is because samples of lower sequencing depth fail to detect rare taxa. Highly sequenced samples will thus appear more similar to each other than to shallow sequenced samples because they are scored as sharing the same rare taxa.

To assess all the normalization methods (Table 3-1), we conducted simulations in the context of results that are highly critical of the rarefying technique [30]. Briefly, only necessary modifications (Methods) were made to the code of McMurdie and Holmes [30], making our

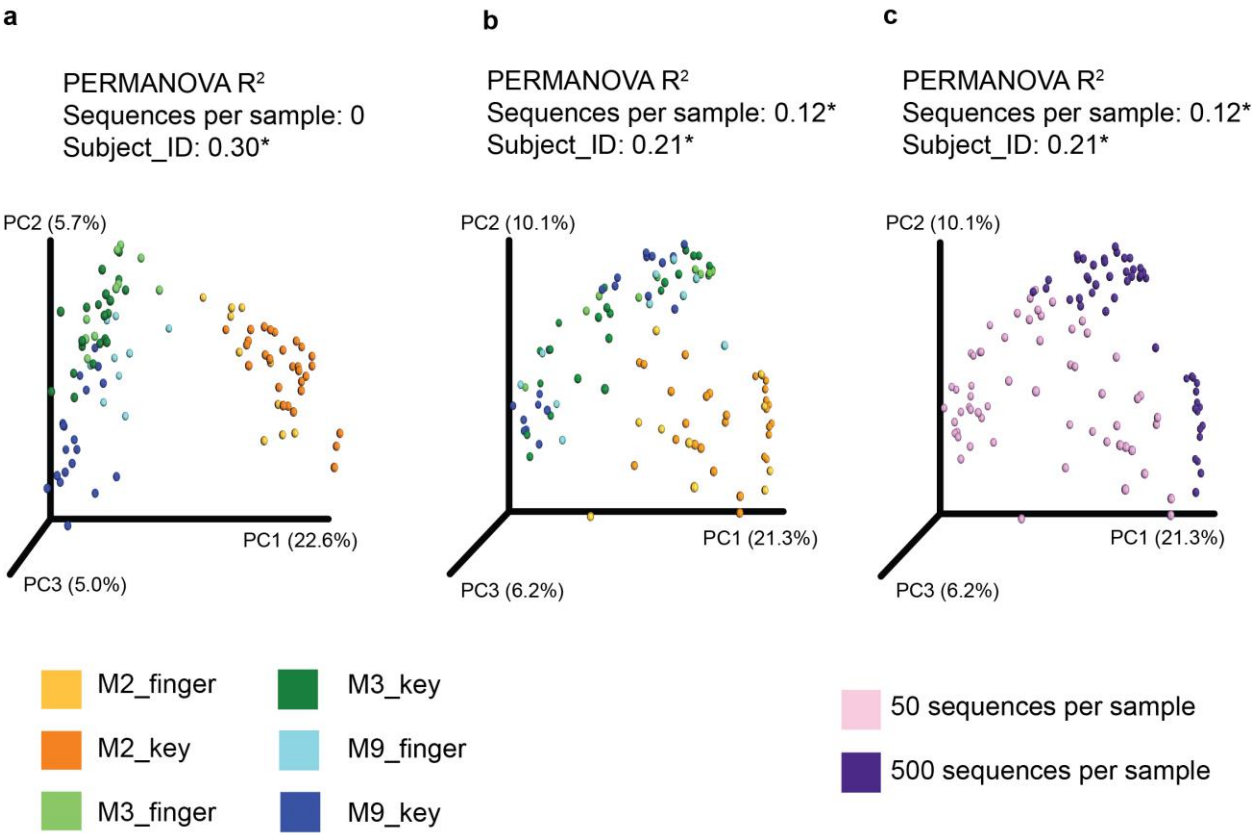


Figure 3-1 Effect of sampling depth on ordination methods. **(a)** Data rarefied at 500 sequences per sample**. (b, c)** Data not normalized, with a random half of the samples subsampled to 500 sequences per sample and the other half to 50 sequences per sample. **(b)** Colored by subject_ID, **(c)** Colored by sequences per sample. Non-parametric ANOVA (PERMANOVA) effect sizes (R^2) roughly represent the percent variance that can be explained by the given variable. Asterisk (*) indicates significance at $p < 0.01$. The distance metric of unweighted UniFrac was used for all panels.

Method	Description
None	No correction for unequal library sizes is applied.
Proportion	Counts in each column are scaled by the column's sum.
Rarefy	Each column is subsampled to even depth without replacement (hypergeometric model).
logUQ	log Upper Quartile - Each sample is scaled by the 75th percentile of its count distribution, then the counts are log transformed.
CSS	Cumulative Sum Scaling - This method is similar to logUQ, except CSS enables a flexible sample distribution-dependent threshold for determining each sample's quantile divisor. Only the segment of each sample's count distribution that is relatively invariant across samples is scaled by CSS. This attempts to mitigate the influence of larger count values in the same matix column.
DESeqVS	Variance Stabilization (VS) - For each column, a scaling factor for each OTU is calculated as that OTU's value divided by its geometric mean across all samples. All of the reads for each column are then divided by the median of the scaling factors for that column. The median is chosen to prevent OTUs with large count values from having undue influence on the values of other OTUs. Then, using the scaled counts for all the OTUs, and assuming a Negative Binomial (NB) distribution, a mean-variance relation is fit. This adjusts the matrix counts using a log-like transformation in the NB generalized linear model (GLM) such that the variance in an OTU's counts across samples is approximately independent of its mean.
edgeR-TMM	Trimmed Mean by M-Values (TMM) - The TMM scaling factor is calculated as the weighted mean of log-ratios between each pair of samples, after excluding the highest count OTUs and OTUs with the largest log-fold change. This minimizes the log-fold change between samples for most OTUs. The TMM scaling factors are usually around 1, since TMM normalization, like DESeqVS, assumes that the majority of OTUs are not differentially abundant. The normalization factors for each sample are the product of the TMM scaling factor and the original library size.

Table 3-1 Normalization methods tested

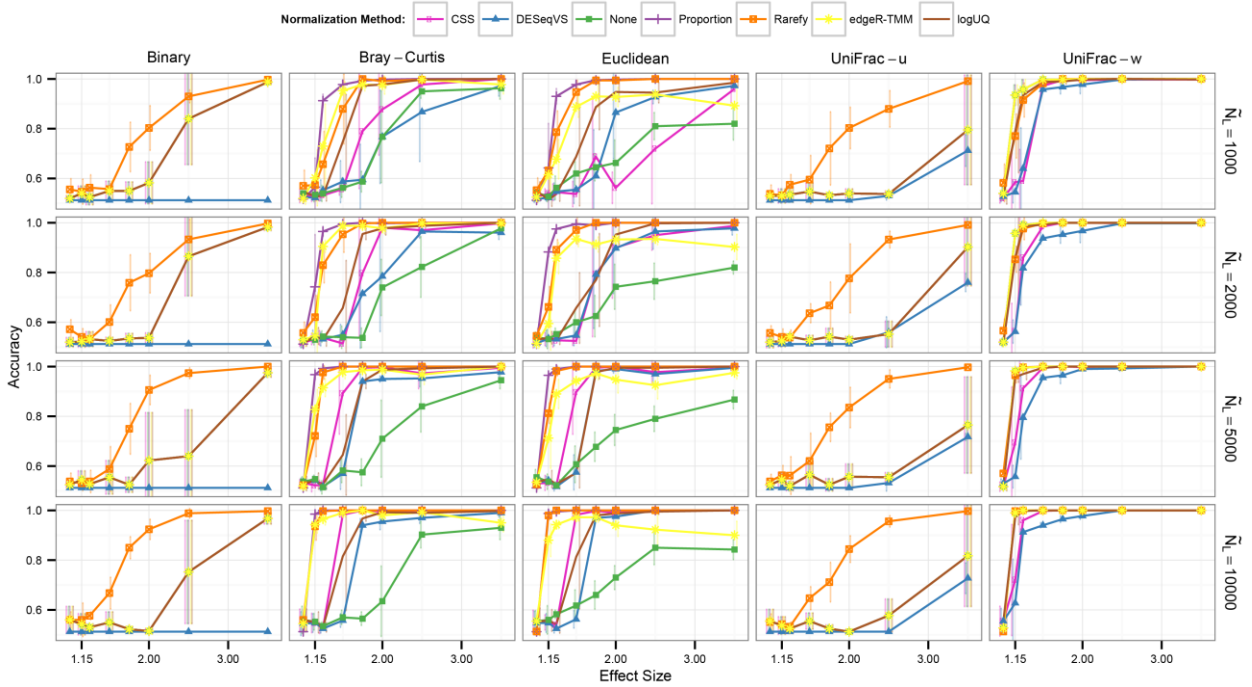


Figure 3-2 Simulated clustering accuracy if rarefying is not penalized for removing the lowest 15th percentile samples. The right axis represents the median library size (N_L), while the x-axis 'effect size' is the multinomial mixing proportions of the two classes of samples, *'Ocean'* and *'Feces'*. See caption for Figure 3-3 for further details.

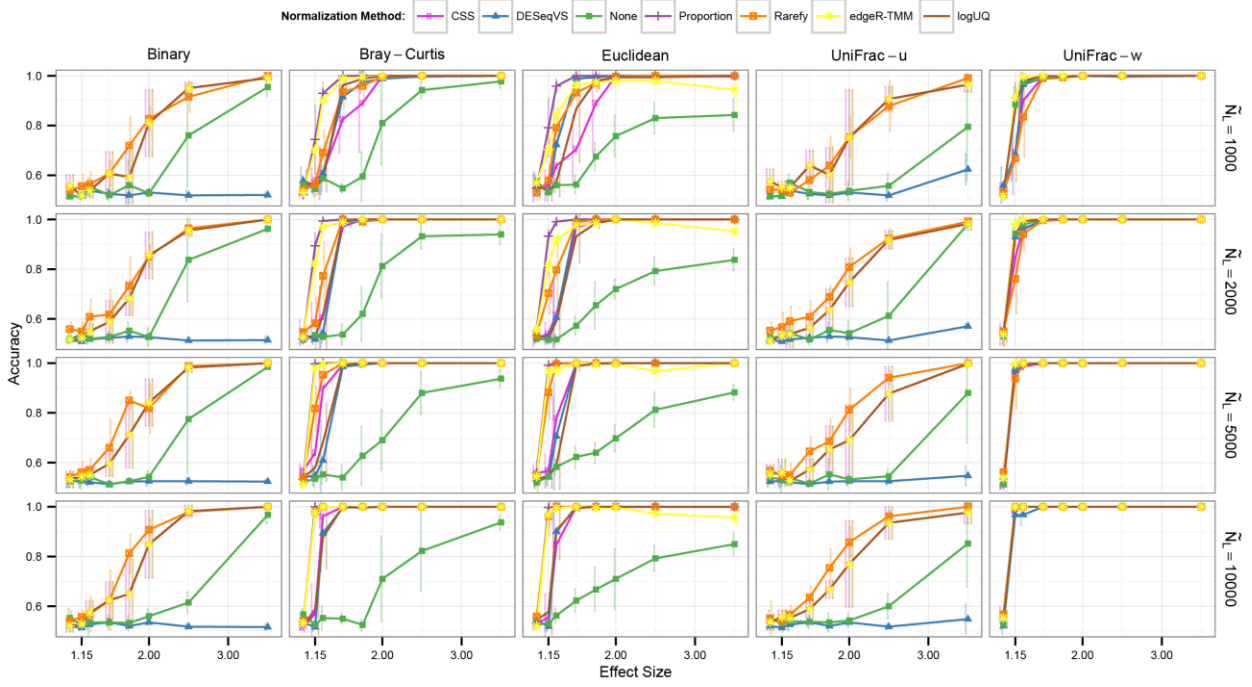


Figure 3-3 Comparison of common distance metrics and normalization methods across library sizes when low-coverage samples are excluded. Clustering accuracy is shown for all combinations of five common distance metrics (panels arranged from left to right) across four library depths (panels arranged from top to bottom; N_L, median library size), six sample normalization methods (series within each panel), and several effect sizes (x-axis within panels). In all cases, samples below the 15th percentile of library size

were dropped from the analysis in order to isolate the effects of rarifying from the effects of dropping low-coverage samples. The x-axis ('effect size') within each panel represents the multinomial mixing proportions of the two sample classes *'Ocean*' and '*Feces*'. A higher effect size represents an easier clustering task. The y-axis ('accuracy') shows the accuracy of each classifier, as assessed by the fraction of simulated samples correctly clustered.

approach easily comparable. If rarefying is not penalized for the fifteenth percentile lowest depth samples that are thrown out, it can do better than other techniques (Figure 3-2). This practice of removing low depth samples from the analysis is supported by the recent discovery that small biomass samples are of poorer quality and may contain contaminating sequences [54, 112]. Furthermore, alternatives to rarefying also recommend discarding low-depth samples, especially if they cluster separately from the rest of the data [28, 91]. If all other techniques are run only on the same samples as rarefying, rarefying still does well (Figure 3-3). These results demonstrate that previous microbiome ordinations using rarefying as a normalization method likely drew correct conclusions, even if some low depth samples were removed. However, these results also suggest that CSS [28] and DESeq's variance-stabilizing transformation [89] are promising alternatives for normalization prior to PCoA analysis, especially for weighted distance metrics. For unweighted metrics that are based on species presence and absence, like binary Jaccard and unweighted UniFrac, DESeq's variance-stabilizing transformation performs poorly. This is because the negatives resulting from DESeq's log-like transformation are set to zero (as in McMurdie and Holmes [30]), which ignores rare species.

No good solution exists for the negatives output by the DESeq technique. DESeq was developed mainly for use with Euclidean metrics [110, 111], for which negatives are not a problem; however, this issue yields misleading results for ecologically useful non-Euclidean measures, like Bray-Curtis [108] dissimilarity. Also, the negatives pose a problem to UniFrac's [110, 111] branch length. The alternative to setting the negatives to zero, or adding the absolute value of the lowest

negative value back to the normalized matrix, will not work with distance metrics that are not Euclidean because it amounts to multiplying the original matrix by a constant due to DESeq's log-like transformation. Also, the addition of a constant (or pseudocount; here, one) to the count matrix prior to CSS [28], DESeq [89], and logUQ [80] transformation as a way to avoid log(0) is not ideal, and clustering results have been shown to be very sensitive to the choice of pseudocount, due to the nonlinear nature of the log transform [86, 87]. This underscores the need for a better solution to the zero problem so that log-like approaches inspired by Aitchison can be used [72], and is especially critical since microbial matrices are almost always much more than half sparse [28].

While simulations are a useful initial check, real datasets are often much more complex. Therefore, all normalization methods were also examined on real data. To perform an initial, detailed comparison of normalization methods, we selected the data set from Caporaso *et al.* [117]. The data included a wide variety of samples, representing both environmental and host-associated sources. To provide an extreme example of differences in sequencing depth, we artificially decreased the library size by 90% for half the samples in the data set. The samples selected for library size reduction were chosen randomly, and the same artificially altered data was used in all normalization comparisons.

Using the data set from Caporaso *et al.* [117], we observed substantial biases/confounding of results due to sequencing depth. In ordination of unweighted UniFrac distance by PCoA, the soil samples were split into two groups along the first principal coordinate when no normalization was used (Figure 3-4A). Soil samples appearing in the group to the left had more reads than those appearing in the group to the right. Similarly, the two stool samples in the data set were arranged close to soil samples with similar library size. When the data was rarefied prior to ordination, soil and stool samples were arranged along the first two coordinates

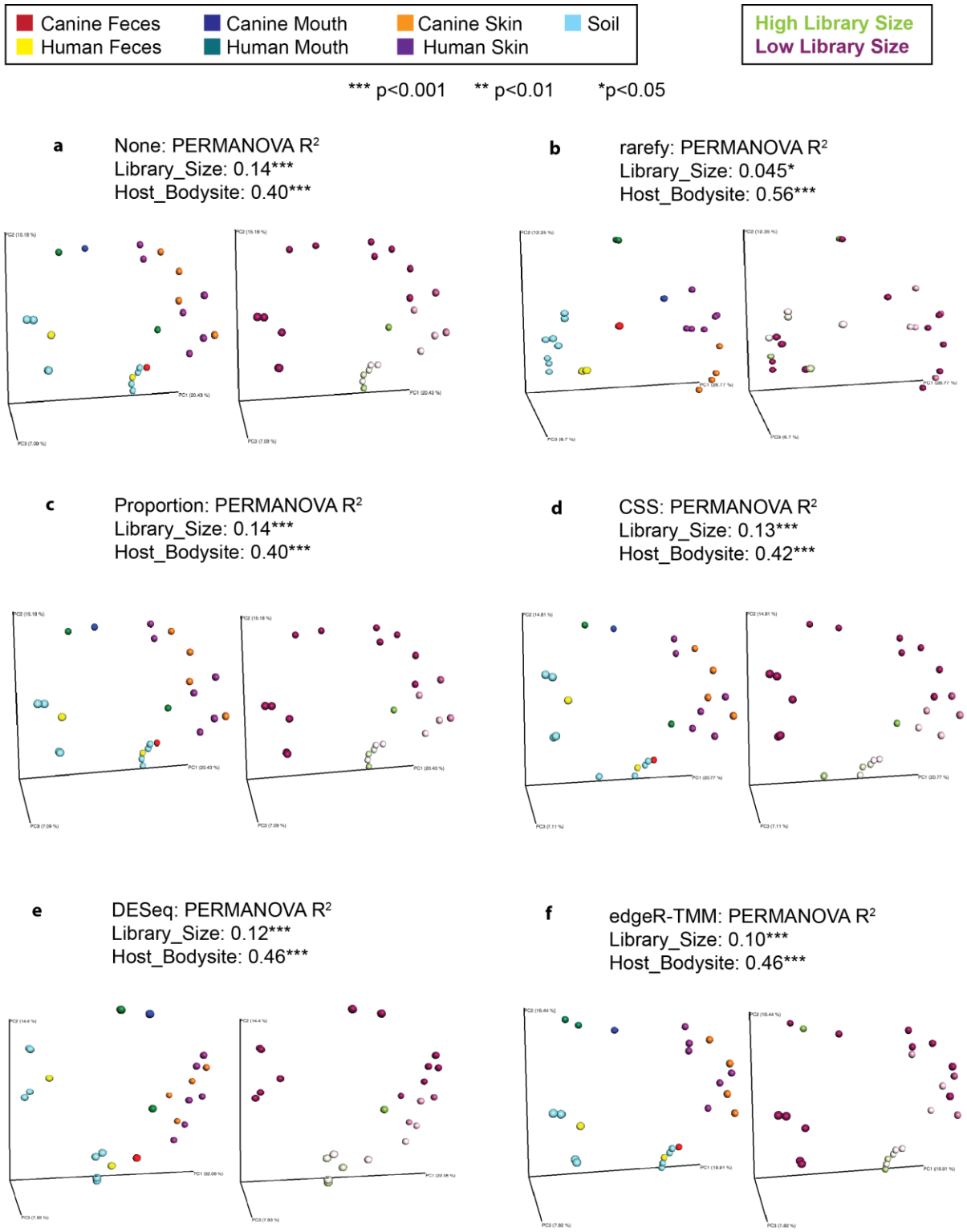


Figure 3-4 Rarefying clusters more according to biological origin, and diminishes the effect of library size. Rarefying exhibits a higher effect size (R^2) for biological origin, and a lower effect size (R^2) of original library size. Unweighted UniFrac was used for clustering, and a random half of samples were subsampled to 10 times fewer sequences per sample. The 45-degree line splits low from high depth samples in all but the rarefying technique.

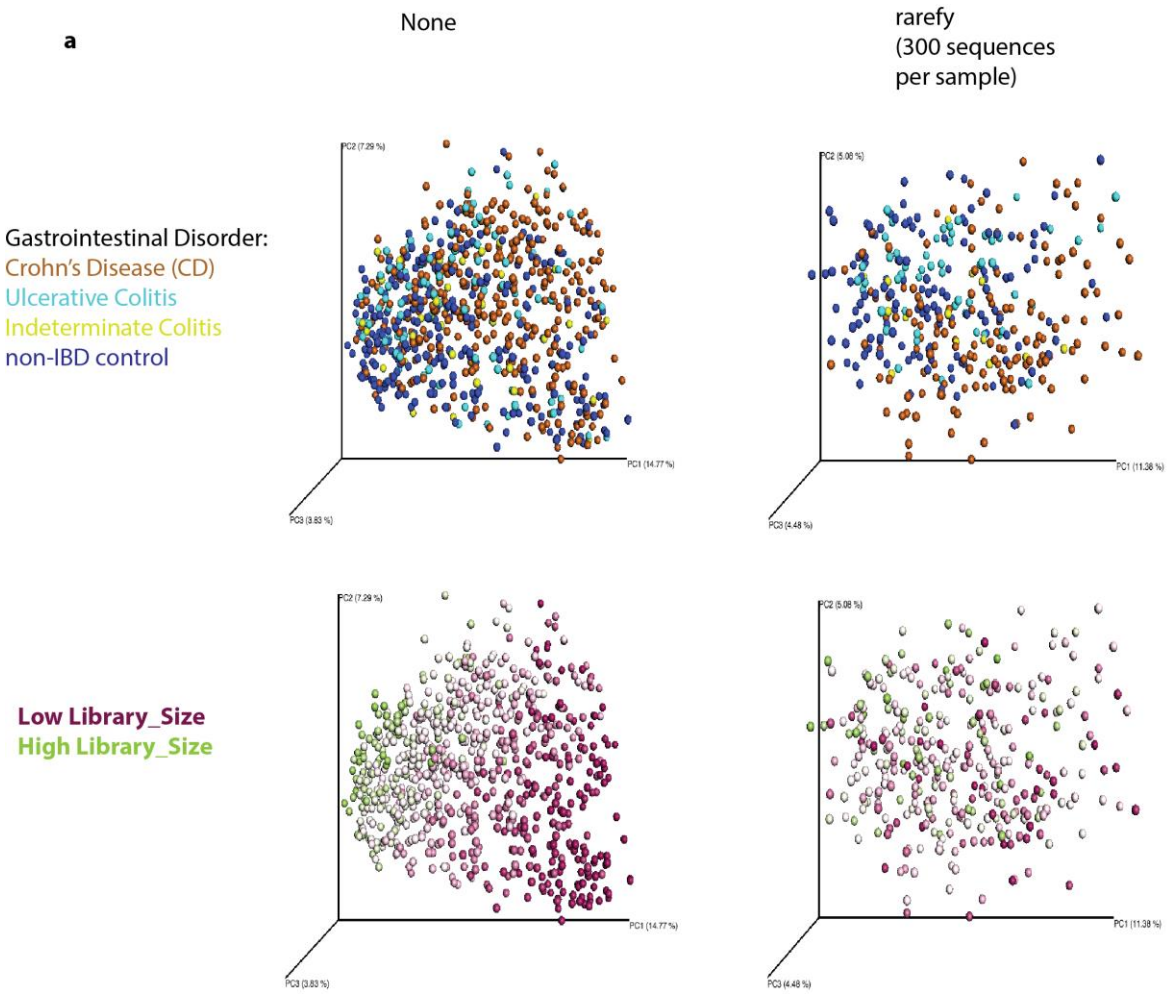
according to sample type rather than library size (Figure 3-4B). Other methods of normalization preserved the characteristic pattern seen in the non-normalized data, where soil and stool samples were separated into groups according to library size (Figure 3-4C-F).

Normalization did not affect conclusions drawn from non-parametric multivariate ANOVA (PERMANOVA) [50], but we did observe differences in the effect size estimated for sample type, and library size (R^2). Without normalization, the estimated effect size of sample type for unweighted UniFrac distance was R^2=0.40. When the data was rarefied prior to computing distances, the estimated effect size increased to R^2=0.56.

Other methods of normalization produced effect sizes similar to the non-normalized result. Although the true effect size is not known for this data set, the environment of origin is known to be a dominant effect in the determination of bacterial species observed [118]. Without normalization, there is a large effect (R^2=0.14) corresponding to original library size, which is a known artifact of the sequencing process. Rarefying helps to remove the effect of sequencing depth (R^2=0.045), whereas other normalization techniques do not remove this signal artifact, again resembling the non-normalized data.

As another example, we selected the inflammatory bowel disease (IBD) data set from Gevers *et al.* [7]. In contrast to the previous data set, all samples here were taken from a single environment type, namely human stool, and were extremely low depth, having an average of 375 sequences per sample. In an ordination of unweighted UniFrac distance with no normalization, there is again strong clustering by library size, with a group of samples with low sequencing depth appearing slightly separate from the other samples (Figure 3-5A). Samples in the low-depth group are either dominated by a lack of species detected due to few sequences, thus artificially inflating the β–diversity, or constitute different bacterial species than the main group

a



b

y ~ Library_Size + Gastrointestinal_Disorder

	PERMANOVA (R^2)	None	rarefy	Proportion	CSS	DESeq	edgeR-TMM
Unweighted UniFrac	Library_Size	0.055***	0.007***	0.055***	0.055***	0.055***	0.055***
	Gastrointestinal Disorder	0.022***	0.036***	0.022***	0.022***	0.022***	0.022***
Weighted UniFrac	Library_Size	0.018***	0.041***	0.018***	0.054***	0.05***	0.018***
	Gastrointestinal Disorder	0.026***	0.078***	0.026***	0.036***	0.033***	0.026***

*** $p < 0.001$

Figure 3-5 Low library size samples can diminish result quality, regardless of normalization technique. The inflammatory bowel disease (IBD) dataset of Gevers et al.**,** average library size 375 sequences per sample. **(a)** Extremely low depth samples cluster in lower right hand corner of PCoA plots with no normalization, or rarefying alternatives, unweighted UniFrac. **(b)** The original library size of samples is a dominant effect, even influencing weighted UniFrac, with low library sizes and subtle biological clustering for rarefying alternatives. This diminishes if low library size samples are removed.

of stool samples, which should raise suspicion of potential problems from contamination or poor quality PCR products. Furthermore, the first principal coordinate in Figure 3-5A is more strongly correlated with library size (R^2=0.055, Fig 3-5B) and poorly correlated with disease state (R^2=0.022), with sampling depth explaining twice the variance of the studied biological effect. Subsampling the data to uniform library size increased the correlation with disease state (R^2=0.036), while other methods did not (R^2=0.022 for proportion, DESeq, and CSS). Because the average library size is so low for this study, the library size also affects weighted UniFrac, where there is again low effect size for this gastrointestinal disorder. Thus, extremely low depth samples still need to be discarded from rarefying alternatives, especially if they are suspected of yielding a poor representation of the true bacterial community due to experimental factors.

PCoA plots using ecologically common metrics for all of the normalization techniques on a few key real datasets representing a gradient [119], distinct body sites [120], and time series [121] are shown in Figures 3-6 and 3-7. Most measures do well in these cases where there is strong separation between the categories. Clustering according to sequence depth is less of a problem in these datasets since they have strong clustering patterns, however, some clustering according to depth persists. For example, in the 'Moving Pictures of the Human Microbiome' dataset [121], there is some clustering by sequence depth within each of the four main clusters when normalization alternatives to rarefying are applied. It is noteworthy that CSS normalization results appear robust to the distance metric used, including even Euclidean distance (results not shown), which have been reported to perform poorly on highly sparse matrices [122].

Thus, both simulations and real data suggest that rarefying remains a strong technique for sample normalization prior to ordination and clustering, especially for presence/absence distance metrics that have historically been very useful (such as binary Jaccard and unweighted

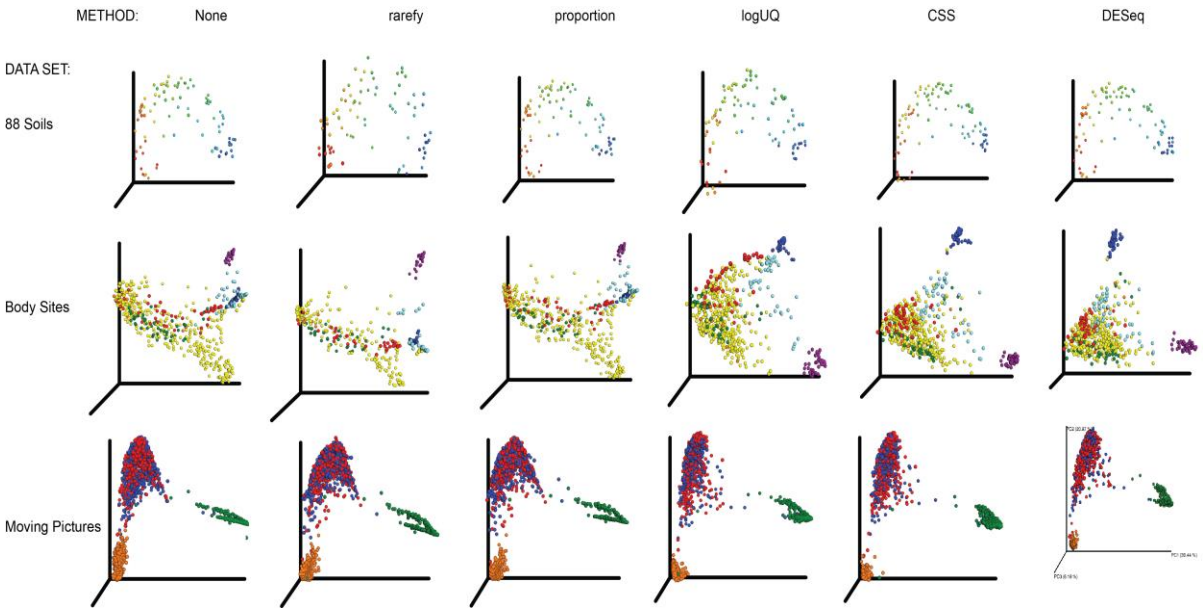


Figure 3-6 All normalization techniques on key microbiome datasets, Bray Curtis distance. Rows of panels show (from top to bottom) data from 88soils, Body Sites, Moving Pictures. 88 soils is colored according to a color gradient from low to high pH. The Costello et al. body sites dataset is colored according to body site: feces (blue), oral cavity (purple), the rest of the colors are external auditory canal, hair, nostril, skin, and urine. Moving Pictures dataset: Left and Right palm (red/blue), tongue (green), feces (orange). It is important to note that all the samples in these datasets are approximately the same depth, and there are very strong driving gradients.

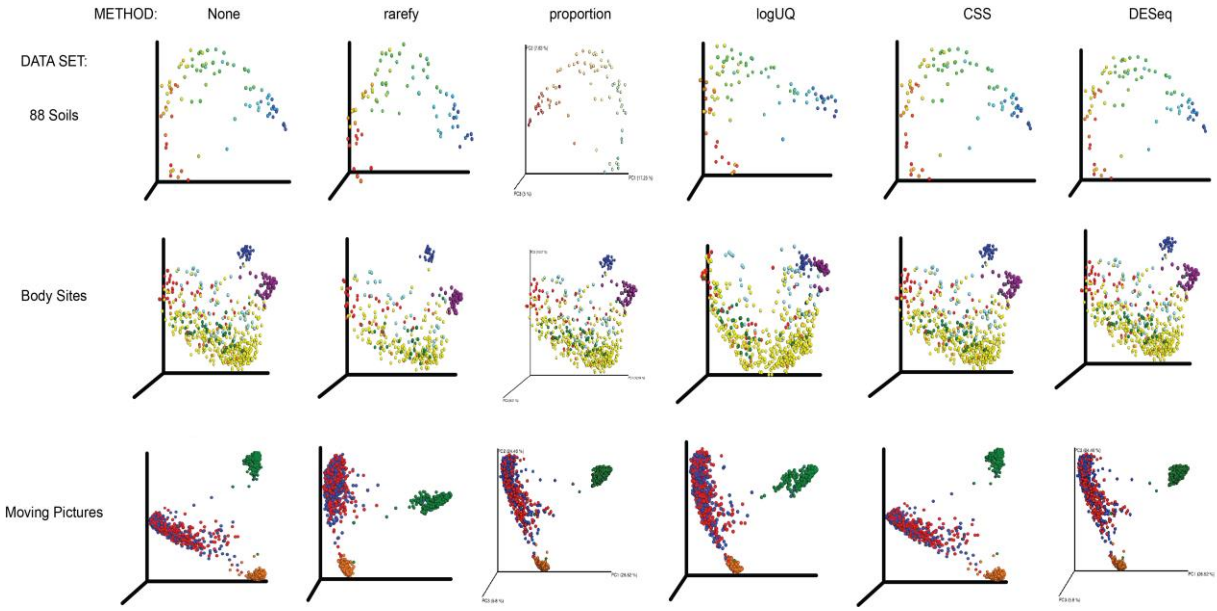


Figure 3-7 All normalization techniques on key microbiome datasets, unweighed UniFrac distance. See Figure 3-6 caption for details.

UniFrac [110] distances), subtle effects, small library sizes, and large differences in library size. Of the other methods, and for weighted distance measures, we recommend metagenomeSeq's CSS [28] or DESeq's variance stabilizing transformation [89]; however, the researcher must check for erroneous clustering according to sequence depth.

3.3.2 Differential Abundance Testing

Differential abundance analysis is useful for testing whether certain microbes have higher relative abundance in one condition vs. another (e.g. healthy vs. diseased patients). More complex statistical methods specifically for RNA-Seq data have been developed and include DESeq [89], DESeq2 [91], edgeR [92, 94], and Voom [93] (Table S2). MetagenomeSeq [28] however, was developed specifically for microbial datasets, which usually contain many more zeros than RNA-Seq data. These five methods incorporate more sensitive statistical tests than the standard non-parametric distributional tests such as the Wilcoxon rank-sum test, and they assume a distribution. Therefore, they hold great potential for better prediction of rare OTU behavior.

Previous work in this area concluded that the newer differential abundance testing models are worthwhile, and that the traditional practice of rarefying causes a high rate of false positives [30]. However, the latter conclusion was due to an artifact within the simulation (see Methods and Figure 3-8A-B). Instead, we found that rarefying does not cause a high rate of false positives, but may lead to false negatives due to the decreased power that results from throwing away some of the data (Figure 3-9). The severity of the power decrease caused by rarifying depends upon how much data has been thrown away. (This problem has been known for a long time, leading to the general guideline to rarefy to the highest depth possible without losing too many samples [123].) In order to determine where the greatest loss in power or information occurs when a dataset is rarefied, we constructed power curves from a simple two-species simulation

Method	Description
Wilcoxon rank-sum test	Also called the Mann-Whitney *U* test. A non-parametric rank test, which is used on the un-normalized ('None'), proportion normalized, and rarefied matrices.
DESeq	nbinomTest - a Negative Binomial model conditioned test. More conservative shrinkage estimates compared to DESeq2; resulting in stricter type I error control.
DESeq2	nbinomWaldTest - The Negative Binomial generalized linear model (GLM) is used to obtain maximum likelihood estimates for an OTU's log-fold change between two conditions. Then Bayesian shrinkage, using a zero-centered Normal distribution as a prior, is used to shrink the log-fold change towards zero for those OTUs of lower mean count and/or with higher dispersion in their count distribution. These shrunken long fold changes are then used with the Wald test for signficance.
edgeR	exactTest - The same normalization method (in R, method = 'RLE') as DESeq is utilized, and for differential abundance testing also assumes the NB model. The main difference is in the estimation of the dispersion, or variance, term. DESeq estimates a higher variance than edgeR, making it more conservative in calling differentially expressed OTUs.
Voom	'variance modeling at the observational level' - library sizes are scaled using the edgeR log counts per million (cpm) normalization factors. Then LOWESS (locally weighted regression) is applied to incorporate the mean-variance trend into precision weights for each OTU.
metagenomeSeq	fitZIG - a zero-inflated Gaussian (ZIG) where the count distribution is modeled as a mixture of two distributions; a point mass at zero, and a normal distribution. Since OTUs are usually sparse, the zero counts are modeled with the former, and the rest of the log ransformed counts are modeled as the latter distribution. The parameters for the mixture model are estimated with an Expectation-Maximization algorithm, which is coupled with a moderated t-statistic.

Table 3-2 Differential abundance methods tested

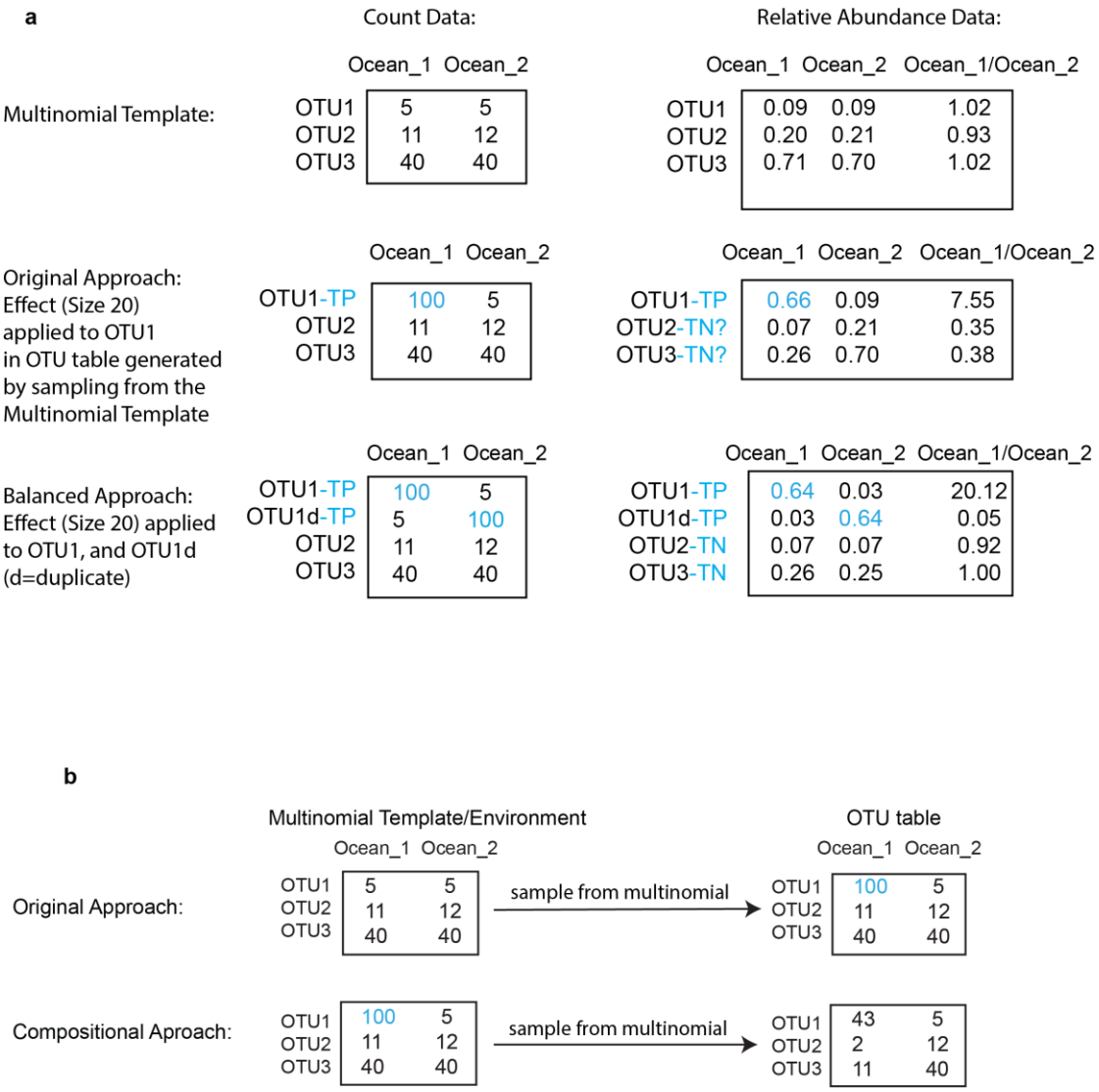


Figure 3-8 Simple example of the reasoning behind differential abundance simulations. **(a)** In actual OTU tables generated from sequencing data, the counts (left column) are already compositional and therefore only relative (left column). Application of the 'effect size' to the original 'Multinomial' template to create fold-change differences disturbs the distinction between true positive (TP) and true negative (TN) OTUs in the 'Original' simulation, but not the 'Balanced' simulation. **(b)** Tracking the sum of average differences in relativized counts between non-differentially abundant (true negative) OTU counts in two sample categories during the simulation. The first panel on the right indicates library size (2000 and 50000 sequences per sample on average). The second panel indicates numbers of samples per category respectively (3 and 100). **(c)** Creation of a 'Compositional' OTU table from the 'Multinomial' template, where the counts/relative abundances are intentionally blurred for the TN OTUs.

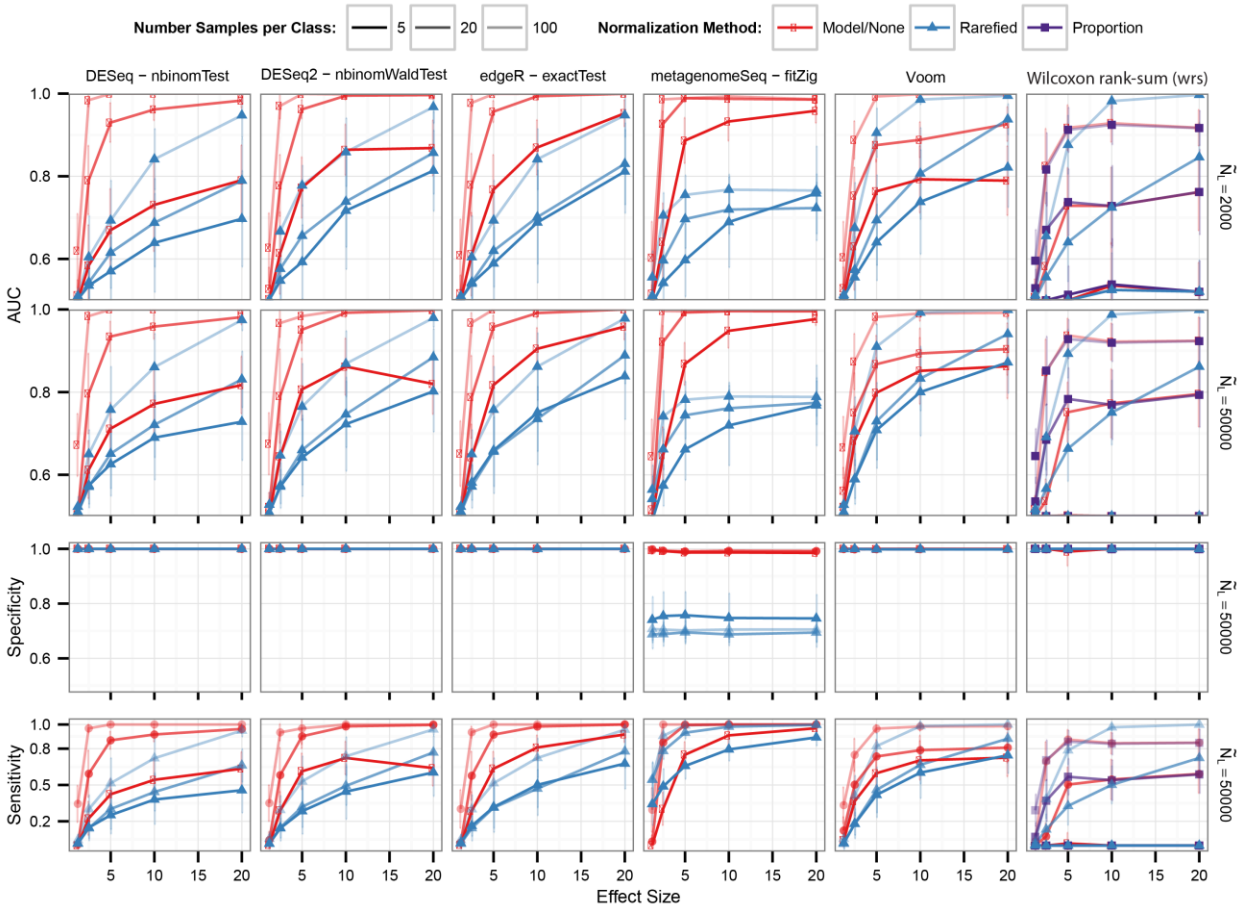


Figure 3-9 Differential abundance detection performance.
The AUC ('Area Under the Curve') version of the ROC ('Receiver Operator Characteristic') curve is the ratio of sensitivity to (1-specificity), or true positive rate vs. false positive rate. A higher AUC indicates better differential abundance detection performance. The 'effect size' represents the fold-change of the 'true positive' OTUs from one condition (e.g. case) to another (e.g. control). The right axis represents the median library size (N_L), while the shading on the graph lines represents the number of samples per class. 'Model/None' represents data analyzed with a parametric statistical model (e.g. DESeq), or no normalization. Blue lines in, e.g. the DESeq column represents the data was rarefied, then DESeq was applied. Since the fitZIG model depends upon original library size information, the model does poorly on rarefied data.

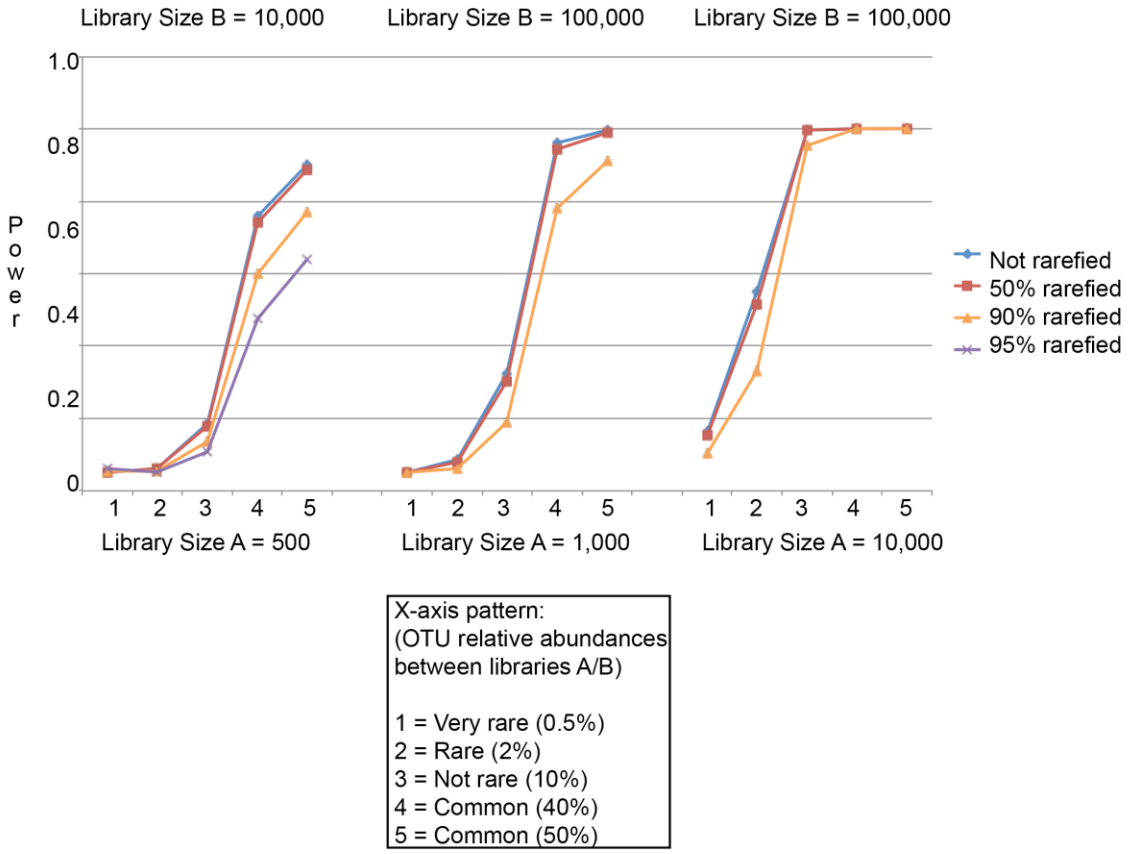


Figure 3-10 The effect of rarefying on power for different OTU relative abundances and library sizes. The detection power of differentially abundant OTUs of varying levels of relative abundance (very rare to common). This is for two samples A and B. For power calculations, we assumed that OTU1 fraction of group B is 85% of the OTU1 fraction of group A. Library type A was fixed, while library size B was subsampled at different percentages, creating the power curves calculated with Fisher's exact test.

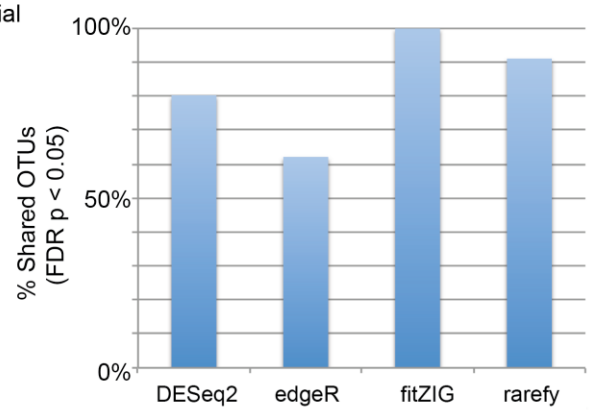
(Figure 3-10). The greatest loss in power occurs for rare to common OTUs (e.g. relative abundance ranging from 0.5% to 50%) depending on the library size. This has also been observed in gene expression studies [124]. Also, consistent with other studies on subsampling [123, 124], subsampling to library sizes close to the original does not have much effect on the results (50% is treated as "close to the original" in this simplified example, but real microbiome studies are much more complex and thus the real threshold is likely lower, and data-dependent). We also observed that the performance of rarefying degrades faster for smaller library sizes.

Since simulations do not necessarily mirror reality, we also investigated the performance of the techniques on real data. This was done for the techniques shown to be most promising in the simulations: DESeq2 [91], edgeR [92, 94], metagenomeSeq [28], and rarefying. Ranges of dataset sizes were analyzed for environments that likely contain differentially abundant OTUs, as evidenced by PCoA plots and significance tests (Figure 3-11). Approximately 6 samples in each of the categories of human skin vs. soil from Caporaso *et al.* [117], 28 samples in each of the lean vs. obese categories from Piombino *et al.* [113], and 500 samples in the tongue vs. left palm categories from Caporaso *et al.* [121] were tested. Although we do not necessarily know which OTUs are true positives in these actual data, it is of interest to investigate how the most promising techniques compare to each other. While rarefying (at the 15th percentile as in McMurdie and Holmes [30]) finds fewer OTUs as significant, the OTUs it does find to be differentially expressed are remarkably stable. Agreeing with our modified simulation, it does not appear that rarefying causes a high type I error. For example, in Figure 3-11 there is high agreement between rarefying and the other techniques. However, edgeR, which is known to be too lenient in its dispersion estimates [28, 91], predicts a large number of significantly differentially abundant OTUs relative to other methods, especially for studies with fewer samples (Figure 3-11A), suggesting a high false positive rate in agreement with RNA-Seq studies [91, 95, 103].

We also used simulated data to investigate the situation in which the average library size between the two categories was not approximately equal (Figure 3-12). We found that of the newer methods, metagenomeSeq's figZIG [28] has a high sensitivity and a low false positive rate (1-specificity) compared to the other techniques. However, the false positive rate is still high. Rarefying achieves the lowest false positive rate, but at a cost to sensitivity. Thus, the method employed by investigators may depend on the sensitivity of the analysis in question to false

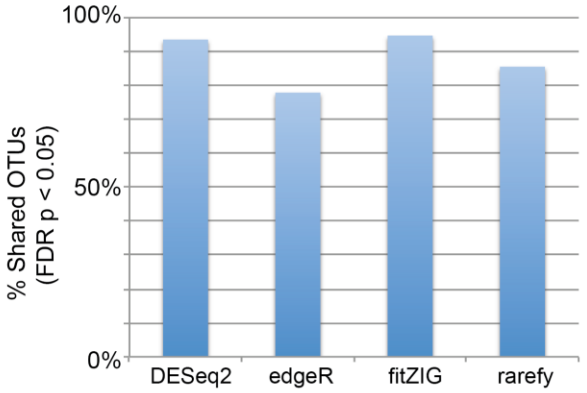
a Caporaso *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms *ISME* (2012).
~ 6 skin samples, 8 soil samples
mean sequences per sample: 1.3 million

	DESeq2	edgeR	fitZIG	rarefy
DESeq2	**1182**	1135	934	376
edgeR		**2357**	1344	706
fitZIG			**1445**	680
rarefy				**763**



b Piombino *et al.* Saliva from Obese Individuals Suppresses the Release of Aroma Compounds from Wine. *PLoS One* (2014).
~28 samples per category (lean vs. obese)
mean sequences per sample: 75,580

	DESeq2	edgeR	fitZIG	rarefy
DESeq2	**232**	206	164	122
edgeR		**267**	159	113
fitZIG			**189**	128
rarefy				**145**



c Caporaso *et al.* Moving Pictures of the Human Microbiome. *Genome Biol.* (2011).
~500 samples per category (tongue vs. left palm)
mean sequences per sample: 25,600

	DESeq2	edgeR	fitZIG	rarefy
DESeq2	**1070**	1051	955	1038
edgeR		**1122**	995	1097
fitZIG			**1038**	1004
rarefy				**1185**

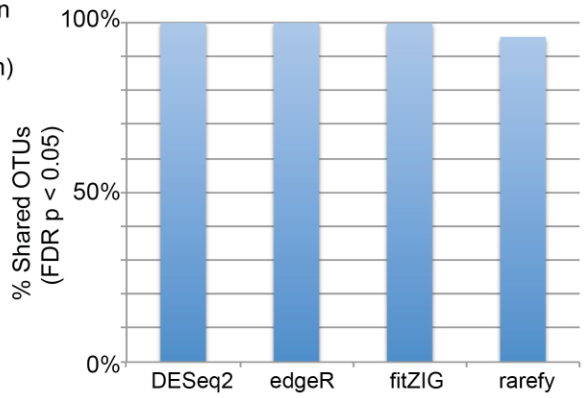


Figure 3-11 Comparison of the most promising differential abundance detection techniques on real datasets. Each table's diagonal represents the number of OTUs found significant (Benjamini & Hochberg FDR < 0.05) by that technique. The off-diagonal entries represent the number of shared differentially abundant OTUs between two techniques. The bar charts represents the percentage of differentially abundant OTUs shared by at least one other technique.

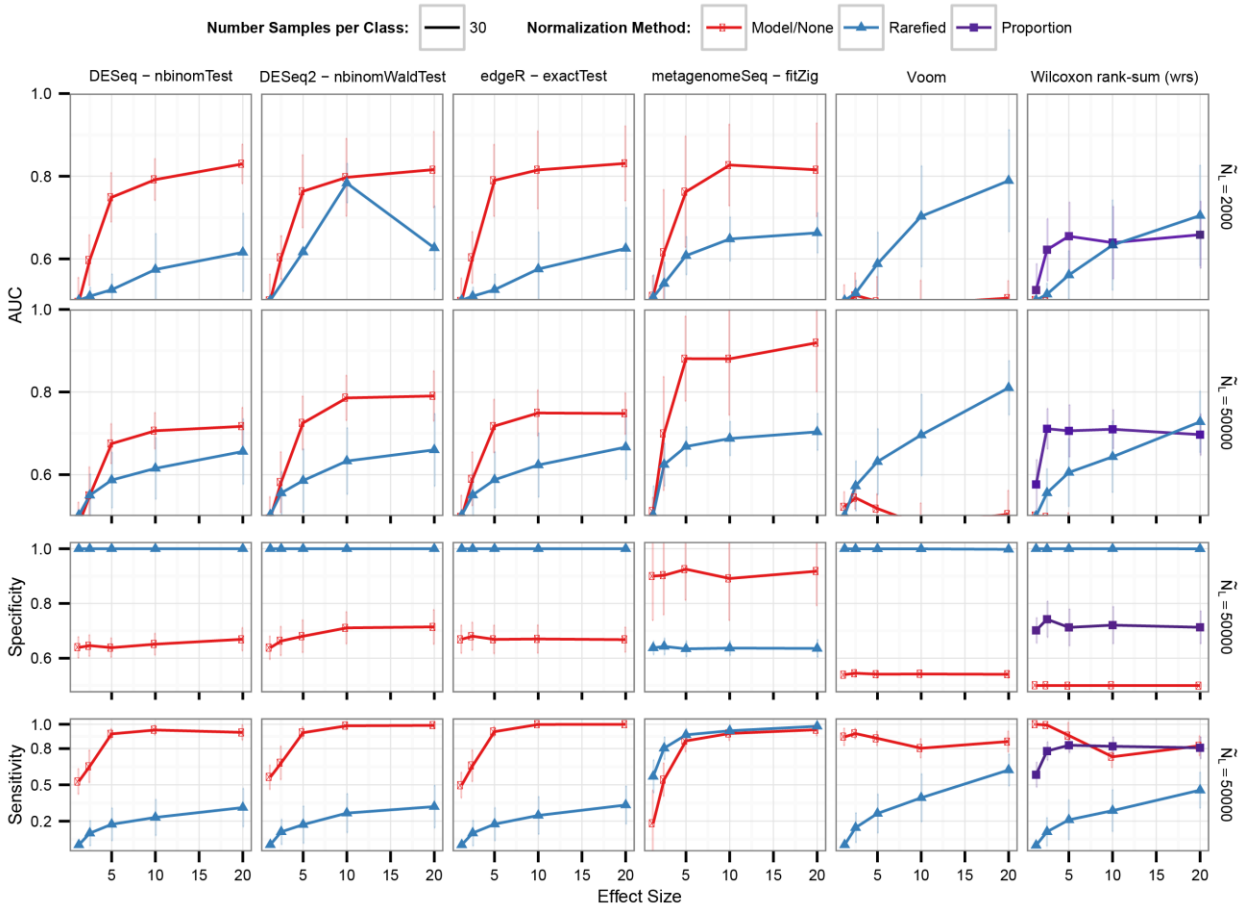


Figure 3-12 Differential abundance detection performance where one sample type library is 10 times the size of the other. Labels are the same as in Fig. 3-9. A significant difference from the results of Fig. 3-9 was first observed at 2-3-fold difference in library sizes.

negatives *vs.* false positives. We often place higher importance in reducing false positives, but this will vary depending on experimental design. For example, study designs in which community analysis is used as a pre-screening, and significant changes will be confirmed in high-throughput follow-up experiments may allow greater tolerance of false positives.

However, while both fitZIG or rarefying followed by Wilcoxon rank sum tests in isolation may be applicable for detecting differential abundance in particular situations, our results caution that fitZIG should not be used on rarified data (Figure 3-10), as this combination of methods caused extremely high false positive rates.

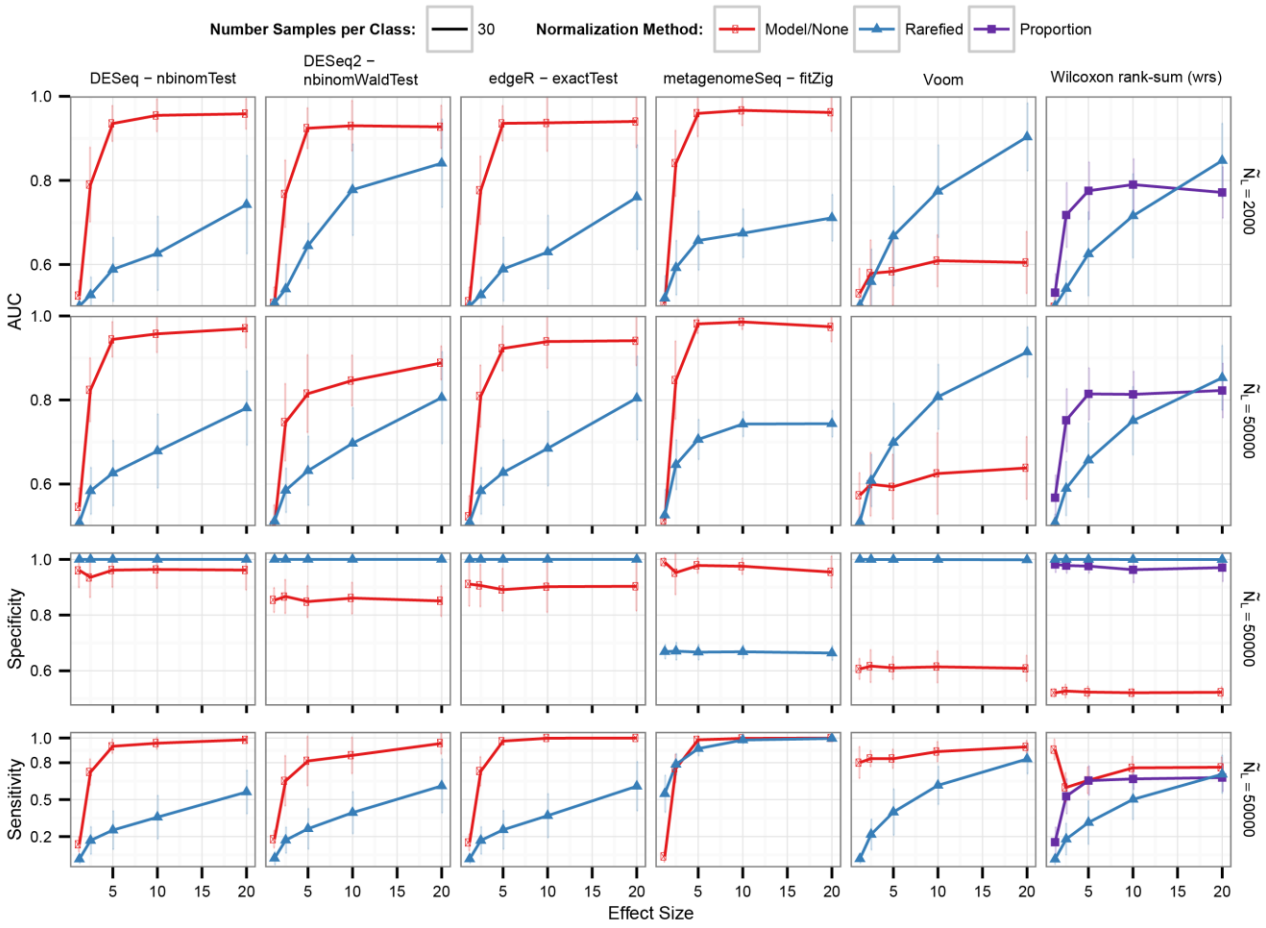


Figure 3-13 Differential abundance detection performance where one sample type library is 3 times the size of the other. Labels are the same as in Fig. 3-9.

While the no-normalization or proportion approaches perform adequately where the average library size is approximately the same between the two groups (Figure 3-9), they do not when one library is 10x larger than the other (Figure 3-12). Therefore, we reiterate that neither the no-normalization nor the naive proportion approach should be used for most statistical analyses. To demonstrate this, we suggest the theoretical example of a data matrix with half the samples derived from diseased patients and half from healthy patients. If the samples from the healthy patients have a 10x larger library size, OTUs of all mean abundance levels will be found to be differentially abundant simply because they may have 10x the number of counts in the

healthy patient samples. (Such systematic bias can happen if, for example, healthy vs. diseased patients are sequenced on separate sequencing runs or are being compared in a meta-analysis). The same warning applies for naive proportions, especially for rare OTUs that could be deemed differentially abundant simply due to differences in sequencing depth. This is seen even with some filtering to remove very rare OTUs (Figure 3-12). We first observed a transition from the results of Figure 3-12 to Figure 3-13 at around 2-3x difference in library sizes (Figure 3-13). Further, we investigated uneven numbers of samples per class, with not much difference in results from Figure 3-9.

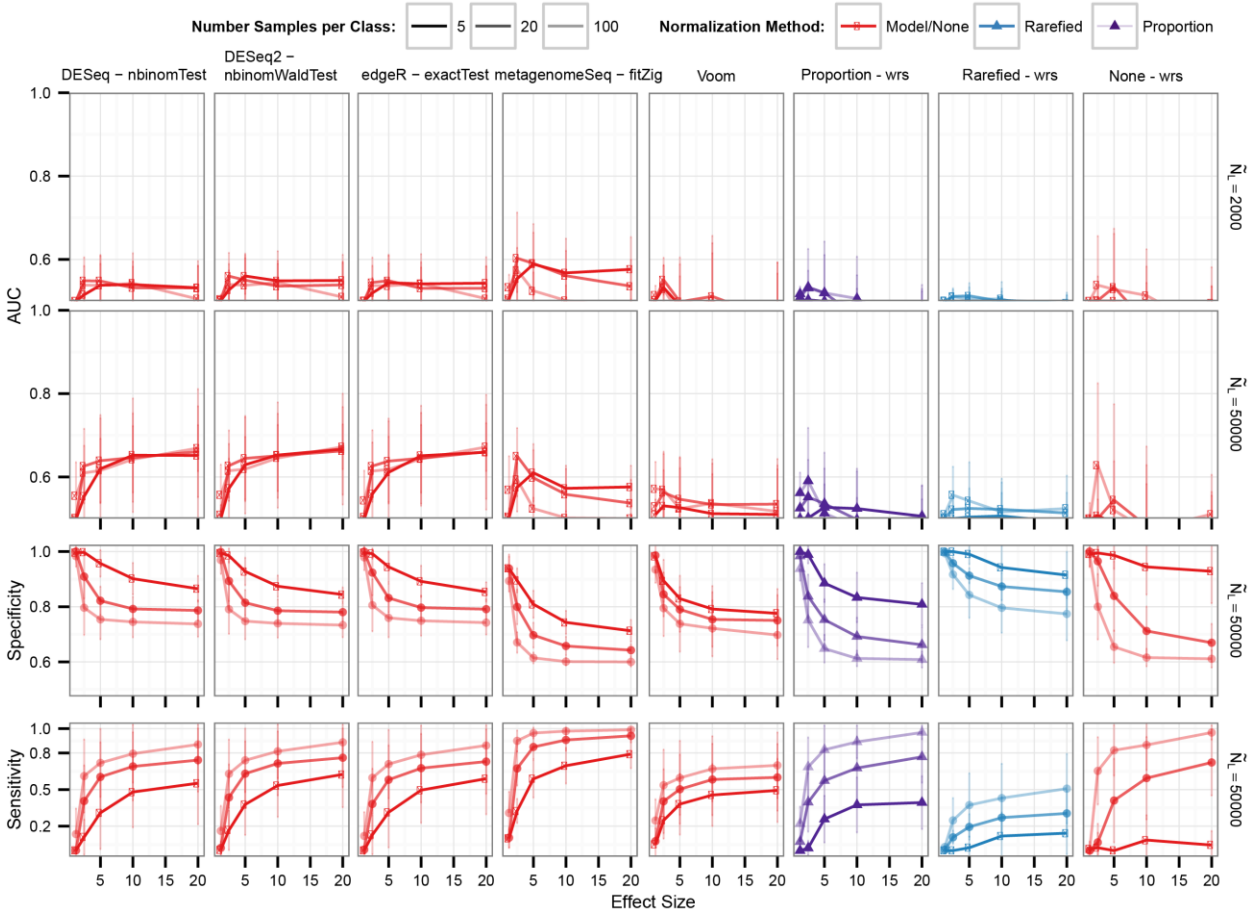


Figure 3-14 Differential abundance detection performance when the dataset is compositional. 25% of OTUs are differentially abundant. Labels the same as in Fig. 3-9.

While our previous simulations did not have compositionality, we next evaluated the performance of the techniques with a compositional OTU table (see Methods, Fig. 3-8B). In simulations where the abundances of 25% of the OTUs increased in one group, no method does well in terms of false positive rate (Fig. S8). Proportion normalization again performs poorly in the face of compositionality, which is present in all realistic datasets. For DESeq/DESeq2, poor performance may be due to the model's assumption that differentially abundant OTUs are not a large portion of the population[81], or the model's overdispersion estimates[28]. Thus, compositionality is still a large unsolved problem in differential abundance testing[27], and we would urge caution in data sets where compositionality may play a large role, e.g. when the alpha diversity of the samples is low [24].

3.4 CONCLUSIONS

More testing of the approaches on experimental data is necessary. Of methods for normalizing microbial data for ordination analysis, we found that DESeq normalization [89, 91], which was developed for RNA-Seq data and makes use of a log-like transformation, does not work well with ecologically useful metrics, except weighted UniFrac [111]. In contrast, MetagenomeSeq's CSS normalization [28] was developed for microbial data and does not result in troublesome negative output values. However, with techniques other than rarefying, library size can be a confounding factor with very low library sizes (under approximately 1000 sequences per sample), or if presence/absence metrics like unweighted UniFrac are used [110]. Extremely low-depth samples should be removed regardless of normalization technique, especially if it is suspected that they contain a higher proportion of contaminants [54, 112]. Also, when using alternatives to rarefying, the researcher must check that clustering by sequence depth does not obscure biologically meaningful results. Therefore, rarefying is still an extremely useful

normalization technique, especially for presence/absence metrics. Rarefying can erase the artifact of sample library size better than other normalization techniques, and results in a higher PERMANOVA effect size (R^2) for the studied biological effect, especially for small (<1000 sequences per sample), and uneven library sizes between groups. For both normalization and differential abundance testing, we stress that no normalization and naive proportion approaches should not be used as they can generate artifactual clusters based on sequencing depth, and may result in mistaken OTU differential abundance significance or insignificance.

For differential abundance testing, we studied the methods using both simulations and real data. The most promising of current techniques are based on GLMs with either the negative binomial or zero-inflated Gaussian distributions. It appears that DESeq2 [91], metagenomeSeq's fitZIG [28], and rarefying are all acceptable techniques for approximately even library sizes and numbers of samples per class. DESeq2 was designed for, and is a good option for, increased sensitivity on smaller datasets; however computation time becomes very slow for larger datasets, especially over 100 samples per category. MetagenomeSeq's fitZIG is a faster option for larger library sizes, although it may have a higher false positive rate. The fitZIG technique is designed for larger sample sizes, since more counts per OTU enables more accurate approximation of a continuous distribution. Rarefying, paired with traditional non-parametric tests to account for the non-normal distribution of microbial data, is useful for all dataset sizes, with sensitivity approaching parametric models in larger datasets. Rarefying yields fewer OTUs as significantly differentially abundant, but those OTUs are robust, in the sense that they are almost always identified as significant by at least one other differential abundance detection model. In the case of highly uneven library sizes per category (greater than 2-3x library size difference), we recommend rarefying, which provides higher specificity at a cost to sensitivity, or

metagenomeSeq's fitZIG, giving higher sensitivity at a cost to specificity, over the DESeq2 technique. In situations with highly compositional data, no technique does well.

Prior to differential abundance analysis, we recommend checking for significant differences in library size means and distribution between categories (e.g. healthy vs. sick); and propose a Mann-Whitney test, although the subject could be investigated further. The Mann-Whitney test works on the library sizes simulated for this study, as well as that of McMurdie and Holmes [30]. To check distributional differences, the library sizes of one sample category can be multiplied by a factor (e.g. 2) to make the means comparable prior to applying the Mann-Whitney test. If there is a significant difference in either mean or distribution, we recommend rarefying paired with a non-parametric test; if not, alternatives to rarefying may be used. For the parametric differential abundance techniques, it is recommended that rare OTUs be filtered out of the matrix prior to differential abundance testing. However, we advise OTU filtering after rarefying, and then applying non-parametric tests. Thanks to McMurdie and Holmes' previous work in this area [30], we recognize the potential of these newer techniques, and have incorporated DESeq2 [91] and metagenomeSeq [28] normalization and differential abundance testing into QIIME version 1.9.0 [29], along with the traditional rarefying and non-parametric testing techniques.

3.5 ACKNOWLEDGEMENTS

S.J.W. was funded by the National Human Genome Research Institute Grant# 3 R01 HG004872-03S2, and the National Institute of Health Grant# 5 U01 HG004866-04. Research of S.D.P. was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences, NIH (Z01 ES101744-04). We also thank Joseph N. Paulson, Joey McMurdie, and Jonathan Friedman for helpful conversations, and Greg Caporaso and Jai Ram Rideout for coding advice.

Chapter 4

Carcass Mass has Little Influence on the Structure of Gravesoil Microbial Communities

as published in *the International Journal of Legal Medicine,* 30 May, 2015

4.1 INTRODUCTION

Postmortem microbial communities are crucial and dynamic contributors to corpse decomposition. The activity of these decomposer microorganisms drives many postmortem changes, such as bloating [125] and ethanol production [126]. The structure of these microbial communities changes as a corpse decomposes because available nutrients are consumed [13]. Postmortem microbial communities have received much interest lately because they change in a predictable way, particularly the microorganisms on the skin [13, 127] and in carcass-associated soils [13]. These developmental shifts are analogous to those associated with insects [128] and have great potential to be developed as a means to estimate postmortem interval [13, 127]. We are particularly interested in utilizing soil microbial communities associated with decomposition, also known as gravesoils, because they host a clock-like succession of microbes [13] and are easily accessible at crime-scenes in outdoor scenarios.

The development of soil microorganisms as physical evidence requires us to answer several fundamental questions about the relationships between corpses, decomposition, and soil microbial communities. It is known that microbial activity in gravesoils increases rapidly and significantly during the early stages of decomposition [129, 130] and that this activity is influenced by several variables including soil texture, temperature, moisture, vegetation, and pH [130-134]. Microbial gravesoil activity is primarily driven by bacteria during the early stages of

decomposition [13, 135], followed by increased activity of eukaryotes such as nematodes [13] and fungi[136, 137] during later stages of decomposition. Yet one variable has received little experimental attention: the mass of the corpse.

Corpse mass is an important variable to understand because it can affect the rate of decomposition. However, this relationship is still under investigation. Many studies have utilized swine carcasses since the decomposition rate and arthropod colonization in *Sus scrofa domesticus* corpses mimics that in humans [138-140]. One of the first studies, which was not replicated, reported that larger mass corpses decay faster than smaller mass corpses [141]. Most of the later studies concluded that smaller corpse masses decay faster [142-146]. The exact functional nature of the decay rate in these studies however is not fully agreed upon [143-146]. Also, the effect of corpse mass and decay rate on the host-associated invertebrate and microbial community is not well understood. Hewadikaram and Goff found that corpse mass did not affect arthropod taxa composition or its succession over time [141]. Simmons et al. [143] concluded that corpses of different masses only decayed at different rates if insects were present. Finally, only a few small studies with tiny carcass masses have investigated the relationship between carrion carcass mass and insect types [147, 148].

To our knowledge, no studies have yet investigated the effect of carcass mass on the associated microbial communities. In this paper, we focus on the samples of Spicka et al. [144], which is a statistically well designed study of swine decomposition during a Nebraska summer using four different mass carcasses in triplicate. Spicka et al. [144] observed that larger swine carcasses (20 kg – 50 kg) released a greater concentration of ninhydrin-reactive nitrogen into gravesoil than neonatal (1 kg) carcasses. An additional mass effect was observed where the largest carcasses (40 kg – 50 kg) released a significant amount of total nitrogen more rapidly

than 20 kg carcasses. This release of nutrients, along with the recent observation that soil microorganisms contribute directly to the breakdown of carcass materials [135], leads to our hypothesis that carcass mass will influence the structure of associated soil microbial communities.

To investigate the effect of carcass mass on the structure of postmortem microbial communities in gravesoil, we sequenced the archaeal, bacterial, and eukaryotic microbial communities of soils collected by Spicka et al. [144]. We used the universal and taxonomically-informative 16S rRNA gene and 18S rRNA gene to analyze the structure of the microbial communities associated with the control soil and with carcasses of mass 1 kg, 20 kg, 40 kg, and 50 kg.

4.2 MATERIALS AND METHODS

4.2.1 Carcasses and Decomposition Site

Swine (*Sus scrofa domesticus*) carcasses of different masses (~1 kg, 20 kg, 40 kg, and 50 kg) were killed by blunt force trauma to the skull with a bolt gun, and placed on a weighing frame (2.5 cm^2 polypropylene mesh bound to a 85 cm x 40 cm PVC frame: Figure 4-1) directly on the surface of a grassland soil near Mead, Nebraska, USA in the summer within 60 minutes of death [144]. The grassland soil was a deep, silty, clay loam with a texture of 15.1% sand, 53.6% silt, and 31.3% clay. The soil surface of the decomposition site was flat so that decomposition fluids released from a carcass would collect around the carcass, but was not influenced by slope. Coyotes (*Canis latrans*) and turkey vultures (*Cathartes aura*) were the primary scavengers in the area, however no scavenger activity was observed at this site for five years [144]. Insect activity was not restricted in the current experiment.

4.2.2 Soil Collection and Storage

Gravesoils and control soils (soils not associated with carcasses) were collected as

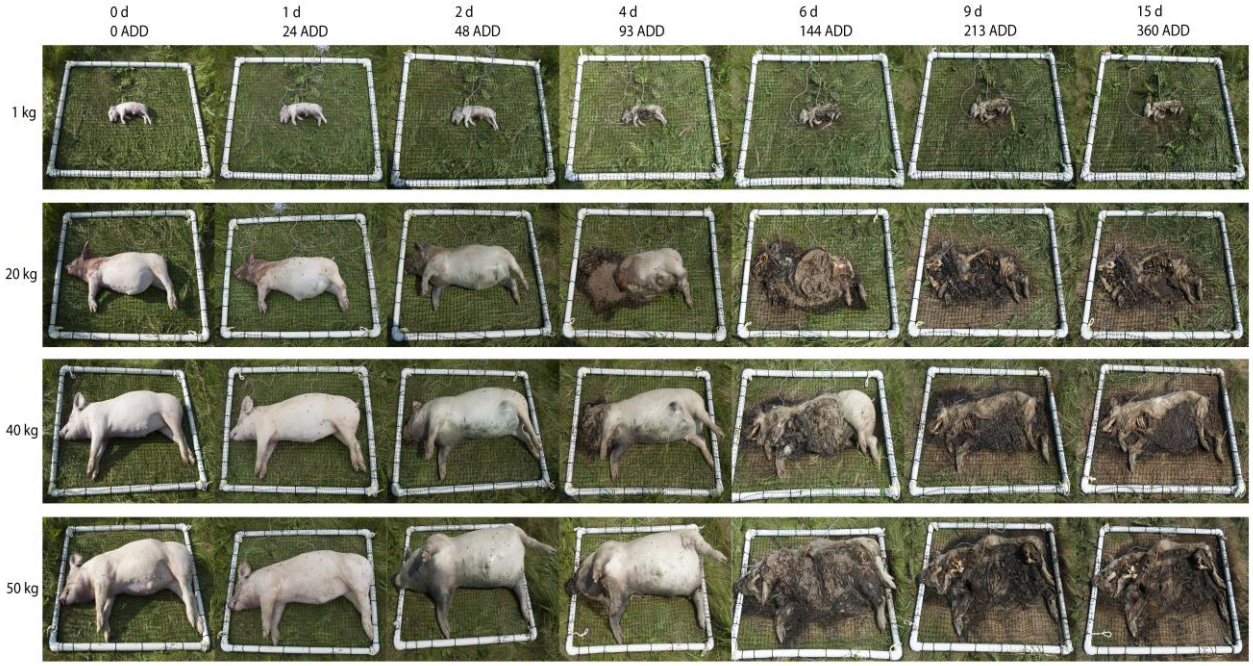


Figure 4-1 The gross decomposition of swine (*Sus scrofa domesticus*) carcasses of contrasting mass (~1 kg, 20 kg, 40 kg, and 50 kg) on the soil surface of a pasture near Mead, Nebraska where postmortem interval was measured as days (d) and Accumulated Degree Days (ADD).

described in Spicka et al. [144]. Gravesoil and control plots were at least 5 m apart. Soil samples were collected from underneath each carcass (0 cm - 5 cm depth) while it was lifted to measure mass loss. Soils were collected from an unsampled location each time using a 2.54 cm diameter KHS soil probe (M&M Supply Company, Clear Lake, Iowa, USA). Probe surfaces were cleaned with ethanol between each sample collection. There was no need to clear plant detritus from the soil surface before each sampling, as it was sparse. Soil samples were collected from the initial time of placement and at 24-hour intervals for 1, 2, 4-6, 9 and 15 days postmortem. Day 3 and day 8 were skipped due to severe thunderstorms. Three carcasses of each weight were placed at once, resulting in a total of 12 carcasses. Daily temperature ranged from 13.7 °C to 32.9 °C. Accumulated Degree Days (ADDs) were calculated as in Arnold [149] using a base temperature of 0 °C [150]. All soils were stored at -20 °C until DNA extraction.

4.2.3 Carcass Decomposition

The mass loss of carcasses followed a sigmoidal curve [144] typically associated with the breakdown of carrion [151]. Adult flies were observed on all carcasses within seconds of placement and larval masses were established on all replicates. Peak volume of larval mass was apparently a function of carcass mass; it took more time for larger carcasses to support peak maggot volume (Figure 4-1). However, the majority of migration was completed by 9 days postmortem (144 ADD) in all replicates. These carcasses did not tend to undergo an abdominal rupture that is often observed with carrion. Rather, fly larvae feeding from the head toward the posterior end typically consumed the carcasses in the current study (Figure 4-1).

4.2.4 Microbiome Analysis

DNA extraction, PCR amplification were conducted as described in Metcalf et al. [13] and following Earth Microbiome Project standard protocols (http://www.earthmicrobiome.org). Archaeal and bacterial 16S rRNA gene amplicons were sequenced using the Illumina HiSeq 2000 (100 basepair reads) and microbial eukaryotic 18S rRNA amplicons were sequenced using the Illumina MiSeq (150 basepair reads). Sequence processing and data analyses were conducted as described in Metcalf et al. [13], except that updated taxonomy databases were used, specifically Greengenes version 13_5 (http://greengenes.secondgenome.com, [152]) for open-reference OTU picking of 16S rRNA sequences, and SILVA version 111 [153] for closed-reference OTU picking of 18S rRNA sequences. Additionally, primer and adapters were removed from the end of the 18S read, resulting in read lengths of approximately 120 basepairs.

For 16S sequences, taxa that were not classified in the Domains Bacteria or Archaea were removed. For 18S sequences, we focused on the microbial community by filtering out taxa classified in groups Craniata, Chloroplastida, Mollusca, and Arthropoda. After these filtering

steps, our 16S and 18S data sets included 9,953,274 sequence reads (mean 82,943 reads per sample) and 567,129 (mean 4,975 reads per sample), respectively. The average number of reads per sample was substantially lower for the 18S data set because of the lower depth of sequencing on the MiSeq platform and because some samples contained a high relative abundance of chloroplast, insect, and host DNA reads, which were filtered out. We rarified the 16S data set to 14,000 sequences per sample and the 18S data set to 430 sequences per sample, which allowed us to include most samples in analyses.

To confirm our rarefying results, and to maximize the statistical power of our data set, we also ran analyses using Cumulative Sum Scaling (CSS) as an alternate normalization technique to rarefying [28]. We only used weighted UniFrac [111] analysis on the CSS transformed data, as rarefying is a more appropriate technique for unweighted UniFrac [110]. Before analysis with CSS, we removed very low depth samples (below 940 and 850 sequences/sample for 16S and 18S datasets) and extreme outliers [65]. This is because low depth samples have a higher proportion of contaminants [112]. We also ensured that the CSS-transformed results did not display clustering based on sample sequencing depth.

Using the QIIME pipeline [29], we explored relative taxon abundances and patterns of community dissimilarity with phylogeny-based UniFrac unweighted and weighted distances. We report p-values and type I sequential sums-of-squares error (R^2) for the strength and statistical significance of sample grouping based on unweighted and weighted UniFrac distances using a non-parametric analysis of variance (PERMANOVA) statistical test [154] with the adonis function in the 'vegan' [155] statistical package for R [156]. We also report Bonferroni-corrected p-values for the distance boxplots using the nonparametric two-sided Student's t-test (999 permutations). Error bars are based on the standard deviation of the UniFrac distance distributions. We report

differentially abundant taxa using the Kruskal-Wallis statistical test.

4.3 RESULTS

The decomposition of a carcass had a significant effect on the structure of gravesoil microbial communities; all gravesoils were significantly (PERMANOVA $p < 0.001$) different compared to control soils by the end of the trial. However, we observed no sustained significant differences between soil microbial communities associated with carcasses of contrasting mass (Figure 4-2). The 1 kg mass was the most different of the masses, although not quite significant (Figure 4-2, Table 4-1). This non-significant finding was supported using an alternate normalization technique (Table 4-2A). Spicka et al. found significant differences in the amount of Ninhydrin-reactive Nitrogen (NRN) released by corpses of contrasting masses. Specifically, Spicka et al. found that the 1kg mass had a greater concentration of NRN per unit carcass (NRN_c) compared to other masses, and the 20kg mass also briefly had greater NRN_c compared to the 40kg and 50kg masses. When NRN differences were controlled for, carcass mass became even less significant (Table 4-3A). Although mass was not a significant factor in determining the 16S microbial community, time was (Table 4-1, Table 4-2A, Table 4-3A).

Archaeal and bacterial groups changed significantly in relative abundance during decomposition. For example, bacterial family "*Candidatus* Chthoniobacteraceae" dominated all soils during the early stages of decomposition but the abundance of these bacteria decreased as carcasses decomposed (Figure 4-3). The abundance of bacteria from taxa Gaiellaceae, Acidobacteria, and *Rhodoplanes* also decreased during decomposition (Bonferroni $p < 0.01$). This was also true for the archaeal taxa "*Candidatus* Nitrososphaera". However, several bacterial taxa significantly increased during decomposition, including those from taxa Planococcaceae, *Sporosarcina* sp., *Ignatzschineria* sp., and Chitinophagaceae (Bonferroni $p < 0.01$).

The dominant soil eukaryotes were fungi and nematodes (Figure 4-4). As observed in the

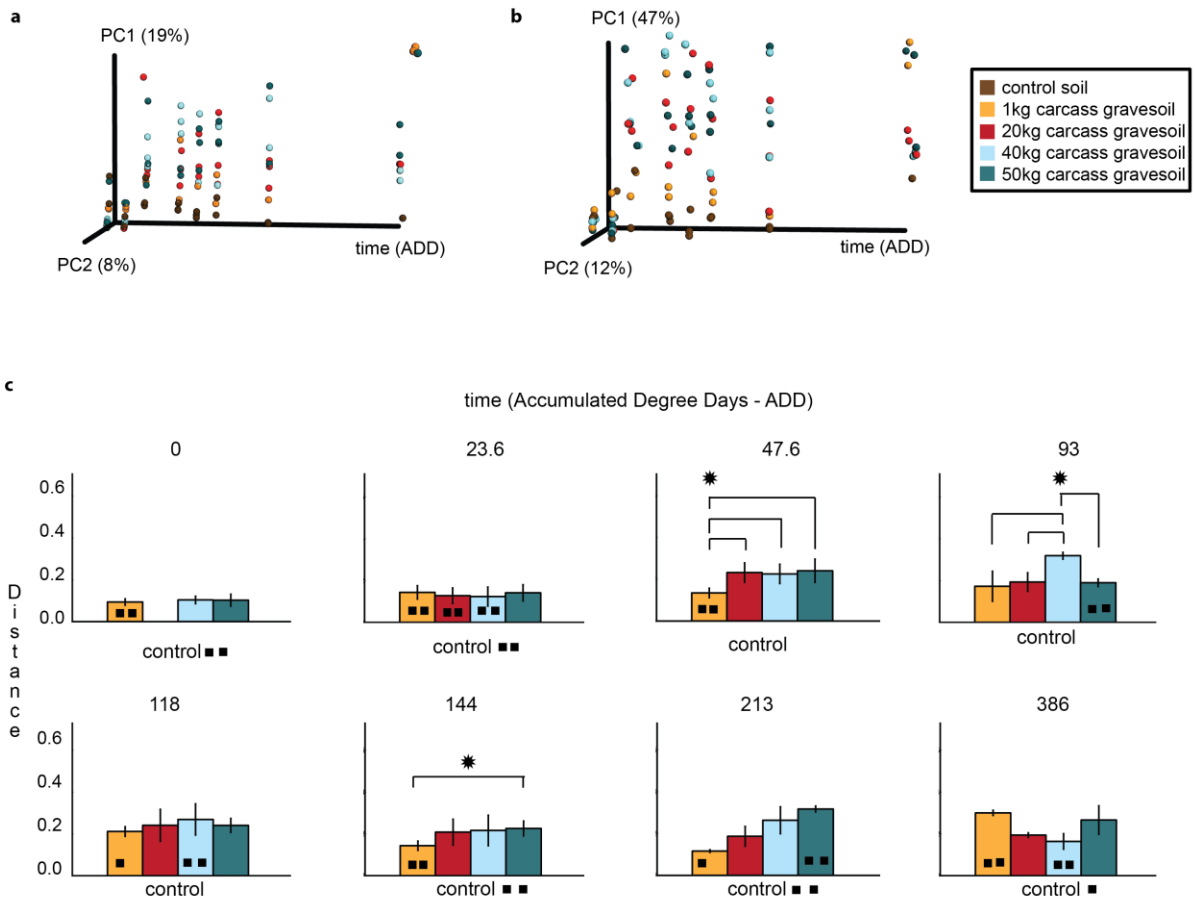


Figure 4-2 Ordination and bar plots to visualize differences between the structure of gravesoil microbial communities during the decomposition of swine (*Sus scrofa domesticus*) carcasses on the soil surface of a pasture near Mead, Nebraska, USA during the summer. **(a)** 16S rarefied unweighted UniFrac Principal Coordinates Analysis (PCoA), **(b)** 16S rarefied weighted UniFrac PCoA, and **(c)** 16S rarefied weighted UniFrac distance comparison bar plots for each mass compared to the control soil on each accumulated degree day (ADD). * indicates a significant nonparametric-t test difference with a Bonferroni-corrected $p < 0.05$. For example, at ADD 144, the weighted UniFrac distance from control soil to 50kg carcass gravesoil is significantly different compared to the distance from control soil to 1kg carcass gravesoil. We use the control soil as a baseline. In cases where less than three samples were in the analysis due to quality concerns, the number of squares (■) indicates how many samples were analyzed. Results for weighted and unweighted UniFrac analyses were nearly identical.

	time (ADD)		mass	
Comparison	**R^2**	**FDR p**	**R^2**	**FDR p**
1kg v. 20kg	0.086	0.01	0.053	0.084
1kg v. 40kg	0.1	0.006	0.042	0.28
1kg v. 50kg	0.16	0.006	0.039	0.3
20k v. 40kg	0.05	0.066	0.025	1
20kg v. 50kg	0.064	0.006	0.033	0.38
40kg v. 50kg	0.078	0.006	0.021	1
control v. all masses	0.052**		0.034*	

**p < 0.001 *p<0.01

Table 4-1 16S rarefied unweighted UniFrac Type 1 sequential sums of squares PERMANOVA. The model y ~ ADD_time + mass was fit to control for differences in the number of replicates at each time point before assessing the effect of carcass mass on gravesoil microbial communities. The FDR procedure is Bonferroni correction.

a

	time (ADD)		mass	
Comparison	**R^2**	**FDR p**	**R^2**	**FDR p**
1kg v. 20kg	0.089	0.012	0.043	0.31
1kg v. 40kg	0.061	0.14	0.061	0.23
1kg v. 50kg	0.13	0.006	0.055	0.13
20k v. 40kg	0.064	0.1	0.035	0.6
20kg v. 50kg	0.011	0.006	0.033	0.72
40kg v. 50kg	0.098	0.018	0.026	1
control v. all masses	0.063**		0.075**	

**p < 0.001

b

	time (ADD)		mass	
Comparison	**R^2**	**FDR p**	**R^2**	**FDR p**
1kg v. 20kg	0.12	0.006	0.044	0.26
1kg v. 40kg	0.13	0.006	0.035	0.66
1kg v. 50kg	0.15	0.006	0.094	0.012
20k v. 40kg	0.14	0.006	0.023	1
20kg v. 50kg	0.14	0.006	0.043	0.31
40kg v. 50kg	0.18	0.006	0.052	0.11
control v. all masses	0.093**		0.1**	

**p < 0.001

Table 4-2 CSS-normalized weighted UniFrac Type 1 sequential sums of squares PERMANOVA. The model y ~ ADD_time + mass was fit to control for differences in the number of replicates at each time point before assessing the effect of carcass mass on gravesoil microbial communities. The FDR procedure is Bonferroni correction. **(a)** 16S **(b)** 18S

a

	time (ADD)		NRN (µg g^{-1} soil)		mass (kg)	
Comparison	**R^2**	**FDR p**	**R^2**	**FDR p**	**R^2**	**FDR p**
1kg v. 20kg	0.086	0.012	0.083	0.006	0.039	0.31
1kg v. 40kg	0.10	0.006	0.063	0.030	0.020	1
1kg v. 50kg	0.16	0.006	0.046	0.21	0.024	1
20k v. 40kg	0.050	0.024	0.043	0.066	0.024	1
20kg v. 50kg	0.064	0.018	0.024	1	0.032	0.59
40kg v. 50kg	0.078	0.006	0.034	0.36	0.022	1
control v. all masses	0.052**		0.035**		0.023*	

**p < 0.001 *p<0.01

b

	time (ADD)		NRN (µg g^{-1} soil)		mass (kg)	
Comparison	**R^2**	**FDR p**	**R^2**	**FDR p**	**R^2**	**FDR p**
1kg v. 20kg	0.072	0.006	0.021	1	0.042	0.072
1kg v. 40kg	0.068	0.006	0.055	0.018	0.034	0.56
1kg v. 50kg	0.060	0.036	0.051	0.054	0.033	0.86
20k v. 40kg	0.095	0.006	0.033	0.44	0.016	1
20kg v. 50kg	0.087	0.006	0.017	1	0.031	0.74
40kg v. 50kg	0.099	0.006	0.024	1	0.029	1
control v. all masses	0.050**		0.029**		0.019*	

**p < 0.001 *p<0.01

Table 4-3 Rarefied unweighted UniFrac Type 1 sequential sums of squares PERMANOVA. The model y ~ ADD_time + NRN + mass was fit to control for differences in the number of replicates at each time point, and the amount of released NRN, before assessing the effect of carcass mass on gravesoil microbial communities. The FDR procedure is Bonferroni correction. **(a)** 16S **(b)** 18S

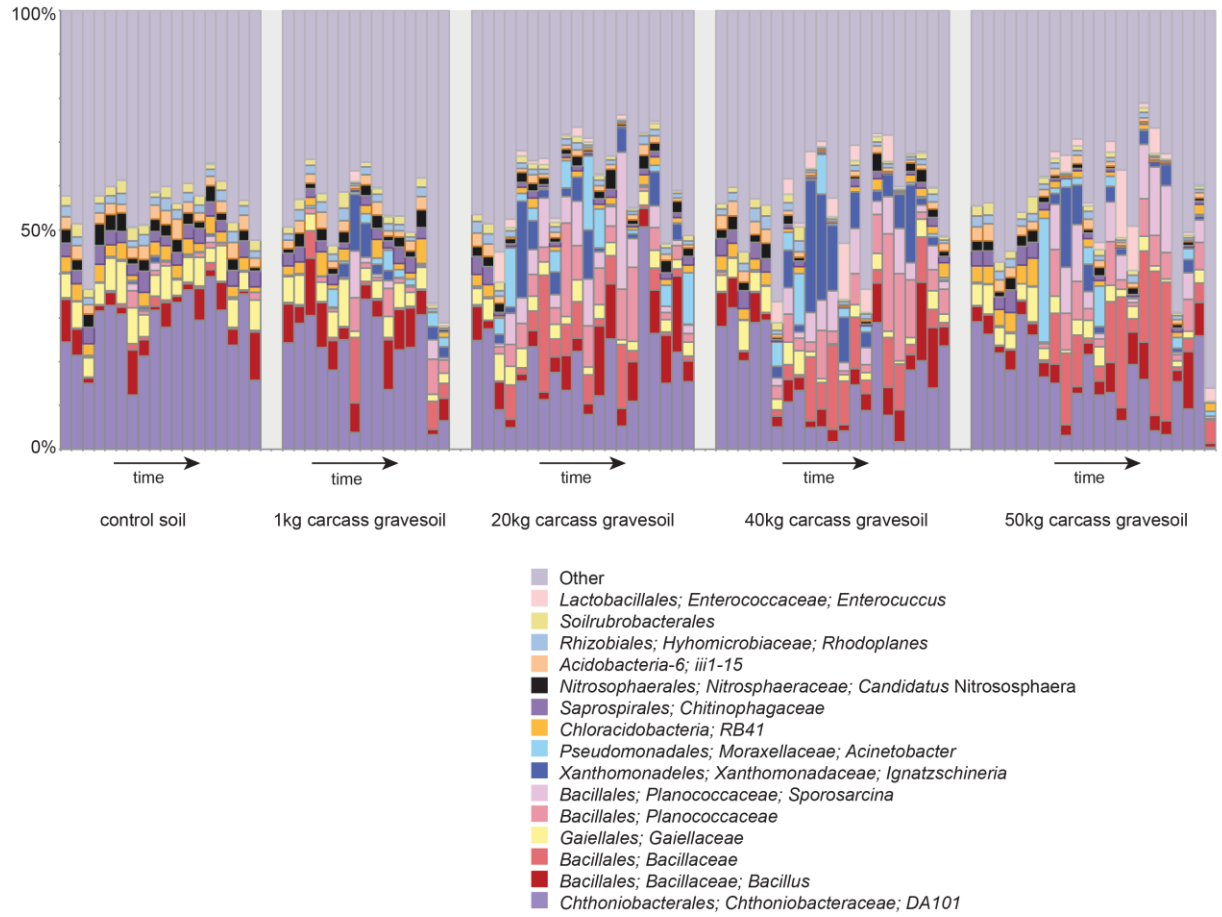


Figure 4-3 Sample relative abundance of control and gravesoil microbial (16S) communities during the decomposition of swine (*Sus scrofa domesticus*) carcasses on the soil surface of a pasture near Mead, Nebraska, USA during the summer. Only the 15 highest relative abundance taxa are shown, starting at the order level. Genus "*Candidatus* Nitrososphaera", is the only archaeal taxa of high abundance in the data set. Additional archaeal and bacterial taxa in each sample are combined into a single 'other' category. Archaea occupy approximately 0.06% of the 'other' category.

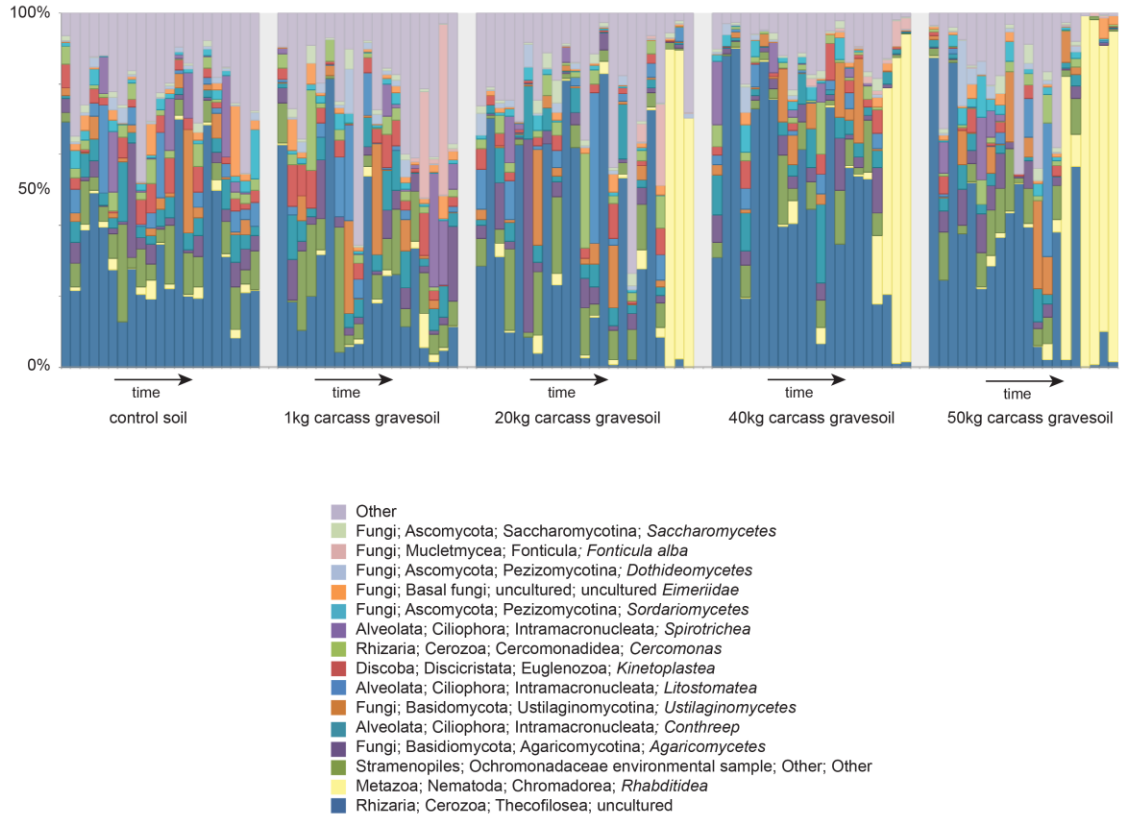


Figure 4-4 Sample relative abundance of control and gravesoil microbial (18S) communities during the decomposition of swine (*Sus scrofa domesticus*) carcasses. Only the 15 highest relative abundance taxa are shown starting at the class level, additional taxa in each sample are combined into a single 'other' category. The apparent lack of nematode bloom in the 1 kg samples at the later time points is because the later time point samples were filtered out due to quality concerns.

	time (ADD)		mass	
Comparison	**R^2**	**FDR p**	**R^2**	**FDR p**
1kg v. 20kg	0.072	0.006	0.041	0.066
1kg v. 40kg	0.068	0.006	0.055	0.012
1kg v. 50kg	0.06	0.048	0.06	0.018
20k v. 40kg	0.095	0.006	0.025	1
20kg v. 50kg	0.087	0.006	0.033	0.54
40kg v. 50kg	0.099	0.006	0.029	1
control v. all masses	0.049**		0.03**	

**$p < 0.001$

Table 4-4 18S rarefied unweighted UniFrac Type 1 sequential sums of squares PERMANOVA. The model y ~ ADD_time + mass was fit to control for differences in the number of replicates at each time point before assessing the effect of carcass mass on gravesoil microbial communities. The FDR procedure is Bonferroni correction.

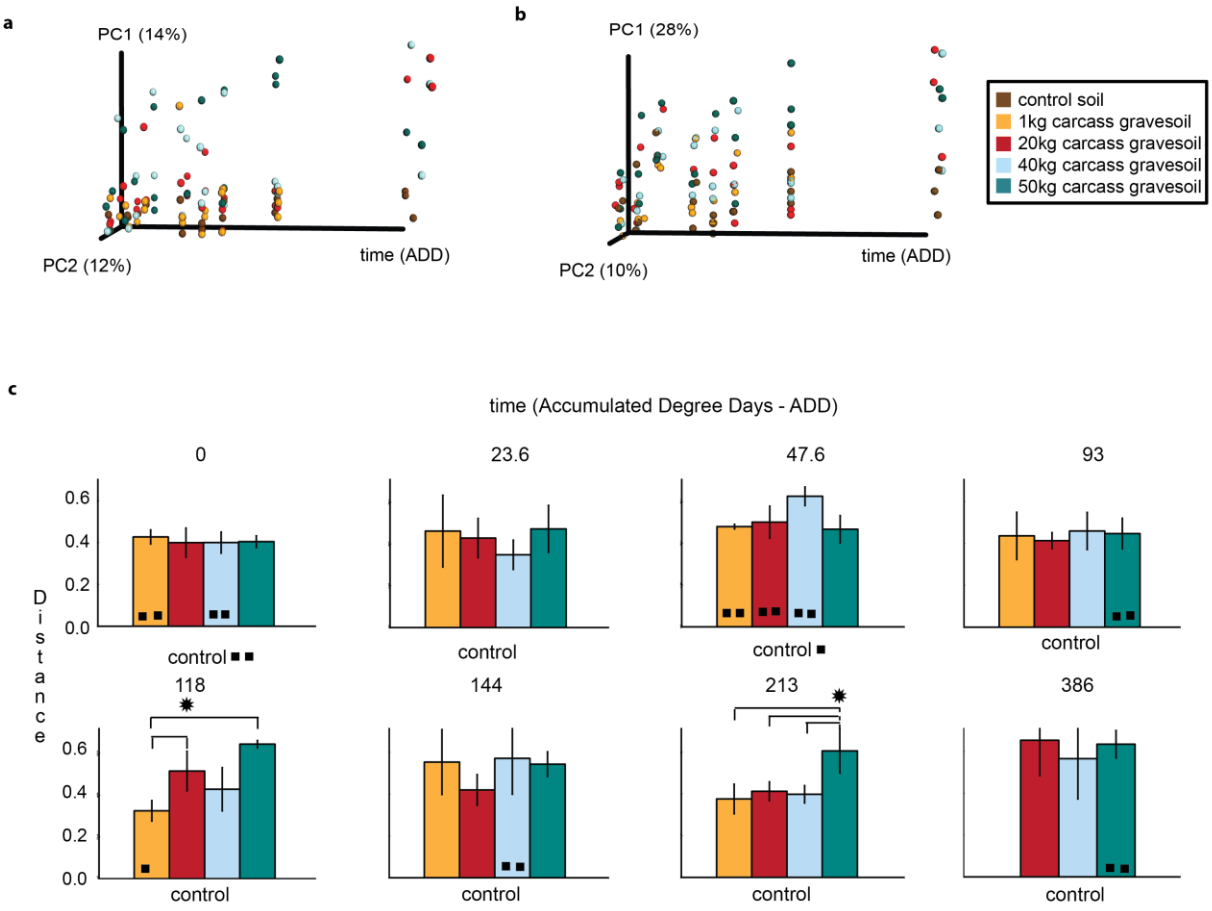


Figure 4-5 Ordination and bar plots to visualize differences between the structure of gravesoil microbial communities during the decomposition of swine (*Sus scrofa domesticus*) carcasses on the soil surface of a pasture near Mead, Nebraska, USA during the summer. **(a)** 18S rarefied unweighted UniFrac Principal Coordinates Analysis (PCoA), **(b)** 18S Cumulative Sum Scaling (CSS) weighted UniFrac PCoA , and **(c)** 18S CSS weighted UniFrac distance comparison bar plots for each mass compared to the control soil on each accumulated degree day (ADD). * indicates a significant nonparametric-t test difference with a Bonferroni-corrected $p < 0.05$. In cases where less than three samples were in the analysis due to quality concerns, the number of squares (■) indicates how many samples were retained. Results for weighted and unweighted UniFrac analyses were nearly identical.

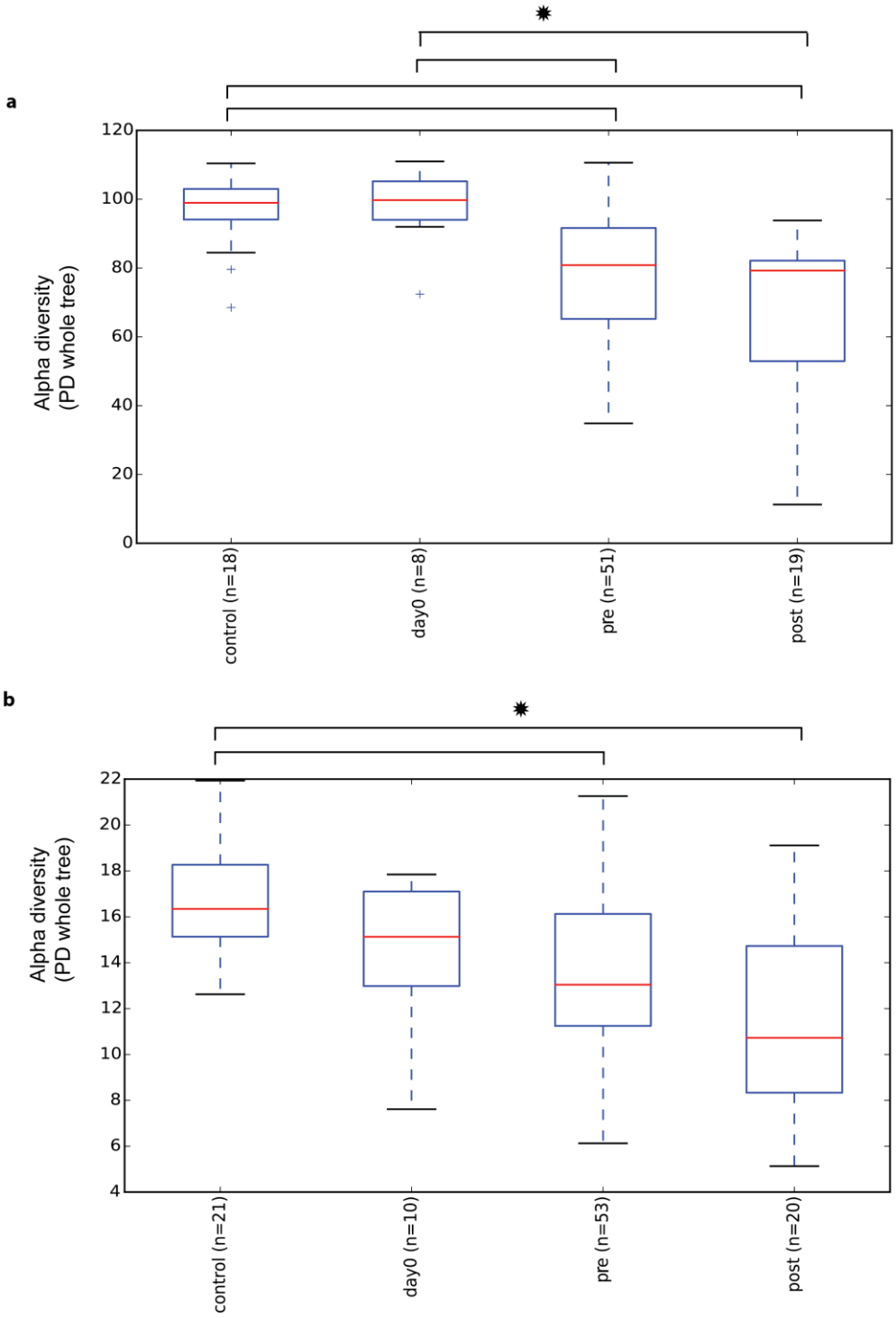


Figure 4-6 Phylogenetic distance (PD) alpha diversity boxplots. This includes control, day0, pre-carcass rupture, and post-carcass rupture samples for all gravesoil masses. * indicates significant differences between boxplots ($p < 0.05$). **(a)** 16S **(b)** 18S

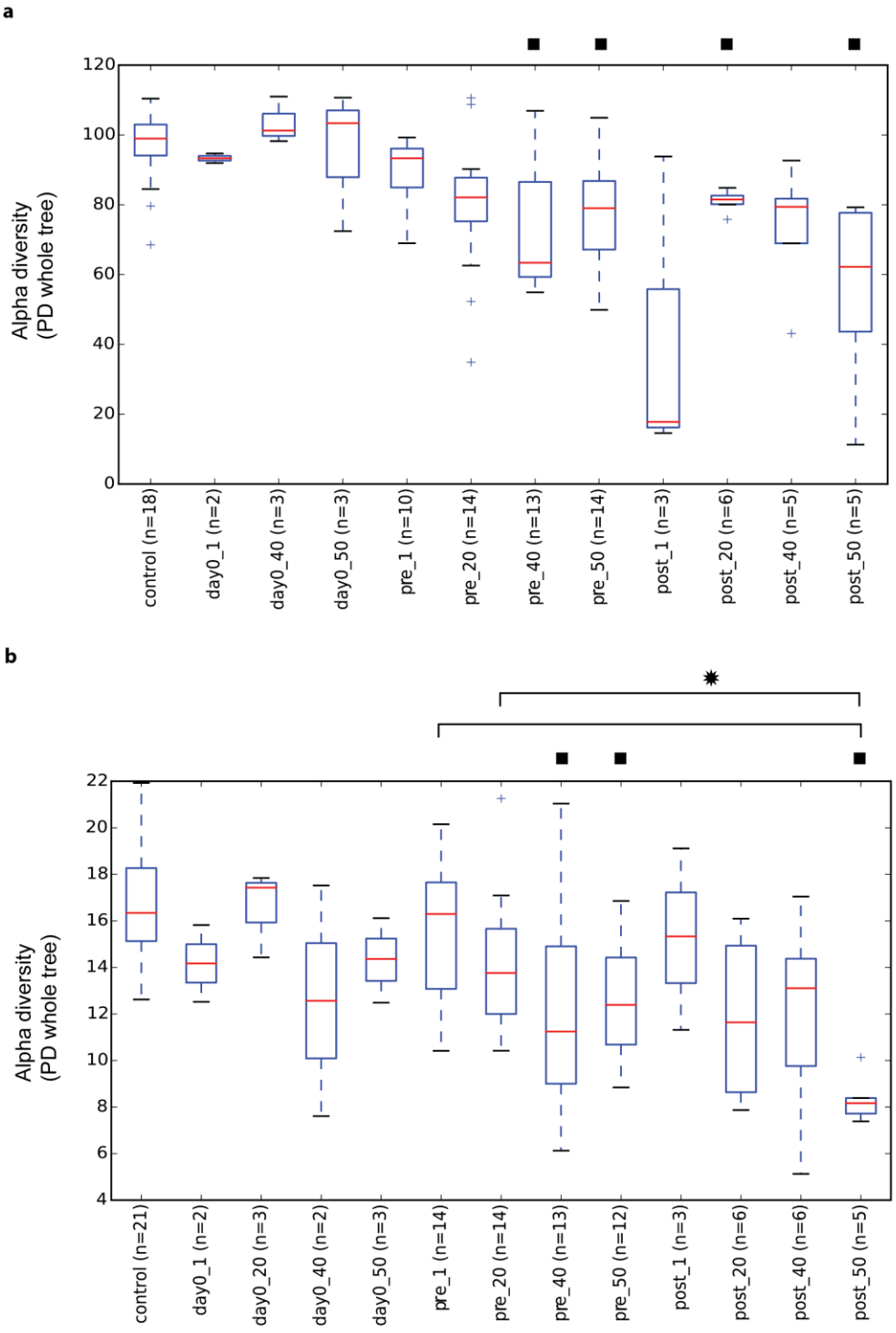


Figure 4-7 Phylogenetic distance (PD) alpha diversity boxplots. This includes day 0, pre-carcass rupture (days 1, 2, 4, 5 and 6), and post-carcass rupture (days 9 and 15) gravesoils for carcasses of masses 1 kg, 20 kg, 40 kg, and 50 kg. For example, label pre_1, signifies alpha diversity of pre-rupture 1kg carcass gravesoils. Squares (■) indicate siginificant differences from the control soils. * indicates possible ($p < 0.08$) weak significant difference between boxplots. Shannon diversity yielded similar results**. (a)** 16S **(b)** 18S.

bacterial communities the structure of these decomposer communities shifted significantly with time compared to the control soils (Table 4-4, Table 4-2B). Similar to archaeal and bacterial communities, the eukaryotic microbial communities associated with different carcass masses shifted similarly over time regardless of mass (Figure 4-5, Table 4-4). Again the 1 kg mass gravesoil was the most different of the masses, and was significantly different compared to the 50 kg mass (Table 4-4, Table 4-2B). However, this difference disappeared when NRN was controlled for (Table 4-3B). The top fifteen relative abundance gravesoil eukaryote communities comprised a large fraction of the total sequences, with the greatest shifts observed as increases in the abundance of nematodes in the family Rhabditidae and slime mold *Fonticula alba* (Bonferroni $p < 0.01$).

For both 16S microbes and 18S eukaryotes, we observed a significant decrease in the alpha diversity, or a measure of how many species are in each sample, of gravesoils during decomposition compared to control soils (Figure 4-6, Figure 4-7). Pre vs. post rupture differences in alpha diversity were present but harder to detect, especially when the masses were analyze separately (Figure 4-7), possibly due to only three replicates per mass. For all analyses, resolution of more subtle effects would require more replicates for increased statistical power.

4.4 DISCUSSION

Our data show that soil microbial communities associated with carcasses greatly differ from control soils during decomposition, but are robust to carcass mass. The 1 kg carcasses were marginally statistically significant compared to other masses, particularly the 50 kg mass. However, all other carcass masses (20 kg, 40 kg, and 50 kg) did not display significant differences in their microbial communities throughout decomposition. This finding suggests that microbial clocks to estimate the postmortem interval may be robust to human cadaver mass, at

least between 20 kg – 50 kg. When NRN was controlled for the 1kg mass moved far away from borderline statistically significant. While the pH of the soil increases during decomposition [13, 157], we observed it to have a smaller effect compared to NRN, as assessed by PERMANOVA R^2 (25% and 50% less on average in 16S and 18S data respectively).

The current findings are similar to those of other recent investigations into the postmortem microbiome. Other studies have identified the smallest mass carcasses as exhibiting the most variable decomposition patterns [144, 145]. Our results also agree that the major variables influencing the structure of the microbial communities is the death of the host [158], and the time since death [13, 127]. Postmortem microbial communities shift when a carcass is decomposing, probably due to rupture, increased resource availability, or the proliferation of insects. It has been show that the cadaver microbial community can influence insect activity and vice versa [159-162]. Added to this is likely an example of 'resource selects community' [163] where the different stages of decomposition offer different nutrients, e.g. pre and post-rupture. We observed a decrease in alpha diversity during decomposition, but the result was only strongly significant when all mass classes were combined (Figure 4-6, Figure 4-7). Metcalf and colleagues found a stronger decrease in alpha diversity of gravesoils during decomposition [13] possibly because their samples were collected for a longer time period and the sample size was larger providing better power.

A striking similarity between the current results and those reported by Metcalf et al. [13] was the increase in nematode abundance. The nematodes in this study were of the same family (Rhabditidae) as those reported by Metcalf et al [13]. This flush of nematodes is probably due to the increase in the abundance of the postmortem bacteria, their primary food source. Nematodes have long been used for environmental monitoring [164] and we find it very interesting that similar

nematode taxa have been observed with decomposing mice [13] and swine in two different soil types. We recommend that the forensic value of nematodes be explored in more detail.

To expand on this research, we also recommend more detailed study into the dynamics between carcass mass, decomposition, and microbial communities with more replicates and over a longer time period to confirm the apparent lack of relationship between carcass mass and the time required for a shift in microbial community structure. Also, additional time points would allow for the use of regression models to estimate PMI. Similarly, we recommend that the decomposition of corpses greater than 50 kg should be investigated in detail to determine if additional trends can be identified, i.e. are corpses greater than 50 kg associated with different gravesoil microbial communities? Also, further investigation would be ideally done on human rather than swine corpses; however, it is difficult to find human donors, and donors are usually older than crime scene victims. While *Sus scrofa* is accepted as a model system most similar to humans because they have similar decay rates, body mass, and skin structure among other factors [137], some differences have been found between the two, for example four times as much stearic acid in swine fat compared to human [165].

The current data contribute to postmortem microbiology, a branch of forensic medicine designed to serve as a useful adjunct to autopsy [166]. The identification of postmortem microorganisms can be used to confirm the presence of a suspected antemortem infection, identify an infection when the cause of death is unknown, and assess the efficacy of antibiotics in treating an infection [167]. Recently, we demonstrated that corpses host a large and diverse microbial community at death [13, 135, 158, 168, 169]. The structure of this microbial community shifts significantly and predictably as a corpse decomposes, and can become less diverse as it decomposes into an increasingly specialized habitat [13, 127, 168]. These are exciting developments

for forensic medicine because they likely foreshadow an expanded use of microorganisms as physical evidence. Indeed, we can envision the development of a postmortem microbiology to aid in establishing cause of death, associating people with objects and locations [170], and estimating postmortem interval [13, 127]. These would be significant developments toward the development of a comprehensive forensic microbiology. The current data add to this fundamental understanding by showing that postmortem microbial communities can be similar regardless of initial carcass mass, which has the potential to simplify initial postmortem analysis. However, we caution that more replicates, time points, and mass types should be investigated; and this experiment was done with swine, therefore results could differ with human cadavers.

4.5 ACKNOWLEDGEMENTS

This research was funded by the Office of Justice Programs National Institute of Justice Award# NIJ-2011-DN-BX-K533. S.W. was funded by the National Human Genome Research Institute Grant# 3 R01 HG004872-03S2, and the National Institute of Health Grant# 5 U01 HG004866-04. Research capacity at Chaminade University of Honolulu was supported by NIH-BRIC P20MD006084.

Chapter 5

Correlation Detection Strategies in Microbial Datasets Vary Widely in Sensitivity and Precision

Nature Methods, under review, 2015

5.1 INTRODUCTION

Microbes interact with their hosts and other microbes in the same microbial community, and these interactions have been implicated in numerous human health conditions including obesity and metabolic syndrome[1-4], cardiovascular disease[5], *C. difficile* colitis[6], inflammatory bowel diseases[7], and HIV[8]. Gut microbial communities are influenced by diet, culture, geography, age, and antibiotic use, among other factors[12], and are also very important in other systems, such as soils, lakes, and oceans[171-173]. An emerging approach to studying microbial communities is 'correlation networks'. Broadly, correlation networks have individual features (e.g. microbial taxa) as nodes and feature-feature pairs as edges. Edges may imply a biologically or biochemically meaningful relationship between taxa. For example, mutualistic microbes, those that benefit each other, should positively correlate across samples. In contrast, microbes with antagonistic relationships such as competing for the same niche should negatively correlate. In practice, microbes also may positively or negatively correlate for indirect reasons, based on their environmental preferences. For example, phylogenetically related microbes have a tendency to positively co-occur[174], perhaps simply because they grow on similar substrates. Recent studies suggest that the microbial relationships revealed by correlation interaction networks can be used to determine drivers in environmental ecology[173, 175, 176] or contribution to disease[172, 177-181]. Correlation networks are also powerful tools for hypothesis generation, such as determining

which interactions might be biologically relevant in their system and can be tested with additional techniques (e.g. through co-culturing to test mutualistic or inhibitory relationships, or whole-genome sequencing to identify genomic signatures of receptor coevolution).

Unfortunately, measuring correlation networks is computationally challenging. One challenge arises from the complexity of microbial communities: many microbial datasets have over 5,000 taxa. Since the number of possible two-taxon interactions for a dataset with n taxa is (n*(n-1))/2, this implies almost 12.5 million possible two-taxon correlations. Also, because microbes live in communities, there are likely 3-taxon interactions, 4-taxon interactions, and more. An additional challenge is that microbial sequence data provide relative abundances based on a fixed total number of sequences, not absolute abundances, introducing the problem of compositionality[24, 90]. Sparsity of features and missing data due to incomplete sampling further complicate statistical analysis[24, 182]. Finally, microbes may display diverse types of relationships, including linear, exponential, or periodic interactions, and most tests are insufficiently general to detect them all; even those that do are unlikely to detect different relationships with the same efficiency[182].

Many approaches for computing these correlation networks have been proposed recently. In theory, any method that quantifies relationships between taxa can be used: for example, standard metrics like Bray-Curtis[183], which measures abundance similarity; the Pearson correlation coefficient, which assesses linear relationships; and the Spearman correlation coefficient, which measures rank relationships are all potentially applicable[25, 26, 184, 185]. Additionally, software programs have been developed and optimized specifically to correct for certain issues with correlation analysis of natural populations. For example, CoNet[179] acknowledges that various techniques have different strengths and weaknesses and/or are

designed to optimally detect different functional relationships, and thus uses an ensemble method with the ReBoot procedure for p-value computation to combine information from several different standard comparison metrics. Local Similarity Analysis (LSA)[171, 173, 175, 186] is optimized to detect non-linear, time-sensitive relationships and can be used to build correlation networks from time-series data. The Maximal Information Coefficient (MIC)[182] is a non-parametric method designed to capture a wide range of associations without limitation to specific function types (such as linear or exponential) and to give similar scores to equally noisy relationships of different types. MENA/RMT[176, 187] adapts Random Matrix Theory from physics to microbiome data, and attempts to be robust to noise and to arbitrary significance thresholds. Finally, SparCC[24] is particularly designed to deal with compositionality in relative abundance data, since it is based on Aitchison's log-ratio analysis.[23]

The performance and limitations of most of these computational methods for inferring correlation networks have not been comparatively evaluated using either real or theoretical datasets, leaving researchers to guess at important properties of their networks such as sensitivity, specificity, precision, and—most importantly—ability to provide interpretable results. Counts of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), and calculations of sensitivity (true positive rate – TP/(TP+FN), specificity (true negative rate – TN/(FP+TN), and precision (TP/(TP+FP)) are among standard benchmark measures. Without an understanding of these important properties, correlation analysis risks diverting attention from meaningful interactions and leading to wasteful pursuit of expensive *in vitro* or *in vivo* validation experiments.

One previous effort in this area tested mainly basic correlation measures for one type of model system[188]. Here, in contrast, we tested the ability of each of these widely used correlation

measures and tools to detect a variety of dependent relationships in both simulated and real microbial datasets. Figure 5-1A outlines the general workflow. Supplementary Table 1 and the Methods section detail how mock data was generated, and all code, test-code, and documentation is available at https://github.com/wdwvt1/correlations. In brief, our simulations comprised 91 different data tables (columns in microbiome data typically represent samples, while microbes/features represent rows) with the number of microbes per table ranging from 200 to 10,000, and generated from eight different sample data generation models: distribution/copula[189], experimental, normalization, feature filtering, null/random, linear and non-linear (Lotka-Volterra) ecological[190], and time series. Within some models, we also introduced the aforementioned compositionality and sparsity challenges.

5.2 RESULTS

5.2.1 Tools Infer Significantly Different Numbers of Edges in Most Datasets

Different tools consistently produce different numbers and types of significant edges for the same data (Figure 5-1B). Tools also generally differ in which edges they detect: on average a pair of tools detects only 31.5% of the same edges across all data sets/models tested (Figure 5-2). This discordance further underscores the need for benchmarking, and suggests that the techniques may have differing strengths and weaknesses in response to the diverse challenges presented by microbiome data.

5.2.2 Different Underlying Distributions Significantly Alter Edge Inferences

We tested the direct impact of different sequencing technologies on OTU distributions in similarly processed sample replicates. Using technical replicates from an arthropod microbiome[191], and Illumina HiSeq vs. MiSeq sample replicates from a gut microbiome[117], we

tested each HiSeq feature compared to the same MiSeq feature using the Kolmogorov-Smirnov (KS) test[192, 193] and found no significant differences in feature count distributions (indicating that

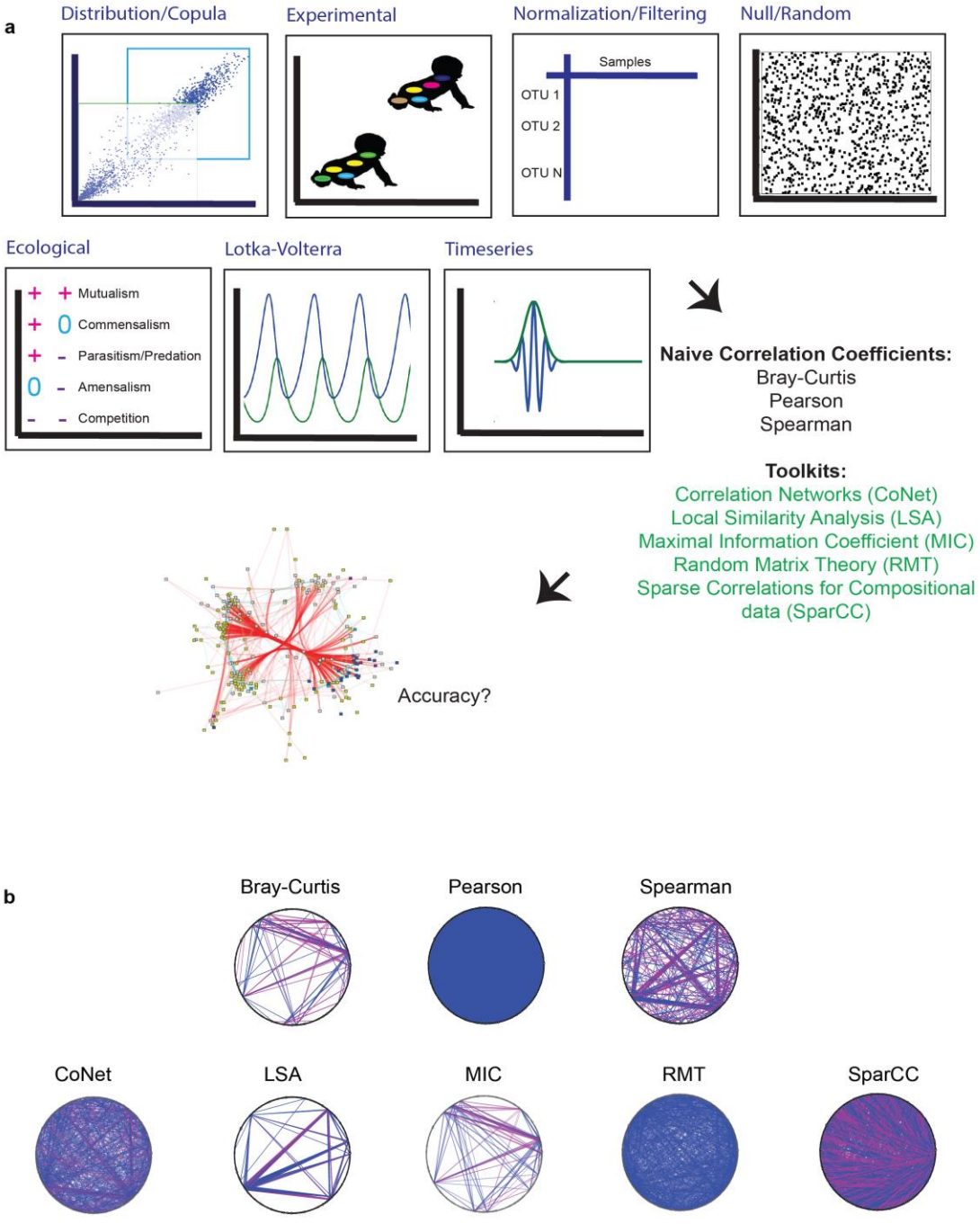


Figure 5-1 Overview and motivation of correlation network technique benchmarking. **(a)** Mathematical properties of microbial communities naturally present in the environment are simulated in different feature x sample tables. These tables are evaluated for significant feature correlation networks by different metrics and toolkits. The networks are then assessed for accuracy. **(b)** Correlation tools find very

different significant pairs on the same data set. A blue (pink) line connects significant positively (negatively) correlated OTU pairs.

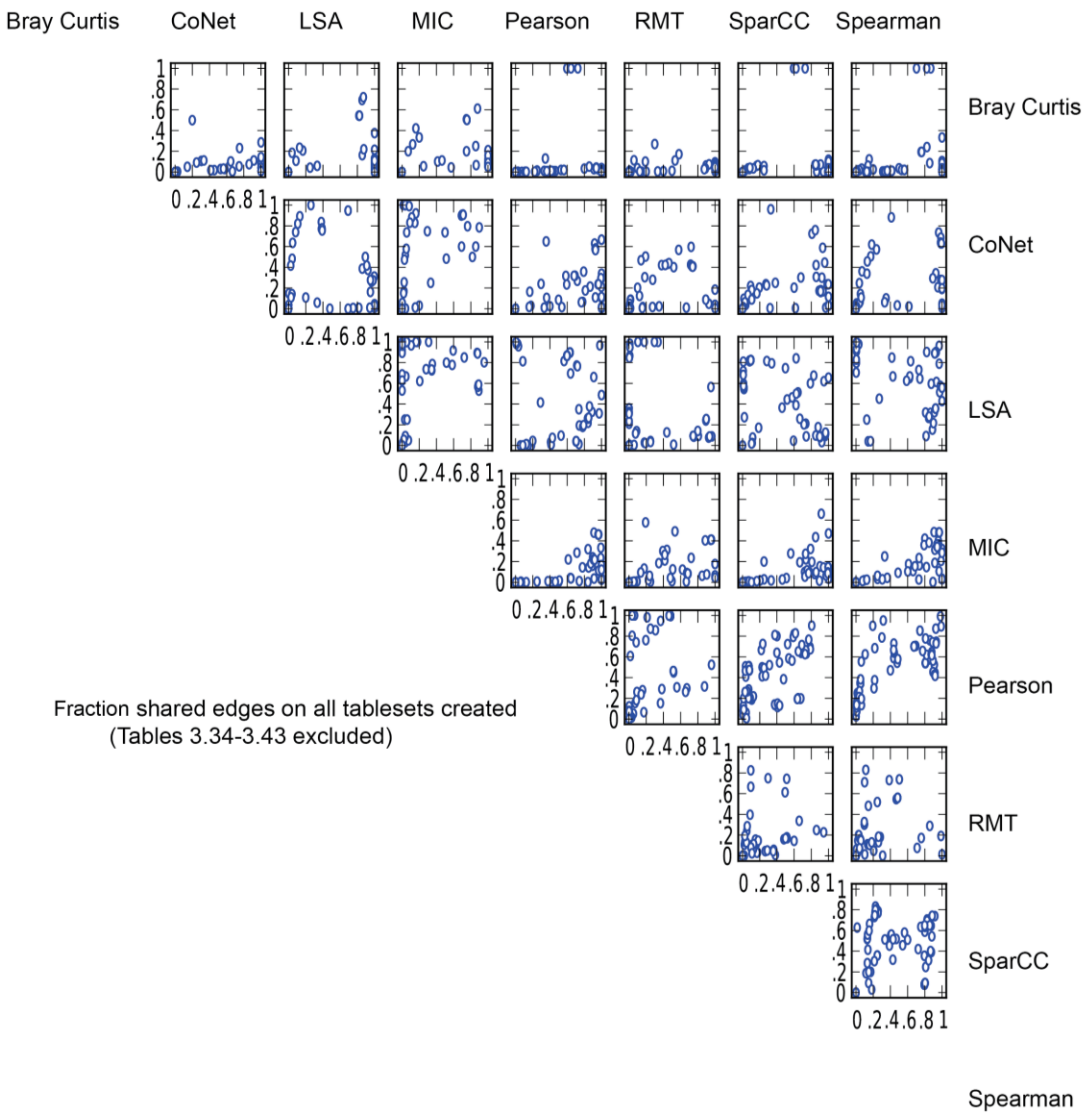


Figure 5-2 The fraction shared edges between the tools on all evaluated tables. Each cell i,j represents the shared edges on all tables (excluding the time series tables 3.34-3.43). The x-axis of each subpanel represents the percentage of tool i's edges that are shared by both tools, and the y-axis represents the percentage of tool j's edges that are shared by both.

these two platforms produce similar results). In contrast, using data generated from 454 and replicates with Illumina[194, 195] we found on average 17% of shared features differed significantly

in count distributions (Benjamini-Hochberg-corrected $p < 0.05$). To investigate further, we processed the 454 and Illumina datasets from Yatsunenko et al.[195] using the same protocol, removed OTUs that were not shared between the technologies, and calculated the fraction of correlated OTU pairs in common between the technologies for all co-occurrence techniques. We found that most techniques (with the exception of Bray-Curtis, which is more robust to differences in the fixed sum of sequences[179]) found < 15% of the same correlated pairs (Figure 5-3A). Given the poor agreement between the networks constructed with 454 and Illumina technologies, we tested the impact of distribution alone on the tool performance using the copula model[189] (Methods). This model simulates contingency tables with the same covariance structure but different marginal distributions so that feature pairs have the same ranked correlation. Employing distributions with many zeros that are often used to model microbiome data[28], such as lognormal, as well as ones mimicking bacterial growth, such as exponential, we found that distribution has less of an impact for those tools that use a rank-based correlation measure like LSA, MIC, and Spearman (Figure 5-3B, Figure 5-4), in agreement with Figure 5-3A.

5.2.3 Different Normalization and Filtering Methods Significantly Alter Edge Inferences

After sequencing and assembling a table of OTU sequence counts (OTU table), the next analysis step is 'normalization' of the data to account for differences in sample sequencing efforts, data sparsity, the limited number of rRNA sequences per sample (compositionality), and extremely rare features whose counts are especially uncertain[24, 28]. Depending on the technique employed, normalization can address some but not all of the first three challenges. It is often paired with 'feature filtration', or selective removal of some features based on certain criteria (e.g. low abundance features), to deal with the last challenge[196]. Here, we quantified the impact

of these approaches on edge inference using real datasets and data generated from a copula model.

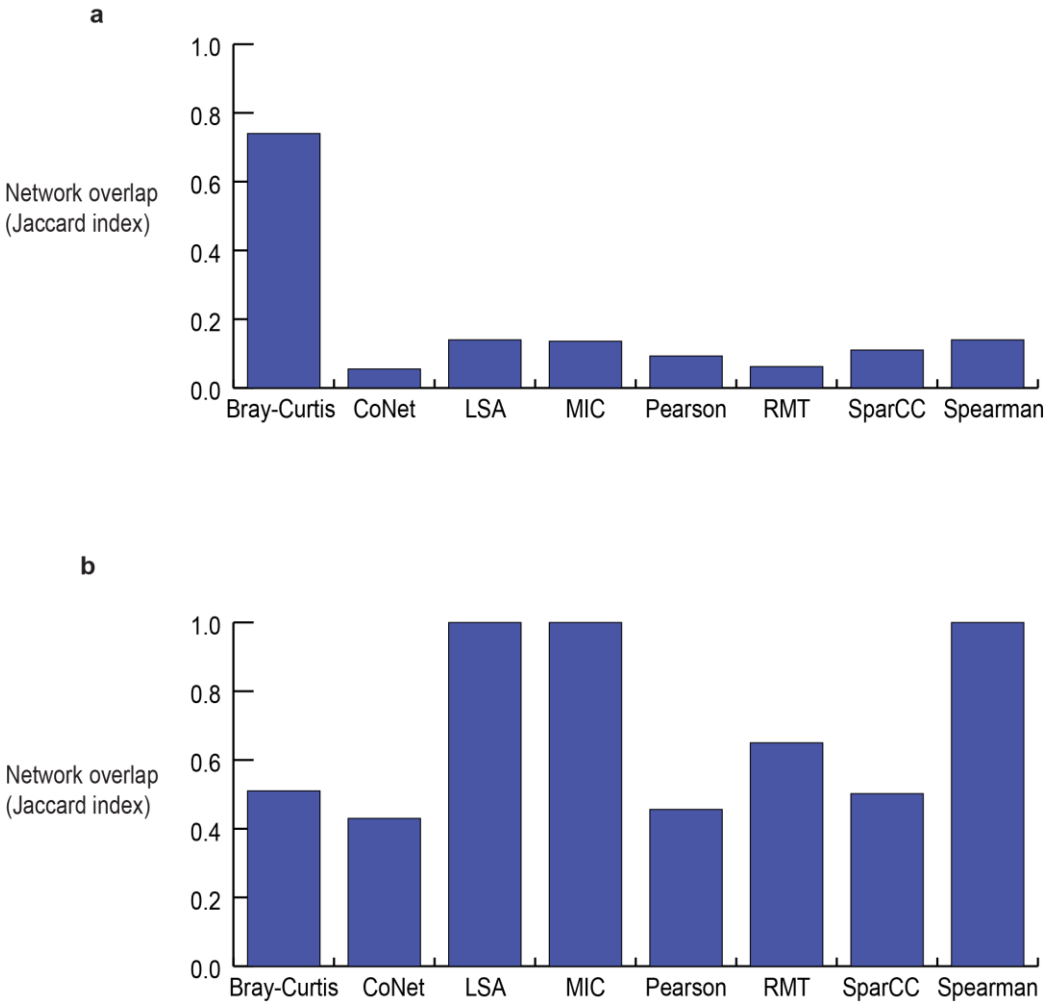


Figure 5-3 Sequencing technology, and therefore distribution significantly affects inferred correlation networks. **(a)** Jaccard index (edge intersection/edge union) showing network overlap by the same technique on 454 sequencing and Illumina datasets. The only difference between the datasets was sequencing technology; they were normalized in the same manner and then filtered to contain only the same OTUs **(b)** Correlation network overlap on datasets with the same rank correlation matrix but different distributions; generated by the copula methodology. Bray-Curtis only detected one or zero edges, causing more variability in the Jaccard index.

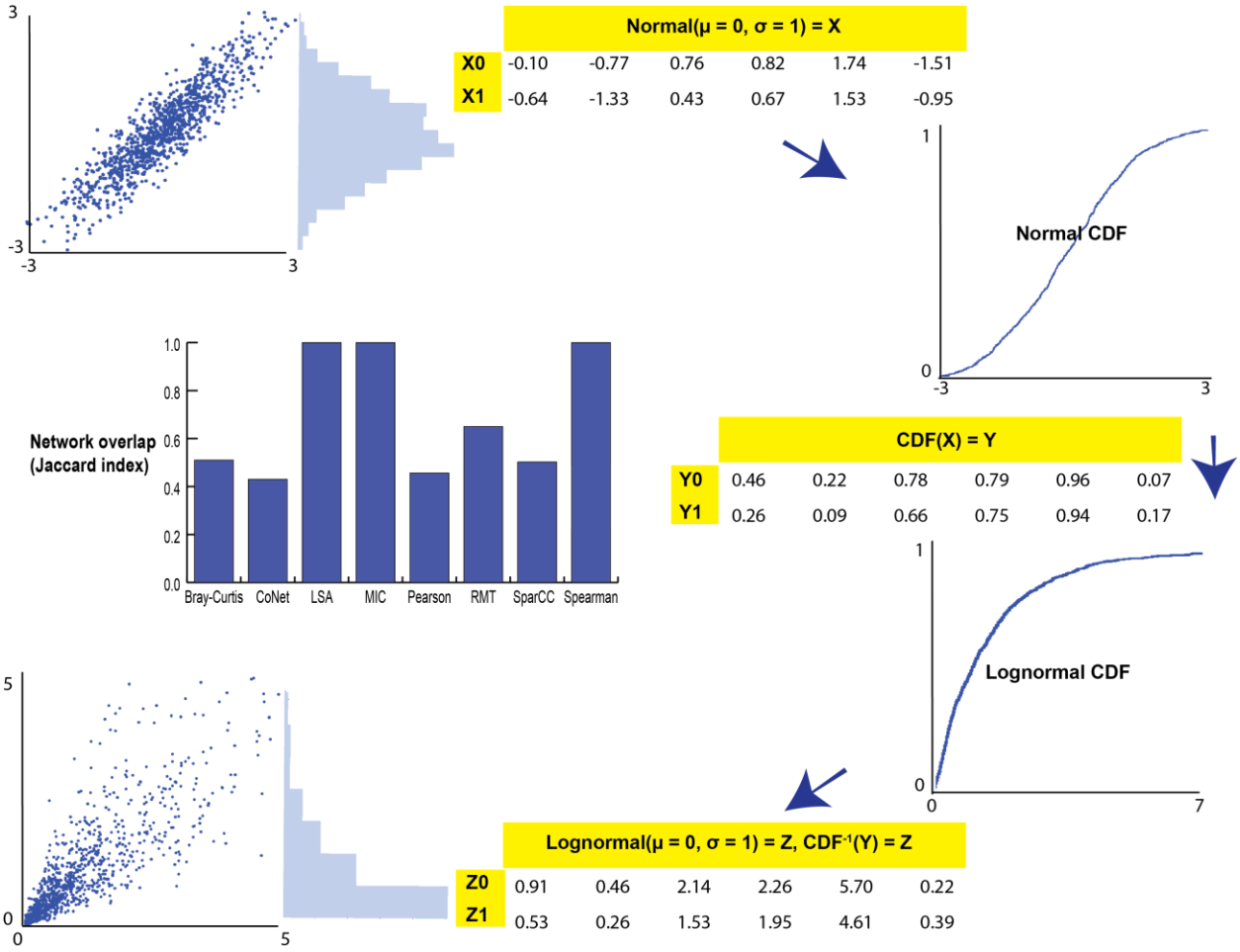


Figure 5-4 Distribution affects some correlation network tools, even if the underlying rank correlation matrix is the same. In the copula methodology, two features with normally distributed scores are converted to their cumulative distribution function (CDF) value for the normal distribution (top). The corresponding CDF score on the lognormal CDF is converted to its lognormal distribution value (bottom). The average fraction of shared feature pairs between copula tables with the same correlation matrix but different distributions is determined (center left) for each tool. This is the same plot as Fig. 2b.

The most common normalization approach is 'rarefying', or drawing without replacement from each sample's distribution until all samples have the same total number of sequences. Rarefying's strength as a normalization technique lies in addressing different column sums and sparse data well. However, due to the random nature of the subsampling, a small amount of variance is introduced into the rarefied data table on different trials[30]. Therefore, we conducted 20 rarefactions (10 at 1000 and 10 at 2000 sequences/sample) and compared the detection profiles of the tools using data from Ridaura et al.[4], who discovered a causal link

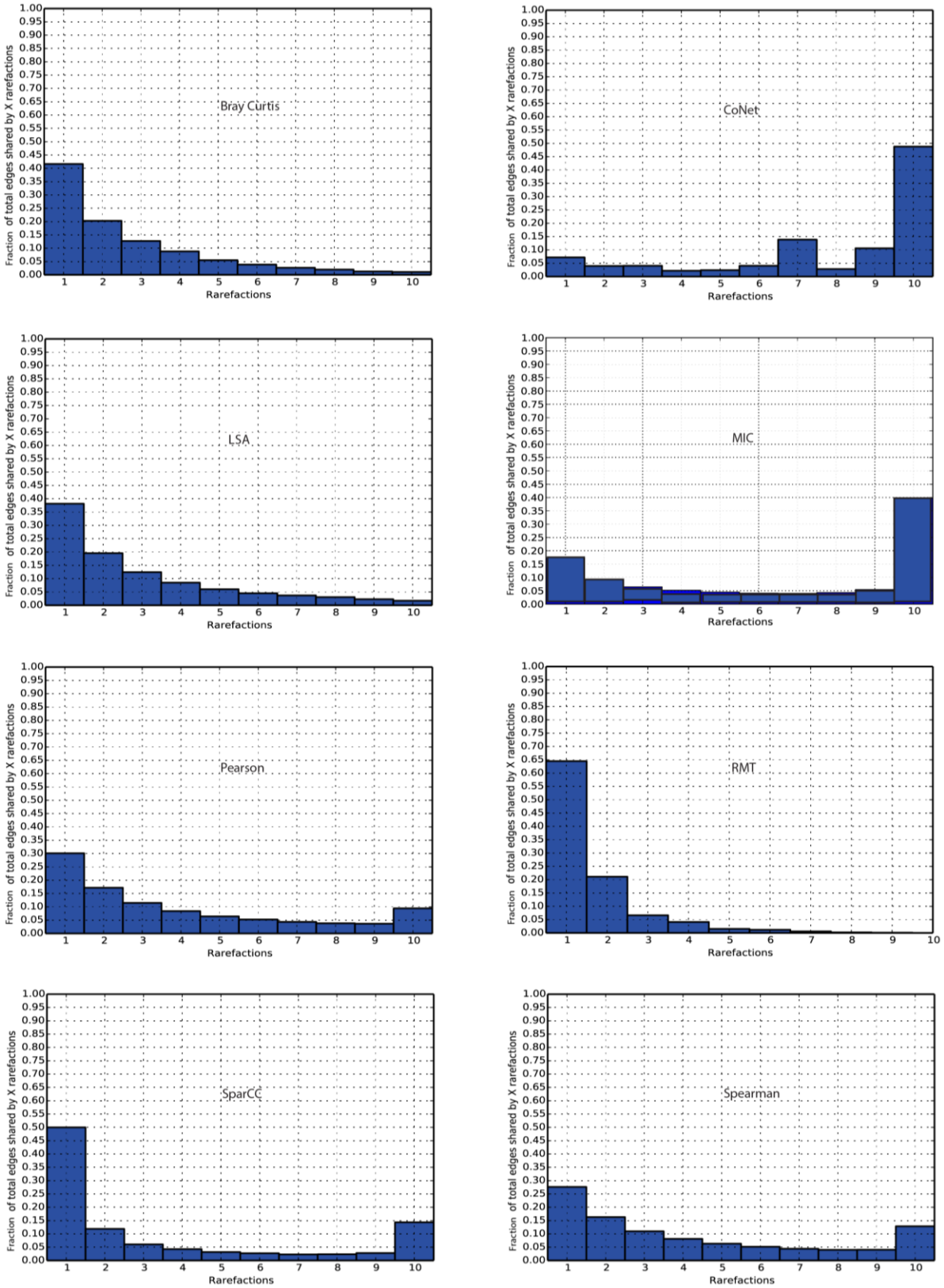


Figure 5-5 The fraction shared feature pairs in X/10 rarefactions of 1000 sample count depth for a given technique.

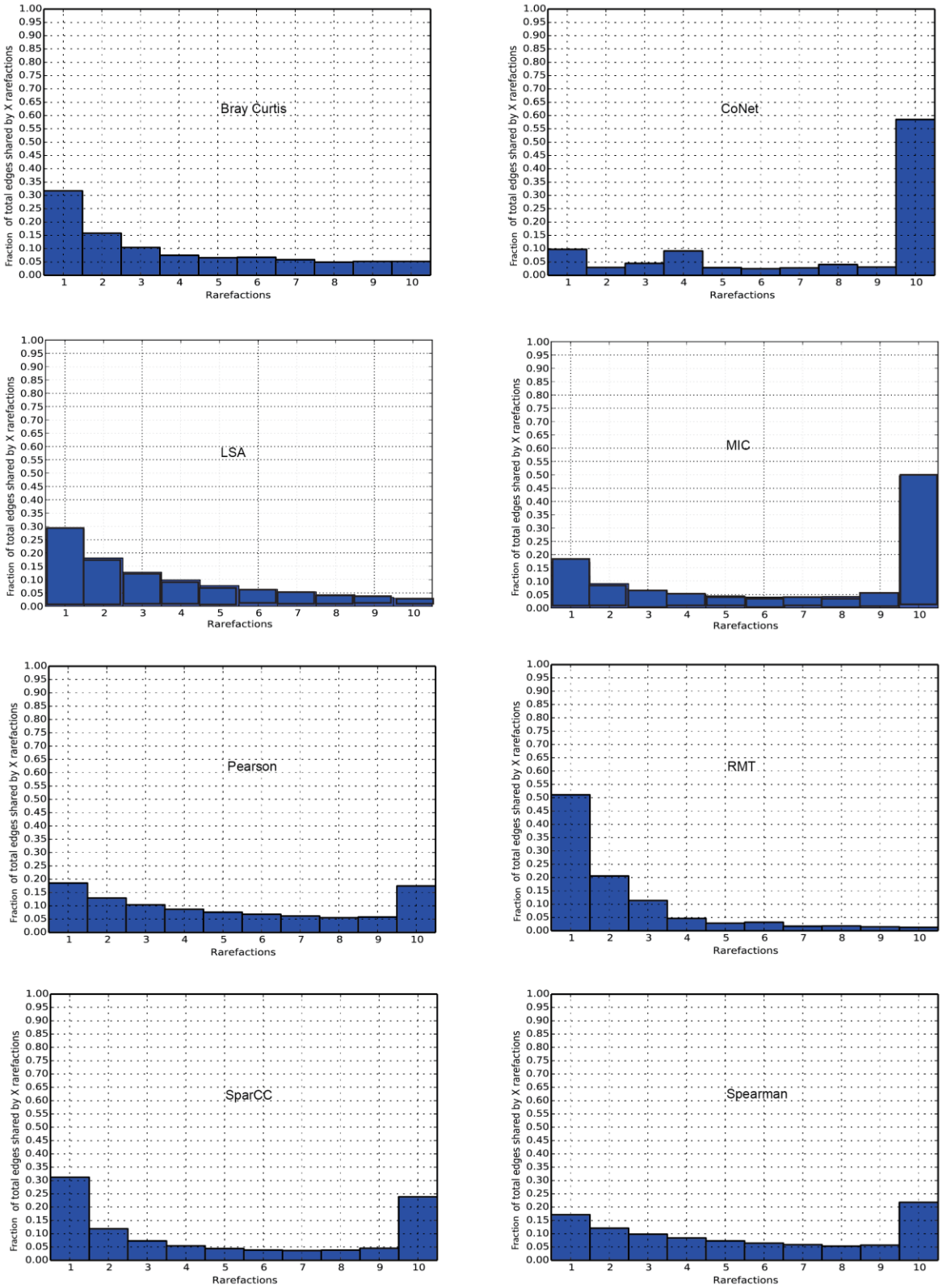


Figure 5-6 The fraction shared feature pairs in X/10 rarefactions of 2000 sample count depth for a given technique.

between gut microbial community composition and the obesity phenotype. The fraction of edges inferred in common in all 10 rarefactions (at a given depth, for a given tool) was under 0.6, suggesting that most tools are very sensitive to small count perturbations (Figure 5-5, Figure 5-6).

Rarefaction-based normalization also does not correct for another serious challenge to correct interpretation of metagenomic data, which is its compositionality (that is, the fact that each sample is composed of a fixed sum of sequences). However, other proposed normalization approaches[30], such as metagenomeSeq's cumulative sum scaling (CSS)[28] and DESeq's log ratio-based variance stabilizing transformation[28], do attempt to correct for this issue. Compositionality can be troublesome to sequencing data interpretation because if the abundance of one species increases while the others do not change, there is less room in the fixed sum for the other species to be counted, thus inducing spurious correlations[24, 89, 90]. Theory suggests that lower numbers of species types should increase the impact of compositionality[24]. We used a set of five copula tables with decreasing numbers of effective species (a measure of microbial diversity) to test how well compositionality is accounted for by each of the correlation and normalization measures (Figure 5-7, Figure 5-8, Figure 5-9, Figure 5-10, Figure 5-11 and Figure 5-12). While the techniques do well on the 'Abundance' tables, we see a dramatic decrease in the number of correct edges for most tools after normalization, which worsens with increased compositionality (smaller n_{eff}). Many edge pairs vary between the same dataset at different n_{eff} (Figure 5-7A), and deviate from the edge predictions based on absolute environmental OTU abundances (Figure 5-7B). Again, rank-based measures such as MIC and Spearman, as well as Bray-Curtis, are less affected by compositionality but still not immune. We did not observe that normalization alternatives to rarefying (CSS[28] and DESeq[28]) ease the compositionality effect; in fact with some

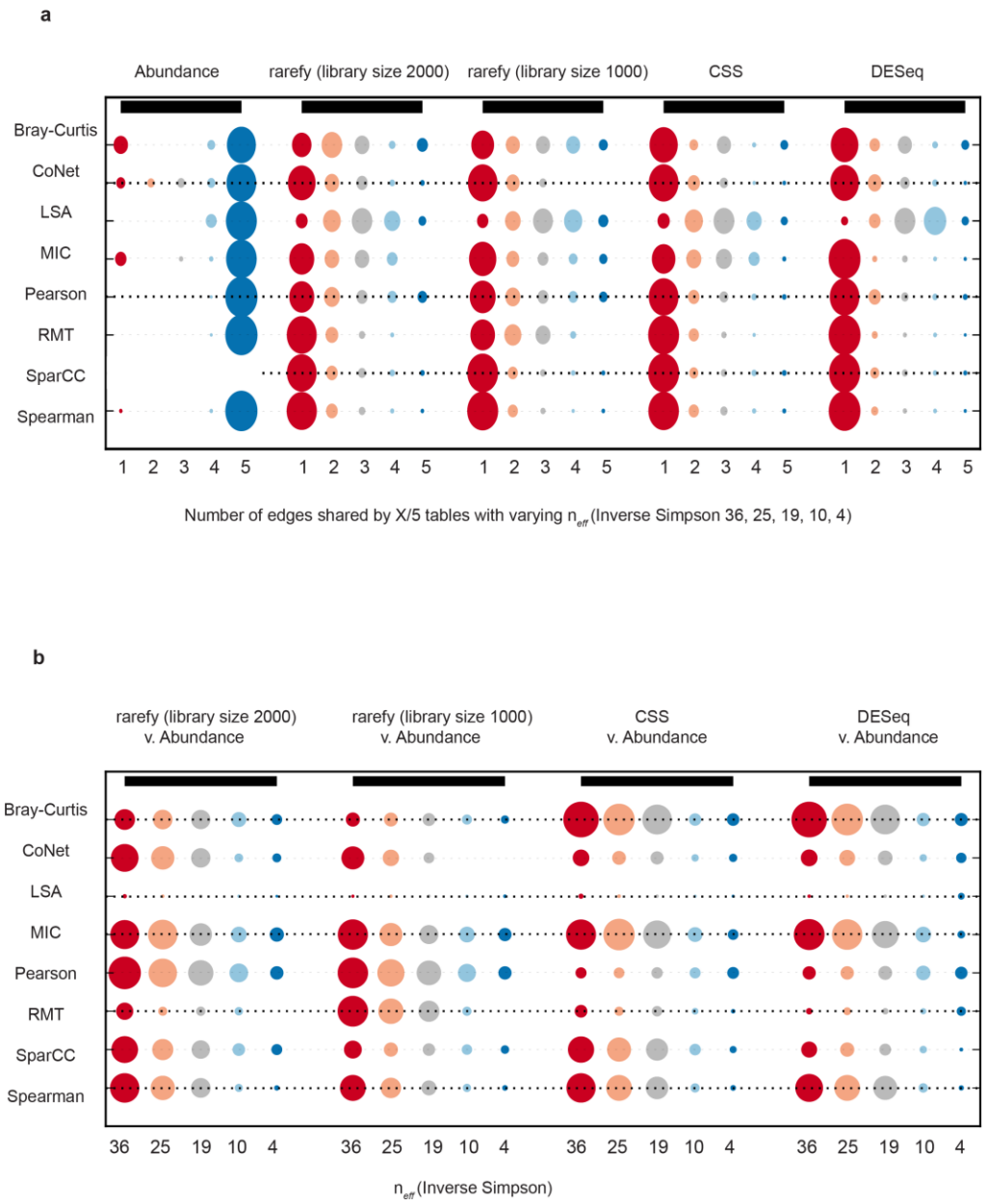


Figure 5-7 The impact of compositionality and normalization strategy on reconstructing actual microbial interactions. Five tables with varying n_{eff} (36, 25, 19, 10, 4) were created by multiplication of the abundances of one OTU pair by a constant; all other OTU abundances remained the same for all tables. These 'Abundance' tables represent the actual OTU abundances in the environment. SparCC assumes the data table is compositional, and hence is not shown. Then, the 'Abundance' tables were sampled without replacement (rarefied), constraining the sum and inducing compositionality, mimicking the experimental sampling process. The rarefy (2000 library size) tables were then either rarefied further (rarefy 1000 library size), CSS normalized, or DESeq normalized. From left to right: **(a)** The five circles within each normalization technique represent, of all the edges found in the five n_{eff} tables the number of edges found 1 (red) - 5 (blue) times. A technique less unaffected by compositionality has a larger circle at point 5, as most tools do in the 'Abundance' tables. **(b)** Network overlap (Jaccard index) between a given normalization technique and the 'Abundance' table for the same tool at a given n_{eff}. A larger circle represents better reconstruction of the true 'Abundance' OTU correlations.

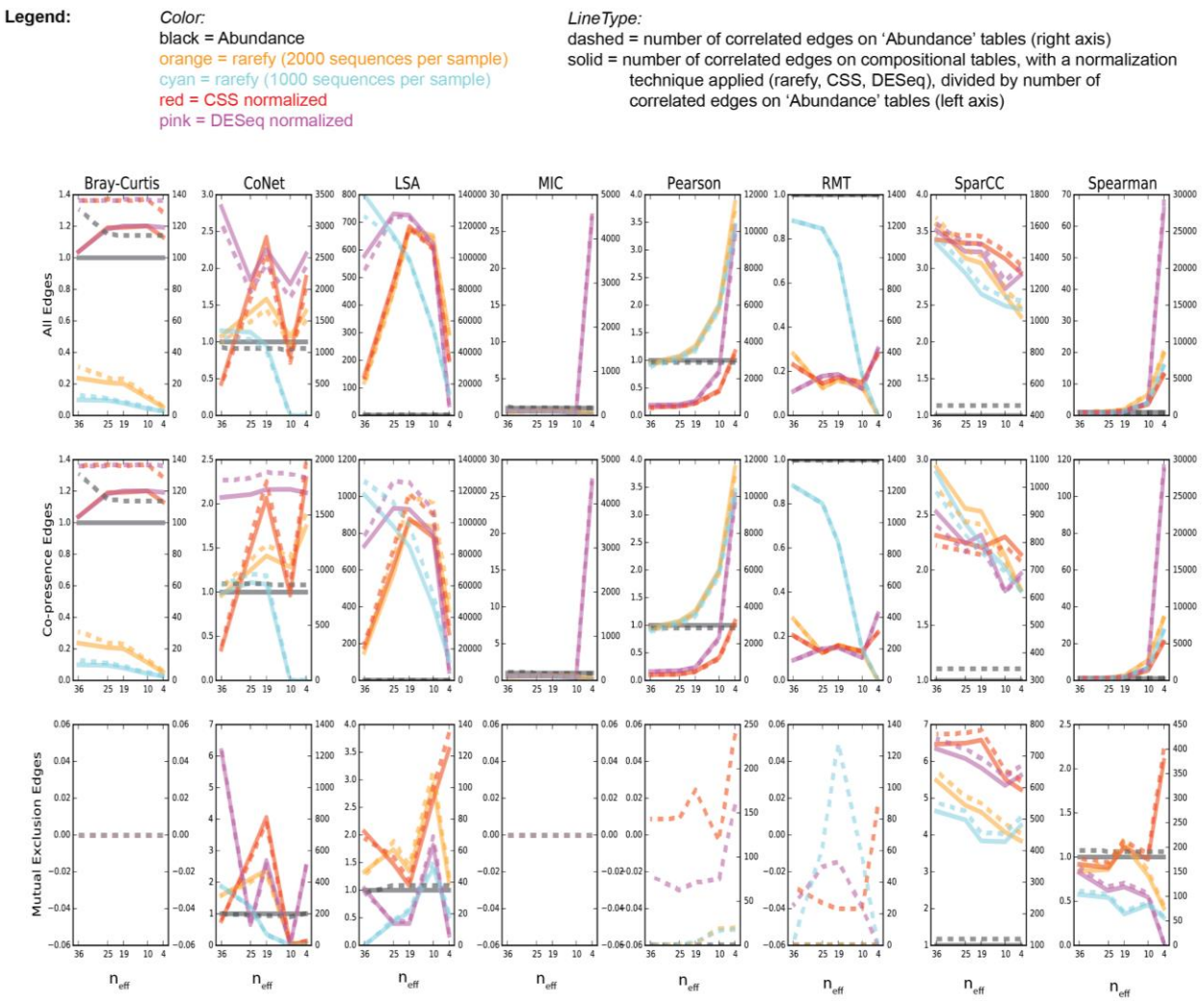


Figure 5-8 The impact of varying n_{eff} and library-size adjustment strategies on resulting significant edges**.** We created five copula tables whose ranked correlation structure and marginal distributions were the same, but where one pair of species was multiplied by an increasing factor to decrease the n_{eff} – these are the same tables as Figure 5-7. The left axis in each plot is fractional and corresponds to the solid lines, thus, the black solid lines are always at 1.

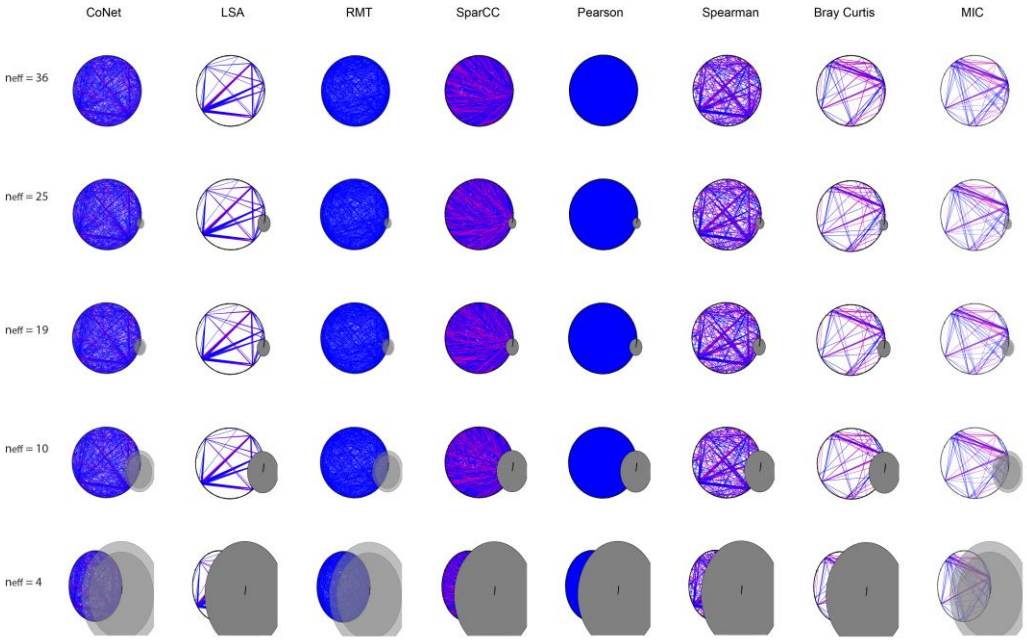


Figure 5-9 Visual depiction of significant (p<0.001) edges found by each technique on raw abundance tables with decreasing n_{eff}: series 1 of 4. This corresponds to the 'Abundance' data of Figure 5-7 and to the black lines in Figure 5-8. The significant edges from each tool graphed (correlated in blue, anti-correlated in pink). p-value thresholds determining a significant edge were set at .001 for all but RMT and CoNet. Nodes are displayed in gray and size is proportional to mean abundance. N_{eff} was calculated using inverse Simpsons.

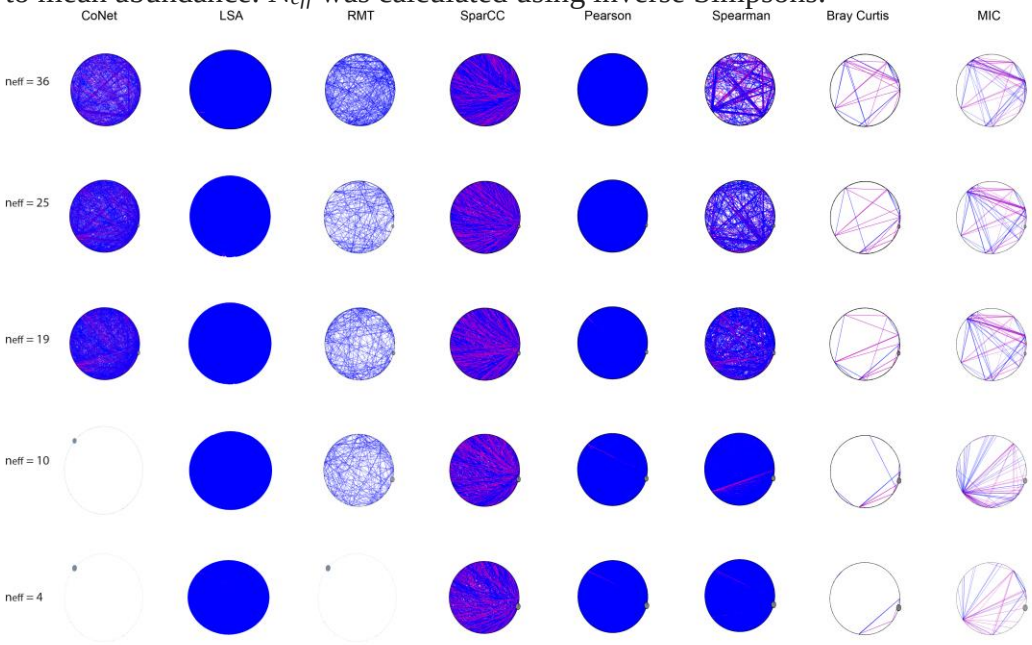


Figure 5-10 Visual depiction of significant (p<0.001) edges found by each technique on compositional tables with decreasing n_{eff}: series 2 of 4. This corresponds to the 'rarefy_2000' data of Figure 5-7 and to the orange lines in Figure 5-8. Label corresponds to Figure 5-8

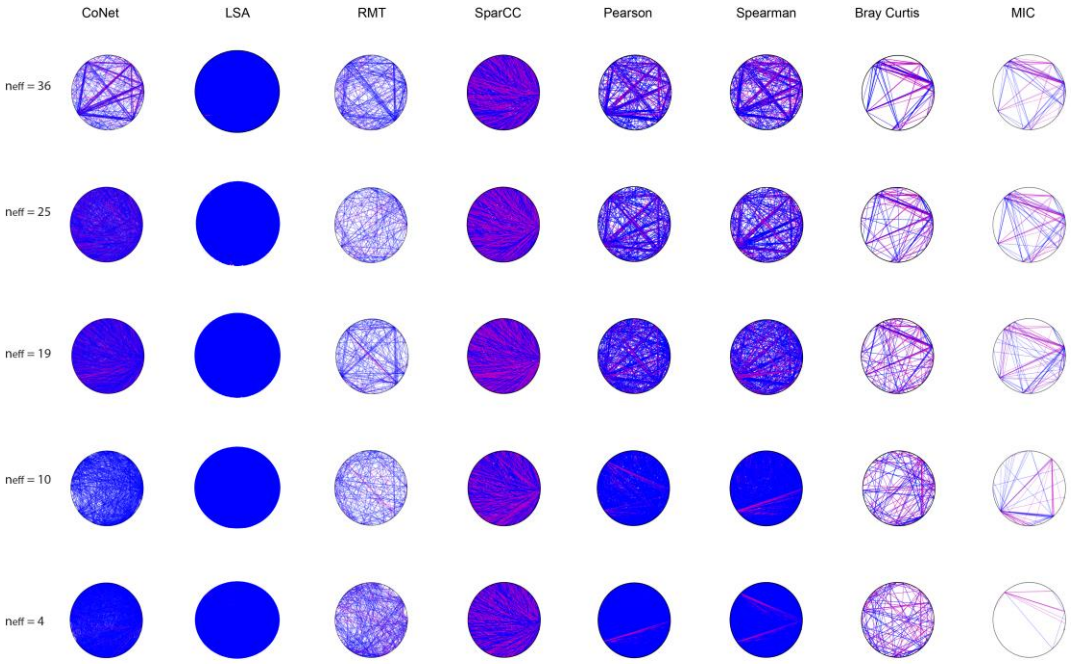


Figure 5-11 Visual depiction of significant ($p<0.001$) edges found by each technique on compositional tables with decreasing n_{eff}: series 3 of 4. This corresponds to the 'CSS' data of Figure 5-7 and to the red lines in Figure 5-8. Label corresponds to Figure 5-8.

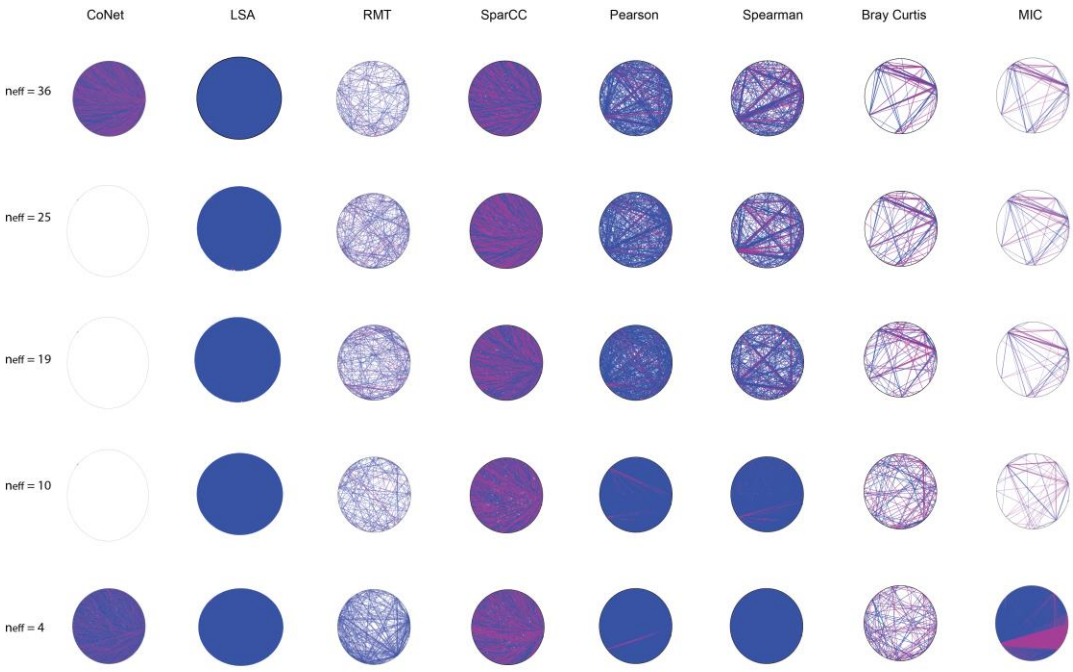


Figure 5-12 Visual depiction of significant ($p<0.001$) edges found by each technique on compositional tables with decreasing n_{eff}: series 4 of 4. This corresponds to the 'DESeq' data of Figure 5-7 and to the pink lines in Figure 5-8. Label corresponds to Figure 5-8.

tools these approaches (especially DESeq) performed worse than rarefying. This is most likely because neither technique accounts for differences in data sparsity well, which rarefying does. In general, across all tools and normalization techniques, the slope of the function describing the number of total edges for a given n_{eff} (Figure 5-8) changes particularly quickly at low n_{eff}, suggesting that the smaller the number of effective species, the larger the impact on edge-inference results. Since rarefaction reduces the number of species present by subsampling, this indicates that rarefying at a lower depth may intensify compositionality effects on correlation networks. However, the main effect is a decrease in the number of edges found (Figure 5-3, Figure 5-8). Notably, while network overlap is affected, many tools such as SparCC maintain high precision compared to predictions on 'Abundance' tables (Figure 5-13). These findings demonstrate that promising work has been done on addressing compositionality as a significant challenge to co-occurrence network inference; however, this problem is still not solved.

Filtering can be performed either on sequences before they are clustered into OTUs, or after an OTU table has been constructed. To investigate the effects of the former on network overlap, we filtered (removed) the raw rRNA reads of Ridaura et al.[4] having relative abundances falling at a few points at the lower end of the recommended[188] range of 0.01% to 0.00001%. This altered the number of edges differentially depending on the technique; however, Bray-Curtis as well as CoNet (possibly because it removes rare OTUs) remained relatively immune (Figure 5-14). After an OTU table is constructed, additional filtering can be done to remove rare OTUs (e.g. those found in 5% or fewer samples) whose low count numbers are less certain, as well as to limit the number of statistical comparisons performed; this is important both for controlling false discovery rate and for minimizing computational time/effort. Again using data

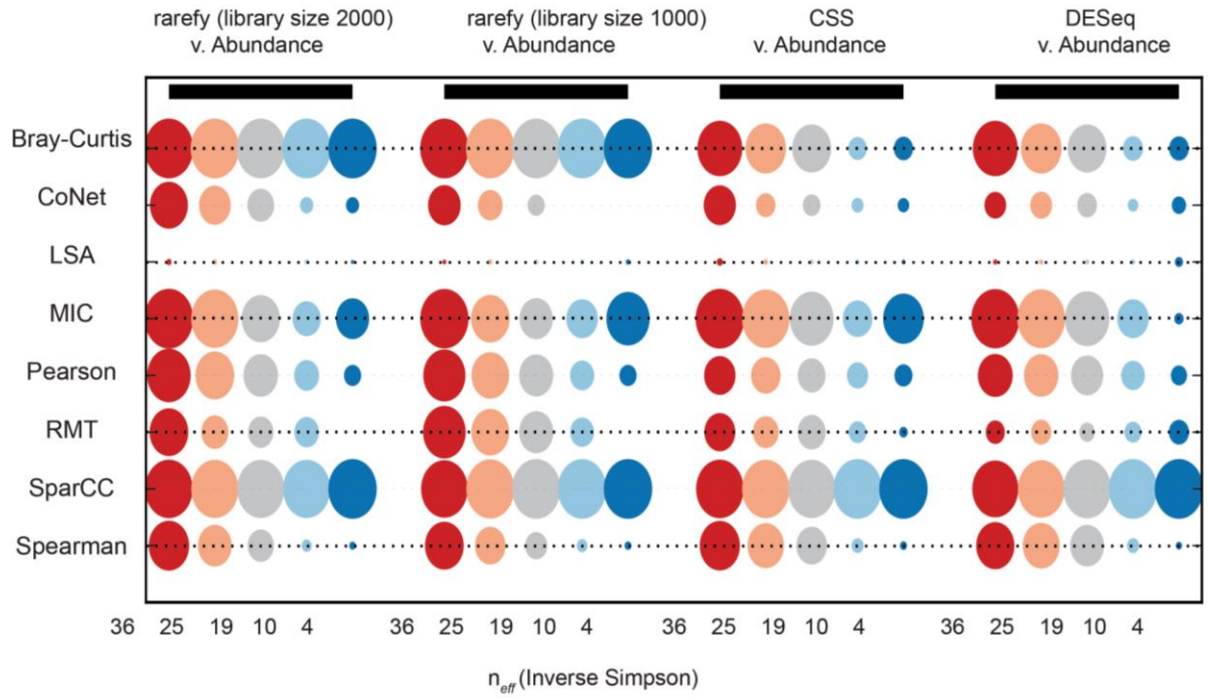


Figure 5-13 Tool precision in response to compositionality. The same data as Figure 5-7b, except using precision instead of Jaccard index as a measure of network overlap. Tool predictions were benchmarked against the 'true positive' edges found by the same tool on absolute 'Abundance' data. SparCC was benchmarked against the log-transformed Pearson correlations on 'Abundance' data, since that is what it seeks to estimate.

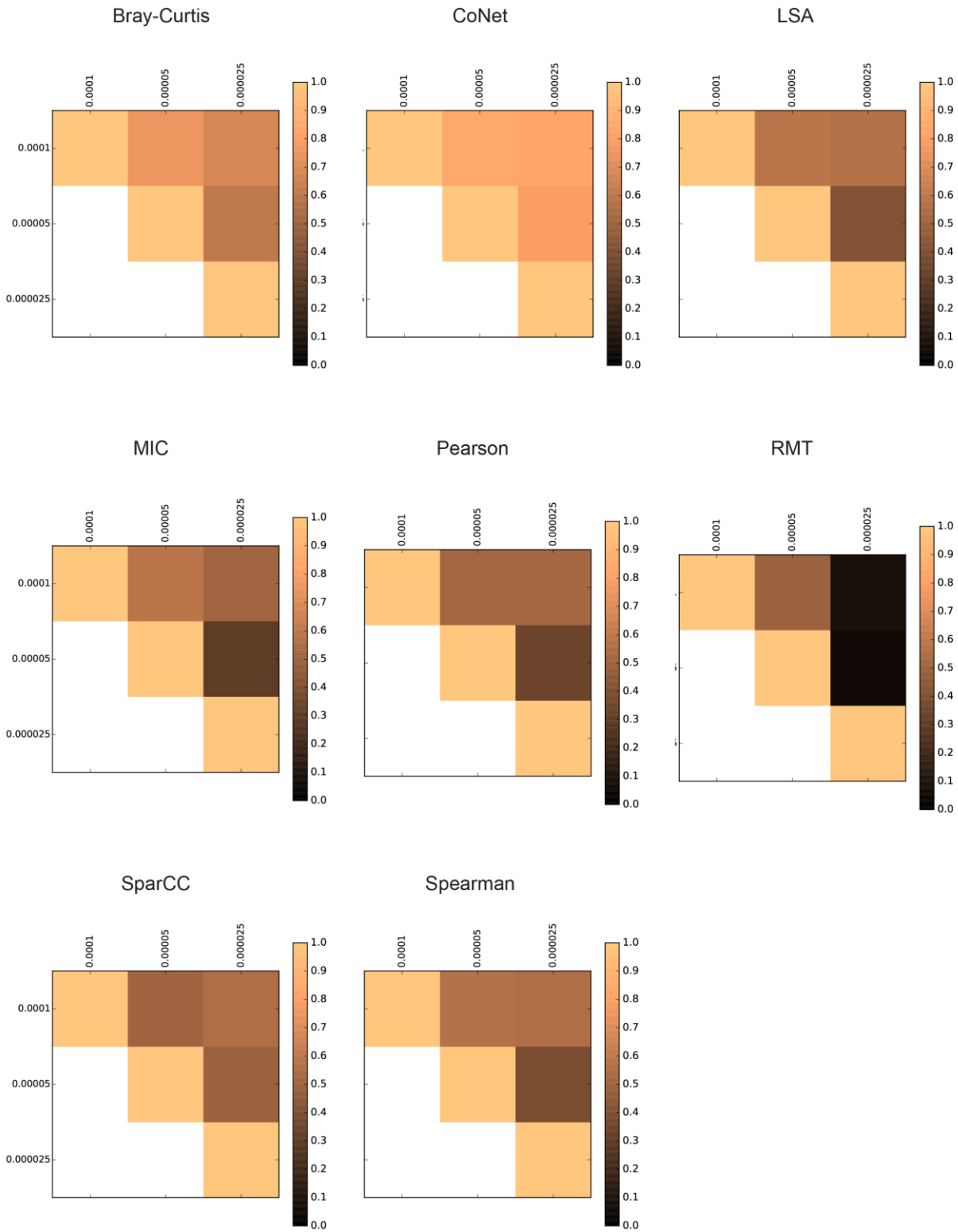


Figure 5-14 Sequence filtering strategy prior to OTU table construction greatly affects resulting correlations. rRNA sequences having percentages of total sequences below the thresholds .00005, .00010, and .000025% were removed. Network overlap calculated with Jaccard index.

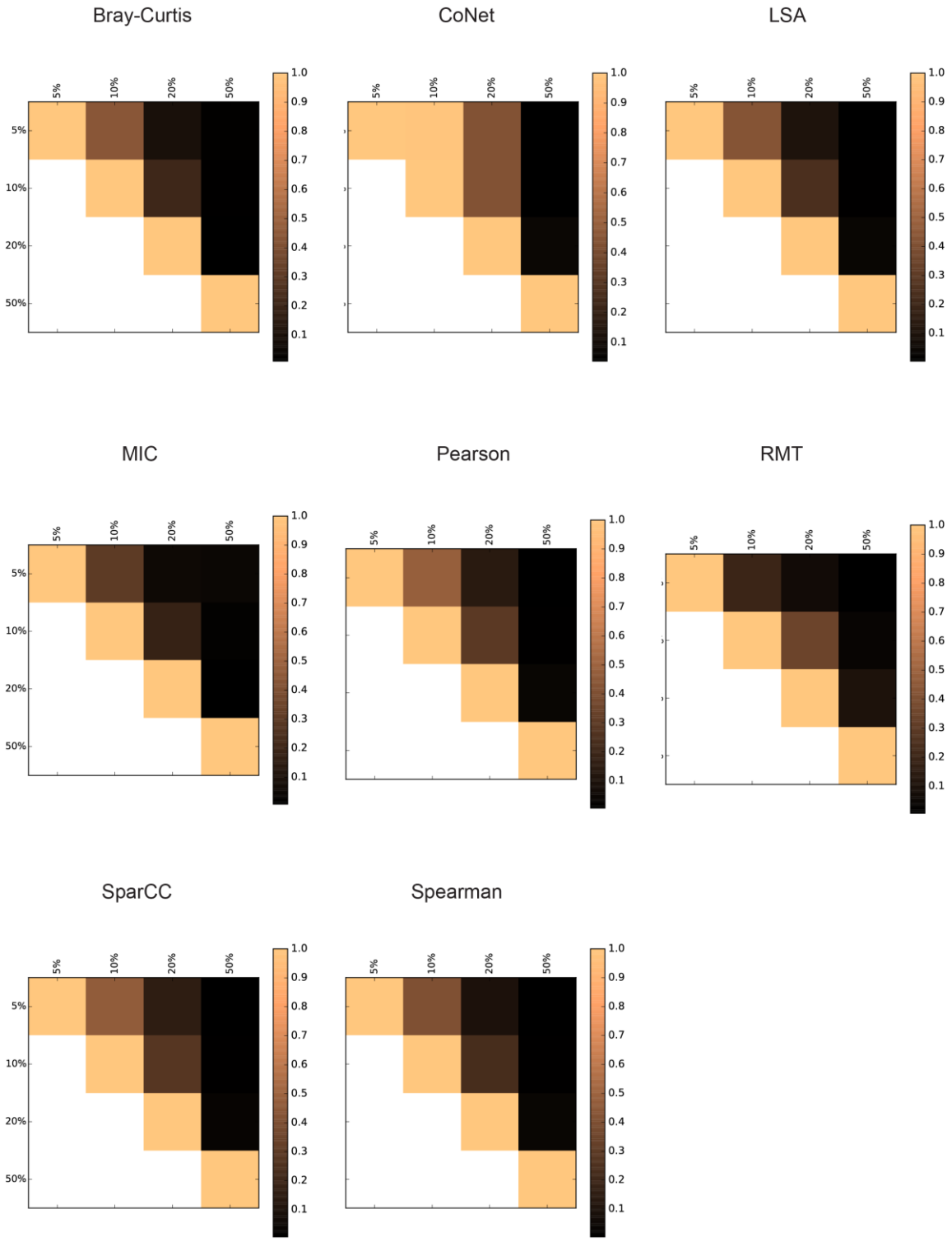


Figure 5-15 Tools are fairly robust to OTU filtering strategy after table normalization by rarefying. OTUs not present in 5%, 10%, 20% and 50% of samples were filtered out after rarefying to 1000 sequences per sample. Network overlap calculated with Jaccard index.

from Ridaura et al.[4], we found that setting a higher filtering threshold, and therefore removing more OTUs, reduced the numbers of overlapping significant correlations (Figure 5-15). We observe that sparsity significantly decreases network inference performance, and that rare OTUs are more likely to be lost stochastically if stringent filtering is used; therefore, we recommend minimal filtering. The exact optimal filtering threshold for each tool requires more investigation and may be data-dependent.

5.2.4 The Number of False Positives in Null Data is Within Expectations but Differs by Tool/Technique and in Some Cases by Distribution

Control of the false positive rate is well established in traditional statistical analysis[197-199] but has not been standardized for correlation inference. RMT allows the method itself to set the correlation threshold, rather than employing an arbitrary user-imposed threshold. LSA, CoNet and SparCC calculate the p-value through permutation-based approaches as well as q-value[199] and Benjamini-Hochberg[198] multiple-hypothesis testing correction. MIC and Bray-Curtis calculate the p-value through distributional approaches, Pearson and Spearman calculate the p-value with Fisher z-transformation, and all of these apply stricter Bonferroni[197] multiple hypothesis testing correction. To enable assessment of the relative performance of these methods, we created two 'null' data tables, one containing random draws from six different zero-heavy distributions and the other from a Dirichlet distribution modeled on real data. (The former simulates differently distributed non-compositional data in which vectors are independent and identically distributed (iid) within a distribution, while the latter simulates compositional data which is not iid, but for which no correlation matrix is specified. Both of these data tables should have no true associations between features.) The performance of the tested tools on these data is generally excellent (Figure 5-16), despite differences in p-value calculation and multiple

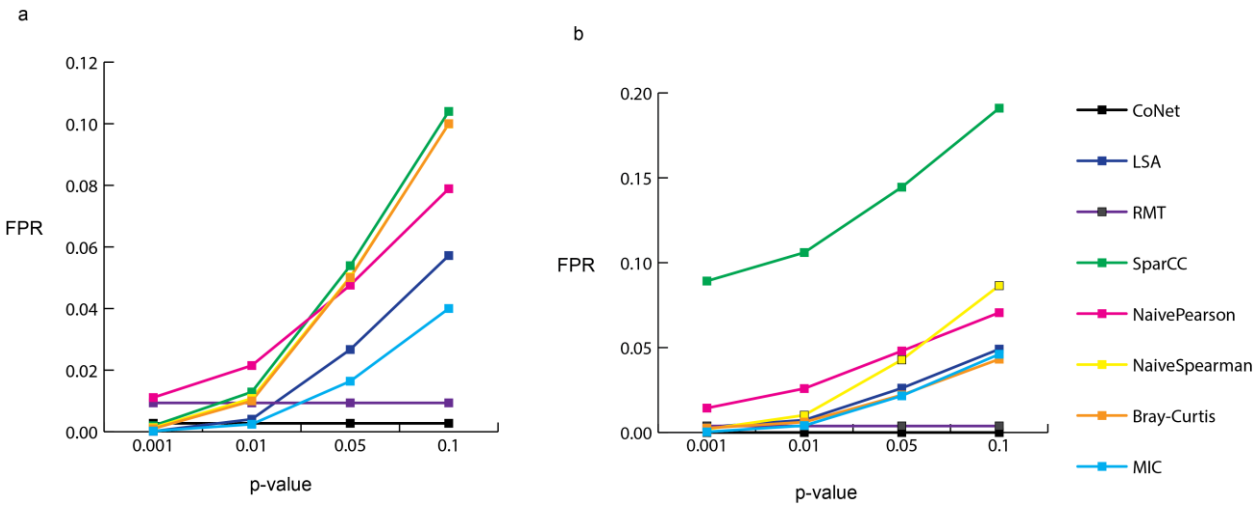


Figure 5-16 The false positive rate (FPR) plotted at different p-values for the different metrics and toolkits. **(a)** The FPR for the null table with features drawn for different distributions. **(b)** The FPR for the null table created from random samplings of a Dirichlet distribution modeled on real data

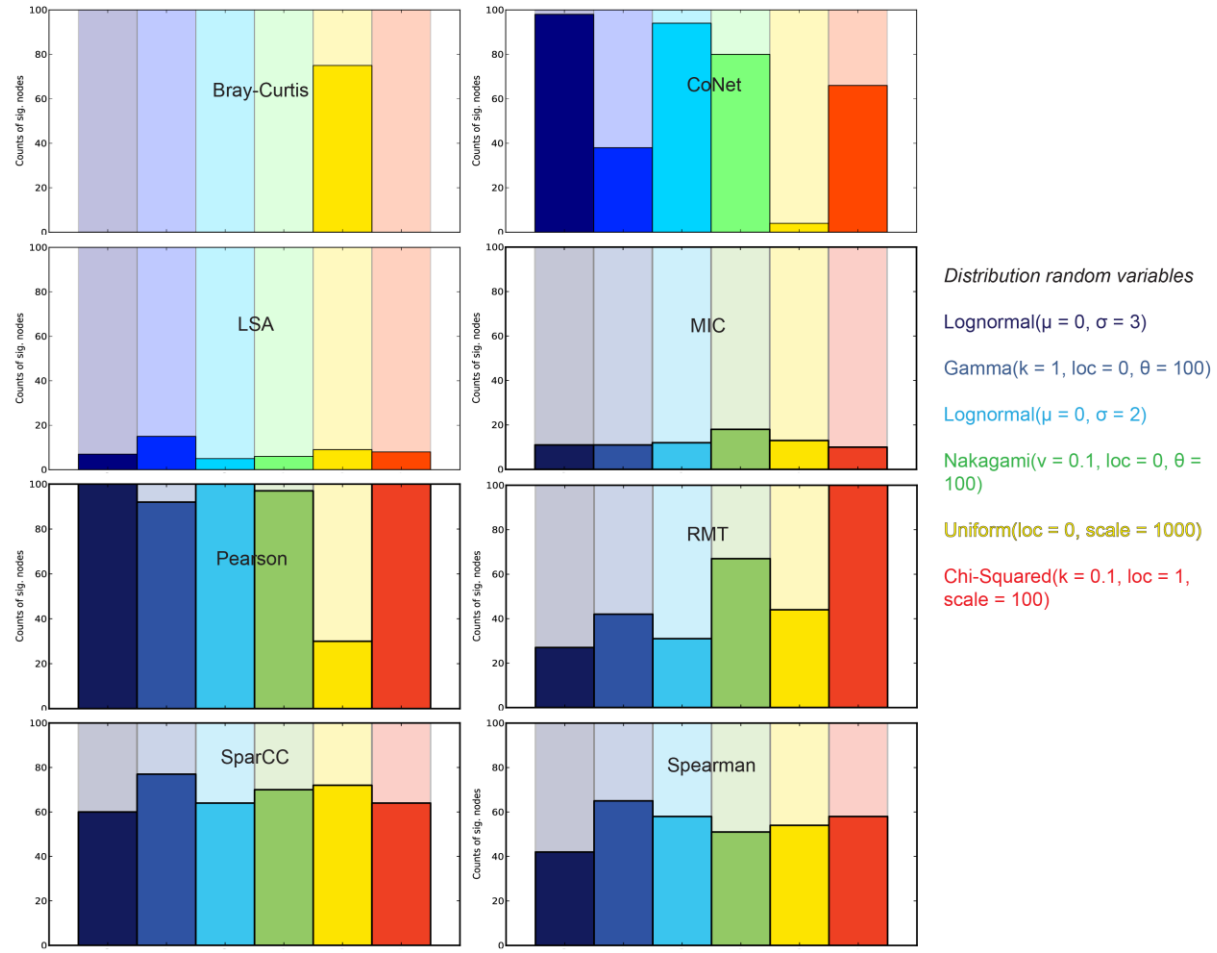


Figure 5-17 Number of features from each null distribution type deemed significant by the given metric or toolkit. One hundred features were drawn from each distribution.

hypothesis testing. However, although the false positive rates (FP/(FP+TN)) are in-line with specified p-values for tools that rely on them, the false discovery rates (FP/(FP+TP)) are not, because TP=0 for these tables. This result suggests that all tools tested may have low precision (below 0.2), suggesting an important area for improvement of future techniques.

Additionally, RMT and CoNet demonstrate an unexpected tendency to preferentially select edges from certain distributions. RMT shows a preference for chi-squared-distributed OTUs, and CoNet prefers OTUs from the chi-squared, Nakagami, and lognormal distributions (Figure 5-17). Bray-Curtis almost exclusively selects edges from the uniform distribution, whereas Pearson finds three times fewer edges from the uniform distribution compared to the other distributions. This means that these tools may preferentially select as correlated the OTUs exhibiting these distributions. For example, if uniform or chi-squared-distributed OTU correlations are preferred, parasitic or predatory relationships, where one species benefits and the other is harmed, may go undetected.

5.2.5 A Subset of Common Linear Ecological Relationships is Detectable by Some Tools

Correctly detecting ecologically meaningful relationships such as competition and mutualism is essential for a correlation tool. In order to test tools' capacity to identify these relationships, we developed simple linear models of the amensal, commensal, competitive, mutual, obligate, parasitic, and partial-obligate-syntrophic ecological relationships (Methods). These ecological relationships manifest as a dependency between the species abundances for a given ecological relationship type. We built tables where the type, strength, and number of OTUs in a linear relationship varied, and introduced compositionality, sparsity, or both. Mutualism and commensalism are well detected by most tools (Figure 5-18A, Supplementary Note), while amensalism and partial-obligate-syntrophy are functionally undetectable. All tools detect

parasitism as a co-presence rather than as mutual exclusion, but three tools (SparCC, Spearman, and LSA) correctly identify competitive relationships as mutual exclusions. As expected, tool performance generally improves with increasing strength of a relationship (i.e., increasing signal/noise ratio). Literature suggests that many biological effects are mediated by more than two species interactions[200]. In tests of data with more than two members, detection profiles were similar to two-species relationships, but considerably attenuated (Figure 5-18B). SparCC and LSA are unique among the tested tools for their ability to correctly infer a competitive 3-member relationship as having components of both co-presence and mutual exclusion. Nonetheless, our results suggest that microbial relationships having more than three members are likely impossible to detect with current approaches.

The features in these data sets were iid unless part of an engineered correlation, which allowed us to accurately assess tool sensitivity and specificity. ROC curves of the ecological data confirm that increasing the complexity of the ecological relationships by mixing three-species relationships with simpler two-species relationships (Figure 5-19A) significantly decreases tool specificity and sensitivity. While tool performance improves on only two-species ecological data even with the addition of compositionality (Figure 5-19B), increasing sparsity (Figure 5-19C) to levels commonly seen in microbiome datasets drastically reduces tool performance to little better than random guessing. In agreement with the above null data, precision of tools is also extremely poor (close to or at zero) under realistic conditions (Figure 5-20A-C).

5.2.6 Non-linear Ecological Relationships are Harder to Detect Than Linear Ones

Lotka-Volterra models are a set of classic ecological models for interacting species based on coupled first-order differential equations[190] that are applicable in a wide range of macro-scale

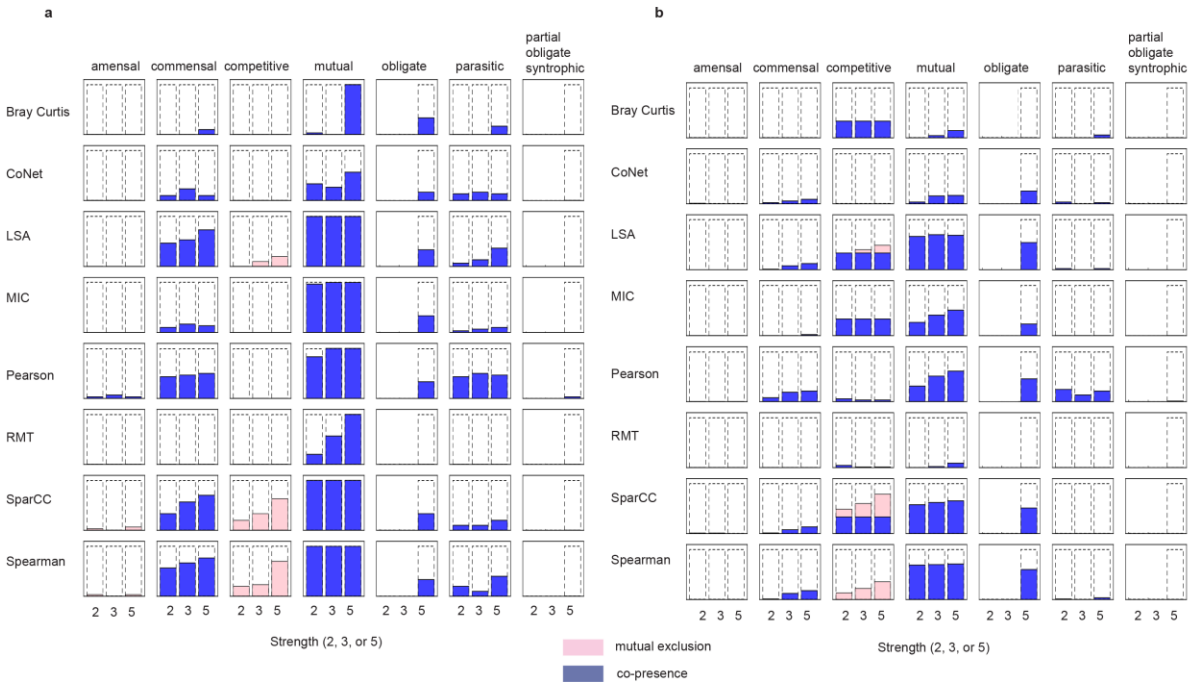


Figure 5-18 Types of linear ecological relationships detected by each correlation technique. The columns represent the eight types of engineered ecological relationships, and the rows the eight tools tested. Each cell contains three histograms with increasing 'strength' of relationship from left to right. The fill in each bar represents the fraction of engineered edges detected as significant when the relationships were composed of **(a)** pairs of features, or **(b)** triples or more.

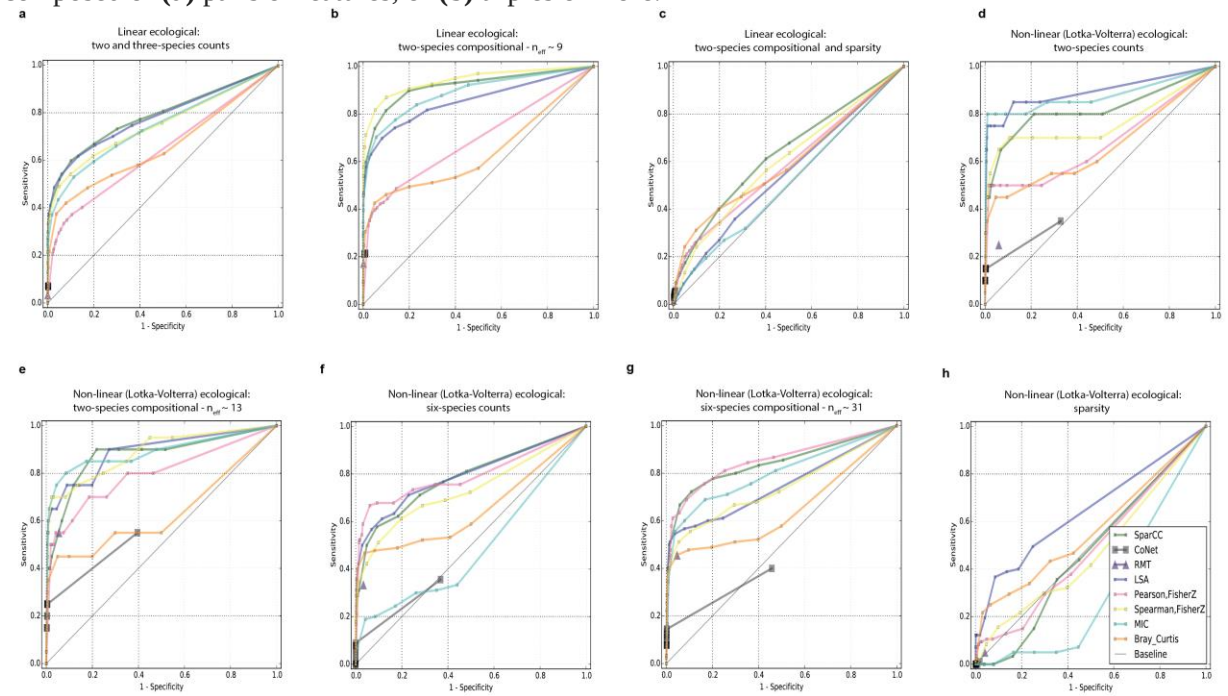


Figure 5-19 Receiver Operating Characteristic (ROC) curves for linear ecological relationships **(a-c)** and non-linear/Lotka-Volterra ecological relationships **(d-h)**. All tables were approximately 40% sparse, except (c) and (h), which were 70% sparse. The CoNet ROC curve does not extend from the bottom left corner to the top right corner of the ROC curves because of the filtering procedure CoNet uses prior to

inferring any correlations. RMT is only a single point since the algorithm sets the p-value, instead of the user imposing a p-value, ROC curves cannot be constructed.

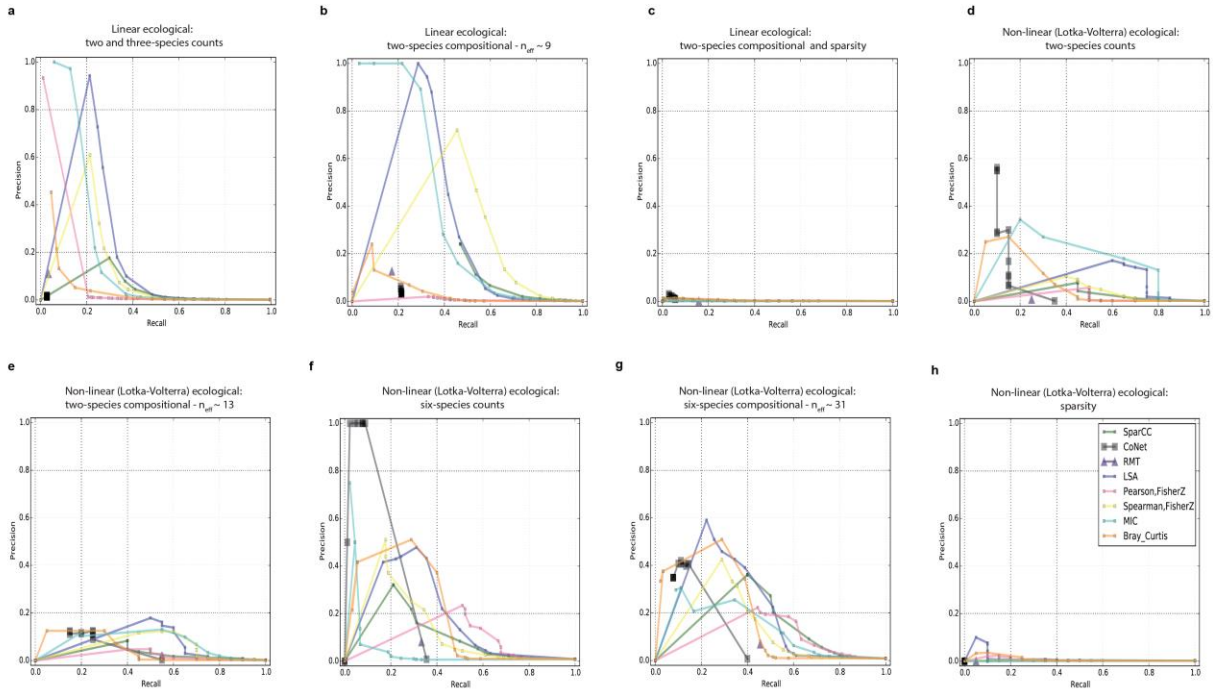


Figure 5-20 Tool precision is extremely low under realistic microbiome dataset conditions. Precision vs. recall (sensitivity) curves for linear **(a-c)** and non-linear ecological relationships **(d-h)**. These are analogous to the ROC curves of Fig 5.

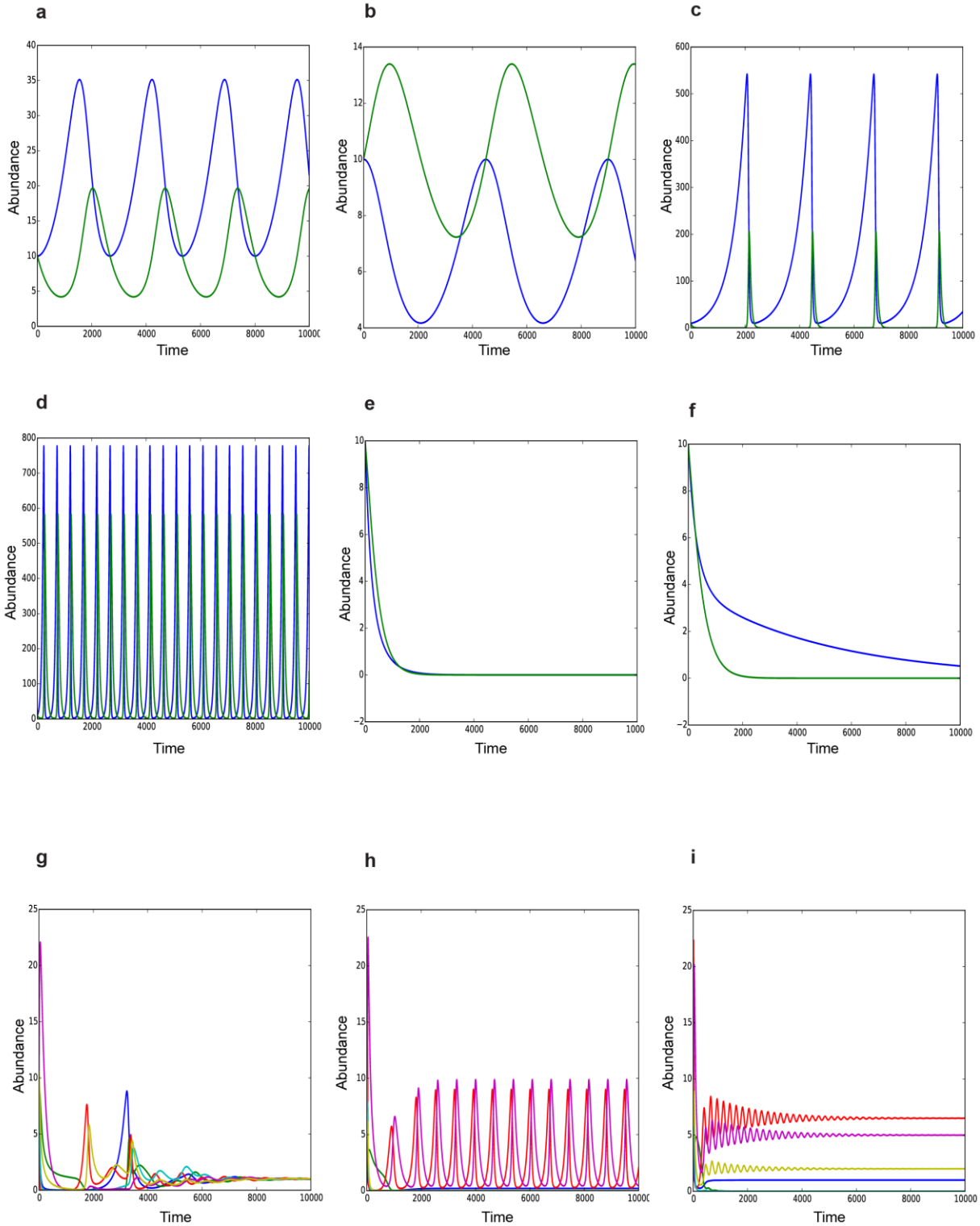


Figure 5-21 Visual depictions of feature count dynamics over time of the ecological Lotka-Volterra equation systems tested. **(a-f)** Two-species Lotka-Volterra relationships. The green line is the values of feature 1, and the blue line is feature 2 generated from the coupled differential equations. **(g-i)** Six-species Lotka-Volterra relationships. The abundance values of the six species over time are shown in different colors.

ecological relationships[200]. Evidence is emerging for their applicability at the micro scale as well—for example, in describing the microbial dynamics in a cheese model community[201] and within individuals, as well as community shifts in response to environmental perturbations[202]. Previous investigation in this area mostly tested standard correlation metrics not developed for microbiome data[188]. We created two- and six-species Lotka-Volterra interactions (Figure 5-19D-H, Figure 5-20D-H, Figure 5-21) and tested whether tools accurately capture these relationships when they are embedded in random noisy signals.

The irregularity of the Lotka-Volterra equations proves difficult for all measures, with an average 10% drop in sensitivity compared to the linear ecological relationships. For the two-species edges, MIC, SparCC, LSA, and Spearman all perform strongly for both count and compositional tables (Figure 5-19D and E), while SparCC consistently performs well on the six-species Lotka-Volterra tables (Figure 5-19F and G). Pearson also performs well on the six-species tables because some of the dissipative relationships display linear correlations. However, again under realistic conditions, when sparsity is boosted from 40% to 70%, performance drops to little better (or even worse) than random guessing (Figure 5-19H). The same is true for precision (Figure 5-19H).

5.2.7 Time-dependent Relationships Vary Based on Signal, Sampling Frequency, and Time Shift

Correlations in time-series data are well studied in other fields, but microbiological research is just beginning to show predictable shifts in microbial communities over time[121, 203, 204]. For example, in Caporaso et al., some fluctuations appear sinusoidal[121]. Here we modeled simple temporal relationships between OTUs as signals with varying noise, amplitude offset, phase shift, frequency, and coupling. In these mixture model tables, composed of sine, cosine,

saw-tooth, and logarithmic patterns, none of the tools was able to detect any specific signal type especially well. For all tools, the most frequently detected co-occurring pairs stemmed from mixed signal types (e.g. co-occurring sin and square wave signals).Furthermore, detected edges varied depending upon at which point in time/how many samples were taken of the fluctuating OTUs. This is important since researchers take discrete samples, and therefore cannot know the abundance of each OTU at every point in time. The time shift of OTU abundance signals also affected resulting correlations (Figure 5-22, Figure 5-23, Figure 5-24, Supplementary Note). This is important insofar as two pulse signals that peak at day zero might be more easily detected as correlated than two signals with the same pulse but offset in phase.

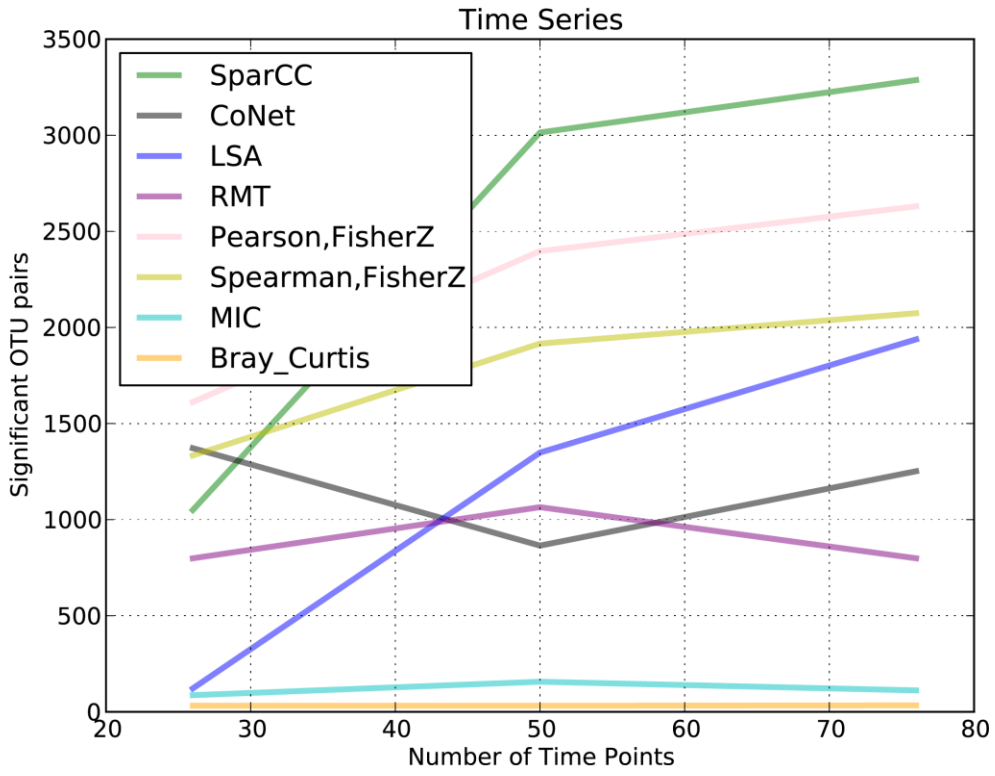


Figure 5-22 The time, or point in the feature signal cycle, at which a sample is taken introduces variability in detected correlations. The number of samples is also a large influence in reconstructing the correct signal, and therefore correlation. The number of co-occurring feature pairs found in 26, 50, and 76 points randomly sampled from a 100 time point time series of features composed of signals with varying noise, amplitude offset, phase shift, frequency, and coupling. These mixture model tables had signals composed of sine, cosine, sawtooth, and logarithmic patterns.

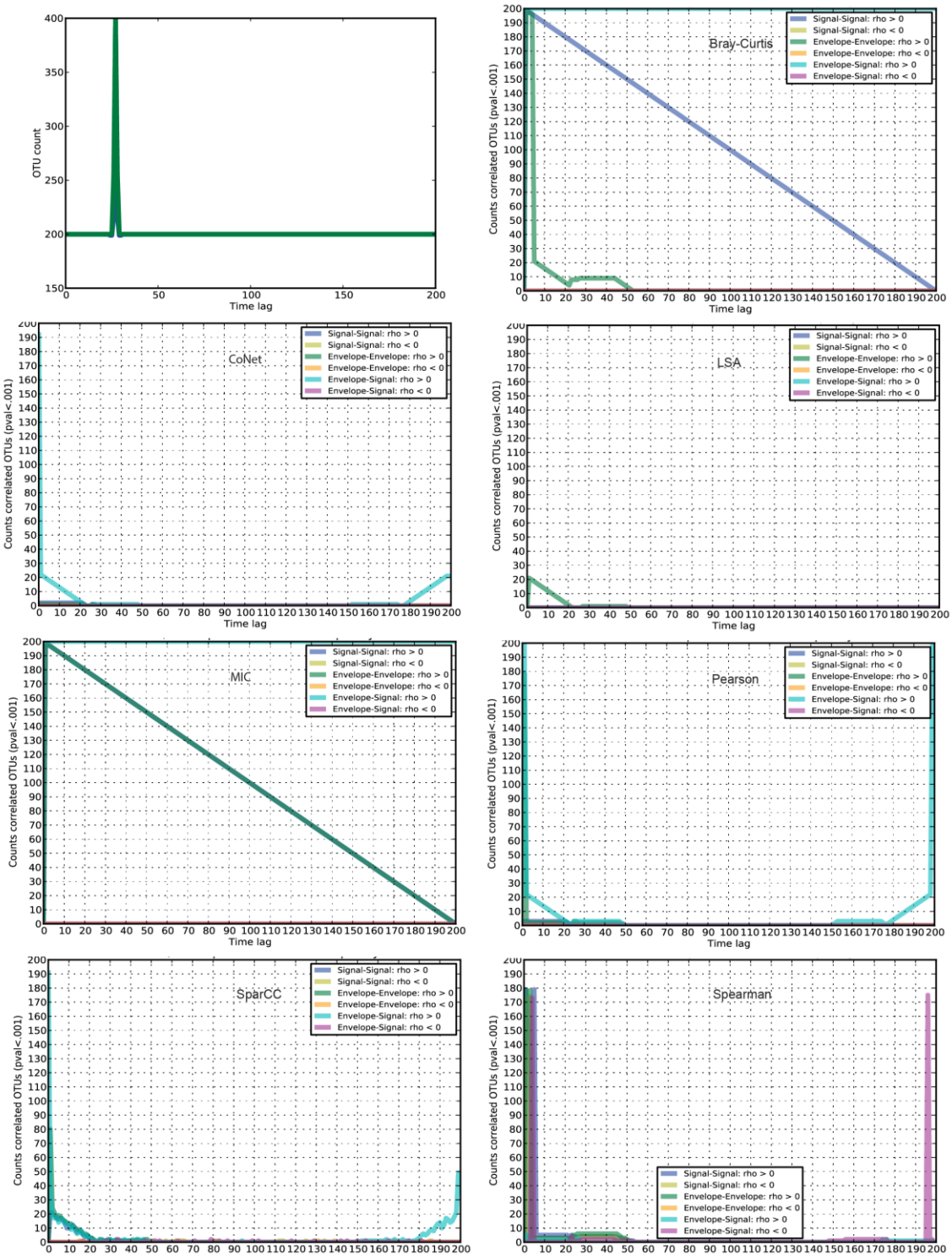


Figure 5-23 Time series correlations: data containing features with 10Hz abundance pulse at some point in time. Top left panel: Example pulse (blue) and envelope (green) features exhibiting a 10Hz spike in abundance. There were 200 pulse features with the abundance peak at times 1-200 of 200 time point samples. There were also 200 envelope features with the same placement throughout time as the

pulses. Other panels show the count of feature pairs (e.g. signal-signal) with a given time lag and correlation sign for each tool. RMT is not included as these signals were too noisy for the technique.

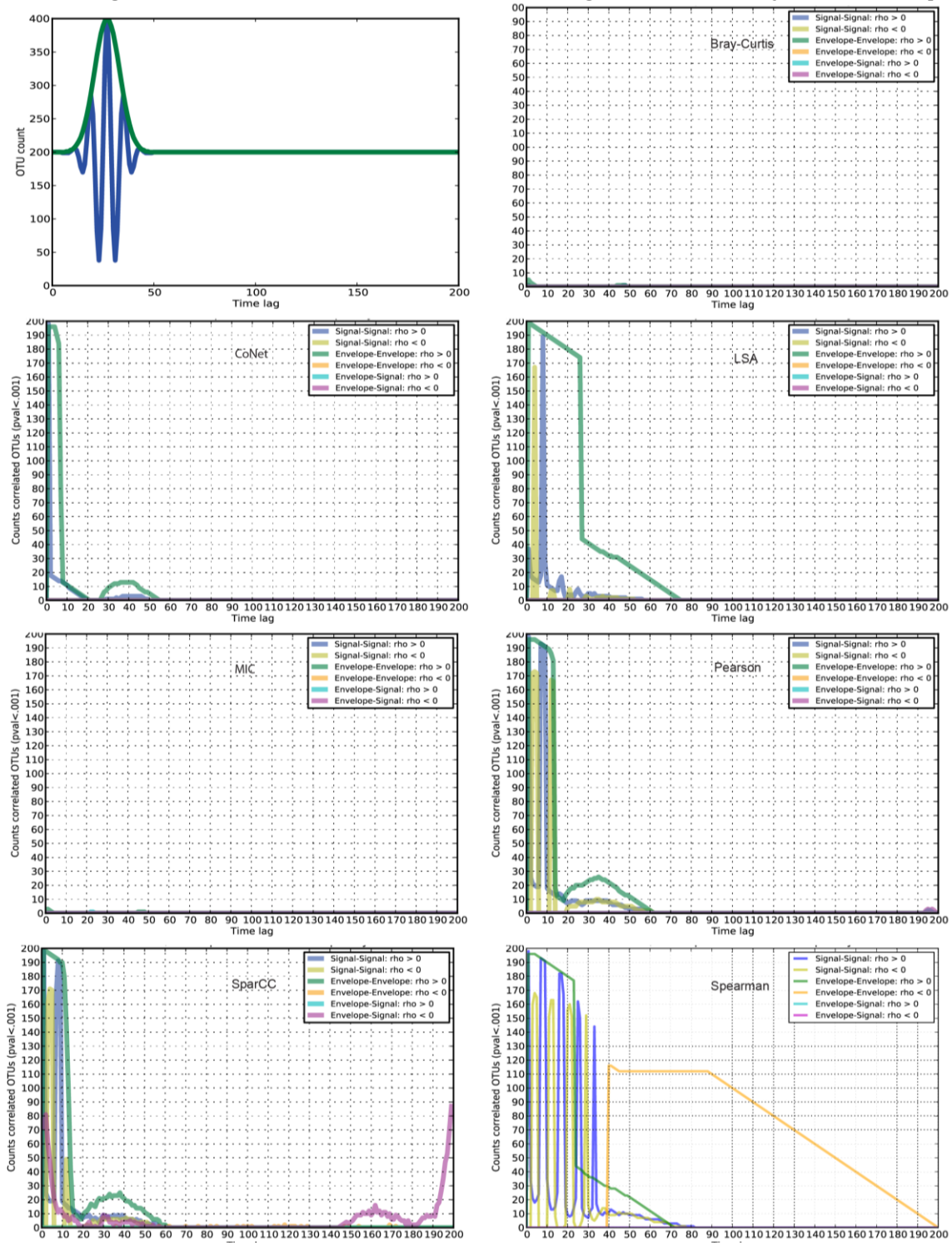


Figure 5-24 Time series correlations: data containing features with 1Hz abundance pulse at some point in time. Top left panel: Example pulse (blue) and envelope (green) features exhibiting a 1Hz pulse in abundance. There were 200 pulse features with the abundance peak at times 1-200 of 200 time point samples. There were also 200 envelope features with the same placement throughout time as the

pulses. Other panels show the count of feature pairs (e.g. signal-signal) with a given time lag and correlation sign for each technique. RMT is not included as these signals were too noisy for the technique.

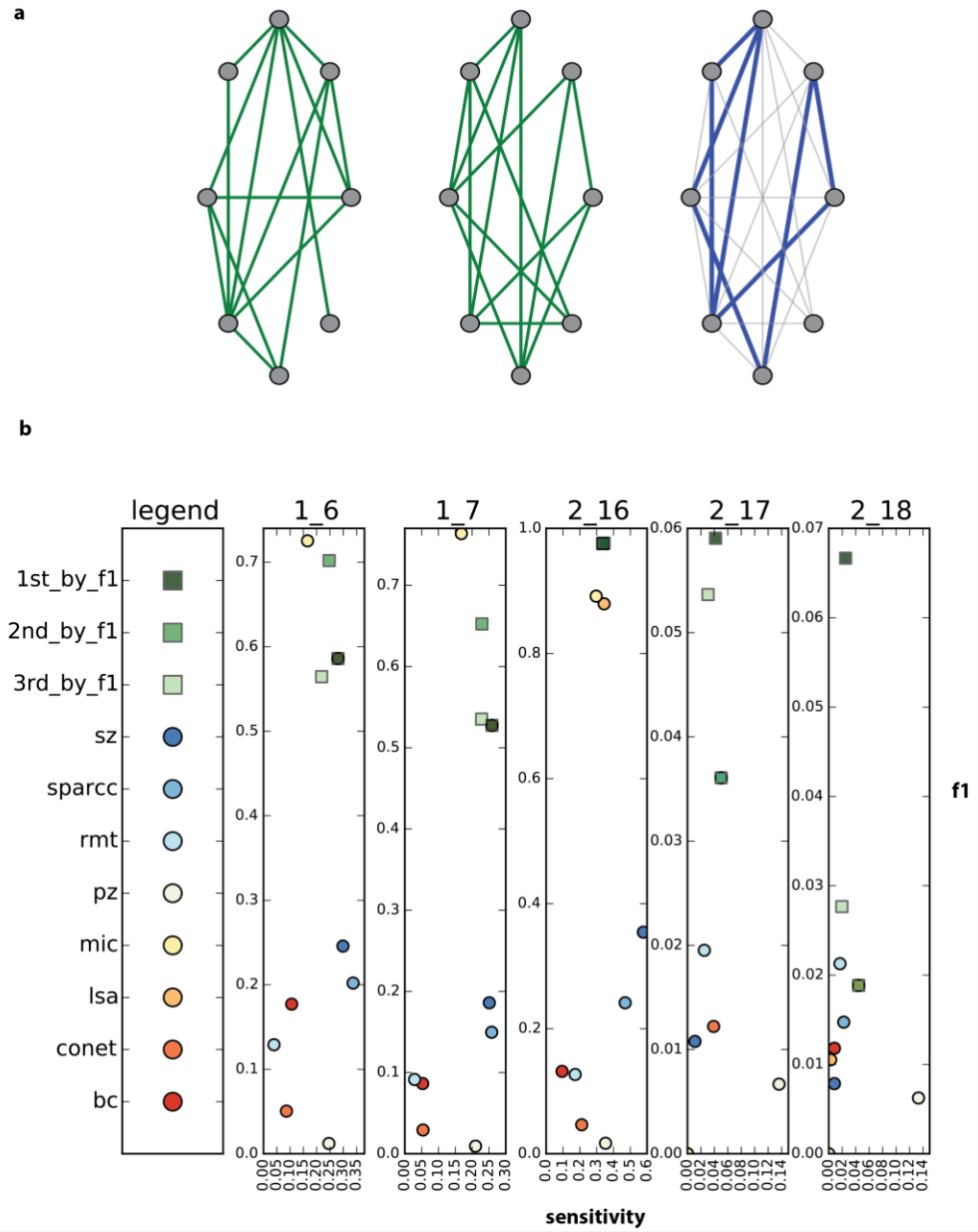


Figure 5-25 Ensemble approach increases precision and the harmonic mean of precision and sensitivity. **(a)** Simple two-tool explanation of ensemble approach. Edges in green are found to be significant by tool one in left network and tool two in middle network. Blue edges in the right network are those edges found by both tool one and tool two. The ensemble approach tested all 2^8 possible one to eight member combinations. **(b)** The top three ensemble approaches ranked by F1 score (harmonic mean of precision and sensitivity) on each linear ecological table type (tables 1.6, 1.7 – two and three species abundance tables - 45% sparse, table 2.16 compositional – 40% sparse, table 2.17 counts - 70% sparsity, table 2.18 compositional - 70% sparse) compared to the tools alone. LSA is hidden beneath the ensemble approaches for the tables 1.6, and 1.7

5.2.8 Ensemble Approaches Boost Precision and the F1 Score

Because different tools detect different edges in the same data, we hypothesized that combining tools for detection purposes in a sensor fusion approach might improve precision. We treat the CoNet approach (Methods), which is itself an ensemble approach of the standard metrics and implements renormalization and permutation (ReBoot) for p-value calculation[179], as one tool. The ensemble approach tested included the toolkits, such as SparCC, and simply calculated the intersection of the edges below a certain p-value, here 0.001, yielded by each technique (Figure 5-25A). In our tests on the linearly ecologically modeled data where engineered correlations are known, the increase in precision is dramatic compared to most tools alone—with many combinations finding zero false positives—at a cost to sensitivity. Although the ensemble shows little gain against MIC or LSA (Figure 5-25B), the gains become larger when sparsity is increased from 40% to a more realistic 70%, although all tools still suffer from drastically decreased sensitivity or hit rate. Our results suggest that an ensemble approach including CoNet and Pearson, or SparCC and Pearson, should be used when precision is required, e.g. for developing hypotheses to test with co-culturing. For non-linear 70% sparse ecological relationships, LSA also has high precision/F1 score.

5.3 DISCUSSION

Correlation detection is an emerging analytical technique that can select biochemically or ecologically relevant pairs of interacting taxa detected using microbial community DNA sequencing. At the highest level, different tools infer markedly different networks from the same input data (Figure 5-1, Figure 5-2). While the potential of this approach is clear, our work shows that current tools have significant limitations that must be accounted for when performing correlation analyses. More specifically, the usual corrected p-value threshold of 0.05 is too

lenient to allow high-precision detection with almost all tools; a stricter threshold such as 0.001 is more useful. Also, processing choices such as sequencing technology type, normalization and filtering have a great impact on which network edges are detected. New strategies must be explored and validated to mitigate the impact of preprocessing on the final inferred network topology. Our results confirm that progress has been made on addressing previously published compositionality effects in the context of low numbers of effective species[24] (meaning that when a few microbes are highly abundant, fluctuations in these dominant abundances changes the resulting correlation networks dramatically because the total number of sequences per sample is constrained by a sum).

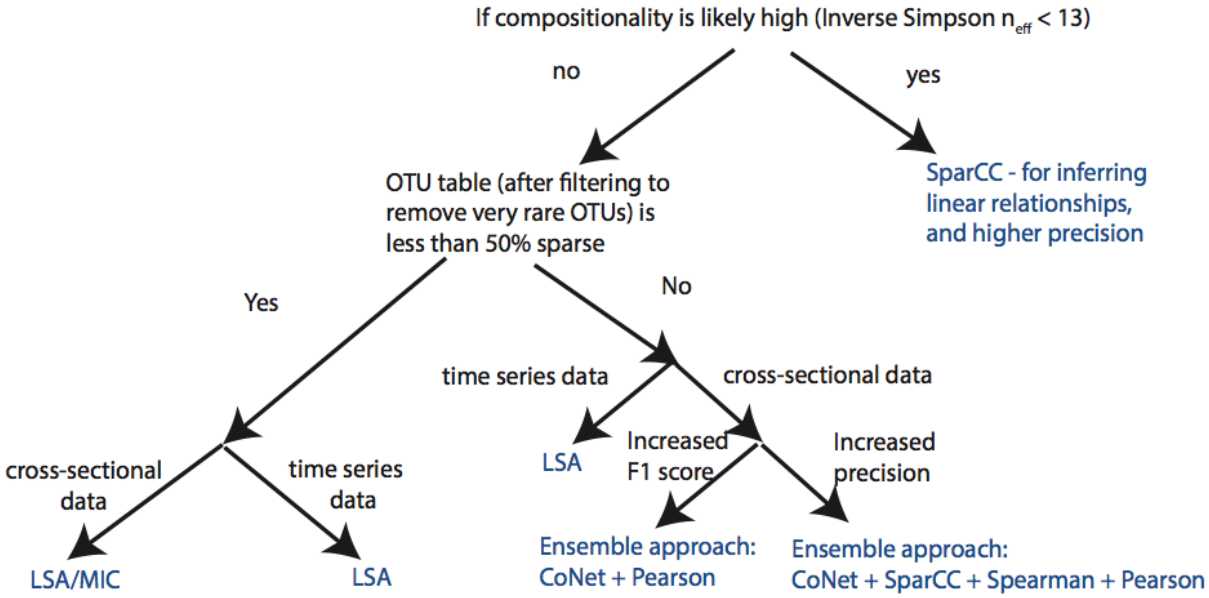


Figure 5-26 Summary of the strengths and weaknesses of each correlation technique.

Encouragingly, all tools have reasonable false positive rates. However, detection of ecological relationships (manifested as abundance dependencies) is very poor for relationships other than commensalism and mutualism (Figure 5-18), and sparsity is perhaps the most significant unaddressed challenge of all (Figures 5-19 and 5-20, panels C and H). No tool performs well with sparsity by any measure, suggesting caution in interpreting the resulting

networks, given that microbiome data sets are almost always more than 70% sparse[28]. Nonetheless, the strengths and weaknesses of each tool are summarized in Figure 5-26, and tool runtime in the Supplementary Note. Under realistic conditions, including both compositionality and 70% sparsity, no technique yet performs well, and an ensemble approach is best for high precision detection of linear relationships in e.g. situations where explicit tests of all hypothesized interactions are prohibitively inefficient. For non-linear sparse relationships LSA alone yields high precision. Finally, while the tools may accurately identify certain overall biological relationships, researchers should be aware of which relationships a given tool is actually capable of detecting: for instance, concluding that a particular microbial community shows no signs of amensal interactions on the basis of a correlation analysis is likely incorrect since none of the tested tools could accurately identify engineered amensal correlations.

Thus, we have identified the strengths and weaknesses of the main microbial correlation analysis techniques, and provide many recommendations for future study and for use of the existing tools, recognizing their limitations. Studies incorporating correlation network analysis will likely continue to increase in number, given the enormous potential significance of identifying interacting taxa. Additional datasets containing experimentally verified microbial interactions will be especially valuable in accelerating progress in this area.

5.4 ACKNOWLEDGEMENTS

W.V.T. and S.J.W. were supported by the National Human Genome Research Institute Grant# 3 R01 HG004872-03S2, and the National Institute of Health Grant# 5 U01 HG004866-04. This work was supported in part by the Howard Hughes Medical Institute (R.K. was an HHMI Early Career Scientist).

5.5 SUPPORTING INFORMATION

5.5.1 Methods

Tools:

CoNet[179]:

All OTUs occurring in fewer than one third of the samples were discarded (except for table set 3, where minimum occurrence across samples was set to 350 for tables 0-22, 28, 30 and 32 – a more lenient threshold because these tables had a large number of samples – and to 50 for table 23, to yield more initial edges). If counts were provided, they were converted into relative abundances by dividing each entry by the total read count of its corresponding sample. For table set 3, a minimum sample sum of 800 was imposed to avoid zero-count samples (except for tables 23 and 34-42, which were much less sparse). If lineages were available, higher-level taxa were assigned up to phylum level by summing relative abundances of lower-level member taxa. Parent-child relationships between taxa were prevented from occurring in all subsequent computations. For each of five similarity measures (Bray-Curtis[183] and Kullback-Leibler[196] dissimilarity, Pearson[184] and Spearman[185] correlation, and mutual information), a distribution of all pair-wise scores was computed. Given these distributions, starting thresholds were selected such that the initial network contained 2,000 positive and 2,000 negative edges supported by all five measures. For each measure and each edge, 1,000 permutation (with renormalization for correlation measures) and bootstrap scores were generated, following the ReBoot routine. The measure-specific p-value was then computed as the probability of the null value (represented by the mean of the null distribution) under a Gauss curve generated from the mean and standard deviation of the bootstrap distribution. Since a one-sided test was carried out, p-values close to one were considered indicative of mutual exclusion and converted into low p-values by subtraction from one. Next, measure-specific p-values were merged using Brown's method[190],

which takes dependencies between measures into account. After applying Benjamini-Hochberg's[48] false discovery rate (FDR) correction, edges with merged p-values below 0.05 were kept. Any edge for which the five measures did not agree on the interaction type (i.e. positive or negative) or whose initial interaction type contradicted the interaction type determined with the p-value was also discarded. Edges with scores outside the 95% confidence interval defined by the bootstrap distribution or not supported by all five measures were discarded as well.

RMT[187]:

Initially proposed by Wigner and Dyson[45, 49] for studying the spectrum of complex nuclei, RMT is a powerful approach for identifying and modeling phase transitions associated with disorder and noise in statistical physics and materials science. It has be successfully used for studying the behavior of different complex systems, such as the spectra of large atoms[45, 49], metal insulator transitions in disordered systems[51, 205], spectra of quasi-periodic systems[206], chaotic systems[207], brain response[208], and the stock market[209]. It was first adopted for delineating gene expression networks[210, 211].

All RMT calculations were implemented through the Molecular Ecological Network Approach Pipeline (MENAP) at http://ieg2.ou.edu/MENA. All OTUs occurring in fewer than half of the samples were discarded except in table set 3 where minimum occurrence across samples was 50 in the 2,000 total samples. Since RMT requires that more than 80 OTUs remain after removing the above OTUs, a few of the tested tables were not analyzed. Thereafter, Pearson correlation coefficient (r value) was calculated between each pair of OTUs and a symmetric similarity matrix was formed after all r-values were calculated. Theoretically, the RMT approach is applicable to any similarity matrix[187], but here it was only used to automatically detect a reliable cutoff for the Pearson correlation matrix based on the χ^2 test with Poisson distribution.

The threshold for defining a network is mathematically determined by calculating the transition from Gaussian orthogonal ensemble (GOE) to Poisson distribution of the nearest neighbor eigenvalues, and hence the network is automatically defined based on the data structure itself. In order to control the false positive rate, the most stringent thresholds (significance of $\chi^2 > 0.05$) were set for the tests.

MIC[182]:

Maximal Information Coefficients (MIC) were calculated with default parameters in minerva, an R wrapper for the cmine implementation of Maximal Information-based Nonparametric Exploration statistics, to quantify the linear or non-linear association between pairs of OTUs. An empirically based approach was taken for p-value calculation; for example, with a p-value threshold of 0.001, we chose the MIC threshold that made the top 0.001 (one-thousandths) of the edges significant. Bonferroni multiple-test correction was applied[197].

LSA[175, 212]:

The eLSA analysis was run with the program's default parameters, i.e., with no delay allowed (delayLimit=0), p-value calculated by theoretical approximation (pvalueMethod=theo), required precision of p-value as 1/1,000 (precision=1,000), and data rank-normalized and Z-transformed (normMethod=robustZ). Multiple-test correction was done using q-values[213].

The theoretical p-values approximate the statistical significance of local similarity analysis based on the tail distribution of excursion range of random walk. The approximation works reasonably well (starting at time points n>10 with no delay) and provides P-values comparable to those from permutations. One significant advantage of theoretical p-values is that it enables constant time calculation of statistical significance for pairwise local similarity analysis, making possible all-to-all comparisons for high-throughput data otherwise prohibitive.

SparCC[24]:

The tool was run with default parameters and 500 bootstraps. Pseudo p-values were calculated as the proportion of simulated bootstrapped datasets with a correlation at least as extreme as the one computed for the original dataset.

Pearson[184] and Spearman[185] correlations:

Fisher z-transformation was used to calculate p-values[194, 214].Bonferroni multiple-test correction was applied[197].

Bray-Curtis[183]:

An empirically based approach was taken for p-value calculation; for example, with a p-value threshold of 0.001, we chose the correlation threshold that made the top 0.001 (one-thousandth) of the edges significant. Bonferroni multiple-test correction was applied[197].

Models:

Copula:

This model enables generation of random variables having a specified covariance matrix from a given distribution[189]. The particular copula method we used is the Gaussian copula, which is founded on the fact that applying the normal cumulative distribution function (CDF) to a standard normal random variable results in a uniform random variable between 0 and 1. Inverse transform sampling then enables the creation of any distribution by applying that distribution's inverse CDF to a uniform random variable between 0 and 1 (Figure 5-4)[189]. The copula function controls the joint distribution of the random variables and their rank correlations. Real covariance matrices are symmetric and positive definite; therefore the Cholesky decomposition is used to test for positive definiteness and so ensure meaningful OTU generation.

Null model:

This model was used to generate data tables from null distributions of several types to support testing the false discovery rates of various tools. Three methods were implemented. In method 1, the OTU table was created by randomly drawing sample vectors from a given distribution and parameters. In method 2, the OTU table was created with compositionality in mind and therefore the sum of each sample is constrained. Tables are either not sum-constrained (providing raw abundance) or sum-constrained (providing relative abundances by dividing each OTU by the total number of sequences in its sample) and were produced by the Dirichlet distribution. In method 3, the OTU table was created with compositionality in mind, similar to model 2, but with higher sparsity than is normally created with the Dirichlet procedure by subtracting the mean value of the table from all entries (entries < 0 = 0).

Ecological:

This model was used to create tables with simple (ecologically based) relationships between OTUs to test if the tools can accurately recapture relationships that are defined by a mechanism rather than by a high correlation score. The reason we chose this method is because we wanted a way to assess if relationships we know to exist in biological contexts can be revealed through correlation analysis as frequently reported. The types of ecological models tested were amensal, commensal, mutual, parasitic, competitive, and partial-obligate-syntrophic. All interactions were linear and dependent upon OTU abundance.

1. The amensal model depresses the abundance of OTU2 when OTU1 is present by strength*OTU1; OTU1 is unaffected by the presence of OTU2.
2. The commensal model increases abundance of OTU2 when OTU1 is present by strength*OTU1; OTU1 is unaffected by the presence of OTU2.

3. The mutualism relationship increases abundance of both OTUs when both are present; the strength of increase in each OTU is proportional to the abundance of the other OTU.

4. The parasitism model increases the abundance of OTU1 and decreases abundance of OTU2 when both present. Thus, OTU1 grows at the expense of OTU2 with strength proportional to the abundance of OTU2.

5. The competitive model depresses the abundance of both OTUs if both are present. This simulates OTU competition for some limiting resource, with the strength of each OTU's decrease proportional to the abundance of the other OTU.

6. The obligate-syntrophy model allows OTU2 only when OTU1 is present, at an abundance proportional to the relationship strength. This mimics a relationship where OTU2 depends on the presence of OTU1 and cannot exist without it.

7. The partial-obligate-syntrophy model allows OTU2 if and only if OTU1 is present. This is similar to obligate syntrophy except the presence of OTU1 does not necessarily mean OTU2 is also present.

Lotka-Volterra:

These are systems of n differential equations that model the dependencies and interactions of the abundances of n species. The most widely used is a simple 2-species system of equations modeling predator-prey (e.g. fox and rabbit) abundances (Figure 5-21A-F), developed by Volterra himself[215]. The behavior of the Lotka-Volterra equations is much less understood for systems larger than two species; for example, starting with the three-species equations, the system dynamics become much more complex[216], chaotic behavior may occur, and a solution may not converge. Therefore, for the six-species equations in this paper, we used only small variations of the six-species systems of equations explored by Idema[216]. Because of

the system complexity, small variations in the interaction matrix lead to very different abundance patterns (Figure 5-21G-I).

Time Series:

This model was used to create OTU tables with simple time-series relationships. All signals take the form of: y_shift + alpha*signal_function(phi(theta+omega)) + noise, where alpha is the amplitude, phi is the frequency, and omega is the phase shift, and theta is the time parameter. Options to subsample the waves at even, randomly selected indices, as well as to add sparsity are included.

table set construction:

Methods for table set 1

All tables can be found at: https://github.com/wdwvt1/correlations. Tables 1 and 2 were created with the copula method with margins from the lognormal ($\mu = 3$, s.d. = 0) and gamma (shape parameter = 1, location = 0, lambda = 100) distributions, and with rho matrix entries ranging from [-0.01, 0.02]. Table 4 was created with the null model and no compositionality. It was created by random calls to the lognormal, gamma, Nakagami, uniform, and chi-squared functions – again, distributions that could mimic bacterial growth and real OTU table sparsity, although the overall sparsity was still lower than in reality. Table 5 was created with OTUs from a Dirichlet distribution where the prior counts were given by random variables with a lognormal distribution. Tables 6 and 7 were ecological tables, having competitive, mutual, commensal, amensal, parasitic, obligate, and partial-oligate-syntrophic relationships of various strengths (2, 3, and 5) as well as two-species (OTU1 acts on OTU2) and three-species (OTU1 and OTU2 together act on OTU3) interactions.

Methods for table set 2

Tables 1-5 are time-series tables with changing frequency, amplitude, phase, noise, and subsampling routine. Table 1 is OTUs with sine wave variations, while table 2 is OTUs with a square wave for half the samples and a cosine wave for the other half of the samples. Table 3 is a half-sampling of the table 2 OTUs, table 4 is OTUs composed of sawtooth/cosine summations, and table 5 is OTUs made of a significantly undersampled sawtooth wave added to a low-frequency wave. Tables 6-10 are two-species Lotka-Volterra relationships, and tables 11-15 are six-species Lotka-Volterra relationships[215, 216]. All Lotka-Volterra relationships are n-species abundances described by n systems of differential equations mimicking interesting ecological relationships, such as predator-prey (Figure 5-21). The Lotka-Volterra relationships in tables 6-15 were padded and confounded with random OTUs from lognormal and gamma distributions. From the values generated with the Lotka-Volterra equations and confounding OTUs, tables 6 and 11 were made into relative abundance tables with points taken at equal intervals, while tables 7 and 12 were the same as 6 and 11 except the values were counts instead. Tables 8 and 13 were relative abundance tables with points taken at random indices, and tables 9 and 14 were the same as tables 8 and 13 except the values were counts instead. Tables 10 and 15 were generated from the same system of differential equations as tables 6-9 and 11-14 respectively, except 60% of the values were randomly set to zero. Tables 16-18 were again ecological tables but with one-dimensional linear relationships only. The values were relative abundance, 50% sparsity, and relative abundance of the 50% sparsity table respectively. Tables 19-21 were copula tables drawn from lognormal ($\mu = 3$, s.d. = 0), gamma (shape parameter = 1, location = 0, lambda = 100), and exponential ($\mu = 0$, lambda = 1,000) distributions using the same generating rho matrix as tables 1 and 2 from table set 1.

Methods for table set 3

Sequences for the study "Cultured gut bacterial consortia from twins discordant for obesity modulate adiposity and metabolic phenotypes in gnotobiotic mice" by Ridaura et. al.[4] were retrieved from the QIIME database[29] and picked with default closed-reference settings (QIIME 1.7-dev, GreenGenes[152] reference database v. 13_5) at 97 percent similarity. Briefly, 10 independent rarefactions were conducted at 1,000 ,sequences/sample and 10 at 2,000 sequences/sample using the QIIME script 'multiple_rarefactions_even_depth.py'[29]. These formed tables 0-9 and 10-19 (respectively) of table set 3. Tables 20-23 were created by taking table 0 (described above) and filtering out OTUs that did not occur in some percentage of samples (table 20, 21, 22, 23; 5, 10, 20, 50%). Tables 24-26 were created by filtering the unprocessed OTU table (described above) to eliminate OTUs whose overall sequence count was below a percentage threshold (a suggested step in Bokulich et al[188]) and then rarefying at 1,000 sequences/sample (table 24, 25, 26; 0.00005, 0.00010, 0.000025%). Table 27 was created by taking table 24 and performing the additional step of removing OTUs found in less than 20 percent of the samples. Table 28 was created by summarizing OTUs from the raw unprocessed table at L6 (genus level) using the QIIME script 'summarize_taxa.py'[29]. The table was then rarefied to 1,000 sequences/sample, and OTUs not found in at least 20% of samples were removed. Table 29 was created by picking from the Ridaura et. al.[4] sequences (described above) using the same parameters except that the similarity threshold for OTU clustering was reduced to 94% (the genus level). The resulting table then underwent the same processing steps as table 28. Tables 30 and 31 were the same as 28 and 29 except the summary was conducted at L5 (family level), and the similarity threshold was reduced to 91 percent (respectively). Tables 32 and 33 were again the same, but with summary at L4 (order level) and similarity threshold of 88% (respectively).

Tables 34-43 were created with the generator methods described in the main text. All of these tables have periodic signals that are composed of sine, cosine, and square waves (superimposed, in some cases) as well a logistic growth curve and a Gaussian pulse and envelope. There are 6 parameters that are varied in these tables (other than the signal function): frequency, amplitude, phase, noise, sampling routine and sparsity. The sampling routine is either to evenly space the points in time, to randomly draw an ordered subset, or to draw an evenly spaced subset and then randomly select a fraction of those samples to be zeroed (abundance = 0). For table 34, frequency is varied from 0.25 to 200 (arbitrary units), phase is varied between 0 and pi/2, and the subsampling routine is varied between even, random, and even with zeroing, while other parameters are held constant. There are sin, square, sin for half into square for half, and logistically growing OTUs. Table 35 is the same as table 34 in all respects except the pseudo-random number generator is set to a different seed and the percentage of subsampling is doubled (50 samples instead of 26). Table 36 is again the same but with subsampling again increased from 50 to 74 samples. Table 37 is a half-sampling (evenly) of table 34, table 38 is a half-sampling of table 35, and table 39 is a half-sampling of table 36. Tables 40-42 have OTUs that are constructed as Gaussian pulses and their envelopes. The frequency of the pulse is varied (table 40, 41, 42; 1, 10, 0.1hz).

Methods for table set 4:

An OTU table was generated with the copula model and lognormal distribution, with the rank correlation matrix specified as having all OTU correlations close to zero. Then six positively correlated OTUs were added, having rank correlations greater than 0.2. Six negatively correlated OTUs were added as well, with rank correlation less than -0.2. The effective number of species (n_{eff}), calculated with the inverse Simpson alpha diversity measure, in this table (table

0) was 36. Four more tables (tables 1-4) were created by replicating table 0, but multiplying one OTU by a constant factor such that the n_{eff} of the resulting tables was 25, 19, 10, and 4, respectively. Tables 0-4 were taken to be the absolute abundances, reflective of the microbial correlations in the natural environment. Compositionality was then induced, reflecting the sampling/sequencing process, by rarefying tables 0-4 at a depth of 2,000 sequences/sample to create tables 5-9. To test the effect of rarefying at a lower depth, tables at 1,000 sequences/sample were created. To test the effect of alternate normalization techniques designed to correct for compositionality, we also created CSS-normalized[28] (tables 10-14) and DESeq-normalized[89, 90] (setting the negatives to zero as in McMurdie and Holmes[215], tables 15-19) versions of tables 5-9.

5.5.2 SUPPLEMENTARY NOTES

Supplementary note for ecological data

We assessed the tools on their ability to detect simple two-species ecological relationships (two features, one edge) when the data were presented as unaltered (tables 1.6 and 1.7), compositional (table 2.16), sparse (table 2.17), or sparse and compositional (table 2.18) , to maximally confound the tools. In general, the tools performed reasonably well but precision was low—on average 0.25 for tables with 40% sparsity, and 0.01 for tables with 70% sparsity (tables 2.17 and 2.18). For instance, using the common p-value threshold of 0.05 for p-values calculated from Spearman correlation with Fisher z-transformation (SZ) resulted in a precision of 0.021 (table 2.16): for every correctly detected edge in this network there would be 50 incorrect edges. For unaltered or compositional data, LSA and MIC were the most precise by far (with precision 0.54, and 0.79 respectively), but this degraded when sparsity was added. A combination of tools was the most precise for tables with realistic sparsity levels (tables 2.17 and 2.18). Specificity

was fairly high with an average close to one across all tools and ecological tables. Sensitivity was relatively low, with an average of 0.22 for tables with approximately 40% sparsity (tables 1.6, 1.7 and 2.16), and 0.03 for tables with 70% sparsity (tables 2.17 and 2.18).

In these ecological comparisons, we also assessed the performance of the tools on different types of ecological relationships. The detection profiles for the different ecological relationships were striking, with amensal and partial obligate-syntrophic relationships virtually undetectable by any tool and mutual relationships detectable by all tools (Figure 5-18A, Figure 5-19B-C, Figure 5-20B-C). To determine if the strength of a relationship played a role in its detectability, our unaltered data (mentioned above) contained 90 interactions for each of the ecological relationships (e.g. 90 different OTU pairs related in a mutualistic way) split into 3 groups of 30 that were each generated with different strengths (higher strength corresponded to more change from the background distribution and a cleaner signal). For amensal edges, only SparCC and SZ with permissive p-value thresholds detected more than ~10% of all available edges. Furthermore, in contrast to the other relationships types, there was no correspondence between the strength of the edge relationship and the detection probability. For competitive edges, SparCC, LSA, and SZ all performed well, and detected more edges as the strength of relationship increased. CoNet, RMT, and Pearson with Fisher z-transformation (PZ) were functionally unable to detect competitive relationships. For commensal or mutually related edges, SparCC, LSA, SZ, and PZ performed well, with CoNet performing at an intermediate level and RMT finding no edges. Parasitic edges were best detected by PZ and SZ, and had intermediate detectability with the other tools except RMT, which did not find any of these edges.

We also tested detection profiles of the tools for more complex (but still linear) ecological relationships (Figure 5-18B, Figure 5-19A, Figure 5-20A). In these relationships, we required two or more OTUs to be present to cause an interaction and a modification to a third OTU (or more). Ecological literature suggests that there are likely important relationships mediated by more than two members[200], and we tested a simple case of this. In general, the detection profiles of the three-species relationships were similar to those in the two-species case. SparCC, LSA, and SZ more easily identified the three-species competitive relationships than their two-species counterparts (the same was true of PZ, but it had minimal detection of either). Parasitic three-species edges were identified well, but the correlation patterns were hard to interpret; edges which we *a priori* assumed would be assigned as negatively co-occurring were positive and vice versa. This suggests that the non-linearity of multiple OTUs interacting in a network can confound assignment. Mutual three-species OTUs were discovered with high efficiency by most tools. However, detection of any of the above two and three-species ecological relationship types decayed to little better than random guessing when the sparsity in the OTU table was raised to realistic levels.

The importance of determining which tool is best at finding which relationships is clear when one considers the post-hoc way in which correlation networks are used. For example, given the knowledge that SparCC can detect competitive relationships more easily than amensal relationships, negative edges (mutual exclusions) in a SparCC-generated correlation network should be interpreted as competitive interactions between taxa rather than amensal ones.

Supplementary note for time series discussion

Work remains to determine the optimal sampling frequency to capture as much of available microbial signals as possible. This can dramatically affect results, as demonstrated in

Figure 5-22. With fewer time points taken for a given 100 time-unit signal (76, 50 and 26 points respectively), the tools generally found fewer edges. The main exceptions to this were CoNet, RMT, Bray-Curtis, and MIC. Bray-Curtis and MIC found very few edges, suggesting that they are not very sensitive to time-series relationships. This implies that different signals are construed from the actual signal depending on the sampling frequency, greatly affecting OTU pairs deemed to be co-occurring. CoNet and RMT were relatively stable across sampling frequencies.

Simple time-shifted OTU relationships were also tested. These data sets were composed of OTUs exhibiting a pulse (a sharp increase in abundance for some of the time points) or envelope (smoothed single wave outlining the maximum values of a pulse). Most measures only considered OTUs displaying pulses at similar times as correlated (Figure 5-23, Figure 5-24). These OTU pulse tables were too noisy for RMT to evaluate, even though in one table the pulse was sustained over 50 samples. Bray-Curtis and MIC did not detect OTUs exhibiting high-frequency pulses, and did not distinguish between lower-frequency time-shifted signals.

Supplementary note for tool timings

Rough estimates of correlation technique were run time on a local 64-bit machine using a1053 feature x 257 column OTU matrix.[195] Pearson with fisher z-transform: 96s*, Spearman with fisher z-transform: 267s*, Bray-Curtis: 10s*, LSA: 6153s**, CoNet: 3826s**, MIC: 457s**, RMT: 284s***, SparCC correlations calculation: 107s* for all iterations and for 589 OTUs after filtering those with relative abundance less than .001 and those not found in at least 5 samples. Note that SparCC p-value calculations scale with the number of permutations, so for 100 permutations it takes 100*107 = 10700s*, or about 3 hours. Thus, it is much faster to threshold only by correlation value of 0.35, with no significant difference in results as a 0.01 or

0.001 p-value threshold (data not shown).

Timings were done with * 8GB, **16GB, ***64GB memory, respectively.

Chapter 6

Parallel Mapping of Antibiotic Resistance Alleles in *Escherichia coli*

PLoS One, submitted, 2015

6.1 INTRODUCTION

Chemical genomics, or the study of the genome-scale response to small molecules, has rapidly advanced thanks to synthetic biology approaches. For example, studies of phenotype mapping of small molecule landscapes have led to elucidation of novel genetic functions and drug mechanisms [15-17]. These pioneering studies took large genomic libraries, usually painstakingly created [217, 218], and characterized them under a range of chemical and physical conditions using DNA microarrays. Studies of chemical tolerance have also used adaptive evolution methods to identify mutations that contribute to fitness [38, 219]. While these studies closely mimic responses to stresses in nature, the extent of genotyping is limited by the throughput of whole-genome sequencing.

The increasing throughput and decreasing cost of multiplex oligonucleotide synthesis [39] and high-throughput sequencing [40] has enabled unprecedented advances in throughput of genome engineering and analysis technologies [18, 41-43]. For example, recent studies have leveraged high-throughput sequencing to expand the characterization of yeast deletion libraries [220]. Along these lines, we recently reported the trackable multiplex recombineering (TRMR) approach [19]: a one-pot construction of a barcoded, genome-scale library simulating overexpression and knockdown of over 4,000 genes in the Gram-negative bacterium *E. coli.* Initial experiments with the library focused on the genomic response to various carbon sources and biofuel-related inhibitory conditions using DNA microarrays and exploratory by-hand analyses [19, 221].

At sub-lethal antibiotic concentrations such as those found in wastewater and agricultural runoff, the contribution to microbial fitness of cellular factors is not nearly as well-studied [35] as horizontal gene transfer of specific resistance effectors [34]. Thus, understanding the response and resistance of microbes to antimicrobial compounds is of critical importance. To isolate gene products contributing to antibiotic resistance, several genomic and proteomic studies have been performed [151, 222-225]. However, previous attempts to characterize genome-scale responses to antibiotic challenges [15, 17, 36-38, 219] relied on either (1) the low-throughput construction of large libraries or (2) many generations of adaptive evolution, where characterization was limited by sequencing surviving colonies.

Here we report a method for the rapid and deep characterization of laboratory population dynamics in response to eight antibiotics by multiplex selection, next-generation sequencing, and multivariate analysis of *E. coli* TRMR libraries. Our findings support the development of multi-drug resistance and susceptibility genes as an important step in the evolution of antibiotic resistance in microbial populations at sub-lethal concentrations. Finally, to expand the throughput and extent of our bioinformatic analysis, we integrate the data gathered into the QIIME multivariate analysis pipeline, with which we examine the response at a pathway level and identify a unique genomic signature for each antibiotic.

6.2 METHODS

6.2.1 Strains and Plasmids

The TRMR library was previously constructed [19]. Briefly, *E. coli* MG1655 cells were subjected to multiplex recombineering using synthetic DNA cassettes containing either an “up” (strong promoter and RBS) or “down” (no promoter or RBS) phenotype with homology regions corresponding to over 4,000 genes in the *E. coli* genome. The synthetic cassettes also contained

unique barcodes for rapid characterization and gene-trait mapping. In this study, a modified version of strain JWKAN, which is MG1655 with the kanamycin resistance gene *neoR* (from pKD13 [218]) inserted in a safe region and barcoded as in the rest of the library, was used as the wild-type control. Expression of FLP recombinase (pCP20 [226]) excised *neoR* from the genome using flanking FRT sites to create a barcoded MG1655 without kanamycin resistance, which we refer to as MG1655-BC. This phenotype was confirmed by replica plating and the genotype confirmed by colony PCR.

6.2.2 Antibiotic MIC Determination

Overnight cultures of MG1655-BC cells were subcultured into various concentrations of antibiotics in MOPS media [227] at 37C to determine the minimum inhibitory concentration (MIC) for each compound. All antibiotics were purchased from Sigma-Aldrich (St. Louis, MO). The MIC for each antibiotic was determined by an iterative process using the procedures and definitions of Andrews [228]. First, an estimate was determined by growing MG1655-BC in 96-well plates in triplicate in 2-fold increments around the MIC found in the literature (if any) [228]. The 2-fold determined MIC was then refined by growth in 1.2 fold increments. The refined MIC was used for liquid culture in MOPS media in 250 mL flasks, inoculated at OD_{600} 0.02 with MG1655-BC or the recovered TRMR library. The final MIC concentration was determined to be the concentration at which MG1655-BC showed no growth and the TRMR library showed significant (OD_{600} > 0.2) growth at 24 hours.

6.2.3 Cell culture and selection conditions

The TRMR “up” and “down” libraries were recovered from frozen stocks by inoculating glycerol stocks of the constructed libraries in low salt LB media with 90µg/mL blasticidin-S to OD_{600} 0.4. The cells were grown at 37C in a shaking incubator to an OD_{600} of approximately 0.8.

When the initial TRMR and MG1655 cultures reached the desired OD_{600}, they were transferred to two identical sets of three selection flasks containing 50 mL MOPS media at 80% of the previously determined MIC (sub-inhibitory selection concentration, SSC) for each of the eight antibiotics (for 48 flasks total) tested to an OD_{600} of approximately 0.02 [19]. TRMR "up" and TRMR "down" libraries were added in equal amounts as determined by OD. These initial cultures were then harvested by centrifugation and frozen as pellets for initial concentration values, which we refer to as time point zero. Growth proceeded under antibiotic selection conditions at 37°C and cells were harvested by centrifugation after 24 hours and upon reaching a 1.5 OD_{600}.

6.2.4 Antibiotic Colony Sequencing

Individual colonies from each selection were amplified including the barcode tags by PCR. All PCRs were performed using Phusion polymerase (NEB). The PCR product was confirmed to correspond to the barcode region by gel electrophoresis. The DNA was then purified using a QIAquick gel extraction kit (Qiagen), and sent for Sanger sequencing (MWG Eurofins Operon). The incorporated tag sequence was compared with Supplementary Table 1 of Warner et al. [19] to identify alleles. For high-throughput sequencing, the genomic DNA from 10^9 cells from all the selections was extracted using the DNeasy Blood & Tissue Kit (Qiagen). Four base-pair tags were appended using PCR near the beginning of each TRMR-unique barcode to further distinguish the samples by replicate. PCR products of roughly 180 bp were gel-extracted and purified using the QIAquick gel extraction kit (Qiagen), and combined in equimolar amounts. The resulting mixture of amplicons from all replicates and time points for each antibiotic sample was assigned a unique Illumina index and prepared for sequencing according to Illumina TruSeq 1x50 guidelines [229] and sequenced on an Illumina HiSeq 2000.

6.2.5 Sequencing Data Analysis

Each FASTQ file produced by the high-throughput sequencing was read and signal quality filtered in parallel using a custom MATLAB script. The 50 base pair reads were matched to 50 base pair DNA sequences in a mapping file corresponding to the expected barcodes in genomic context. These sequences included a four base pair tag for replicate and experiment identification, as well as the unique TRMR tag sequences for each gene as found in Supplemental Table 1 of Warner et al [19]. Any FASTQ sequence not matching those in the mapping file within 1 bp was discarded to allow distinguishing between the replicates while minimizing spurious mapping of sequences to genes. This strict quality filtering meant only 10-40% of the sequences in each FASTQ file were retained.

Inherent bias in construction and limited sampling meant that not every allele appeared in the naïve (unselected) cases. Thus enrichment (fitness) in this study was defined as the relative increase in a particular allele after selection with respect to the naïve population according to the following formula for enrichment of a given allele A.

$$enrichment_A = \left(\frac{counts_A}{\sum_n counts_n} \right)_{selection} - \left(\frac{counts_A}{\sum_n counts_n} \right)_{naive}$$

The "top" alleles described are the alleles in each selection case with the highest fitness over the naïve case.

6.2.6 Bioinformatic Analysis

Most analyses were performed using the QIIME (Quantitative Insights into Microbial Ecology) pipeline, version 1.7.0 [29]. The open-source QIIME pipeline was built using the PyCogent libraries [230] and the Python programming language. QIIME analyses are performed though a simple command-line interface, where the input and output file paths are specified, as

well as any method parameters. QIIME was used for all of the following analyses: normalization, formation of a distance matrix, principal coordinates analysis (PCoA), Procrustes analysis, supervised learning, part of the network analysis, COG relative abundance plots, and ANOSIM. The QIIME scripts used for the above and below list of analyses were: single_rarefaction.py, normalize_table.py, beta_diversity.py, principal_coordinates.py, transform_coordinate_matrices.py, supervised_learning.py, make_otu_network.py, summarize_taxa_through_plots.py, and compare_categories.py. All of these QIIME scripts use as input the table of gene counts in each sample, and corresponding metadata, or products from previously used scripts (e.g. beta_diversity.py should be used before principal_coordinates.py).

First, the sequencing tables were normalized. Normalization is necessary to correct for uneven library sizes, as well as other artifacts of the sequencing process. [112] [28] The tables were subsampled (rarefied) to a depth of 2000 sequences per sample. Another normalization method implemented in R and QIIME, metagenomeSeq's cumulative sum scaling (CSS), was performed in order to ensure robustness of results [28]. Next, a distance matrix was formed using Bray-Curtis dissimilarity [183, 231], since antibiotics selecting for the same genes should be deemed more similar, and because Bray-Curtis is less sensitive to data sparsity and compositionality [23, 24, 179]. Then, PCoA was performed on the distance matrices. We also assessed the results using Euclidean and binary Jaccard metrics with similar results.

Procrustes analysis, which enables comparison of the relative distances between points in two multivariate datasets, [232] was also performed on the gene and COG distance matrices. The measure of fit (M^2) was calculated as the sum of the squared distances between corresponding sample points after the data is translated, rotated, and scaled to minimize the distance between the two datasets. The p-value was calculated by 1000 Monte-Carlo permutations in which the

sample labels were randomly permuted; the number of iterations in which the M^2 value was lower than the actual was divided by 1000 to yield the p-value.

Supervised learning was performed in QIIME using the random forest machine learning method [46], with 5,000 sequences per sample, 500 trees, and leave-one-out cross-validation to estimate the generalization error and feature importance [47, 233]. Plots of alleles based on genomic location were generated using Circos software [234]. Genes were annotated with their corresponding Clusters of Orthologous Groups (COGs) [235]. Relativized counts were plotted using the summarize taxa scripts in QIIME [29]. Networks were constructed using Cytoscape [236]. ANOSIM was also carried out in QIIME [237] using the 'vegan' package in R {Dixon, 2003 #398}.

6.2.7 Databank Submission

Raw .fastq files have been uploaded to the NCBI Sequence Read Archive (SRA), accession number SRP047041.

6.3 RESULTS AND DISCUSSION

6.3.1 Selection of Antibiotic-Resistant Alleles From a Genome-Scale Library

We subjected our genome-scale, barcoded library to selection on eight different antibiotics with three different mechanisms of action (Table 6-1). Pairs of antibiotics were selected for chemical similarity (e.g., ticarcillin differs from carbenicillin only by the substitution of a five-membered thiophenyl moiety for a benzyl moiety) (Figure 6-1A). Briefly, *E. coli*

Antibiotic	Class	Method of action	Determined MIC (μg/mL)	Sub-inhibitory selection concentration (μg/mL)
Gentamycin*	Aminoglycoside	Binds 30S ribosomal subunit inhibiting translation	0.38	0.3
Kanamycin*			1.88	1.5
Doxycycline*	Tetracycline	Binds 30S ribosomal subunit inhibiting aminoacyl-tRNA transfer	4.5	3.6
Tetracycline			4.5	3.6
Carbenicillin	β-Lactam (Carboxypenicillin)	Inhibits cell wall synthesis	44.4	35.5
Ticarcillin			31.2	25.0
Cefixime*	β-Lactam (3G-cephalosporin)		3.25	2.6
Ceftazidime*			6.5	5.2

Table 6-1 Antibiotics and concentrations used in this study

MG1655 cells were previously subjected to multiplex recombineering using synthetic DNA cassettes containing either an "up" (strong promoter and ribosome binding sequence [RBS]) or "down" (no promoter or RBS) sequence along with homology regions corresponding to 4,077 genes in the *E. coli* genome. The synthetic cassettes also contained unique barcodes for rapid quantification of each of the approximately 8,000 TRMR mutants by microarray or pyrosequencing technologies (Figure 6-2).

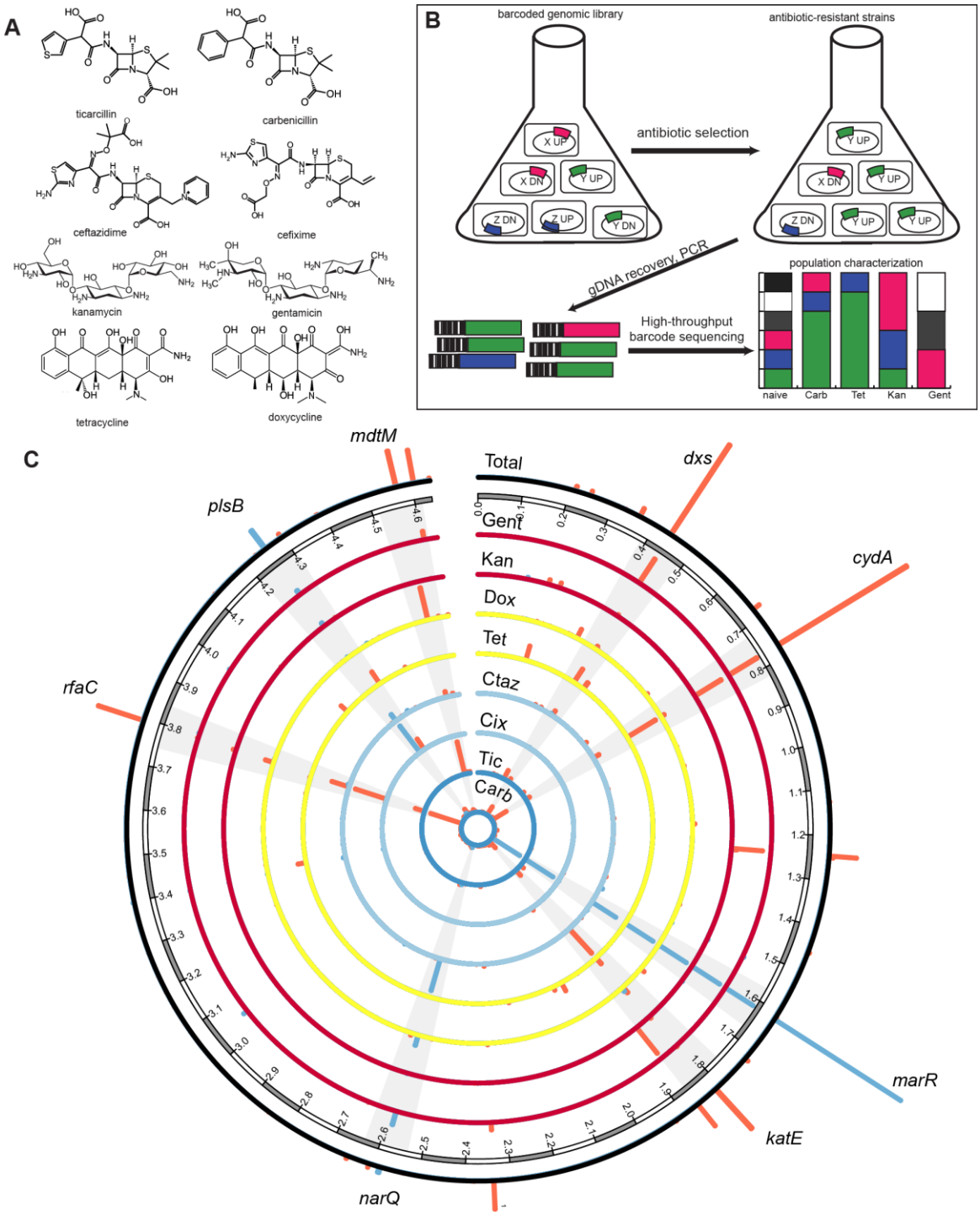


Figure 6-1 Selection of a genome-scale library on several antibiotics yields multi-drug resistant genes. **(a)** Chemical structures of the eight antibiotics used in this study. **(b)** The TRMR library containing strains simulating "up" or "down" expression phenotypes in *E. coli* is grown in selective conditions. The genomic DNA of the survivors is harvested and amplified by PCR and the amplicon is sent to high-throughput sequencing, after which it is analyzed. **(c)** Fitnesses for TRMR "up" (blue) or "down" (red) alleles for particular antibiotics are plotted relative to their location in the *E. coli* genome (in Mb). Alleles enriched in many or all selections are highlighted. The outside ring represents a linear combination of all eight antibiotic trials.

To design our growth selections, we first measured the minimum inhibitory concentration (MIC) for each antibiotic of interest in a strain equivalent to the parent strain of the TRMR library. MG1655-BC, a version of MG1655 with a barcode inserted at a silent site (the *attn7* site), was grown in liquid culture in triplicate at varying amounts of antibiotic to determine the concentration at which growth of the wild-type strain was inhibited (Table 6-1). Once the MIC was determined, the TRMR library was inoculated in triplicate in two identical sets of flasks containing MOPS rich defined media and one of eight antibiotics of interest at 80% of the MIC (48 flasks total). We performed selections at these concentrations in an attempt to normalize the selective pressure across all antibiotics. These flasks were grown until the late exponential phase with samples extracted at 24 hours and upon reaching late exponential phase. Genomic DNA was extracted and used as a template for PCR amplification and preparation for Illumina HiSeq sequencing of the barcode region (Figure 6-1B). More than 22 million barcode reads were counted and assigned to individual clones and fitnesses (see **Methods**) were calculated for all 8,077 TRMR mutants in each of the selections performed (Figure 6-1C). This analysis identified alleles enriched in TRMR libraries after selection that are consistent with previous studies on antibiotic resistance, plus uncharacterized genes potentially involved in resistance that could be important for further study (Table 6-2). In addition, we use fitness measurements to report alleles that may confer hypersensitivity (Table 6-3, Figure 6-3).

6.3.2 Alleles Contributing to Antibiotic Resistance

Our data suggest that multi-drug resistance alleles are consistently enriched regardless of the antibiotic selection performed (Figure 6-1C), and comprise a large fraction (10-90%) of each of the selected populations (Figure 6-4). Specifically, we found five alleles that were enriched in all eight selections and six that were enriched in all but one case. These 11 alleles comprised

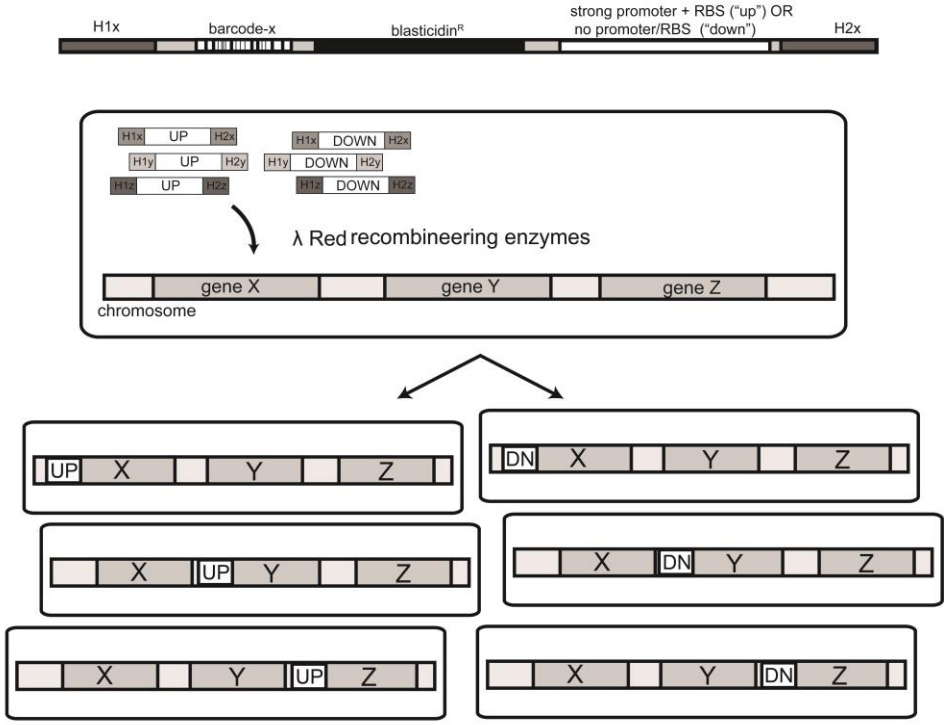


Figure 6-2 Schematic of the inserted cassette in TRMR library mutants.

	Ticarcillin	Carbenicillin	Cefixime	Ceftazidime	Doxycycline	Tetracycline	Gentamicin	Kanamycin
1	marR_up	marR_up	rfaC_dn	katE_dn	marR_up	marR_up	cydA_dn	yeaE_dn
2	mdtM_dn	rfaC_dn	plsB_up	narQ_up	mdtM_dn	cydA_dn	dxs_dn	ycgY_dn
3	narQ_up	narQ_up	yjcO_dn	marR_up	narQ_up	dxs_dn	rsmC_dn	cydA_dn
4	rfaC_dn	plsB_up	yebY_dn	rfaC_dn	lacY_dn	accA_dn	ubiG_dn	marR_up
5	lacY_dn	cydA_dn	mreC_up	plsB_up	rfaC_dn	yjcO_dn	marR_up	dhaM_dn
6	katE_dn	dxs_dn	fdnG_dn	cydA_dn	chbC_up	dcyD_dn	katE_dn	yafL_dn
7	cydA_dn	katE_dn	marR_up	mdtM_dn	katE_dn	yiiR_dn	narQ_up	tilS_dn
8	sodC_dn	yjcO_dn	secD_up	dxs_dn	sodC_dn	envR_dn	rfaC_dn	rfaC_dn
9	dxs_dn	adiA_up	yibG_dn	ybeT_dn	cydA_dn	katE_dn	iscS_dn	plsB_up
10	hyfJ_dn	yebY_dn	lldD_dn	yncH_up	dxs_dn	kefA_dn	plsB_up	katE_dn

Table 6-2 Top 10 high fitness alleles in OD selections for each antibiotic.

	Ticarcillin	Carbenicillin	Cefixime	Ceftazidime	Doxycycline	Tetracycline	Gentamicin	Kanamycin
1	yncE_dn	chbC_up	chbC_up	yncE_dn	yncE_dn	chbC_up	chbC_up	chbC_up
2	chbC_up	yncE_dn	yncE_dn	chbC_up	motA_up	yncE_dn	yncE_dn	yncE_dn
3	motA_up	motA_up	motA_up	motA_up	ansP_dn	motA_up	motA_up	ycfS_dn
4	yadE_up	ycfS_dn	ycfS_dn	ycfS_dn	flgH_dn	ycfS_dn	ycfS_dn	nikB_up
5	nikB_up	yadE_up	yadE_up	yjgN_dn	yjgN_dn	yadE_up	yadE_up	yagM_up
6	flgH_dn	ansP_dn	nikB_up	ansP_dn	iclR_dn	nikB_up	nikB_up	flgH_dn
7	ydcI_dn	nikB_up	yncA_up	puuA_up	yagM_up	flgH_dn	yncA_up	fdoH_dn
8	yqeH_dn	puuA_up	flgH_dn	nikB_up	ydcI_dn	yncA_up	flgH_dn	tnaC_dn
9	yjgN_dn	yagM_up	ydcI_dn	ydcI_dn	wzb_dn	yqeH_dn	yjgN_dn	yqeH_dn
10	ansP_dn	wzb_dn	ansP_dn	yagM_up	copA_dn	ydcI_dn	ansP_dn	yjgN_dn

Table 6-3 Top 10 low fitness alleles in OD selections for each antibiotic.

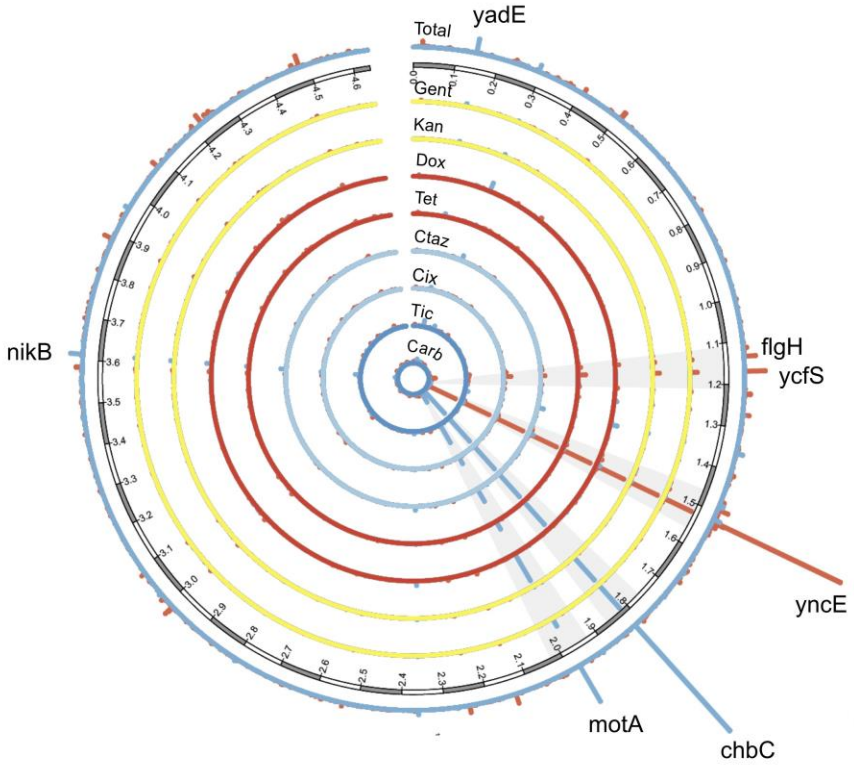


Figure 6-3 Multi-drug sensitivity genes selected for across all eight antibiotics.
TRMR "up" (blue) or "down" (red) alleles conveying the lowest fitness for particular antibiotics are plotted relative to their location in the E. coli genome (in Mb). Alleles enriched in many or all selections are highlighted. The outer ring represents a linear combination of all eight antibiotic trials.

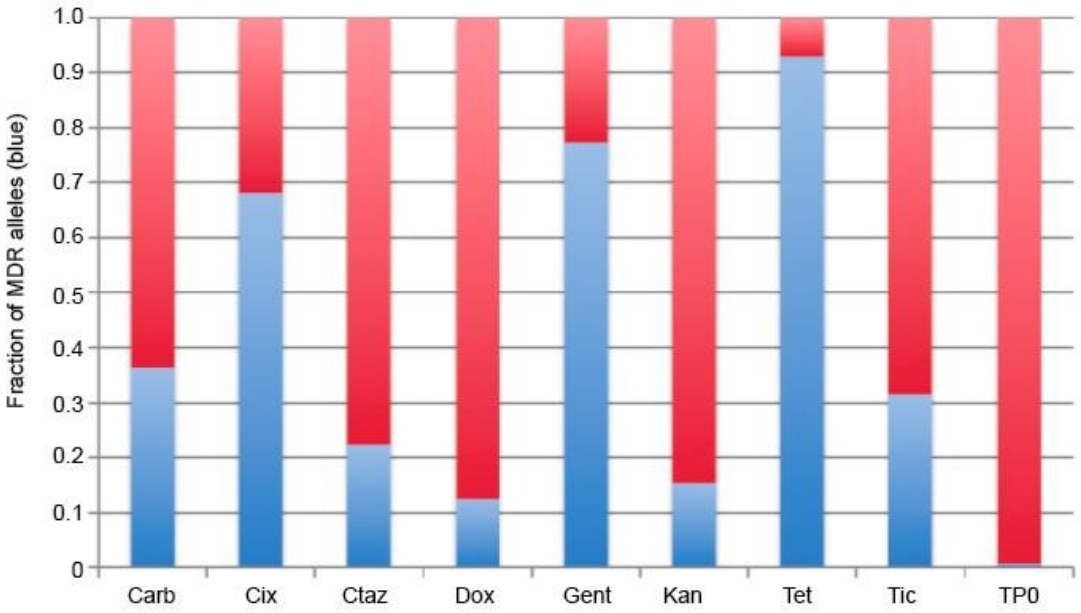


Figure 6-4 Percent of selected populations comprising multi-drug resistant genes.

over 30% of the selected population in six cases, but comprised only 0.6% of the population before selection. These results suggest that laboratory selections enrich for MDR alleles (generalists), and not only for distinct sets of individual antibiotic resistance alleles (specialists). It is important to note that previous selections of the TRMR library on the same media without antibiotics [19] did not result in significant enrichment of any of the below noted alleles (i.e., all rank below the 100 most highly enriched in fitness in MOPS media alone).

One of the most prevalent alleles, occurring in the ten most highly enriched alleles in all cases (Figure 6-1C) is *marR*_up. In this construct, the *marRAB* (where *mar* stands for "multiple antibiotic resistance"), which is normally negatively autoregulated by *marR* [238], is under control of the TRMR strong and constitutive promoter (pL_{tetO}). *MarA* is known to regulate several genes involved in resistance to antibiotics and multidrug efflux [239]. The *rfaC_down* strain occurs in the ten most highly enriched alleles in seven of the eight cases. In this mutant (and all other TRMR "down" mutants), the native RBS has been removed to minimize translation. *RfaC* catalyzes a key step in lipopolysaccharide synthesis [240]. *RfaC* mutants in several pathogenic bacteria including *E. coli* show increased resistance to various antibiotics [241]. It is not clear why the "down" mutation was selected (as opposed to the "up" mutation). However, because the blasticidin resistance cassette contains a strong EM7 promoter 5' of the gene of interest (Figure 6-2), it is possible that some read-through may occur, leading to constitutive downstream expression.

Other alleles consistently enriched by selection with several antibiotics and previously associated with antibiotic resistance included genes related to (1) managing oxidative stress: *katE* [242] and *sodC* [243] , (2) transport and efflux: *cydA* [41] and *mdtM* [244] , and (3) other metabolic processes: *dxs* [245], and *plsB* [246]. We then confirmed that apparent increased antibiotic resistance

led to increased MIC on many antibiotics. The *katE_down* allele all showed at least a 2-fold increase in MIC in all antibiotics with greater than 16-fold increases observed in some cases (Figure 6-5), while *marR_up* showed a 2-fold or greater increase in MIC on all antibiotics except gentamicin. In addition to genes enriched in multiple selections, we identified a range of genes enriched in individual selections and several genes of unknown or uncharacterized function (Table 6-2).

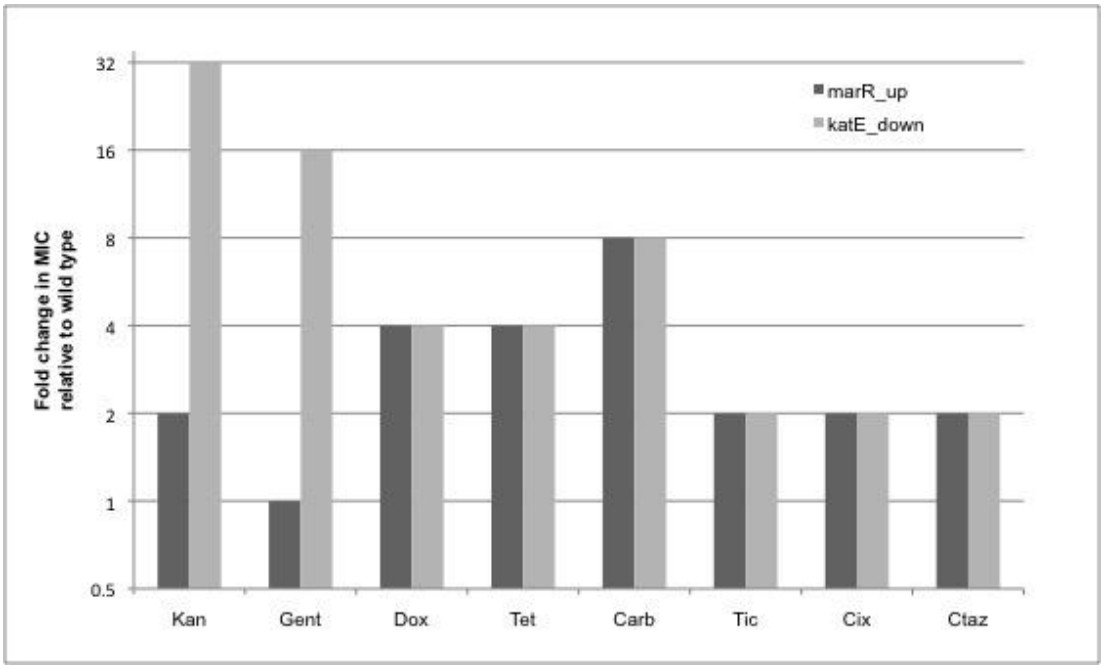


Figure 6-5 Minimum inhibitory concentrations (MICs) of antibiotics for TRMR mutants. MICs for JWKAN, marR_up, and katE_down were determined in liquid culture as in Methods. MICs were plotted relative to the JWKAN (wild-type) MIC for marR_up (black) and katE_down (grey).

Among the genes enriched in individual selections was *rsmC*, a 16S ribosomal subunit nucleotide methylase. The *rsmC_down* allele was highly enriched in the gentamicin selection. Interestingly, a recent study implicates 16S ribosomal RNA methylases in aminoglycoside resistance in *Enterobacteriaceae* [247]. A highly enriched allele for cefixime resistance was *mreC_up*. *MreC* is a rod-shape determining protein involved in peptidoglycan synthesis that has

been associated with β-lactam resistance in *Helicobacter pylori* [248]. *SecD*, another allele isolated in the cefixime selection, has also been linked to β-lactam resistance in *E. coli* [243].

Unexpectedly, several enriched fitness alleles for cefixime, ticarcillin, and gentamicin selection(s) corresponded to hydrogen production and formate processing including *fdnG, hyfJ,* and *narQ*. It is possible that the actions of these proteins affect the proton motive force, either facilitating increased drug efflux by increased PMF or decreasing drug uptake by reducing PMF (as is well known to affect the toxicity of charged compounds such as aminoglycoside antibiotics [249]). Several alleles were isolated that correspond to genes with unknown functions. They include: *ycjO* (putative ABC transporter), *yiiR*, *ybeT* (conserved outer membrane protein), *yafL* (inner membrane protein), *ycgY*, *yeaE* (methylglyoxal reductase), *yebY*, *yigB*, *yiiR*, and *yncH*. The contribution of these genes to antibiotic resistance warrants further investigation.

Finally, the targeting of antibiotic sensitivity genes provides a possible mechanism to treat resistant infections. To determine genes that might convey sensitivity to the antibiotics of interest, we also recorded the alleles with the lowest fitness (i.e., largest decrease in frequency throughout a selection) (Table 6-3). Our analysis suggested considerable overlap in susceptibility genes across the antibiotics investigated (Figure 6-3). Many of the proteins encoded by these alleles are targeted to the inner membrane. Previous experiments also showed that these specific alleles grew well on non-selective rich MOPS media [19]. Thus, it is possible that changes in expression of these inner membrane proteins alter the overall inner membrane fluidity or porosity, allowing antibiotics to traverse membrane more easily. While this possibility should not be discounted, it should also be noted that all of the above susceptible alleles were present in large quantities at time point zero. Given the strength of each selection, it is possible that these alleles were simply diluted down to the limit of detection. This is an issue of selection design; in

particular our designs were targeted at enrichment for resistance phenotypes as opposed to identification of susceptibility phenotypes.

6.3.3 Allelic Responses to Chemically Similar Antibiotics are Weakly Dissimilar

Our data suggested that sub-lethal antibiotic treatment strategies selected for multi-drug resistance alleles. To explore this suggestion in more depth, we performed principal coordinate analysis (PCoA) on all replicates from each selection. PCoA allows for visualization of multi-dimensional variables in 3D space by condensing distance metrics into the most important coordinates while minimizing the loss of information. We specifically hypothesized that antibiotics with similar chemical structure and belonging to the same class (e.g., gentamicin and kanamycin) would present a similar allelic response and therefore cluster together in PCoA space, and that antibiotics having similar mechanisms of action (e.g., the aminoglycosides and the tetracyclines, which both act by binding 30S ribosomal subunits) would as well. Although some patterns appear at 24 hours and after reaching the late exponential phase (Figure 6-6A and B), such as the location of gentamicin and kanamycin in the upper half of PCoA space, other patterns are unexplained. For example, doxycycline, carbenicillin, and ceftazidime cluster near time point zero. This finding is supported by a weaker ANOSIM R value for antibiotic class or mechanism of action (Figure 6-6A and B). ANOSIM R-values near zero indicate random grouping. Network analysis, in which samples sharing similar genes are drawn together, confirms that the subtle antibiotic PCoA clustering patterns, as there are no large differences between antibiotic types (Figure 6-7). However, differences between antibiotics are discernable by ANOSIM [237], which is an extremely sensitive test (Table 6-4). These results are robust to normalization technique, replicate, and distance metric (Figure 6-8, Figure 6-9).

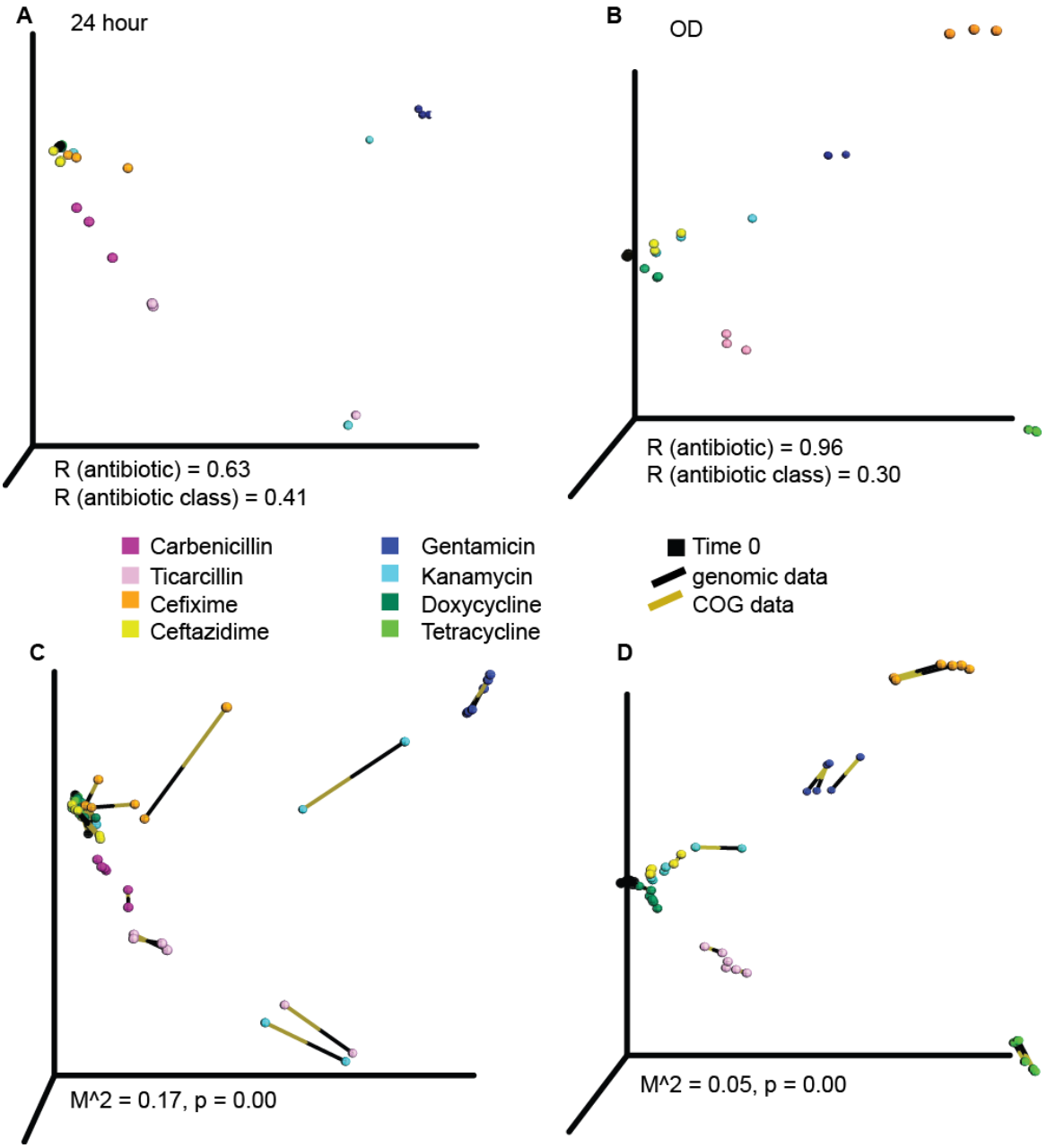


Figure 6-6 Separation of antibiotic classes in PCoA space is weak across multiple levels of functional hierarchy. **(a,b)** PCoA analysis, using Bray-Curtis distance, of the antibiotics at **(a)** 24 hours **(b)** upon reaching the late exponential phase (OD). ANOSIM R-values are plotted for separation by antibiotic or by mechanism of action. **(c,d)** Procrustes analysis indicates significant alignment between the COG (gold end of the line) and gene (black end of the line) PCoA profiles in the 24 h and OD selections. The longer the line connecting the COG and gene points, the less aligned the two points are in PCoA space, increasing the stress value (M^2).

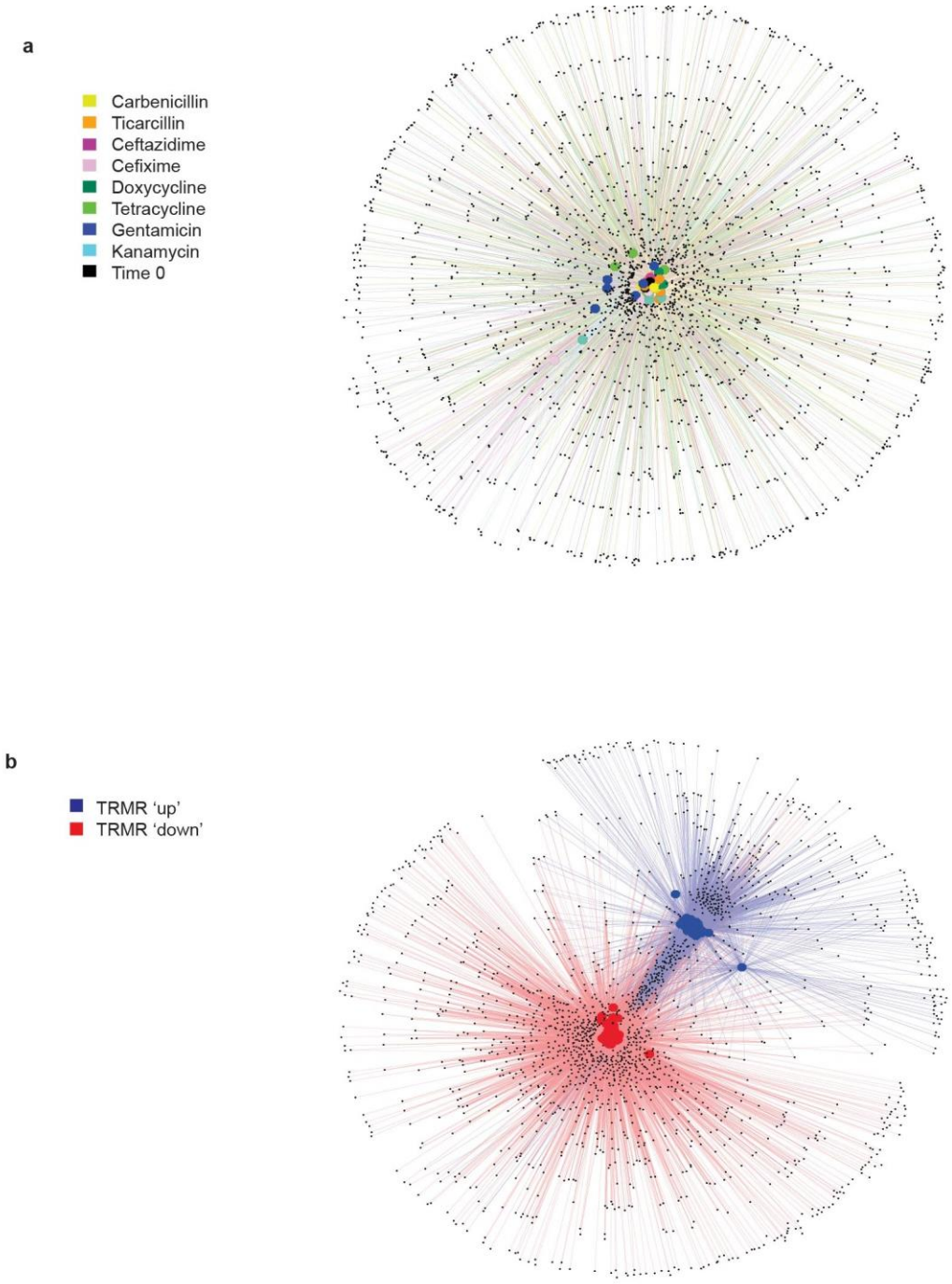


Figure 6-7 Network analysis of antibiotic selections. **(a)** Nodes are the antibiotic type, while the black dots are the genes. If a gene is shared between the antibiotics, it pulls those nodes closer at an amount weighted by the gene's abundance. If a gene is not shared between the antibiotics, it pulls the antibiotic sample node it is attached to towards the outside of the diagram, separating the nodes. The close clustering of the antibiotic nodes indicates many shared genes. **(b)** The separate clustering of the TRMR 'up' vs. the TRMR 'down' antibiotic selections indicates that very different up/down genes are selected for.

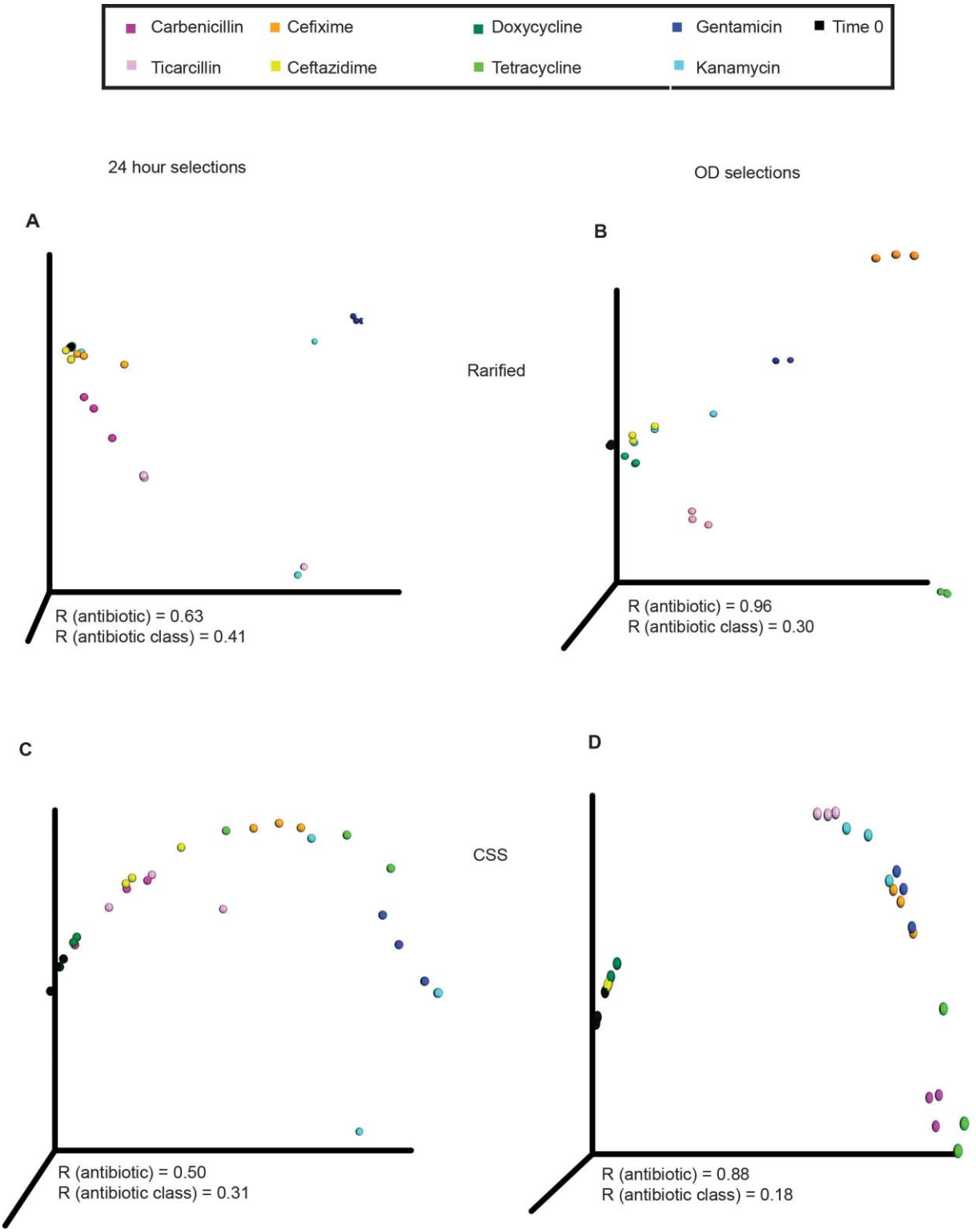


Figure 6-8 Clustering by antibiotic class is consistent regardless of normalization technique. 24 hour time point (left) and late exponential phase (OD) selections (right). The rows are the normalization methods used, which are rarefying or cumulative-sum scaling (CSS)

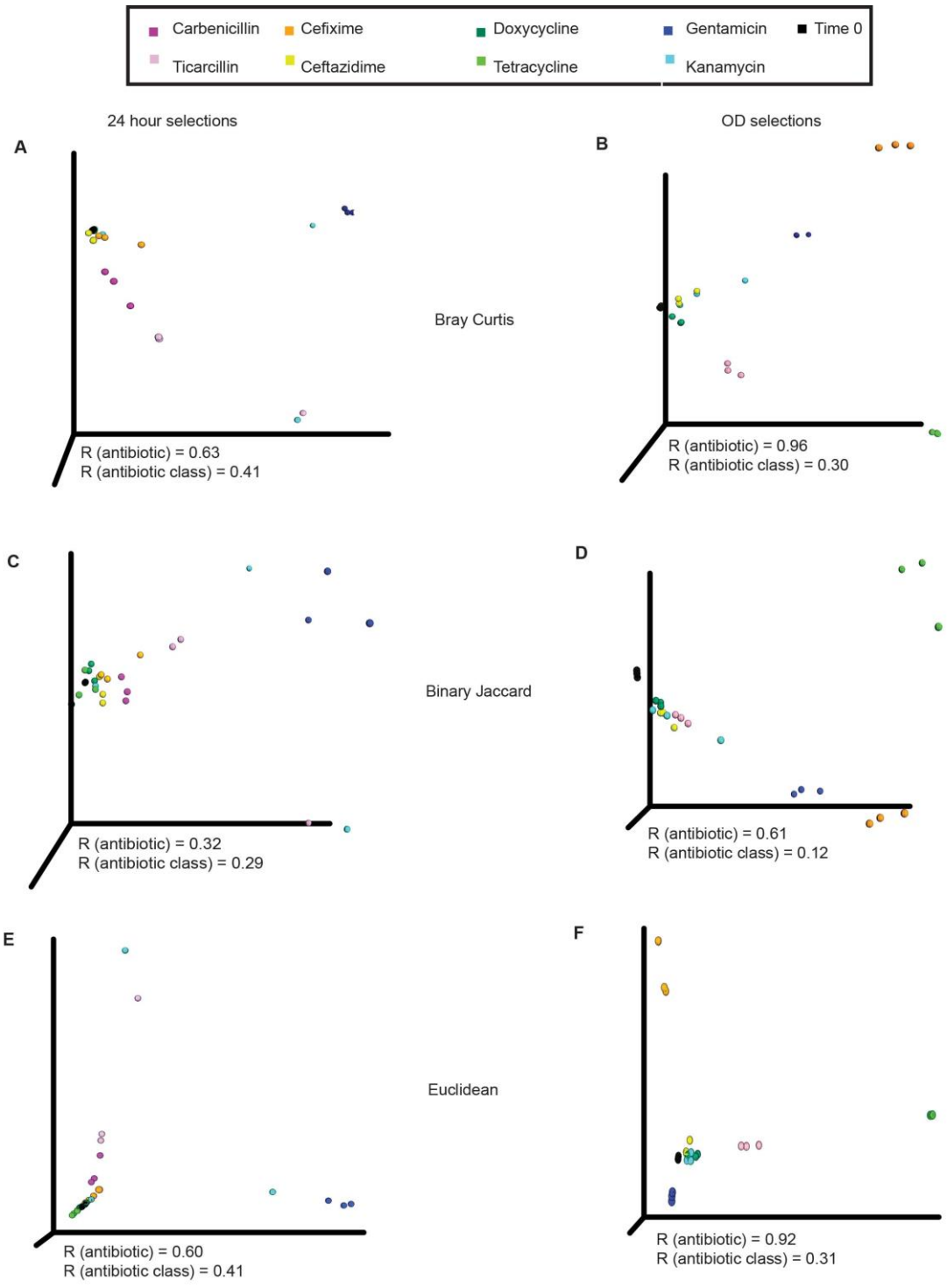


Figure 6-9 ANOSIM R-values are consistent regardless of distance metrics. 24 hour (left) and late exponential phase (OD) selections (right). Each row represents clustering with a different distance metric. The much smaller ANOSIM R-value for the binary Jaccard selections supports the hypothesis of Figure 6-4: that differences in allelic population abundances, rather than the alleles themselves, are the main variable driving the antibiotic separation.

dataset A1:				
Category	R	p-value	Permutations	Category Details
TRMR	0.96	0.001	999	TRMR "up" vs. "down" regulated
Antibiotic	0.49	0.001	999	Kan, Gent, Doxy, Tet, Carb, Tic, Cix, Ctaz
Antibiotic_Replicate1	0.19	0.09	999	Antibiotics samples in replicate 1
Antibiotic_Replicate2	0.41	0.005	999	Antibiotics samples in replicate 2
Antibiotic_Replicate3	0.25	0.05	999	Antibiotics samples in replicate 3
Antibiotic Class	0.32	0.001	999	Aminoglycoside, Tetracycline,
				Carboxypenicillin, third generation Cepham
Method	0.06	0.13	999	24hr, OD,Time_0
Replicate	-0.05	1	999	1, 2, or 3
dataset A2:				
Category	R	p-value	Permutations	Category Details
TRMR	0.96	0.001	999	TRMR "up" vs. "down" regulated
Antibiotic	0.99	0.001	999	Kanamicin and Gentamycin
Method	0.01	0.33	999	24hr, CT, SD, Time_0
MethodType	-0.02	0.58	999	24hr,CT1,CT2,CT3,SD1,SD2,SD3,Time_0
Replicate	0.09	0.032	999	1, 2, or 3

Table 6-4 Nonparametric ANOSIM values for important categories in this study.
The R statistic represents the how different the tested categories are, with a value near zero indicating no significant difference between the groups, and a value near 1 indicating difference. Dataset A1 is a Kanamycin and Gentamicin detailed time course study; Dataset A2 is of all eight antibiotics, with either 24 hr selection, three constant transfers (CT) or three serial dilutions (SD) of antibiotics.

6.3.4 Clusters of Orthologous Groups Analysis Elucidates Functional Hierarchy

Although antibiotics of similar classes or targeting the same complex did not exhibit significant clustering in PCoA space at the specific allele level, we speculated that clearer patterns might be revealed when the PCoA analysis was performed at the level of encoded functions. To gain an understanding of mechanisms of action on a pathway level [250], a matrix of clusters of orthologous groups (COG) [235] was formed by summing the counts of genes belonging to the same COG in the same sample. We then performed Procrustes analysis to analyze the similarity of the gene [45]and COG distributions in PCoA space (Figure 6-6 C and D). Procrustes analysis stretches, rotates, and scales two datasets to determine if similar conclusions could be drawn [45]. The p-values are less than 0.001, suggesting that the functional profiles could be predicted from the TRMR alleles enriched by selection because both matrices display similar PCoA clustering patterns. This match between COG and gene distributions implies that selection

acts in broadly similar ways at multiple levels of the functional hierarchy. Figure 6-10 shows the similar COG distribution of the antibiotic samples over time.

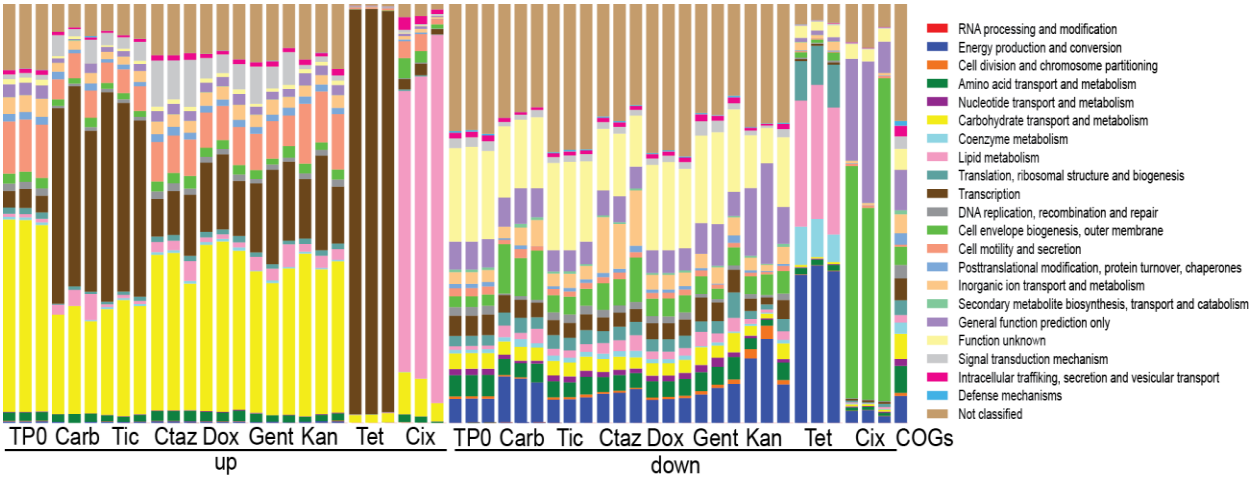


Figure 6-10 Clusters of orthologous groups (COGs) analysis of selected populations. Label format is Antibiotic_replicate_up/down, e.g., CarbOD_1_up means COG counts for carbenicillin, replicate 1 of 3, and TRMR "up" alleles. Antibiotic abbreviations: Carbenicillin (Carb), Ticarcillin (Tic), Ceftazidime (Ctaz), Cefixime (Cix), Gentamicin (Gent), Kanamycin (Kan), Doxycycline (Dox), and Tetracycline (Tet). Far right bar: COG distribution as represented in the wild-type *E. coli* genome.

6.3.5 Supervised Learning Distinguishes Resistance "Fingerprints"

Given that MDR alleles were a significant fraction of every selection (Figure 6-4), we wanted to understand whether the final antibiotic populations could be distinguished. To do so, we used supervised learning to identify combinations of genes that may be unique to individual antibiotics, and thus represent a genomic "resistance fingerprint" for each antibiotic. We used the random forest classifier [46] to generate confusion matrices from 48 samples (24 hour and late-exponential phase selections, in triplicate, on each of eight antibiotics), which indicate true vs. predicted classifications when a portion of the dataset is withheld from model training (Figure 6-11). At the level of individual alleles, it was difficult to distinguish between some antibiotics (as shown by shading off of the diagonal), especially between antibiotics of the same class or mechanism of action (Figure 6-11A). The random forest classifier returns a ratio of baseline

error to observed error of 2.2, indicating that the classifications are estimated to be 2.2 times more accurate than random guessing, a statistically significant but weak effect.

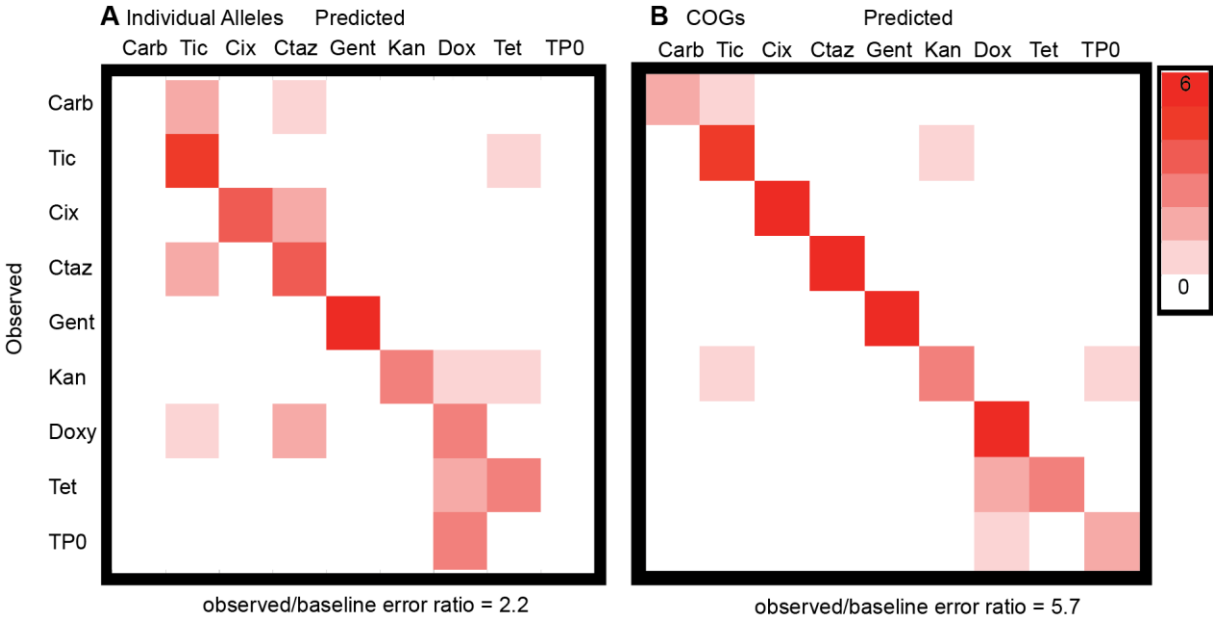


Figure 6-11 Supervised learning is able to distinguish between the antibiotics at both the allele and COG level. Confusion matrices for random forest classifiers. Off diagonal classification represents classifier error. Antibiotic abbreviations as in Figure 6-10.

However, when alleles are grouped by COG category, supervised learning improves substantially. There is excellent classification of antibiotics with a baseline error ratio of 5.7 (Figure 6-11B). This indicates that each antibiotic has a unique signature at the COG level. Classification between the antibiotics may further improve by adding more antibiotics within each class. This hypothesis is supported by perfect distinction (baseline error ratio 24.0) between gentamicin and kanamycin, antibiotics of the same class with similar chemical structure, in a separate detailed time course selection (Figure 6-12A and C). Furthermore, the subtle genetic differences arise at the first antibiotic application, independent of selection length (Figure 6-12B).

The genes, COGs, and the enrichment patterns the random forest classifier uses most to distinguish between the antibiotics are found in Figures 6-13 and 6-14. Interestingly, most of the

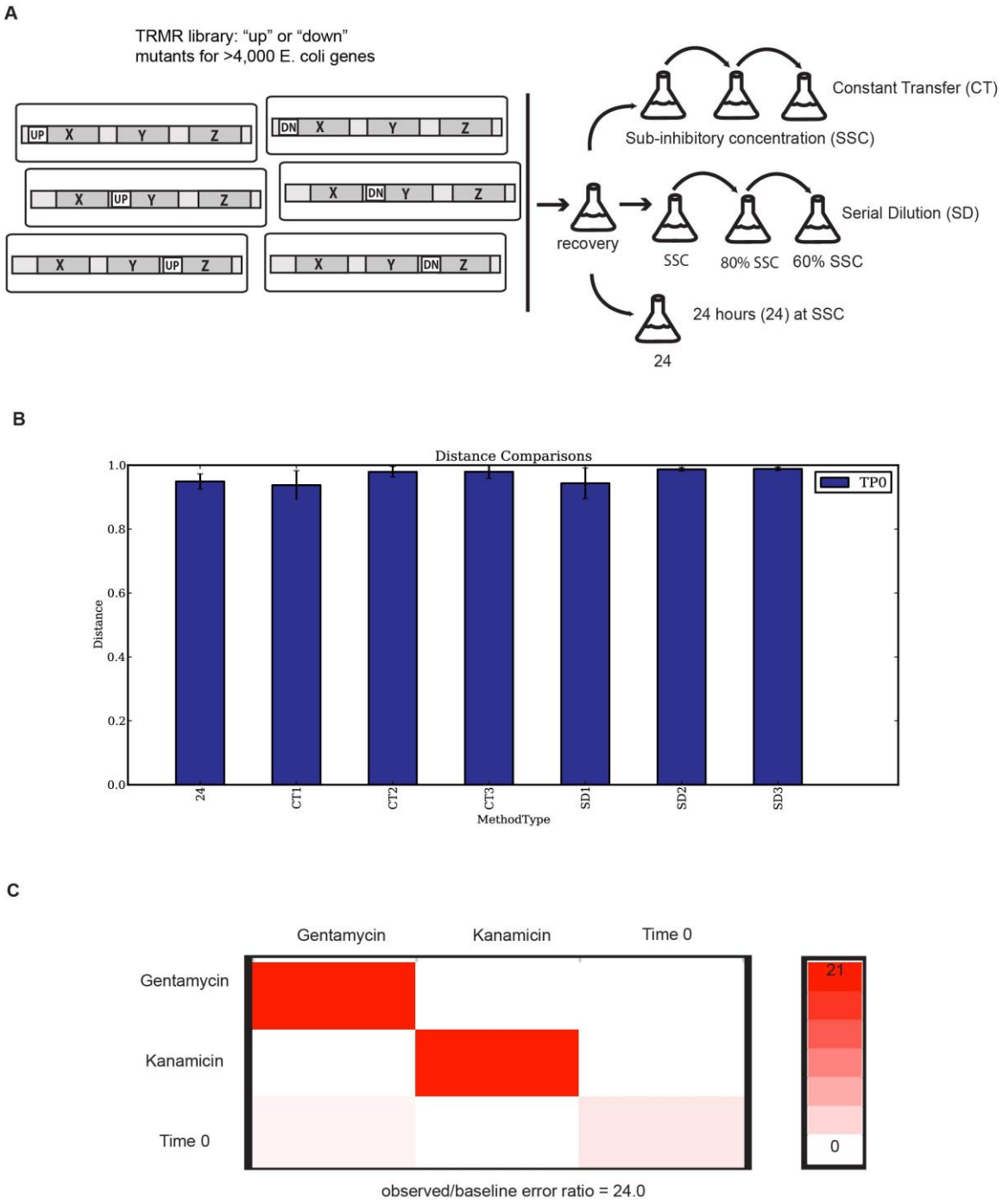


Figure 6-12 A detailed time course selection, with many more samples, on gentamicin and kanamycin results in near perfect supervised learning classification. (A) Schematic of the types of time-course experimental setups. All of these selections were done in triplicate to control for experimental variations. (B) The majority of change to the allele population occurs in the first selection, regardless of selection type. This is shown by the large Bray-Curtis distance between Time_0 (TP0) and the selection types. Also, the second and third constant transfers (CT2, CT3) or serial dilutions (SD2, SD3) do not have much higher bars than CT1or ST1. (C) Supervised learning confusion matrix for the detailed Gentamicin and Kanamycin time course study shows no error (off diagonal classification) between the two antibiotics.

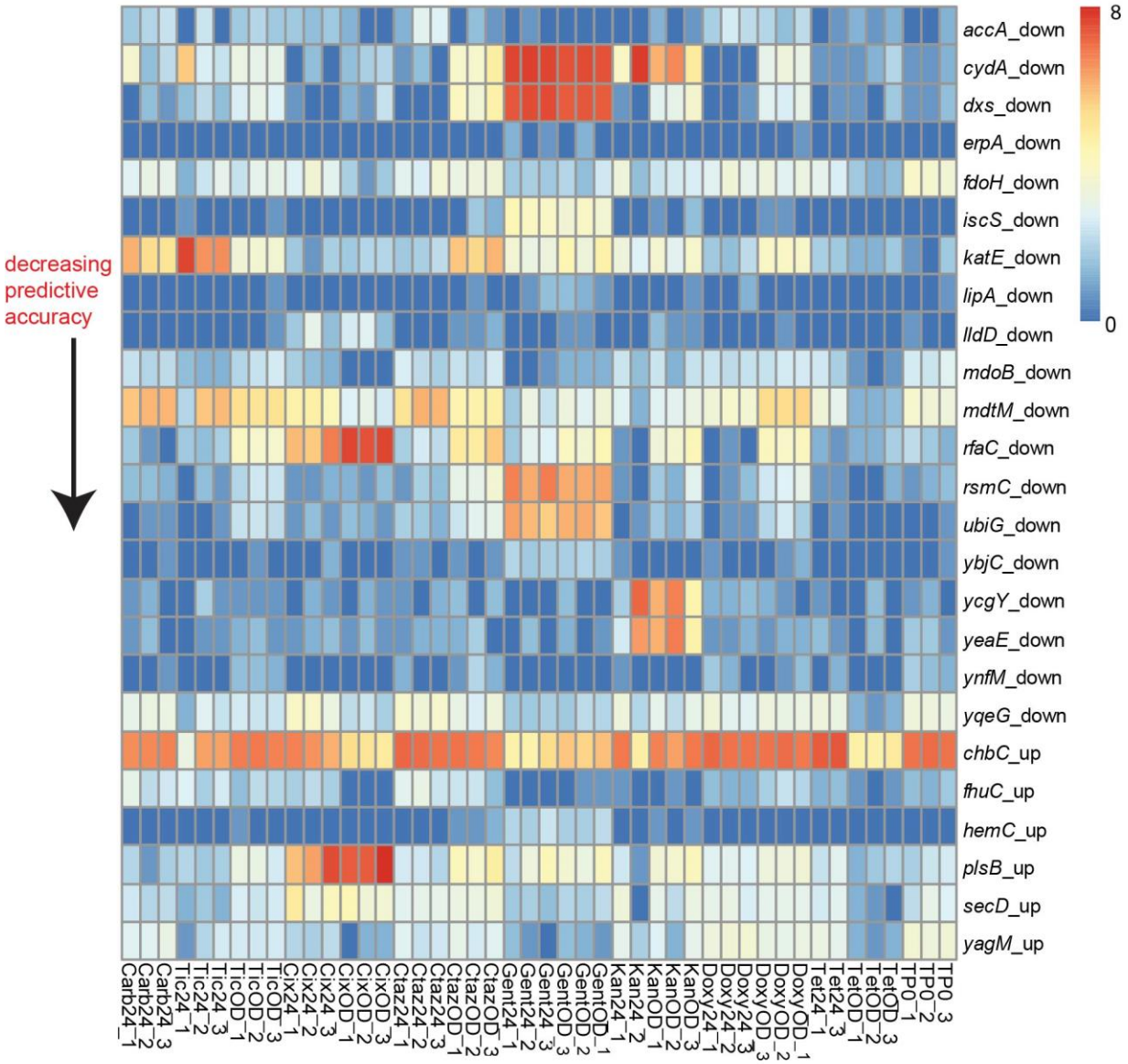


Figure 6-13 Heatmap of the log10 counts of the top 25 genes that distinguish antibiotic categories in the supervised learning classifier. Label format is Antibiotic/Method_replicate. For example, Carb24_1 means Carbenicillin was used, it is the 24- hour selection, and it is replicate 1 of 3.

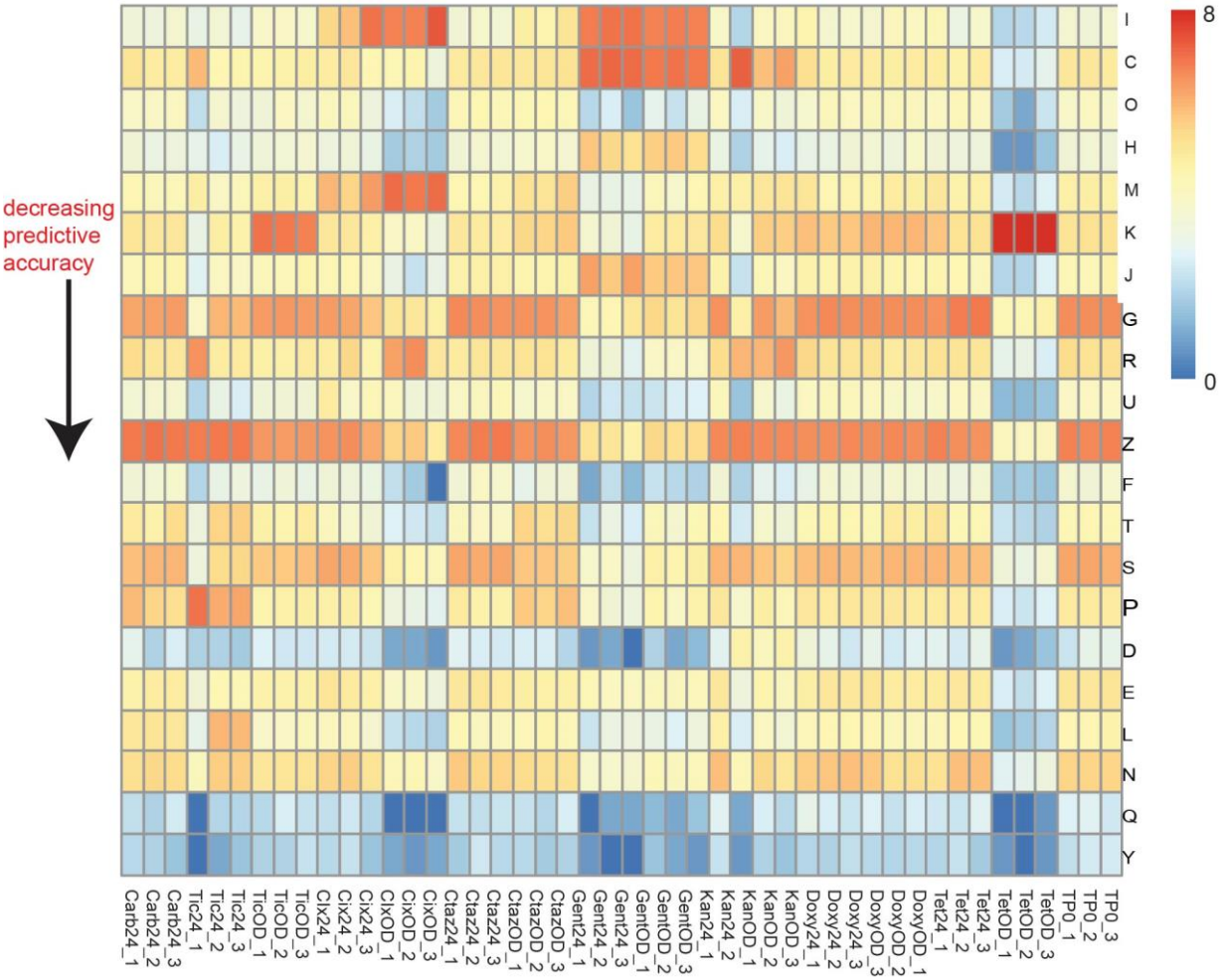


Figure 6-14 Heatmap of the log10 counts of the COG categories used in the supervised learning classifier. Labeling as in Figure 6-13. COG category symbol and meaning: C – Energy production and conversion, D – Cell cycle control and mitosis, E –Amino Acid metabolism and transport, F – Nucleotide metabolism and transport, G – Carbohydrate metabolism and transport, H – Coenzyme metabolism, I – Lipid metabolism, J – Translation, K – Transcription, L – Replication and repair, M – Cell wall/membrane/envelope biogenesis, N – Cell motility, O – Post-translational modification, protein turnover, chaperone functions, P – Inorganic ion transport and metabolism, Q – Secondary structure, T – Signal transduction, U – Intracellular trafficking and secretion, Y – Nuclear structure, Z – Cytoskeleton, R – General function prediction only, S – Function unknown.

genes that are key in building the antibiotic classifier, which examines the prediction strengths of individual genes, are also identified as the high/low fitness alleles analyzed in the above genomic plots (Figure 6-1C, Figure 6-3). Also, the distinction between antibiotics and their classes diminishes when using the binary Jaccard distance metric, which operates on a gene presence/absence basis (Figure 6-7C and D). This strengthens the conclusion that while alleles

conferring multi-drug resistance are found in many cases, variation in the degree of enrichment of these MDR alleles for a particular antibiotic is a predictor of the genetic fingerprint of a particular antibiotic.

6.4 CONCLUSION

We have presented a model pipeline for the analysis of gene products leading to antimicrobial resistance in *E. coli*. We discovered that many alleles isolated from treatment with low levels of single antibiotics conferred resistance to many antibiotics. This lends support to the hypothesis that low-dose antibiotics as used in livestock growth promotion and found in wastewater likely promote resistance to a wide range of antimicrobial compounds including last-resort therapeutics [35]. The rise of antimicrobial resistance is also important in microbial ecology, including soil [251] and human gut [252] bacteria.

Chemical tolerance in microbes is often a complex phenotype conferred by a range of genetic factors that are often not intuitive or obvious. A seminal work in chemical genomics in *E. coli* was recently published in which a library of over 4,000 strains including the Keio deletion library was screened under many different chemical and physical conditions [17]. In that work, individual strains were plated robotically in 1,536-well format, and colony size was investigated to determine fitness. A similar work examined the effect of a library of 4,000 *E. coli* genes overexpressed on plasmids challenged by a variety of chemicals [37]. While the library was assayed in multiplex in microtiter plates, characterization of alleles (by the nature of Sanger sequencing) was limited to less than 10 colonies per condition. A recent study focused on aminoglycoside antibiotics used adaptive evolution over hundreds of generations to examine beneficial mutations and characterized by whole genome sequencing of 240 parallel-evolved lines [38]. The study concluded that mutations that affected efflux pumps such as AcrAB

contributed to multiple-drug resistance. Our observation of the *marR* allele observed agrees with this result, but as the scope of our search was much broader we were also able to determine multi-drug resistant alleles with mechanisms which do not necessarily have to do with efflux pump regulation as well as alleles with unknown function without whole-genome sequencing.

The original application of the TRMR library used DNA microarrays and exploratory, not multivariate, analyses to characterize the genome-level responses to various conditions. However, this application required custom-made arrays corresponding to the barcodes. In addition, as demonstrated by the application of Bar-seq to a yeast deletion library [220], sequencing has many advantages over microarrays for rapid analysis of barcoded libraries, including but not limited to cost, the ability to pursue many biological replicates under various conditions in one sequencing lane, reduced crosstalk, and increased resolution on the low and high ends of detection [253]. Previously, DNA sequencing data from barcoded libraries was analyzed using packages with specialized analyses, and for smaller, number dense datasets [92]. In contrast, sparse datasets (containing many zeroes) like the one presented in this work make metagenomic techniques like the analyses in QIIME more appropriate [28]. QIIME also contains many analysis types in one package, streamlining analyses, and can easily analyze dataset sizes from small to massive [21]. Overall, our approach allows such analyses in multiplex at the level of growth selections (over roughly 24 to 48 hours) and now in the sequencing steps as well, allowing considerably faster, deeper, and larger laboratory population genomic dynamics studies in bacteria. Barcoding maximizes the usefulness of short reads and allows for the use of HiSeq technology to generate millions of times more data points than Sanger sequencing would allow. In addition, the barcoded and pre-defined nature of the library circumvents the need for long adaptation cycles (10-100 times fewer generations required) and whole genome sequencing.

Thus, the combination of a method to map the specific effect of genes to selectable traits (TRMR), high-throughput sequencing, and streamlined bioinformatics analysis software (QIIME) provides a powerful toolbox for exploring the genetic basis of a broad variety of complex phenotypes [29]. Finally, the same methodologies of selection, high-throughput sequencing, and bioinformatic analysis are broadly applicable to experiments on chemical tolerance for any inhibitory chemical, from antibiotics to toxic metabolites to next-generation biofuels.

6.5 ACKNOWLEDGEMENTS

The authors thank Anis Karimpour-Fard for providing the COG gene assignments, and Tirzah Glebes and Lauren Woodruff for helpful discussions. SJW was funded by the NIH/CU Molecular Biophysics Training Scholarship (T32 GM-065103).

Chapter 7

Conclusions and Future Directions

Microbial community analysis is important in a vast range of application areas, some of which we have contributed to in this thesis. Major outstanding challenges when we started this work were contamination, uneven library sizes, compositionality, and sparsity. In Chapter 2, we recommend using a positive control, randomly assigning samples to DNA extraction batches, PCR batches, and sequencing runs, while keeping track of these variables during analysis. This will help avoid erroneous conclusions when, for example, the batch effect coincides with the biological variable of interest [53, 56]. In Chapter 3, we move to the first analysis step after an OTU table is constructed, where the effect of contaminants is hopefully minimized. We evaluate many different normalization methods that attempt to address some, but not all of the three challenges, and highlight the strengths and weaknesses of all the methods. We propose which normalization method to use depending on the distance metric of interest, and the distribution of library sizes. To visualize the normalization effects, we use PCoA, which is a common next analysis step, as a proxy to other analysis types. For example, researchers with data that contains low library sizes, and subtle effects, may come to spurious conclusions if they use an inappropriate normalization method/distance metric. We then look at statistical tests for determining the taxa driving the PCoA clustering patterns, and again make recommendations for which test to use when. Statistical power can be increased in the correct situation by using specially developed parametric tests instead of the traditional non-parametric techniques. However, when the difference in library sizes between the two categories of interest (e.g. ‘Case’ vs. ‘Control’) is greater than 3x (increasing matrix sparsity), or when the data is suspected of being highly compositional, it is better to revert back to non-parametric techniques. We then apply these

normalization and differential abundance testing techniques in Chapter 4 to elucidate the effect of carcass mass on the associated microbial communities.

In Chapter 5, we move from understanding single taxa behavior to understanding how they behave in communities. We investigate the behavior of co-occurrence tools designed specifically for microbiome data, as well as more standard tools, in response to ten computational challenges: sequencing technology choice, distribution, normalization, feature filtering, null data, linear and non-linear ecological relationships, time series, compositionality and sparsity. We make correlation method recommendations in each case and overall. We also propose an ensemble approach for dramatically increased precision, although at a cost to sensitivity.

As demonstrations of our work in Chapters 3-5, we have contributed to a variety of topics in microbiome research. In 'A Universal Microbial Clock for Estimating Postmortem Interval' (submitted to *Science*), Metcalf *et al.* found that time of death can be estimated based on the microbial composition of gravesoils for up to three months after the subject has died. Also, the signature of decomposition persists in gravesoil for up to thirty days after a body is removed, which is helpful for clandestine, or unmarked, grave location. This represents a significant advance for forensic science because while insects are currently used in death investigations, insects are not always available as evidence, and conclusions depend on a regional knowledge of entomology [157]. We also identify key taxa that become significantly differentially abundant in post-rupture corpse skin and gravesoil microbial communities. This finding is robust to both human and mice species, and across a wide range of climates/environments including winter and spring, desert, soil, and grassland. 50% of these decomposer taxa that increase significantly as

decomposition progresses were also present before decomposition begins, which may explain why some of them are universal across host species and environments.

We have also contributed to studies showing that, when antibiotics fail, the best way to cure a *Clostridium difficile* intestinal infection is by transplant of a healthy donor microbial community via a fecal mass transfer (FMT). *Clostridium difficile* infections kill approximately 29,000 people per year in the USA. *Proteobacteria*, primarily of order *Enterobacteriales*, decrease significantly following treatment, while *Bacteroidetes* and *Firmicutes* increase significantly following treatment [64] (Kashyap et al. *in preparation*). We also show that, within days after FMT, the patient microbial profile is much better correlated with the healthy donor. However, at long times (~100 days) after the FMT, patients diverges from the donors, to develop a unique microbiome [64]. This is consistent with other research showing that there is no core 'healthy microbiome', with much variation between individuals [55]. We also contributed to a study assessing the microbiome of the Lone Star Tick, and its pathogenic and non-pathogenic species. Ticks are vectors for many pathogens, and the causal agents for diseases such as southern tick-associated rash illness (STARI) have not been identified or well characterized [191]. Bacteria of the order *Rickettsiales* commonly infect ticks, some species of which are harmful to humans. We show using CoNet correlation networks that certain species of *Rickettsiales*, as well as other bacteria, positively co-occur in the Lone Star Tick.

We have also applied other techniques, such as machine learning, to help show that the house a family lives in can be matched to the family based on microbial samples from the house and the people living in the house. Interestingly, the house microbiome changes to match the family's microbiome in less than a day, and also shifts significantly when the family leaves the house [254]. This again has forensic implications, because what location a person has lived in, and

how recently, can be predicted with good accuracy. Unrelated humans living in the same house are more similar in microbial content than average, and outdoor pets drag in plant and soil microbes. Since PCoA is such an important tool for visualizing multi-dimensional data, we have also helped adapt faster calculation algorithms [255, 256], since microbiome datasets are rapidly expanding (Gonzales et al., *in preparation*). The behavior of bacterial communities over time is also poorly understood, since most microbiome studies are surveys where participants are sampled a few times, usually spanning weeks, at best. We have contributed to the first study where participants are sampled every 30 minutes for five days. We show that some oral bacteria have distinct daily, repeatable, abundance fluctuations (Amir et al., *in preparation*).

Additionally, these techniques work well for sparse, high-throughput, biological engineering data, such as that from TRMR. We show through PCoA and network analysis that many antibiotic resistant genes are the same across many classes of antibiotics, and antibiotics of similar classes do not necessarily cluster together. We also identify positively and negatively differentially abundant genes for each antibiotic, and show though Procrustes analysis that the allelic and proteomic profiles are similar. Finally, machine learning helps to discover a unique proteomic signature for each antibiotic. In another study, we used clustering analysis to help identify and compare the genes contributing to ethanol tolerance and production in *E. coli*, a commercially valuable phenotype [257].

Along the way, we have found that this field is immensely interdisciplinary, and that there is tremendous potential for introducing quantitative analyses that can have a large impact. For example, the introduction of machine learning from computer science to microbial community analysis was a tremendous advance for a field that badly needed it [47-49]. Using machine learning, which is able to build a predictive model from high dimensional data, we can

distinguish human diseases based on microbial profile. For example, a classifier can tell with 90% accuracy whether a person is lean or obese; however, using gene data, the classifier is only a few percent accurate[258]. Also, Procrustes analysis[45], which was originally developed in psychology, has proved immensely useful in high-throughput sequencing analysis PCoA for comparing, e.g. whether the same results would be derived using different sequencing technologies[117]. On the experimental side, neither microbial community analysis nor approaches like TRMR would be possible without tremendous advances in DNA synthesis throughput at the border between chemistry and biology.

In the future, there are still substantial limitations in several of the techniques proposed in Chapters 3-6. In Chapter 3, if the effect size of carcass mass on the gravesoil microbial communities is subtle, then three replicates are not enough to resolve it. Also, time points beyond 15 days need to be collected and analyzed; we stopped at 15 days since that is often the latest time point by which most death scenes are responded to by law enforcement. Finally, while swine are widely accepted to be an ideal model system for humans, they are not exactly like humans in body composition. In Chapter 4, a fatal problem with techniques like DESeq, for both normalization and differential abundance testing, is the pseudocount question. While Aitchison's log ratio techniques [23] are an excellent fix for compositionality, no good solution for data sparsity has been found [27], and severe compositionality is still a major challenge facing differential abundance testing. Once the zero-problem has been solved, Aitchison's test for complete subcompositional independence [259] is worth pursuing for testing severity of compositionality in a dataset, rather than the alpha-diversity rule of thumb. In Chapter 5, while good progress has been made in addressing compositionality for correlations, the problem is by no means solved. Other authors have highlighted concerns with the CoNet and SparCC assumptions [27]. We found that

sparsity is the most significant barrier to correlation interpretation. At realistic sparsity levels, all techniques have extremely low F1 scores (harmonic mean of precision and sensitivity), highlighting the need for tools that are more robust to sparsity. Finally, in Chapter 6 more replicates and antibiotics should be investigated, as well as further characterization of TRMR mutants. A significant barrier not attempted in Chapter 6, but an extremely important future extension of this work, is computational methods for navigating the fitness landscape of combinatorial genetic changes for improved trait engineering of biofuels, industrial chemicals, etc. The number of possible combinations of genetic modifications to be explored exceeds the number of atoms in the universe. This is a number matching the complexity of microbial community correlation analysis.

Resolving the above limitations will also enable better understanding of microbial communities, particularly rare species. Rare species that are difficult to reliably detect even with rapidly improving sequencing technology can be critically important in a healthy microbial community. For example, SparCC co-occurrence analysis, used because it performs very well in the evaluations of Chapter 5, revealed that very rare taxa in the family *Christensenellaceae* are a hub species upon which many other taxa depended in lean humans but not obese humans [260]. When the microbiome of obese mice lacking *Christensenellaceae* was amended with *Christensenellaceae minuta*, the mice lost a significant amount of weight. Taken together, the techniques we recommend for normalization, differential abundance analysis, and correlation networks, as well as the discoveries we make regarding forensics and TRMR, are important for current analyses and provide a basis for future research.

Bibliography

1. Ley, R.E. et al. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11070-11075 (2005).
2. Turnbaugh, P.J. et al. A core gut microbiome in obese and lean twins. *Nature* **457**, 480-484 (2009).
3. Vrieze, A. et al. Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology* **143**, 913-916 e917 (2012).
4. Ridaura, V.K. et al. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**, 1241214 (2013).
5. Wang, Z. et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57-63 (2011).
6. Gough, E., Shaikh, H. & Manges, A.R. Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent Clostridium difficile infection. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **53**, 994-1002 (2011).
7. Gevers, D. et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell host & microbe* **15**, 382-392 (2014).
8. Lozupone, C.A. et al. Alterations in the gut microbiota associated with HIV-1 infection. *Cell host & microbe* **14**, 329-339 (2013).
9. Berer, K. et al. Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination. *Nature* **479**, 538-541 (2011).
10. Gilbert, J.A., Krajmalnik-Brown, R., Porazinska, D.L., Weiss, S.J. & Knight, R. Toward effective probiotics for autism and other neurodevelopmental disorders. *Cell* **155**, 1446-1448 (2013).
11. Hsiao, E.Y. et al. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* **155**, 1451-1463 (2013).
12. Lozupone, C.A. et al. Meta-analyses of studies of the human microbiota. *Genome research* **23**, 1704-1714 (2013).
13. Metcalf, J.L. et al. A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. *eLife* **2**, e01104 (2013).
14. Pechal, J.L. et al. The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. *International journal of legal medicine* **128**, 193-205 (2014).
15. Hillenmeyer, M.E. et al. Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. *Genome Biol* **11**, R30 (2010).
16. Hillenmeyer, M.E. et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362-365 (2008).
17. Nichols, R.J. et al. Phenotypic landscape of a bacterial cell. *Cell* **144**, 143-156 (2011).
18. Lynch, M.D., Warnecke, T. & Gill, R.T. SCALEs: multiscale analysis of library enrichment. *Nat Methods* **4**, 87-93 (2007).
19. Warner, J.R., Reeder, P.J., Karimpour-Fard, A., Woodruff, L.B. & Gill, R.T. Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. *Nature biotechnology* **28**, 856-862 (2010).

20. Woodruff, L.B.A. & Gill, R.T. Engineering genomes in multiplex. *Curr Opin Biotech* **22**, 576-583 (2011).
21. Gilbert, J.A., Jansson, J.K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol* **12**, 69 (2014).
22. K, P. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* **60**, 489-502 (1897).
23. Aitchison, J. The statistical analysis of compositional data. (Chapman and Hall, London ; New York; 1986).
24. Friedman, J. & Alm, E.J. Inferring correlation networks from genomic survey data. *PLoS computational biology* **8**, e1002687 (2012).
25. Buccianti, A., Mateu-Figueras, G. & Pawlowsky-Glahn, V. Compositional data analysis in the geosciences : from theory to practice. (The Geological Society, London; 2006).
26. Lovell D, u.W., Taylor J, Zwart A, Helliwell C Caution! compositions! can constraints on omics data lead analyses astray? *CSIRO*, 1-44 (2010).
27. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S. & Bahler, J. Proportionality: a valid alternative to correlation for relative data. *PLoS computational biology* **11**, e1004075 (2015).
,28. Paulson, J.N., Stine, O.C., Bravo, H.C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature methods* **10**, 1200-1202 (2013).
29. Caporaso, J.G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**, 335-336 (2010).
30. McMurdie, P.J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS computational biology* **10** (2014).
31. Schloss, P.D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75**, 7537-7541 (2009).
32. Walsh, C. Molecular mechanisms that confer antibacterial drug resistance. *Nature* **406**, 775-781 (2000).
33. Boucher, H.W. Challenges in anti-infective development in the era of bad bugs, no drugs: a regulatory perspective using the example of bloodstream infection as an indication. *Clin Infect Dis* **50 Suppl 1**, S4-9 (2010).
34. Andersson, D.I. & Hughes, D. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nature reviews. Microbiology* **8**, 260-271 (2010).
35. Andersson, D.I. & Hughes, D. Evolution of antibiotic resistance at non-lethal drug concentrations. *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy* **15**, 162-172 (2012).
36. Giaever, G. et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-391 (2002).
37. Soo, V.W., Hanson-Manful, P. & Patrick, W.M. Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli. *Proc Natl Acad Sci U S A* **108**, 1484-1489 (2011).
38. Lazar, V. et al. Bacterial evolution of antibiotic hypersensitivity. *Mol Syst Biol* **9**, 700 (2013).
39. Kosuri, S. & Church, G.M. Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods* **11**, 499-507 (2014).

40. Sims, D., Sudbery, I., Ilott, N.E., Heger, A. & Ponting, C.P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**, 121-132 (2014).
41. Gill, R.T., Wildt, S., Yang, Y.T., Ziesman, S. & Stephanopoulos, G. Genome-wide screening for trait conferring genes using DNA microarrays. *Proc Natl Acad Sci U S A* **99**, 7033-7038 (2002).
42. Jin, Y.S. & Stephanopoulos, G. Multi-dimensional gene target search for improving lycopene biosynthesis in Escherichia coli. *Metab Eng* **9**, 337-347 (2007).
43. Mazurkiewicz, P., Tang, C.M., Boone, C. & Holden, D.W. Signature-tagged mutagenesis: barcoding mutants for genome-wide screens. *Nat Rev Genet* **7**, 929-939 (2006).
44. Wang, H.H. et al. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894-898 (2009).
45. Mehta, M.L. Random Matrices, 2nd edition. (Academic Press, 1990).
46. Breiman, L. Random forests. *Machine Learning* **45**, 5-32 (2001).
47. Knights, D., Costello, E.K. & Knight, R. Supervised classification of human microbiota. *Fems Microbiol Rev* **35**, 343-359 (2011).
48. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).
49. Wigner, E.P. Random Matrices in Physics. *Siam Rev* **9**, 1-& (1967).
50. Anderson, M.J. A new method for non-parametric multivariate analysis of variance. *Austral Ecol* **26**, 32-46 (2001).
51. Hofstetter, E. & Schreiber, M. Statistical properties of the eigenvalue spectrum of the three-dimensional Anderson Hamiltonian. *Physical review. B, Condensed matter* **48**, 16979-16985 (1993).
52. Vijay-Kumar, M. et al. Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science* **328**, 228-231 (2010).
53. Salter, S.J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12**, 87 (2014).
54. Kennedy, K., Hall, M.W., Lynch, M.D., Moreno-Hagelsieb, G. & Neufeld, J.D. Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. *Applied and environmental microbiology* **80**, 5717-5722 (2014).
55. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214 (2012).
56. Turner, P. et al. A longitudinal study of Streptococcus pneumoniae carriage in a cohort of infants and their mothers on the Thailand-Myanmar border. *PloS one* **7**, e38271 (2012).
57. Leek, J.T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics* **11**, 733-739 (2010).
58. Tanner, M.A., Goebel, B.M., Dojka, M.A. & Pace, N.R. Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Applied and environmental microbiology* **64**, 3110-3113 (1998).
59. Wu, G.D. et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105-108 (2011).

60. Henderson, G. et al. Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. *PloS one* **8**, e74787 (2013).
61. Gower, J.C. Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika* **53**, 325-& (1966).
62. Kelly, C.R. et al. Fecal microbiota transplant for treatment of Clostridium difficile infection in immunocompromised patients. *The American journal of gastroenterology* **109**, 1065-1071 (2014).
63. Shankar, V. et al. Species and genus level resolution analysis of gut microbiota in Clostridium difficile patients following fecal microbiota transplantation. *Microbiome* **2**, 13 (2014).
64. Weingarden, A. et al. Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection. *Microbiome* **3**, 10 (2015).
65. Pop, M. et al. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome biology* **15**, R76 (2014).
66. David, L.A. et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559-563 (2014).
67. Lozupone, C.A. et al. Meta-analyses of studies of the human microbiota. *Genome Res* **23**, 1704-1714 (2013).
68. Qin, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65 (2010).
69. Rodriguez, R.L. & Konstantinidis, K.T. Estimating coverage in metagenomic data sets and why it matters. *The ISME journal* **8**, 2349-2351 (2014).
70. Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: an effective distance metric for microbial community comparison. *The ISME journal* **5**, 169-172 (2011).
71. Lovell D, e.a. Caution! compositions! can constraints on omics data lead analyses astray? *CSIRO*, 1-44 (2010).
72. Aitchison, J. The Statistical-Analysis of Compositional Data. *J Roy Stat Soc B Met* **44**, 139-177 (1982).
73. Gotelli, N.J. & Colwell, R.K. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* **4**, 379-391 (2001).
74. Brewer, A. & Williamson, M. A New Relationship for Rarefaction. *Biodivers Conserv* **3**, 373-379 (1994).
75. Horner-Devine, M.C., Lage, M., Hughes, J.B. & Bohannan, B.J. A taxa-area relationship for bacteria. *Nature* **432**, 750-753 (2004).
76. Jernvall, J. & Wright, P.C. Diversity components of impending primate extinctions. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 11279-11283 (1998).
77. Jari Oksanen, F.G.B., Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens and Helene Wagner vegan: Community Ecology Package. *R package version 2.2-1* (2015).
78. McMurdie, P.J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PloS one* **8** (2013).

79. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, R25 (2010).
80. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* **11**, 94 (2010).
81. Dillies, M.A. et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics* **14**, 671-683 (2013).
82. Agresti, A. & Hitchcock, D.B. Bayesian inference for categorical data analysis : a survey. (University of Florida, Gainesville, Fla.?).
83. Pearson, K. Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proc. Roy. Soc.* **60**, 489-498 (1896).
84. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barcelo-Vidal, C. Isometric logratio transformations for compositional data analysis. *Math Geol* **35**, 279-300 (2003).
85. Greenacre, M. Measuring Subcompositional Incoherence. *Math Geosci* **43**, 681-693 (2011).
86. Costea, P.I., Zeller, G., Sunagawa, S. & Bork, P. A fair comparison. *Nature methods* **11**, 359 (2014).
87. Paulson, J.N., Bravo, H.C. & Pop, M. Reply to: "a fair comparison". *Nature methods* **11**, 359-360 (2014).
88. Wagner, B.D., Robertson, C.E. & Harris, J.K. Application of two-part statistics for comparison of sequence variant counts. *PloS one* **6**, e20296 (2011).
89. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11** (2010).
90. Anders, S. et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* **8**, 1765-1786 (2013).
91. Love MI, H.W.a.A.S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15** (2014).
92. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
93. Law, C.W., Chen, Y.S., Shi, W. & Smyth, G.K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* **15** (2014).
94. Robinson, M.D. & Smyth, G.K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**, 321-332 (2008).
95. Rapaport, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology* **14**, R95 (2013).
96. Cameron, A.C. & Trivedi, P.K. Regression analysis of count data, Edn. Second edition.
97. White, J.R., Nagarajan, N. & Pop, M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS computational biology* **5**, e1000352 (2009).
98. Connolly, S.R., Dornelas, M., Bellwood, D.R. & Hughes, T.P. Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. *Ecology* **90**, 3138-3149 (2009).

99. Cheung, Y.B. Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in medicine* **21**, 1461-1469 (2002).
100. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one* **7**, e30126 (2012).
101. Auer, P.L. & Doerge, R.W. Statistical design and analysis of RNA sequencing data. *Genetics* **185**, 405-416 (2010).
102. Yu, D., Huber, W. & Vitek, O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics* **29**, 1275-1282 (2013).
103. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics* **14**, 91 (2013).
104. Caporaso, J.G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* **108 Suppl 1**, 4516-4522 (2011).
105. Kaufman L., R.P. Finding Groups in Data: An introduction to Cluster Analysis. (JohnWiley & Sons, 1990).
106. Reynolds A, R.G., Iglesia B, Rayward-Smith V Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* **5**, 475-504 (2006).
107. Colwell, R.K. et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J Plant Ecol-Uk* **5**, 3-21 (2012).
108. Bray, J.R. & Curtis, J.T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol Monogr* **27**, 326-349 (1957).
109. Witten, D.M. Classification and Clustering of Sequencing Data Using a Poisson Model. *Ann Appl Stat* **5**, 2493-2518 (2011).
110. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* **71**, 8228-8235 (2005).
111. Lozupone, C.A., Hamady, M., Kelley, S.T. & Knight, R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology* **73**, 1576-1585 (2007).
112. Salter, S.J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology* **12**, 87 (2014).
113. Piombino, P. et al. Saliva from obese individuals suppresses the release of aroma compounds from wine. *PloS one* **9**, e85611 (2014).
114. Team, R.C. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* (2014).
115. Gentleman, R.C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5** (2004).
116. Vazquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: a tool for visualizing high-throughput microbial community data. *GigaScience* **2**, 16 (2013).
117. Caporaso, J.G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal* **6**, 1621-1624 (2012).

118. Lozupone, C.A. & Knight, R. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 11436-11440 (2007).
119. Lauber, C.L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and environmental microbiology* **75**, 5111-5120 (2009).
120. Costello, E.K. et al. Bacterial community variation in human body habitats across space and time. *Science* **326**, 1694-1697 (2009).
121. Caporaso, J.G. et al. Moving pictures of the human microbiome. *Genome biology* **12**, R50 (2011).
122. Legendre, P. & Gallagher, E.D. Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**, 271-280 (2001).
123. Carcer, D.A., Denman, S.E., McSweeney, C. & Morrison, M. Evaluation of subsampling-based normalization strategies for tagged high-throughput sequencing data sets from gut microbiomes. *Applied and environmental microbiology* **77**, 8795-8798 (2011).
124. Robles, J.A. et al. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC genomics* **13**, 484 (2012).
125. Gill-King, H. in Forensic Taphonomy: The Postmortem Fate of Human Remains. (ed. e. W. D. Haglund and M. H. Sorg) (CRC Press, Boca Raton, FL.; 1997).
126. Correy, J.E.L. Possible sources of ethanol ante- and post-mortem: its relationships to the biochemistry and microbiology of decomposition. *Journal of Applied Bacteriology* **44**, 1-56 (1978).
127. Pechal, J.L., T.L. Crippen, M.E. Benbow, A.M. Tarone, S. Dowd, and J.K. Tomberlin The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. *International Journal of Legal Medicine* **128**, 193-205 (2014).
128. Haskell, N.H.a.E.P.C., editors. Entomology and Death: A Procedural Guide., Edn. Second. (East Park Printing, 2008).
129. Putman, R.J. Patterns of carbon dioxide evolution from decaying carrion. l. Decomposition of small mamman carrion in temperate systems. *Oikos* **31**, 47-57 (1978).
130. Carter, D.O., D. Yellowlees, and M. Tibbett. Temperature affects microbial decomposition of cadavers (Rattus rattus) in contrasting soils. *Applied Soil Ecology* **40**, 129-137 (2008).
131. Carter, D.O., D. Yellowlees, and M. Tibbett. Moisture can be the dominant environmental parameter governing cadaver decomposition in soil. *Forensic Science International* **200**, 60-66 (2010).
132. Ibekwe, A.M. et al. Microbial diversity along a transect of agronomic zones. *FEMS microbiology ecology* **39**, 183-191 (2002).
133. Kuske, C.R. et al. Comparison of soil bacterial communities in rhizospheres of three plant species and the interspaces in an arid grassland. *Applied and environmental microbiology* **68**, 1854-1863 (2002).

134. Haslam, T.C. & Tibbett, M. Soils of contrasting pH affect the decomposition of buried mammalian (Ovis aries) skeletal muscle tissue. *Journal of forensic sciences* **54**, 900-904 (2009).
135. Lauber, C.L. et al. Vertebrate decomposition is accelerated by soil microbes. *Applied and environmental microbiology* **80**, 4920-4929 (2014).
136. Sagara, N., T. Yamanaka, and M. Tibbet in Soil Analysis in Forensic Taphonomy: Chemical and Biological Effects of Buried Human Remains. (ed. e. M. Tibbett and D.O. Carter) (CRC Press, Boca Raton, FL, USA; 2008).
137. Tibbett, M. & Carter, D.O. Soil analysis in forensic taphonomy : chemical and biological effects of buried human remains. (CRC Press, Boca Raton; 2008).
138. Anderson, G.S. The use of insects in death investigations: an analysis of cases in British Columbia over a five year period. *Canadian Society of Forensic Science Journal* **28**, 277-292 (1995).
139. Anderson, G.S.a.S.L.V. Initial studies on insect succession on carrion in southwestern British Columbia. *Journal Forensic Science* **41**, 617-625 (1996).
140. Campobasso, C.P. & Introna, F. The forensic entomologist in the context of the forensic pathologist's role. *Forensic Sci Int* **120**, 132-139 (2001).
141. Hewadikaram KA, a.G.M. Effect of carcass size on rate of decomposition and arthropod succession patterns. *American Journal of Forensic Medicine and Pathology* **12**, 235-240 (1991).
142. Komar D, B.O. Effects of carcass size on decay rates of shade and sun exposed carrion. *Can Soc Forensic Sci J* **31**, 35-43 (1998).
143. Simmons, T., Adlam, R.E. & Moffatt, C. Debugging decomposition data--comparative taphonomic studies and the influence of insects and carcass size on decomposition rate. *Journal of forensic sciences* **55**, 8-13 (2010).
144. Spicka, A., R. Johnson, J. Bushing, L. G. Higley, and D. O. Carter Carcass mass can influence rate of decomposition and release of ninhydrin-reactive nitrogen. *Forensic Science International* **209**, 80-85 (2011).
145. Sutherland, A., Myburgh, J., Steyn, M. & Becker, P.J. The effect of body size on the rate of decomposition in a temperate region of South Africa. *Forensic Sci Int* **231**, 257-262 (2013).
146. Matuszewski, S., Konwerski, S., Fratczak, K. & Szafalowicz, M. Effect of body mass and clothing on decomposition of pig carcasses. *Int J Legal Med* **128**, 1039-1048 (2014).
147. Kuusela, S. & Hanski, I. The Structure of Carrion Fly Communities - the Size and the Type of Carrion. *Holarctic Ecol* **5**, 337-348 (1982).
148. Kneidel, K.A. Influence of Carcass Taxon and Size on Species Composition of Carrion-Breeding Diptera. *Am Midl Nat* **111**, 57-63 (1984).
149. Arnold, C.Y. The determination and significance of the base temperature in a linear heat unit system. *Proc. Am. Soc. Hortic. Sci.* **74**, 430-445 (1959).
150. A. A. Vass, W.M.B., J.D. Wolt, J.E. Foss, J.T. Ammons Time since death determinations of human cadavers using soil solution. *Journal of forensic sciences* **37**, 1236-1253 (1992).
151. Carter, D.O., Yellowlees, D. & Tibbett, M. Cadaver decomposition in terrestrial ecosystems. *Die Naturwissenschaften* **94**, 12-24 (2007).

152. McDonald, D. et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *Isme Journal* **6**, 610-618 (2012).
153. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research* **41**, D590-596 (2013).
154. Anderson, M.J. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32-46 (2001).
155. Jari Oksanen, F.G.B., Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens and Helene Wagner vegan: Community Ecology Package. (2015).
156. Team, R.C. (Vienna, Austria; 2014).
157. Meyer, J., Anderson, B. & Carter, D.O. Seasonal variation of carcass decomposition and gravesoil chemistry in a cold (Dfa) climate. *Journal of forensic sciences* **58**, 1175-1182 (2013).
158. Pechal, J.L. et al. Microbial community functional change during vertebrate carrion decomposition. *PloS one* **8**, e79035 (2013).
159. LeBlanc, H.N. & Logan, J.G. Exploiting Insect Olfaction in Forensic Entomology. *Current Concepts in Forensic Entomology*, 205-221 (2010).
160. Tomberlin, J.K., Benbow, M.E., Tarone, A.M. & Mohr, R.M. Basic research in evolution and ecology enhances forensics. *Trends in ecology & evolution* **26**, 53-55 (2011).
161. Ma, Q. et al. Proteus mirabilis interkingdom swarming signals attract blow flies. *Isme Journal* **6**, 1356-1366 (2012).
162. Tomberlin, J.K. et al. Interkingdom responses of flies to bacteria mediated by fly physiology and bacterial quorum sensing. *Anim Behav* **84**, 1449-1456 (2012).
163. De Wit, R.a.T.B. 'Everything is everywhere, but the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology* **8**, 755-758 (2006).
164. Bongers, T. & Ferris, H. Nematode community structure as a bioindicator in environmental monitoring. *Trends in ecology & evolution* **14**, 224-228 (1999).
165. Gunstone, F.D. An Introduction to the Chemistry and Biochemistry of Fatty Acids and their Glycerides. (Chapman Hall, London, UK; 1967).
166. Ridgway, E.J.a.D.J.H. in The Hospital Autopsy: A Manual of Fundamental Autopsy Practice. (ed. e. J. L. Burton and G. Rutty) (CRC Press, Boca Raton, FL, USA; 2010).
167. Koontz, C.M.J.a.F.P. Cumitech 35 Postmortem Microbiology. . *ASM Press, Washington D.C.* (2001).
168. Hyde, E.R., Haarmann, D.P., Lynne, A.M., Bucheli, S.R. & Petrosino, J.F. The living dead: bacterial community structure of a cadaver at the onset and end of the bloat stage of decomposition. *PloS one* **8**, e77733 (2013).
169. Maujean, G., Guinet, T., Fanton, L. & Malicier, D. The interest of postmortem bacteriology in putrefied bodies. *Journal of forensic sciences* **58**, 1069-1070 (2013).
170. Fierer, N. et al. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 6477-6481 (2010).
171. Beman, J.M., Steele, J.A. & Fuhrman, J.A. Co-occurrence patterns for abundant marine archaeal and bacterial lineages in the deep chlorophyll maximum of coastal California. *The ISME journal* **5**, 1077-1085 (2011).

172. Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome research* **20**, 947-959 (2010).
173. Steele, J.A. et al. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME journal* **5**, 1414-1425 (2011).
174. Lozupone, C. et al. Identifying genomic and metabolic features that can underline early successional and opportunistic lifestyles of human gut symbionts. *Genome research* **22**, 1974-1984 (2012).
175. Ruan, Q. et al. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* **22**, 2532-2538 (2006).
176. Zhou, J., Deng, Y., Luo, F., He, Z. & Yang, Y. Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO2. *mBio* **2** (2011).
177. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174-180 (2011).
178. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature reviews. Microbiology* **10**, 538-550 (2012).
179. Faust, K. et al. Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology* **8**, e1002606 (2012).
180. Greenblum, S., Turnbaugh, P.J. & Borenstein, E. 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 of the United States of America* **109**, 594-599 (2012).
181. Oakley, B.B. et al. The poultry-associated microbiome: network analysis and farm-to-fork characterizations. *PloS one* **8**, e57190 (2013).
182. Reshef, D.N. et al. Detecting novel associations in large data sets. *Science* **334**, 1518-1524 (2011).
183. Bray, J.R., Curtis, J.T. An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs* **27**, 325-349 (1957).
184. Pearson, K. Determination of the Coefficient of Correlation. *Science* **30**, 23-25 (1909).
185. Spearman, C. The proof and measurement of association between two things. *The American journal of psychology* **15**, 72-101 (1904).
186. Tarca, A.L. et al. Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge. *Bioinformatics* **29**, 2892-2899 (2013).
187. Deng, Y. et al. Molecular ecological network analyses. *BMC bioinformatics* **13**, 113 (2012).
188. Bokulich, N.A. et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature methods* **10**, 57-59 (2013).
189. Trivedi, P.K. & Zimmer, D.M. Copula modeling : an introduction for practitioners. (Now, Boston; 2007).
190. Brown, M. A method for combining non-independent, one-sided tests of significance. *Biometrics* **31**, 987-992 (1975).

191. Ponnusamy, L. et al. Diversity of Rickettsiales in the microbiome of the lone star tick, Amblyomma americanum. *Applied and environmental microbiology* **80**, 354-359 (2014).
192. Kolmogorov, A.N. Sulla Determinazione Empirica di una Legge di Distribuzione. *Giornale dell' Instituto Italiano delgi Attuari* **4**, 83-91 (1933).
193. N, S. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* **19**, 279-281 (1948).
194. Fisher, R.A. On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron* **1**, 3-32 (1921).
195. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222-227 (2012).
196. Kullback, S. & Leibler, R.A. On Information and Sufficiency. *Annals of Mathematical Statistics* **22**, 79-86 (1951).
197. Dunn, O.J. Multiple Comparisons Among Means. *Journal of the American Statistical Association* **56**, 52-64 (1961).
198. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Statistics in medicine* **9**, 811-818 (1990).
199. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445 (2003).
200. Shade, A. et al. Fundamentals of microbial community resistance and resilience. *Frontiers in microbiology* **3**, 417 (2012).
201. Mounier, J. et al. Microbial interactions within a cheese microbial community. *Applied and environmental microbiology* **74**, 172-181 (2008).
202. Pepper, J.W. & Rosenfeld, S. The emerging medical ecology of the human gut microbiome. *Trends in ecology & evolution* **27**, 381-384 (2012).
203. Gonzalez, A. et al. Characterizing microbial communities through space and time. *Current opinion in biotechnology* **23**, 431-436 (2012).
204. Shade, A., Caporaso, J.G., Handelsman, J., Knight, R. & Fierer, N. A meta-analysis of changes in bacterial and archaeal communities with time. *The ISME journal* **7**, 1493-1506 (2013).
205. Altshuler, B.L. & Shklovskii, B.I. Repulsion of Energy-Levels and the Conductance of Small Metallic Samples. *Zh Eksp Teor Fiz+* **91**, 220-234 (1986).
206. Zhong, J.X. & Geisel, T. Level fluctuations in quantum systems with multifractal eigenstates. *Phys Rev E* **59**, 4071-4074 (1999).
207. Bohigas, O., Giannoni, M.J. & Schmit, C. Spectral Properties of the Laplacian and Random Matrix Theories. *J Phys Lett-Paris* **45**, 1015-1022 (1984).
208. Seba, P. Random matrix analysis of human EEG data. *Physical review letters* **91**, 198104 (2003).
209. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N. & Stanley, H.E. Universal and nonuniversal properties of cross correlations in financial time series. *Physical review letters* **83**, 1471-1474 (1999).
210. Luo, F. et al. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC bioinformatics* **8** (2007).
211. Zhou, A.F. et al. Characterization of NaCl tolerance in Desulfovibrio vulgaris Hildenborough through experimental evolution. *Isme Journal* **7**, 1790-1802 (2013).

212. Xia, L.C., Ai, D., Cram, J., Fuhrman, J.A. & Sun, F. Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. *Bioinformatics* **29**, 230-237 (2013).
213. Storey, J.D. A direct approach to false discovery rates. *J Roy Stat Soc B* **64**, 479-498 (2002).
214. Fisher, R.A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507-521 (1915).
215. Volterra, V. Le¸cons sur la th´eorie math´ematique de la lutte pour la vie. (Gauthier-Villars, 1931).
216. Idema, T. in Mathematics, Vol. Doctorate 66 (Leiden University, 2005).
217. Winzeler, E.A. et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-906 (1999).
218. Baba, T. et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008 (2006).
219. Toprak, E. et al. Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. *Nat Protoc* **8**, 555-567 (2013).
220. Smith, A.M. et al. Quantitative phenotyping via deep barcode sequencing. *Genome Res* **19**, 1836-1842 (2009).
221. Sandoval, N.R. et al. Strategy for directing combinatorial genome engineering in Escherichia coli. *Proc Natl Acad Sci U S A* **109**, 10540-10545 (2012).
222. Diep, B.A. et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. *Lancet* **367**, 731-739 (2006).
223. Comas, I. et al. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet* **44**, 106-110 (2012).
224. Piras, C. et al. Comparative proteomics to evaluate multi drug resistance in Escherichia coli. *Mol Biosyst* **8**, 1060-1067 (2012).
225. Karatzas, K.A. et al. Phenotypic and proteomic characterization of multiply antibiotic-resistant variants of Salmonella enterica serovar Typhimurium selected following exposure to disinfectants. *Appl Environ Microbiol* **74**, 1508-1516 (2008).
226. Datsenko, K.A. & Wanner, B.L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A* **97**, 6640-6645 (2000).
227. Neidhardt, F.C., Bloch, P.L. & Smith, D.F. Culture medium for enterobacteria. *J Bacteriol* **119**, 736-747 (1974).
228. Andrews, J.M. Determination of minimum inhibitory concentrations. *J Antimicrob Chemother* **48 Suppl 1**, 5-16 (2001).
229. Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nature Methods* **7**, 119-U147 (2010).
230. Knight, R. et al. PyCogent: a toolkit for making sense from sequence. *Genome Biol* **8**, R171 (2007).
231. Bloom, S.A. Similarity Indexes in Community Studies - Potential Pitfalls. *Mar Ecol Prog Ser* **5**, 125-128 (1981).

232. Hurley, J.R. & Cattell, R.B. The Procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science* **7**, 258-262 (1962).
233. Knights, D. et al. Supervised classification of microbiota mitigates mislabeling errors. *The ISME journal* **5**, 570-573 (2011).
234. Krzywinski, M. et al. Circos: An information aesthetic for comparative genomics. *Genome Research* **19**, 1639-1645 (2009).
235. Tatusov, R.L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
236. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).
237. Clarke, K.R. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* **18**, 117-143 (1993).
238. Martin, R.G. & Rosner, J.L. Transcriptional and translational regulation of the marRAB multiple antibiotic resistance operon in Escherichia coli. *Mol Microbiol* **53**, 183-191 (2004).
239. Ruiz, C. & Levy, S.B. Many chromosomal genes modulate MarA-mediated multidrug resistance in Escherichia coli. *Antimicrob Agents Chemother* **54**, 2125-2134 (2010).
240. Kadrmas, J.L. & Raetz, C.R. Enzymatic synthesis of lipopolysaccharide in Escherichia coli. Purification and properties of heptosyltransferase i. *J Biol Chem* **273**, 2799-2807 (1998).
241. Tamaki, S., Sato, T. & Matsuhashi, M. Role of lipopolysaccharides in antibiotic resistance and bacteriophage adsorption of Escherichia coli K-12. *J Bacteriol* **105**, 968-975 (1971).
242. Goswami, M., Mangoli, S.H. & Jawali, N. Involvement of reactive oxygen species in the action of ciprofloxacin against Escherichia coli. *Antimicrob Agents Chemother* **50**, 949-954 (2006).
243. Kaldalu, N., Mei, R. & Lewis, K. Killing by ampicillin and ofloxacin induces overlapping changes in Escherichia coli transcription profile. *Antimicrob Agents Chemother* **48**, 890-896 (2004).
244. Nishino, K. & Yamaguchi, A. Analysis of a complete library of putative drug transporter genes in Escherichia coli. *J Bacteriol* **183**, 5803-5812 (2001).
245. Lawhorn, B.G., Gerdes, S.Y. & Begley, T.P. A genetic screen for the identification of thiamin metabolic genes. *J Biol Chem* **279**, 43555-43559 (2004).
246. Spoering, A.L., Vulic, M. & Lewis, K. GlpD and PlsB participate in persister cell formation in Escherichia coli. *J Bacteriol* **188**, 5136-5144 (2006).
247. Doi, Y. & Arakawa, Y. 16S ribosomal RNA methylation: emerging resistance mechanism against aminoglycosides. *Clin Infect Dis* **45**, 88-94 (2007).
248. Kwon, D.H. et al. High-level beta-lactam resistance associated with acquired multidrug resistance in Helicobacter pylori. *Antimicrob Agents Chemother* **47**, 2169-2178 (2003).
249. Webber, M.A. & Piddock, L.J. The importance of efflux pumps in bacterial antibiotic resistance. *J Antimicrob Chemother* **51**, 9-11 (2003).
250. Warnecke, T.E. et al. Rapid dissection of a complex phenotype through genomic-scale mapping of fitness altering genes. *Metab Eng* **12**, 241-250 (2010).
251. Forsberg, K.J. et al. Bacterial phylogeny structures soil resistomes across habitats. *Nature* **509**, 612-616 (2014).

252. Sommer, M.O., Dantas, G. & Church, G.M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325**, 1128-1131 (2009).
253. Robinson, D.G., Chen, W., Storey, J.D. & Gresham, D. Design and analysis of Bar-seq experiments. *G3 (Bethesda)* **4**, 11-18 (2014).
254. Lax, S. et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345**, 1048-1052 (2014).
255. Halko, N., Martinsson, P.G., Shkolnisky, Y. & Tygert, M. An Algorithm for the Principal Component Analysis of Large Data Sets. *Siam J Sci Comput* **33**, 2580-2594 (2011).
256. Halko, N., Martinsson, P.G. & Tropp, J.A. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *Siam Rev* **53**, 217-288 (2011).
257. Woodruff, L.B. et al. Genome-scale identification and characterization of ethanol tolerance genes in Escherichia coli. *Metabolic engineering* **15**, 124-133 (2013).
258. Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-+ (2013).
259. Aitchison, J. A New Approach to Null Correlations of Proportions. *J Int Ass Math Geol* **13**, 175-189 (1981).
260. Goodrich, J.K. et al. Human genetics shape the gut microbiome. *Cell* **159**, 789-799 (2014).