

## ABSTRACT

Title of Document: NOVEL METHODS FOR METAGENOMIC ANALYSIS

James Robert White, Doctor of Philosophy, 2010

Directed By: Professor Mihai Pop, Department of Computer Science, Applied Mathematics, Statistics, and Scientific Computation Program Affiliate

By sampling the genetic content of microbes at the nucleotide level, metagenomics has rapidly established itself as the standard in characterizing the taxonomic diversity and functional capacity of microbial populations throughout nature. The decreasing cost of sequencing technologies and the simultaneous increase of throughput per run has given scientists the ability to deeply sample highly diverse communities on a reasonable budget. The Human Microbiome Project is representative of the flood of sequence data that will arrive in the coming years. Despite these advancements, there remains the significant challenge of analyzing massive metagenomic datasets to make appropriate biological conclusions. This dissertation is a collection of novel methods developed for improved analysis of metagenomic data: (1) We begin with Figaro, a statistical algorithm that quickly and accurately infers and trims vector sequence from large Sanger-based read sets without prior knowledge of the vector used in library construction. (2) Next, we perform a rigorous evaluation of methodologies used to

cluster environmental 16S rRNA sequences into species-level operational taxonomic units, and discover that many published studies utilize highly stringent parameters, resulting in overestimation of microbial diversity. (3) To assist in comparative metagenomics studies, we have created Metastats, a robust statistical methodology for comparing large-scale clinical datasets with up to thousands of subjects. Given a collection of annotated metagenomic features (e.g. taxa, COGs, or pathways), Metastats determines which features are differentially abundant between two populations. (4) Finally, we report on a new methodology that employs the generalized Lotka-Volterra model to infer microbe-microbe interactions from longitudinal 16S rRNA data. It is our hope that these methods will enhance standard metagenomic analysis techniques to provide better insight into the human microbiome and microbial communities throughout our world. To assist metagenomics researchers and those developing methods, all software described in this thesis is open-source and available online.

NOVEL METHODS FOR METAGENOMIC ANALYSIS

By

James Robert White

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2010

Advisory Committee:  
Professor Mihai Pop, Chair  
Professor Steven Salzberg  
Professor James A. Yorke  
Professor Najib El-Sayed  
Professor Charles Delwiche (Dean's Representative)

© Copyright by  
James Robert White  
2010

## Preface

The four works in this thesis have either been published in peer-reviewed journals or are currently in preparation for publication. I am indebted to each of my co-authors on these studies – their dedication and expertise resulted in much stronger scientific papers both in statistical rigor and biological scope. At the time of this writing, Chapters 2 and 4 have appeared in print and are slightly modified here. Chapter 3 has been accepted for publication. Permission for republication of this material has been granted and is available upon request.

### Chapter 2:

**White, J.R., Roberts, M., Yorke, J.A. and M. Pop, *Figaro: a novel statistical method for vector sequence removal*. *Bioinformatics*, 2008. 24(4): p. 462-7.**

*We thank Art Delcher and Aleksey Zimin for their suggestions and feedback. We are grateful to Steven Salzberg for help naming our method. M.P. was supported in part by grant R01-LM006845 from NIH and grant HU001-06-1-0015 from the Uniformed Services University of the Health Sciences administered by the Henry Jackson Foundation. J.Y. and M.R. were supported in part under NSF grant DMS 0616585 and NIH grant 1R01HG0294501. J.W. was supported in part by University of Maryland NSF VIGRE fellowship DMS0240049.*

### Chapter 3:

**White, J.R., Navlakha, S., Nagarajan, N., Ghodsi, M., Kingsford, C. and M. Pop. *Alignment and clustering of phylogenetic markers – implications for microbial diversity studies*. *BMC Bioinformatics*, 2010. 11(152).**

### Chapter 4:

**White, J.R., Nagarajan, N. and M. Pop, *Statistical methods for detecting differentially abundant features in clinical metagenomic samples*. *PLoS Comput Biol*, 2009. 5(4): p. e1000352.**

*Funded in part by Bill and Melinda Gates Foundation and the Henry Jackson Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are very grateful to Jeff Gordon, Peter Turnbaugh, and Ruth Ley for their input and assistance with data. We thank Frank Siewerdt, Aleksey Zimin, Radu Balan, Kunmi Ayanbule, and Kenneth Ryals for helpful discussions. Finally we would like to thank the anonymous reviewers for their constructive comments. Our manuscript was greatly improved by their contribution.*

Chapter 5:

**White, J.R., Turnbaugh, P., Paulson, J., Gordon, J.I. and M. Pop. *Inferring microbial interaction webs from time-series metagenomic data.* In preparation.**

## Dedication

In early 2008, I experienced a profound life-changing epiphany.

I dedicate this dissertation to the person I was before then.

I finished this for you, I miss you, and I will not forget you.

## Acknowledgements

First and foremost I wish to acknowledge the dedication of my advisor Mihai Pop. Mihai introduced me to the field of metagenomics and helped to shape me into a meticulous researcher. I am indebted to Jim Yorke, my Coach, who has had a larger impact on my professional direction than anyone else. Jim guided me from pure to applied mathematics, a path that resulted in more personal satisfaction and excitement than I could have imagined. I must also thank other members of my defense committee (Steven Salzberg, Chuck Delwiche and Najib El-Sayed) for their input and appreciation of this work. I have had many excellent teachers over the years, but Linda Barnes and Dianne O’leary hold a special place – both of your approaches to teaching matched my learning style perfectly.

I undoubtedly would not have been able to finish this thesis if I had not been in such wonderful company. In the throws of this challenging experience, I met some of the best people I will ever know. Mike, Adam, Dave, and Saket – working alongside you has been fantastic, I wouldn’t trade it for anything, and I can’t wait to see where we all end up. Smeds, Tom, Rachel, and Jeremy – thank you so much for your friendship and support. Andrea Ottesen – my travel buddy and collaborator – you have made me appreciate the natural world so much more. A special thanks to Denise Cross and Alverda McCoy for helping me sort through sometimes frustrating university policies and keeping me on track to finish.

I finally must express my gratitude to my family. They have supported me through thick and thin, and kept me balanced in times of stress. I've learned how lucky I am to have unconditional love from my parents and sisters, and I hope I continue to cherish this throughout the remaining chapters of my life.

# Table of Contents

Preface .....	ii
Dedication .....	iv
Acknowledgements .....	v
List of Tables .....	ix
List of Figures .....	x
Chapter 1: Introduction.....	1
<u>Early microbiology</u> .....	1
<u>16S rRNA gene surveys and metagenomics</u> .....	2
<u>The Human Microbiome Project</u> .....	6
<u>This work</u> .....	7
<i>Mathematical and computational contributions</i> .....	8
Chapter 2: Figaro – a novel statistical method for vector sequence removal.....	11
<u>Background</u> .....	11
<u>Methods</u> .....	15
Detection of vectormers.....	16
Vector clip estimation.....	19
<u>Results</u> .....	21
Vector trimming sensitivity and specificity.....	21
Improving assemblies with Figaro.....	25
<u>Discussion</u> .....	28
Chapter 3: Alignment and clustering of phylogenetic markers – implications for microbial diversity studies.....	31
<u>Background</u> .....	31
<u>Results</u> .....	33
Simulated environments.....	33
Comprehensive search of OTU methodologies.....	34
OTU variability.....	34
Nonparametric estimators of richness and diversity.....	40
Partial masking of MSAs.....	42
Pairwise versus multiple sequence alignments.....	42
Supervised clustering alternatives.....	45
Consistency of methods across multiple datasets.....	48
<u>Materials &amp; Methods</u> .....	49
Creation of simulated datasets.....	49
Multiple sequence alignment.....	50
Distance corrections and clustering methods.....	50
Measures of similarity for clusterings.....	51
Computation of the Rand index and variation of information for clusterings....	52
VI-cut method for defining OTUs.....	53
ANOVA of methodology components.....	54
Chapter 4: Statistical methods for detecting differentially abundant features in clinical metagenomic samples.....	56
<u>Background</u> .....	56

<u>Materials &amp; Methods</u> .....	58
Data normalization .....	59
Analysis of differential abundance .....	60
Assessing significance .....	61
Multiple hypothesis testing correction .....	63
Handling sparse counts .....	64
Creating the Feature Abundance Matrix .....	66
Data used in this paper .....	66
<u>Results</u> .....	67
Comparison with other statistical methods .....	67
Taxa associated with human obesity .....	75
Differentially abundant COGs between mature and infant human gut microbiomes .....	78
Differentially abundant metabolic subsystems in microbial and viral metagenomes .....	84
<u>Discussion</u> .....	85
Chapter 5: Inferring microbial interaction webs from time-series metagenomic data	87
<u>Background</u> .....	87
<u>Materials &amp; Methods</u> .....	89
Modeling microbial communities .....	89
Learning a model from the data .....	95
An inference methodology with confidence values .....	97
Small interaction network simulation design .....	98
Humanized gnotobiotic mouse gut dataset .....	99
<u>Results</u> .....	101
Prediction of small interaction webs .....	101
Validation of regression approach .....	104
Microbial dynamics of mice on a Western diet .....	108
<u>Discussion</u> .....	113
Chapter 6: Conclusions and further study .....	124
Appendices .....	127
Bibliography .....	140

## List of Tables

Table 1.....	14
Table 2.....	24
Table 3.....	25
Table 4.....	27
Table 5.....	39
Table 6.....	49
Table 7.....	52
Table 8.....	55
Table 9.....	74
Table 10.....	77
Table 11.....	83
Table 12.....	85
Table 13.....	95
Table 14.....	102
Table 15.....	111
Table 16.....	111
Table 17.....	123
Table 18.....	139

## List of Figures

Figure 1.....	12
Figure 2.....	13
Figure 3.....	18
Figure 4.....	18
Figure 5.....	21
Figure 6.....	37
Figure 7.....	38
Figure 8.....	41
Figure 9.....	44
Figure 10.....	47
Figure 11.....	59
Figure 12.....	65
Figure 13.....	69
Figure 14.....	71
Figure 15.....	73
Figure 16.....	103
Figure 17.....	107
Figure 18.....	110
Figure 19.....	118
Figure 20.....	122

## Chapter 1: Introduction

### Early microbiology

As with most new fields of science, microbiology was born out of a major technological innovation – the microscope. Developed in the early 1600s, microscopes had limited application in biology until Antonie van Leeuwenhoek, a clothing merchant and amateur scientist, used a simple magnifying lens of exceptionally high quality to examine water from a lake near his home. Leeuwenhoek discovered a world of “little animalcules” many of which were actual bacteria (though this term did not appear for another 150 years [1]). In 1674 he shared his findings and sketches with the British Royal Society [2], revealing a mysterious and complex world hidden from our view.

Improvements in microscopy through the late 19<sup>th</sup> century (e.g. staining) helped to encourage the field, but observation alone was not sufficient to infer the composition of these organisms or their natural functions. Novel techniques were needed to isolate and study each microorganism independently, but it was impossible to interrogate one cell at a time. Rather, the more practical approach was to study large populations of identical cells – a concept known as *pure culture*. Louis Pasteur and Robert Koch, two of the founders of modern microbiology, designed methods for the isolation, cultivation, and study of pure cultures, and these techniques have defined the field for more than a century.

As researchers discovered many more microbes through cultivation, taxonomic classification posed a significant challenge. Morphological characteristics of bacterial cells were simple and limited, the cells reproduced asexually, and common metabolic

properties were not trustworthy indicators of phylogeny. In 1923, the Society of American Bacteriologists published the first edition of Bergey's Manual of Determinative Bacteriology [3], a reference for the classification of culturable bacteria using morphological, physiological, and experimental characteristics. This reference has expanded dramatically into several large volumes, and now employs some of the molecular techniques I discuss below.

Despite the rapid accumulation of information on culturable microbes in the first half of the 20<sup>th</sup> century, there was a glaring problem - most of the microbial world could not be cultured. This was evidenced by the "great plate count anomaly" in which population abundance estimates determined through microscope density measurements and dilution plating differed by several orders of magnitude [4, 5]. These differences were particularly extreme in soil environments, where it was estimated that less than 1% of the microbial community could be cultured using standard techniques [6]. The challenge of learning anything about this sizable majority of microbes seemed insurmountable, and most scientists focused on microorganisms that could be cultivated.

### *16S rRNA gene surveys and metagenomics*

More than 300 years after the first observation of microbes a new technological innovation exposed microbiology to the unculturable majority. In the late 1970s and early 80s, Carl Woese discovered that the 16S rRNA gene was an excellent phylogenetic marker due to its high information content, structurally conservative nature, and ubiquitous presence among prokaryotes [7-10]. Motivated by this result, Norman Pace and colleagues devised a method for rapidly sequencing 16S rRNA genes to

phylogenetically classify organisms [11]. Augmented by the development of universal PCR primers for 16S gene amplification, Pace's sequencing methodology enabled scientists to sample microbial populations in virtually any habitat without culturing-bias.

Since then, the 16S gene has proven to be one of the most important tools in microbial ecology [12], revealing a vast biodiversity of prokaryotes in many environments such as the ocean [13], soil [14], food products [15, 16], crude oil [17], and even the human gut [18-21]. Analysis of 16S markers now employs high-throughput sequencing technologies (e.g. Sanger and 454 pyrosequencing), which provide deeper sampling to observe community members that make up tiny fractions of the total population. Basic sequence analysis is easily automated, so much so that large computational infrastructures are already in place – webservers such as the Ribosomal Database Project (RDP) [22], GreenGenes [23], and MG-RAST [24] give researchers superior computing power and informative analysis of their data.

After the paradigm shift to 16S rRNA surveys, there were new efforts to obtain more information about these microbes than simply their phylogeny. Researchers now sought to study environmental DNA samples with multiple species using shotgun sequencing. By 1997, the term 'metagenomics' was coined to describe this new approach to environmental microbiology [25]. Over the last decade, several pioneering studies have generated a great deal of interest and set precedents for future projects [26]. Here I discuss three of these landmark studies.

*Acid mine drainage.* Published in 2004, the Acid Mine Drainage project (AMD) sampled biofilms growing on the acidic outflows located deep in the Richmond mine of Iron Mountain, California [27]. This environment had been carefully studied prior to

metagenomic analysis, and preliminary experiments indicated a low-complexity native microbial community with only five dominant species (three Bacteria and two Archaea). Despite this low diversity and a total over 100,000 Sanger reads generated from shotgun sequencing, assemblies of three of the species were largely incomplete. The AMD project illustrated the difficulty in acquiring sufficient genome coverage for organisms with lower relative abundances in a community, but also demonstrated how metagenomic data could be analyzed to infer how microbes potentially interact biochemically in a specific environment.

*The Sargasso Sea.* This study, led by Craig Venter, sought to characterize the microbial diversity in the Sargasso Sea, which represents the middle of the North Atlantic Ocean, east of the Gulf Stream and south of North Atlantic current. Using a series of filters to isolate bacterial and archaeal cells from ocean water, researchers took surface samples at multiple sites and performed extensive shotgun sequencing. Over 1.66 million reads were generated totaling 1.36 billion base-pairs of DNA sequence, far more than any other previous metagenomic study [28]. This amazing volume of data resulted in 1.2 million predicted genes, roughly an order of magnitude greater than the entire SwissProt database at the time. Examining the depth of coverage distribution across the metagenomic assembly, Venter and his team found high phylogenetic diversity in the Sargasso Sea with estimates of at least 1800 species in the environment. Unfortunately, it was later determined that the largest sample taken was contaminated by *Shewanella* and *Burkholderia* species, rendering it useless for ecological analysis [29]. This work was the first to perform significant deep sequencing of a high-complexity microbial population, and through its technical problems, affirmed the importance of sampling techniques,

quality control experiments and validation. Moreover, the Sargasso Sea served as a pilot study for the Global Ocean Sampling project, an around-the-world voyage that collected ocean samples approximately every 200 nautical miles, resulting in 7.7 million shotgun sequences, the largest raw metagenomic dataset to date.

*The Obese Gut Microbiome.* While some scientists used metagenomics to investigate traditional environments like soil and water, others were interested in the structure and function of microbial communities within a host. At Washington University in St. Louis, Jeff Gordon and Peter Turnbaugh wanted to characterize microbes inhabiting the distal gut of obese and lean mice to determine if and how gut microbiota contribute to the pathology of obesity [30]. Using metagenomic and biochemical analyses to compare samples taken from genetically obese mice and their lean littermates, Gordon and Turnbaugh discovered that the obese gut microbiome maintained increased capacity for energy harvest and furthermore, that this trait was transmissible to germ-free mice. This work established an important application of metagenomics: characterization prokaryotic communities in a clinical setting to study how human diseases correlate with microflora.

The field of metagenomics has quickly expanded from microbial ecology to other disciplines including medical microbiology, food safety, and wastewater treatment. The next section details the most comprehensive metagenomics collaboration currently underway.

### *The Human Microbiome Project*

A pinnacle achievement in human knowledge, the Human Genome Project represented the largest scientific collaboration in biology ever, spanning areas in molecular biology, computer science and statistics, engineering, and biotechnology. 3.3 billion base-pairs later, as results from the analysis of the genome reached the scientific community, it became clear that the genomic differences between humans and other distantly related eukaryotes were more subtle than anyone had ever anticipated. The human genome contains roughly 20,000-25,000 protein-coding genes [31, 32], remarkably close to the mouse genome [33], and about 40% more than a fruit fly [34].

However, if we think of humans as superorganisms that house thousands of microbial species, then the total number of genes increases to over 100,000. It is estimated that the microbial cells inhabiting a person outnumber somatic and germ cells by an order of magnitude. While bacteria are not present throughout the entire body, they are essential in many of our functions including digestion of complex carbohydrates, synthesis of helpful vitamins, defense against pathogens, and the production of fat cells. Therefore, we must alter our model of human beings to capture the fundamental symbiosis between our microbiota and ourselves. This new perspective on human genetic variation combined with advancements in parallel DNA sequencing technology has led to a natural extension to human genomic research: the Human Microbiome Project.

Initiated in 2007, the Human Microbiome Project (HMP) is a large interdisciplinary collection of metagenomics projects which as a whole aim to characterize the microbial communities inhabiting humans across the globe. The HMP

focuses on five major areas: the gut, oral cavity, sinus cavity, skin, and the female urogenital tract [35, 36]. Each region presents unique challenges to microbiologists. Some bacterial communities are incredibly diverse, whereas others are small and difficult to extract from human tissue.

The HMP will help to develop an infrastructure for clinical studies of human microbiota, and it is hoped that we will find ways to identify bacterial factors associated with human disease and learn how to modify our microbiota to improve our overall health. Major goals of the project include [35]:

- I. Developing a reference set of microbial genome sequences and preliminary characterization of the human microbiome
- II. Examining the relationship between disease and changes in the human microbiome
- III. Developing new technologies and tools for computational analysis
- IV. Establishing a Data Analysis and Coordinating Center (DACC)
- V. Assessing ethical, legal and social implications of HMP research

### *This work*

This dissertation is a series of projects targeted toward achieving goal **III**, the development of new tools for computational analysis of large and complex metagenomic datasets. The direction of my research has been a function of the HMP and metagenomic datasets currently available and others in production. I have organized these projects based on their location in a sequence analysis pipeline: preprocessing (Chapter 2), processing (Chapter 3), and post-processing/modeling (Chapters 4 and 5). It is my hope

that these ideas and associated software packages will be used to improve the analysis of data not only from the HMP, but future metagenomics studies of any environment.

We begin with Figaro, a novel algorithm for trimming vector and other contaminant sequence from genomic and metagenomic datasets (generated by Sanger or potentially pyrosequencing technology) without prior knowledge of the artificial sequence itself (Chapter 2). The second study (Chapter 3) is a rigorous analysis of computational methodologies employed to cluster 16S rRNA sequences into species-like groups called operational taxonomic units (OTUs). In Chapter 4, we address challenges in post-processing large clinical metagenomic datasets with Metastats, a statistical methodology for detecting differentially abundant features between two populations. Finally, in an effort to push HMP data as far as possible, I have designed and validated a method for inferring microbe-microbe interactions using only longitudinal 16S rRNA data (Chapter 5). To close, Chapter 6 summarizes these works and discusses future research directions of considerable importance.

### ***Mathematical and computational contributions***

The following outlines my original mathematical and computational contributions made in each study:

Chapter 2. I developed Figaro to utilize a novel statistical approach that infers unknown vector sequences from the data by examining overrepresented  $k$ mers at the beginnings of reads. Specifically, I model the frequency of each  $k$ mer as a Poisson process, and weight  $k$ mers according to the likelihood that they are indeed part of a vector sequence.

Moreover, I carefully selected statistics to model which *k*mers are most likely to represent the end of each vector sequence, providing more accurate trim points and thus maximizing overall read length. I implemented Figaro and supporting scripts in Perl and C++ to run quickly on millions of reads. In testing, Figaro trimmed 1.5 million Sanger reads in ~11 minutes.

Chapter 3. I collaborated with Saket Navlakha to extend the semi-supervised clustering algorithm *VI-cut* in order to improve its performance when clustering 16S sequences into OTUs. Specifically, we incorporate the concept of *forbidden nodes* – nodes in a hierarchical decomposition that cannot be cut to create clusters. In the context of OTUs, this prevents the creation of large ambiguous clusters when parts of a tree lack sufficient taxonomic annotation.

Chapter 4.

I designed the Metastats statistical methodology to specifically suit the changing characteristics of annotated metagenomic data. Each component of the methodology is well known in statistical analysis (the nonparametric t-test, Fisher's exact test, the false discovery rate), but to my knowledge the unique combination of these tests for large-scale analysis of count data has not been employed, and certainly not in the context of comparative metagenomics.

Chapter 5.

The generalized Lotka-Volterra model has been used in traditional ecology for many years. However, because the number of parameters in this model scales quadratically with the number of taxa, most studies only fit the gLV model to datasets involving two or three organisms. My original computational contribution is a comprehensive Monte Carlo optimization procedure that finds many gLV model fits of suboptimal quality and infers fundamental ecological interactions between members of a community based on culling the resulting distributions of parameter estimates. To my knowledge, this approach has never been taken before in the context of fitting gLV models or in metagenomics. I implemented this computationally intensive optimization procedure both in Matlab and multi-threaded C++ code to improve efficiency.

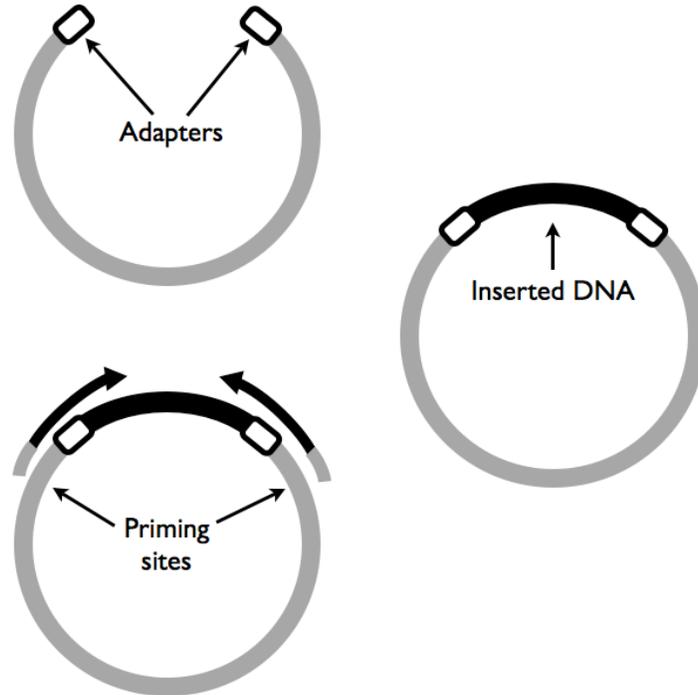
## Chapter 2: Figaro – a novel statistical method for vector sequence removal

### Background

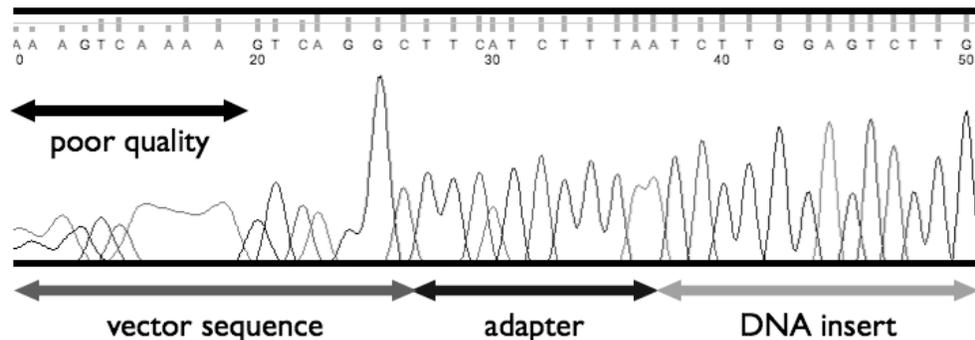
Even as new sequencing technologies become increasingly available [37], Sanger sequencing remains the most widely used technique for decoding the DNA of organisms [38]. High-throughput Sanger sequencing begins by cloning a DNA fragment into a vector (usually a plasmid) that is then transfected into *Escherichia coli* in order to amplify the original DNA fragment. Short adapter sequences are often attached to the ends of the fragment to improve the efficiency of the cloning process [39]. The sequencing reaction is usually performed using universal sequencing primers that anneal within the vector in the vicinity of the fragment insertion site (splice site). As a result of this process (highlighted in Figure 1), each sequence contains a small section of the vector, as well as the adapters used during cloning, in addition to the original DNA fragment. For the purpose of this paper, we will refer to any such artifacts as *vector sequence*. These sequences must be flagged prior to further analysis of the data, in a process called *vector trimming* or *vector clipping*.

Several software tools are available for vector removal: Lucy [40], Crossmatch ([www.phrap.org/phredphrapconsed.html](http://www.phrap.org/phredphrapconsed.html)), and VecScreen ([www.ncbi.nlm.nih.gov/VecScreen](http://www.ncbi.nlm.nih.gov/VecScreen)). These programs compare each read to the sequence of the cloning vector, then flag sections of the read that have strong similarity to the vector (Crossmatch replaces vector sequence with Xs, Lucy provides a list of clipping coordinates in the fasta header, and VecScreen provides a BLAST-like output). The alignments are performed

with relaxed parameters in order to account for the higher error rates at the beginning of reads (see Figure 2). Furthermore this approach requires three sets of information: (i) the sequence of the cloning vector; (ii) the splice site used for sequencing; and (iii) the sequence of the cloning adapters (if used—information that is often lost when the sequences are deposited in public databases). Note that the NCBI Trace Archive provides a mechanism for recording the location within the read where the vector ends (*vector clip point*), however this information is often missing or incorrect.



**Figure 1.** DNA from a sample (black) is cloned into a small circular piece of DNA called a vector (light gray). Short adapters (white) are used to improve efficiency of cloning the sample DNA. The molecule is then transfected into *E. coli*, amplified, and then sequenced from both ends starting from priming sites inside the vector.



**Figure 2.** Raw output from sequencing machines contains poor quality sequence on the ends as well as vector and adapter sequence, in addition to the DNA being sequenced.

As an example, at the beginning of September, 2007, approximately 60% (735 million out of 1.24 billion) of all shotgun reads from the NCBI Trace Archive had either no vector clip information, or a vector clip point of 0 or 1, indicating the vector clipping information was not provided (`clip_vector_left` = 0) or was arbitrarily set to the beginning of the read (`clip_vector_left` = 1). Even when a vector coordinate is provided it is often incorrect, as described below.

We examined the shotgun reads used to assemble the *Xanthomonas oryzae px099a* genome, a dataset for which both vector and quality clipping coordinates had been submitted to the Trace Archive by the sequencing center. We considered all reads whose vector clip coordinate occurred at least 8 base pairs (bp) inside the high-quality region, then tallied the final 8 bp (*8mer*) of the supposed vector sequence. These 8mers should represent the end of the vector sequence; therefore, they should be virtually identical across all reads with the exception of differences caused by sequencing errors. We examined 7,997 reads originating from a single sequencing library (library id

1041054961988). Furthermore, we separately examined reads sequenced with the “Forward”, and “Reverse” trace direction in order to avoid any variability due to differences between the vector sequences flanking the splice site. The results, summarized in table 1, highlight a much higher variability in the set of 8mers than can be explained by sequencing error alone, suggesting the vector clip points are incorrectly assigned.

<b>Trace direction</b>	<b>Forward</b>	<b>Reverse</b>
<b>Number of reads</b>	3,687	4,310
<b>Four most frequent 8mers and frequency</b>	GCGCAGCG 40	GCGCAGCG 46
	GCCGCAGC 29	GTGCTGGA 42
	GATCCATT 29	GGCGATCG 37
	GTGCTGGA 26	TGGCCGAT 35
<b>Number of distinct 8mers</b>	1,679 (45.5%)	1,858 (43.1%)

**Table 1.** Frequency of 8mers extracted upstream from the annotated vector clip point in shotgun reads from *Xanthomonas oryzae px099a*. We only considered reads from the library where the 5' vector clip point was at least 8 bp to the right of the 5' quality clip point. The reads were further binned by sequencing direction. The four most frequent 8mers are shown together with their frequency. The high level of variability indicates errors in the reported clipping coordinates.

In this work, we present an algorithm for detecting and removing the vector sequence from the 5' end of reads without prior knowledge of the vector sequences used. This algorithm can, therefore, be used to correctly identify the vector clipping points for sequences obtained from public databases. The code was implemented as a single streamlined module, named Figaro, which can be easily integrated into a high-throughput computational pipeline. The code is distributed under an open-source license through the AMOS package (<http://amos.sourceforge.net>).

Below we provide a detailed description of the trimming algorithm, and highlight its performance on three datasets: ~1.5 million *Drosophila pseudoobscura* reads; and in the *de novo* assembly of two bacterial genomes.

### Methods

For a set of shotgun reads, Figaro infers the vector sequence from the frequency of occurrence of *k*mers (DNA segments of length *k*). Under the assumption that the vector DNA flanking the inserted sequences is the same for all the sequences in a dataset, the most frequent *k*mers in the data likely represent vector DNA. This assumption is generally true for shotgun sequencing data, with the following exceptions: (i) different sequencing libraries may use different vectors; (ii) the vector sequences upstream and downstream the splice-site are often different (hence “Forward” and “Reverse” reads are prefixed by different vector DNA); and (iii) when cloning adapters are used, two different strings, corresponding to distinct adapter sequence, may prefix the reads even from a single library and orientation. To improve accuracy, the reads are partitioned by library and sequencing direction, if such information is available.

Figaro operates in two phases: (i) identification of frequent *kmers* likely to represent vector DNA (called *vectormers* throughout the text); and (ii) estimation of the vector clip point for every read, on the basis of the *vectormers* identified in step (i). These two components of the algorithm are described in detail below.

### **Detection of vectormers**

The vector sequence can be recognized by identifying *kmers* that are more frequent at the beginning of reads than anywhere else. Intuitively, the beginning of reads represents the DNA from the vector which is shared by the majority of reads in a dataset. The remaining section of each read should be randomly sampled from the genome, leading to few commonalities between distinct reads in the dataset.

A *kmer* frequency table is created which records the number of occurrences of each word of length  $k$  within adjacent windows of length  $L$  over the first  $E$  bases of all reads (a *kmer* is assigned to the window in which it starts, thus allowing us to count *kmers* that cross window boundaries). We truncate all reads to a same length  $E$  in order to avoid artifacts due to the increased error rates at the ends of reads. Given a maximum vector cut length,  $M$ , we declare the *safe zone* of the reads to be the region from base  $M$  to  $E$  (Figure 3). For each *kmer*  $K_i$ , if  $s_i$  is the number of occurrences of  $K_i$  in the safe zone across all reads, then we define its arrival rate  $\alpha_i$  to be:

$$\alpha_i = \frac{s_i}{E - M}$$

Given  $\alpha_i$ , we model the number of occurrences of  $K_i$  as a Poisson process. Letting  $X$  be the frequency of  $K_i$  in a window of length  $t$ ,  $X$  follows a Poisson distribution with parameter  $\lambda = t\alpha_i$ . Considering  $f_j$ , the frequency of occurrence of  $K_i$  within the  $j^{\text{th}}$  window

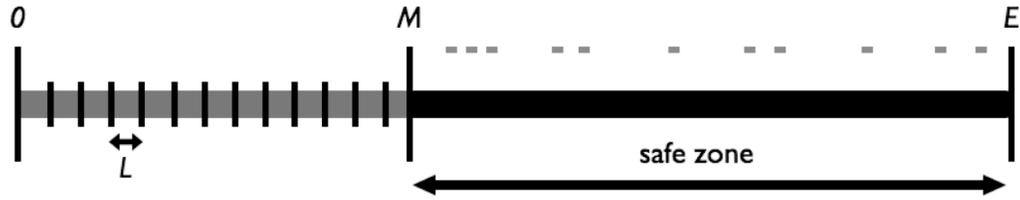
of length  $L$  (Fig. 4), we can estimate the likelihood of observing at least  $f_j$  occurrences of  $K_i$  in  $L$  base pairs given  $\alpha_i$ . Mathematically,

$$P(X \geq f_j) = 1 - P(X < f_j) = 1 - \sum_{y=0}^{f_j-1} \frac{e^{-\lambda} \lambda^y}{y!}$$

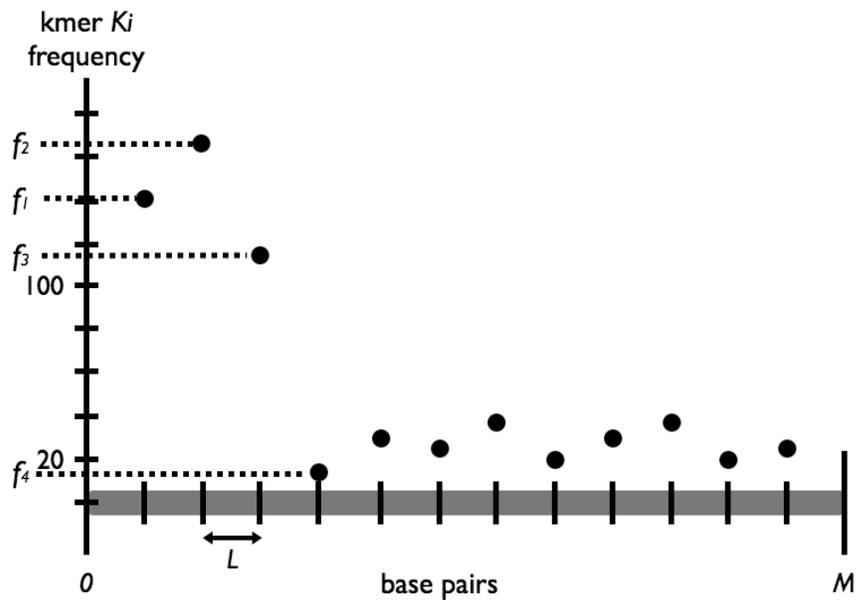
where  $\lambda = L\alpha_i$ . A  $k$ mer is declared to be a vectormer if  $P(X \geq f_j) < 0.001$  for a window within the first  $M$  base pairs of a read. By definition, we expect that  $0.001 * M/L$  of all  $k$ mers are incorrectly classified as vectormers. For example, assuming the average length of a read is 800 bp, four false vectormers are expected within any read for  $M = 100$  and  $L = 20$ .

In large datasets we observed that our algorithm produced many false positives due to statistical noise and common sequencing errors. To correct for this phenomenon, we retain only the most abundant vectormers, specifically, for a user-selected threshold  $T$ , we retain the  $T \times 100$  most frequent vectormers. This simple heuristic significantly reduces overtrimming.

The implementation of Figaro uses  $k = 8$  and  $L = 20$ . By default  $M = 100$  and  $E = 500$ , but these parameters may be modified by the user. A reasonable setting for the threshold  $T$  is automatically computed by Figaro depending on the number of reads in the dataset, however this value can also be controlled by the user.



**Figure 3.** Within the safe zone of all reads, we consider the number of occurrences of each  $k$ mer  $K_i$ , and calculate its average arrival rate. The beginning of the read is separated into bins of length  $L$  and the frequency of each  $k$ mer within each bin is recorded.



**Figure 4.** Frequency distribution for  $k$ mer  $K_i$  across first  $M$  bases of all reads. High frequency counts at the beginning of reads indicate that  $K_i$  is a likely vectormer.

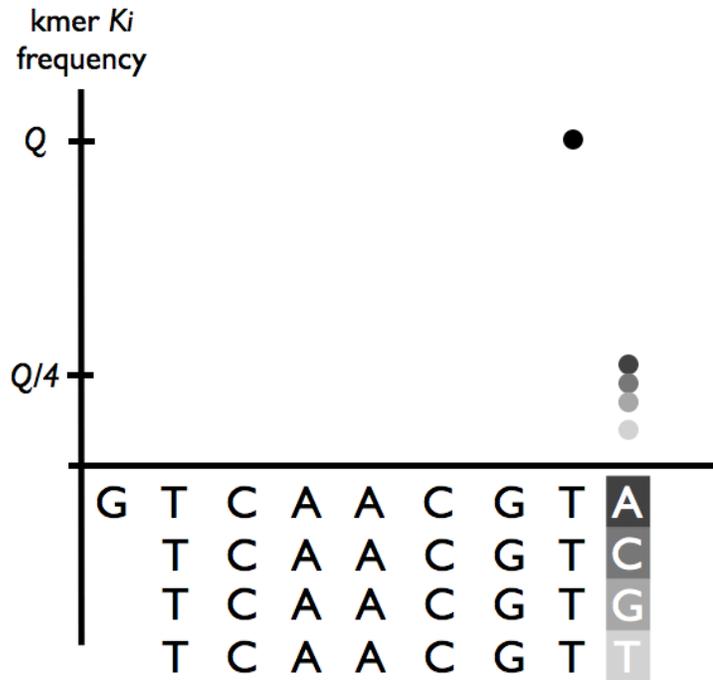
## Vector clip estimation

Once vectormers are computed, the algorithm first attempts to determine which vectormers are most likely to represent the ends of the vector sequences. We call these vectormers *endmers*. Assume a vectormer  $K$  has frequency of occurrence  $Q$ . If it is the true end of the vector, all  $k$ mers directly to the right of this vectormer ( $k$ mers whose prefix is the  $(k-1)$  suffix of  $K$ ) should have a frequency of roughly  $\frac{1}{4} \times Q$  (Fig. 5). The  $\frac{1}{4}$  parameter assumes equal distribution of the A, C, T, and G nucleotides in the genome. To account for the non-uniform distribution of nucleotides, we first estimate the G/C content of the organism being sequenced and adjust this threshold accordingly. Suppose the calculated G/C content is  $\delta$  and the A/T content is  $\varepsilon = 1 - \delta$ . We declare a vectormer to be an endmer if the adjacent  $k$ mers ending in G and C both have frequency  $< Q \times (\delta/2 + 0.1)$ , and if the  $k$ mers ending in A and T both have frequency  $< Q \times (\varepsilon/2 + 0.1)$ . Furthermore, to prevent many spurious endmer declarations when a large number of vectormers are allowed, we only consider the 100 most frequent vectormers as possible endmer candidates. Note that within these 100 vectormers, we only expect to find a small number of endmers (ideally four, however, sequencing errors might lead to a few more).

Once endmers are computed, we trim every sequence using the following algorithm. The first  $M$  base pairs of each sequence are examined right to left, using a 17 bp (10 adjacent 8mers) moving window. We consider we have encountered the end of the vector, and set the clip point accordingly, once we encounter a window containing 7 or more vectormers that ends in an endmer. To improve vector detection in the presence of

sequencing errors, all  $k$ mers within one substitution of an endmer are also labeled as endmers.

Frequent sequencing errors can cause our algorithm to miss the end of the vector sequence (no window contains an endmer). To account for this situation, we simply select the rightmost window containing 7 or more vectormers. Within this window, we identify a rightmost kmer whose distance from the end of the vector is known, then adjust the clip point accordingly. Note, that a side effect of our vectormer detection algorithm is that we can construct a de Bruijn graph (Pevzner et al., 2001) from the set of vectormers. Specifically, every vectormer represents a node in this graph, and two nodes are connected if the corresponding vectormers share a  $k-1$  substring (e.g. TAAAAAAA and AAAAAAAG are neighbors in this graph). Within this graph we mark the location of the endmers, and label each node with its distance (number of edges that need to be traversed) from the nearest endmer, i.e. its distance from the end of the vector. This information is used, as described above, to correctly identify the end of the vector even if an endmer cannot be detected. In the rare case where we cannot identify any vectormer whose distance to the end of the vector is known we simply use the position of the rightmost window with 7 or more vectormers as the vector clip point. Note that the specific parameters of this process were set heuristically to values that performed well in our experiments. It is possible that in some cases they may need to be tuned for specific characteristics of the data being analyzed. We clearly mark these parameters at the beginning of the Figaro source code to allow their easy modification, as we have not yet to identify a suitable automated procedure for estimating these parameters.



**Figure 5.** A conceptual example of identifying endmers (i.e. a vectormer that is likely to be the end of the vector sequence.) Note that the  $k$ mer GTCAAGCT has a frequency of  $Q$  (black dot). Frequencies of adjacent  $k$ mers ending in A, C, G, and T (represented in different shades of gray) are significantly lower than  $Q$ .

## Results

### **Vector trimming sensitivity and specificity**

To create a test in which we know exactly where the true vector ends, we have generated a set of artificial sequences based on shotgun reads from the *Chlamydomophila caviae gpic* genome project [41] containing variable length vector sequence on their ends. We trimmed off the first 300 bases from each of the 19,633 reads, and attached a vector sequence of random length ranging from 10 to 50 bp generated from the SmaI cloning

site of the pUC18 vector (GenBank accession L09136). No vector sequence was attached to about 20% of the reads. Finally, we introduced a varying amount of error within the vector sequence to assess the performance of Figaro in the presence of sequencing errors.

We ran Figaro on datasets with error rates ranging from 0% to 5%, and then compared the sensitivity and specificity of the results taking into account overtrimming and undertrimming. The same parameters were used for all trials:  $T = 30$ ,  $M = 60$ , and  $E = 500$ . For each value of the parameter  $m$ , we denote a true positive ( $TP_m$ ) whenever the identified trimpoint is within  $m$  bases of the true trimpoint. Similarly overtrimming or undertrimming by more than  $m$  bases is denoted as a false positive ( $FP_m$ ) and false negative ( $FN_m$ ), respectively. Sensitivity and specificity are defined as follows:

$$SN_m = \frac{TP_m}{TP_m + FN_m}$$

$$SP_m = \frac{TP_m}{TP_m + FP_m}$$

Table 2 displays the sensitivity and specificity of Figaro for all trials. In the absence of errors, Figaro finds the vector sequence with 100% sensitivity, and rarely overtrims. The sensitivity and specificity remain high, even after introducing errors as high as 5% (higher than commonly encountered in practice). The fact that Figaro overtrims even in the error-less test warrants further discussion. We examined the reads that were overtrimmed by Figaro and found that the majority of these contained little or no vector (approx. 90% of these reads contained less than 15 bp of vector and 56% contained no vector). In such situations our algorithm is unable to identify a clear vector boundary and resorts to an aggressive trimming strategy designed to avoid undertrimming. In very few cases we found that overtrimming was due to significant homology between a section of the read and the end of the cloning vector. Note that such

situations also cause overtrimming when using established, similarity-based, trimming software. Furthermore Figaro is intentionally aggressive as a small amount of overtrimming is preferable to undertrimming.

In order to evaluate our approach on real data, we used as a test set reads from the *Drosophila pseudoobscura* genome sequencing project [42]. We chose these particular data because the sequencing adapters used in the project are known [39]. Searching for the two adapter sequences (16 bp each) using *nucmer* [43, 44], we collected 1,506,679 reads that matched at least 8 bp of an adapter with at least 90% identity. The 3' end of the vector was required to match within the first 50 bp of the read, and was labeled as the true vector trimpoint. We ran Figaro with  $T = 30$ , and  $M = 50$  (maximum vector cut length of 50 bp). Figaro found the exact end of the vector sequences with 99.98% sensitivity and 99.15% specificity (table 3). Without prior knowledge of the vector sequence, Figaro was able to detect and remove virtually all vector with negligible overtrimming. About 0.4% of the reads were overtrimmed by more than 3 bp and 0.01% of the reads were undertrimmed by more than 3 bp. Furthermore, the running time for this test was just short of 11 minutes, indicating that Figaro is efficient even for large eukaryotic projects.

We also tested Figaro on a highly repetitive genome (maize, *Zea mays* [45]). The results on 9,738 sequences from this genome were similar to those obtained for *Drosophila* - we achieved 100%  $SN_I$  and 99.6%  $SP_I$  - indicating our method is robust in the presence of repeats.

Error rate	$SN_0$	$SP_0$	$SN_3$	$SP_3$	$SN_5$	$SP_5$
0%	100%	99.5%	100%	99.7%	100%	99.7%
1%	99.6%	99.3%	99.9%	99.7%	99.9%	99.7%
3%	98.0%	98.9%	99.0%	99.7%	99.1%	99.7%
5%	96.5%	98.0%	98.3%	99.6%	98.6%	99.6%

**Table 2.** Sensitivity and specificity results of Figaro on simulated vector contaminant sequence with different error rates. For each value of the parameter  $m$ , a true positive ( $TP_m$ ) is counted whenever the identified trimpoint is within  $m$  bases of the true trimpoint. Similarly, overtrimming or undertrimming by more than  $m$  bases is denoted as a false positive ( $FP_m$ ) and false negative ( $FN_m$ ), respectively. We define sensitivity,  $SN_m = (TP_m / (TP_m + FN_m))$ , and specificity,  $SP_m = (TP_m / (TP_m + FP_m))$ . Introducing higher error rates reduces the program’s ability to detect the vector sequence boundary, but even with an error rate of 5%, Figaro performs well, effectively removing nearly all of the vector sequence without significantly overtrimming reads.

Max distance <i>m</i>	<i>SN<sub>m</sub></i>	<i>SP<sub>m</sub></i>	<i>TP<sub>m</sub></i>	<i>FN<sub>m</sub></i>	<i>FP<sub>m</sub></i>
0	99.98%	99.15%	1,493,582	316	12,781
3	99.99%	99.29%	1,500,662	186	5,831
5	~100%	99.72%	1,502,428	67	4,184
10	~100%	99.79%	1,503,481	54	3,144

**Table 3.** Sensitivity and specificity results of Figaro on *Drosophila pseudoobscura* shotgun reads. Using a threshold of 30, Figaro is able to remove virtually all vector sequence and only overtrims a small proportion of reads by more than 3 bp. Note false positives and false negatives are computed only if they occur in the high-quality region of a read.

### Improving assemblies with Figaro

To illustrate how Figaro can help to improve high-throughput genomic studies, we used the Celera Assembler [31, 46] to assemble the genomes of *Chlamydophila caviae* GPIC [41] and *Coxiella burnetii* RSA 493 [47], and compared these assemblies to available finished sequence. These genomes were chosen because they have been recently finished, and full quality and vector trimming information is available in the NCBI Trace Archive.

We constructed “Official” assemblies using the provided vector and quality trimming points explicitly; and “Base quality” assemblies using only the quality trimming information. Additional assemblies were created using the output of Figaro

combined with the official quality trimming information. Figaro was run separately for each sequencing library with  $T = 30$ ,  $M = 200$ , and  $E = 500$ .

Table 4 reveals that not only were the Figaro assemblies far superior to the “Base quality” assemblies, but they improved upon the “Official” assemblies. The Figaro assemblies of *C. caviae* and *C. brunetii* produced contigs with a higher N50 size covering more of the reference sequence than their “Official” counterparts. Furthermore, our trimming did not result in any additional mis-assemblies. The *C. brunetii* “Base quality” assembly is a particularly good example of the need for accurate vector trimming. By using Figaro the resulting assembly increased the N50 contig size nearly seven fold over the “Base quality” assembly and by nearly 30% over the “Official” assembly.

Assembly Run	# of contigs	Contig N50 (bp)	% coverage	# of errors in contigs
<b><i>Chlamydophila caviae</i> GPIC</b>				
Base quality	252	9,466	93.0	0
Official	209	11,731	95.0	1
$T = 30$	203	13,044	96.1	1
<b><i>Coxiella brunetii</i> RSA 493</b>				
Base quality	1,535	1,232	77.9	0
Official	719	6,713	94.8	0
$T = 30$	643	8,118	95.6	0

**Table 4.** Assembly results using Figaro on two microbial genomes. The “Official” assemblies used the quality and vector trims provided with the read sets. The “Base quality” assemblies only used the quality trims provided. Assemblies were performed after trimming with Figaro using  $T = 30$ ,  $M = 200$  and  $E = 500$ . Assemblies created using Figaro improve upon their “Official” counterparts by increasing overall contig size without introducing more errors or losing coverage. The “coverage” column denotes the percent of finished sequence covered by assembled contigs; note assembly errors are not accounted for, i.e., partial contig matches are counted toward the coverage. The ContigN50 column denotes that half the bases in the assembly are contained in contigs of the given length or greater.

## Discussion

Figaro is only intended as a tool for identifying and removing vector from the 5' end of reads. Often, entire reads consist of vector sequence (e.g. no fragment was inserted in the vector), while in short libraries vector sequence may also occur at the 3' end of reads. In such situations, our algorithm cannot detect the 3' vector sequence due to the large variation in the amount of vector included in each sequence (at the 5' end the vector ends roughly at the same location in every read), thus Figaro must be augmented with traditional vector trimming software. Furthermore, since Figaro does not trim based on quality values, our software should be used in conjunction with a quality trimming program such as Lucy [40]. The software distribution includes several scripts that automate this process for common types of sequence data. We also provide tools for actually trimming or masking the vector sequence in the dataset.

Note that many sequencing projects use more than one library, and therefore, more than one vector. When the number of libraries is large, Figaro may incur difficulties due to the statistical nature of its algorithm. To avoid such problems, the scripts provided in the Figaro package automatically run our code on each library separately when library information is provided (e.g. NCBI Trace Archive XML file).

In addition, the algorithms implemented in Figaro implicitly assume the randomness of a typical shotgun process. Therefore, Figaro cannot be used for targeted sequencing experiments where a same gene is sequenced across multiple samples. Also, in EST sequencing projects, the use of Figaro may result in the incorrect removal of the polyA tail. Figaro is capable of removing relatively short adapter and vector sequences and therefore may be applicable to data generated using pyrosequencing technologies

such as 454 and Illumina. Though next-generation sequencing does not employ standard vector-based techniques, artificial sequences such as linkers, adapters and barcodes are often used to tag DNA fragments for pooled sequencing. Without identifying and removing these artificial sequences, pyrosequencing datasets would be extremely difficult to assemble and analyze. Figaro may prove useful in detecting and trimming these barcodes, but additional validation is required to sufficiently assess Figaro's sensitivity in these circumstances.

The various parameters controlling the execution of our code are automatically set to reasonable default values. These values can also be controlled by the users if the default values are inappropriate for the data being processed. For example, the parameter  $E$ , marking the end of the "good quality" section of a read, is usually set to 500, however its value should be increased or decreased depending on the average read length being analyzed. Similarly, our code performs best if the parameter  $M$  (the window within which Figaro searches for the vector sequence) is set to a value close to the expected length of the vector. This parameter should, therefore, be adjusted if additional information is available regarding the distance of the sequencing primers from the cloning site. Note, however, that  $M$  should be set conservatively (greater than the expected length of the vector) in order to avoid undertrimming.

Raw shotgun sequences are placed in the NCBI Trace Archive at an ever increasing rate, rapidly outpacing the availability of current assemblies for many genomes. Constructing independent assemblies from these data is complicated by the often incomplete or incorrect vector trimming information reported in the public databases. The program described in this paper, Figaro, provides scientists with the

means to automatically detect and remove the vector sequence from shotgun reads without prior knowledge about the sequencing protocol, thereby enabling the large-scale re-assembly of public data. Furthermore, even if the vector sequence is known, Figaro provides an efficient and effective alternative to commonly used vector removal programs.

## Chapter 3: Alignment and clustering of phylogenetic markers – implications for microbial diversity studies

*Note in this chapter, my contributions include running the comprehensive OTU methodology evaluation, sensitivity analysis of each methodology component, and modification of the VI-cut algorithm for OTU clustering. Saket Navlakha exclusively wrote the description of the modified VI-cut algorithm below.*

### Background

The human body is host to a massive ecosystem with thousands of commensal microbial species. Microbial diversity within the human body has recently been quantified through 16S rRNA surveys [19, 48-50] and metagenomic methods. The latter provide a detailed view of the genomic composition and functional potential of human-associated microbial communities [21]. However this level of resolution comes with a high price-tag — billions of base-pairs need to be sequenced to ensure a sufficient level of sampling of complex communities [51] such as those found in the human gastrointestinal tract. 16S rRNA surveys provide limited insight into the composition of the commensal microbiome, however due to substantially lower costs, such studies are currently the only practical approach for studying large numbers of samples (such as those generated in a clinical setting). In this paper we explore the limits of the methods used to analyze 16S rRNA data, particularly the large impact of small changes in the parameters of the analysis process. We specifically focus on the most common strategy — the clustering of 16S rRNA sequences into a collection of *operational taxonomic units* (OTUs) or *phylotypes* on the basis of sequence similarity. Taxonomic classification through database searches [50] or other fully supervised classification methods [52] are inherently limited

due to the current undersampling of the global microbial population, only allowing accurate classification of a fraction of sequences (as low as 20% in some studies [19]).

The OTU clustering process begins by constructing a multiple alignment (MSA) of the 16S rRNA sequences. The MSA is then used to estimate pairwise distances between individual sequences, expressed as the fraction of nucleotides that have changed as the sequences have evolved from their most recent common ancestor. To accurately reflect evolutionary processes, the distances inferred from the MSA are corrected using one of several models of evolution [53]. The distances are provided as input to a hierarchical clustering algorithm (nearest neighbor, furthest neighbor, or average neighbor/UPGMA are commonly used). Sub-clusters or OTUs are defined by applying a distance threshold, selected to roughly approximate a specific taxonomic level: thresholds between 1-3% are typically used to approximate individual species, 5% for individual genera, 15% for classes, etc. [14, 54, 55]. The first steps of this process (MSA – distance correction – distance matrix) are also the first steps in the phylogenetic analysis of a set of sequences. In this context an accurate MSA (often achieved through painstaking manual curation) and precise estimation of evolutionary distances is necessary. As we will discuss below, however, these steps might be unnecessary if the goal is the determination of the OTU structure of a community.

The choice of MSA, the distance correction, clustering algorithm, and distance threshold varies considerably between studies, and, to our knowledge, there have been no comprehensive evaluations of the impact of methodological choices on the ecological conclusions of the analysis process. In this study, we rely on simulated datasets to

provide a comprehensive assessment of the extent to which individual parameters of the evaluation process affect the analysis of 16S rRNA data.

We evaluate methodological choices in terms of how well the clustering of the sequences into a set of OTUs matches the clustering imposed by the known membership of the sequences to individual bacterial species. As a measure of similarity between clusterings we use the Variation of Information (VI) metric. VI measures the amount of information lost or gained by changing from one clustering to another [56] and (in contrast to other methods for comparing clusterings, e.g. the Rand index) is based on a rigorous mathematical foundation (see Materials & Methods in this chapter).

## Results

### **Simulated environments**

To construct a simulated environment of known composition, we collected 1677 full and partial 16S rRNA gene sequences from the Ribosomal Database Project II (release 9.57; RDP) [22] with complete taxonomic identification. The majority of these sequences were obtained from isolate genomes (96.2%) and had unambiguous taxonomic assignment, as defined by the fact that three independent databases (RDP, NCBI, and GreenGenes) agreed on their identity at the species level (see Methods). The simulated environment spans 49 species, 46 genera, 37 families, 21 orders, 12 classes, and seven phyla including Proteobacteria, Bacteroidetes, Firmicutes, and Actinobacteria. Alpha-, Beta-, and Gammaproteobacteria make up 66% of the sequences in roughly equal proportions. A similar class distribution has been reported for microbial communities found in the

phyllosphere of the Atlantic rainforest [57]. Simulated datasets have previously been successfully used to evaluate methods for the assembly, gene finding, and binning of metagenomic data [58].

### **Comprehensive search of OTU methodologies**

We explored the parameter space of OTU methodologies by varying the MSA, distance correction, clustering algorithm, and distance threshold. Sequences were first aligned using three different MSA programs commonly employed for 16S analyses: NAST [59], MUSCLE [60], and ClustalW [61]. Each program successfully aligned all 1677 sequences, and alignments were subsequently trimmed to within the span of all sequences. We then calculated distance matrices from each alignment using the Jukes Cantor (JC), Kimura-2 (K2P), and Felsenstein 84 (F84) distance corrections using the DNADIST program from the PHYLIP package [53] then clustered the sequences according to three hierarchical clustering strategies (nearest, average, and furthest neighbor) using DOTUR [62]. We finally determined phylotypes using a series of distance thresholds ( $D = 0.00$  to  $0.45$  in  $0.01$  increments), producing a total of 749 distinct OTU sets that were then compared against the known species-level clustering of our data.

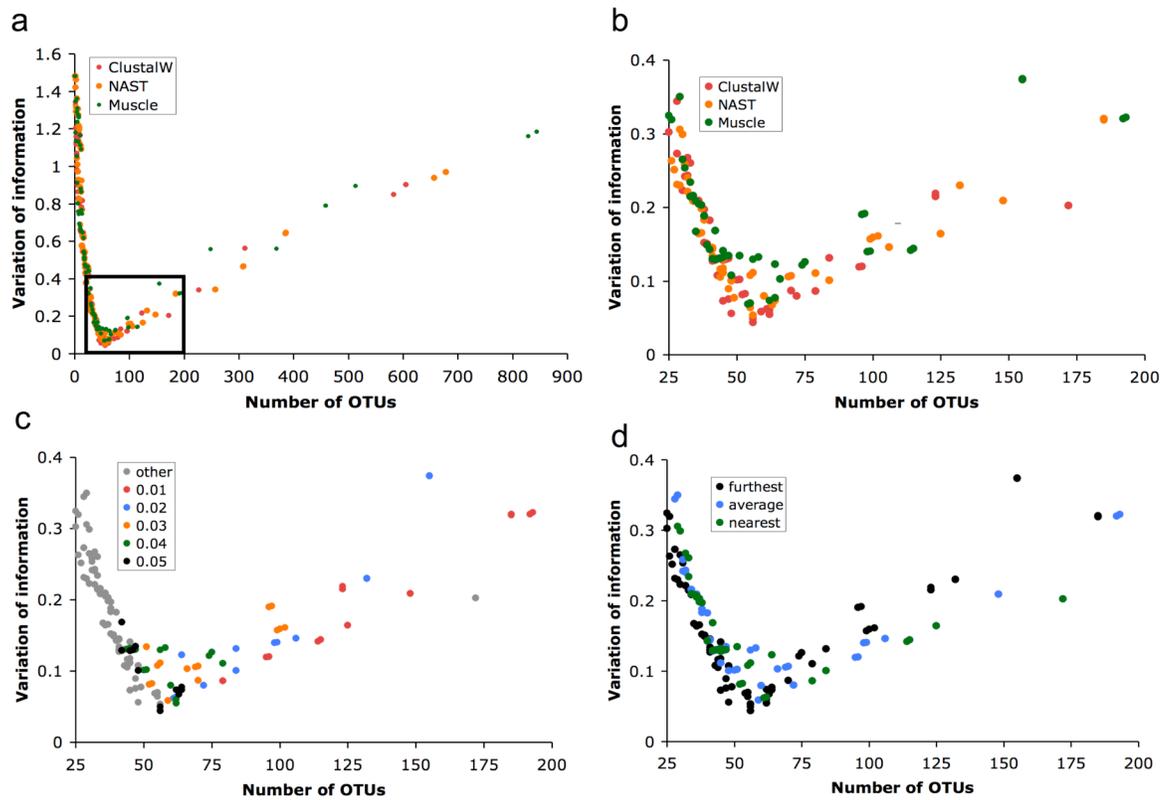
### **OTU variability**

The results of our analysis (summarized in Figure 6) reveal a large variation in the level of concordance of the OTU clustering with the true species-level composition of the

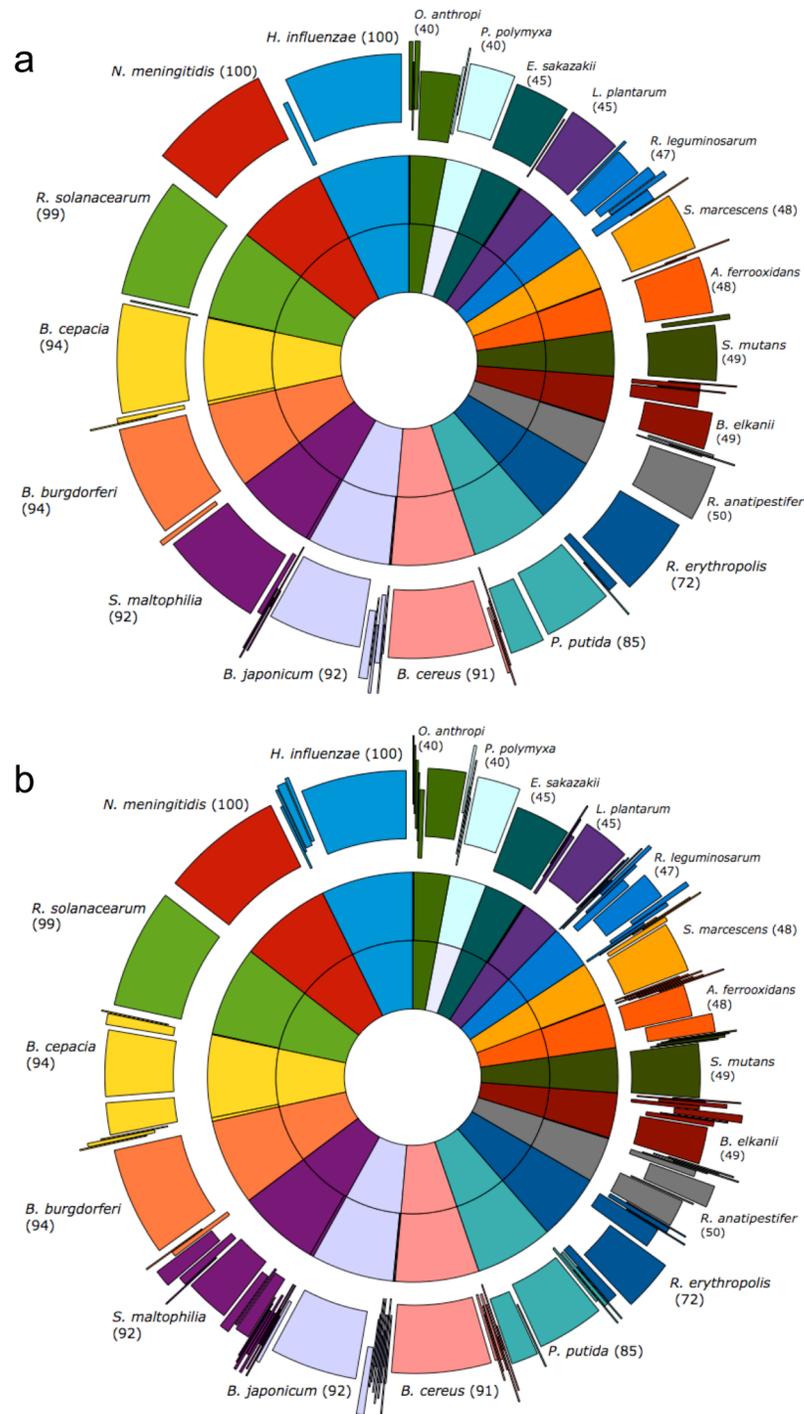
environment when varying methodology parameters. We specifically highlight the parameters with most impact — distance threshold (panel c), MSA (panels a and b), and clustering strategy (panel d) — parameters that accounted for 56, 33, and 7 percent of the variation, respectively, confirmed by ANOVA (see Materials & Methods). We did not observe a significant impact from the use of different distance correction measures (see Figure 9a and Methods). The corresponding number of OTUs generated by the different methodologies varied significantly (from 36 to 257) and none of the parameter combinations managed to capture the true species composition (49 OTUs, VI distance=0). A large variation in the OTU content is observed even when we fix the similarity threshold to 0.01 (approximately strain-level) — the number of OTUs ranges from 79 to 248 at this similarity level. Surprisingly, the best OTU clustering was obtained at a similarity threshold of 0.05 (Figure 6c) — a value larger than the thresholds usually used to approximate the species-level composition of an environment (0.01-0.03 [14, 19, 49]). In terms of alignment, methodologies employing ClustalW or NAST were roughly similar and performed better than those using MUSCLE (Figure 6b). The performance of ClustalW is somewhat surprising as MUSCLE was previously reported to outperform ClustalW when aligning protein sequences [60] and NAST is specifically designed for the alignment of 16S rRNA sequences. In terms of clustering strategy, furthest neighbor resulted in the best agreement with the true species structure of our simulated environment (Figure 6d). Even the best combination of analysis parameters (ClustalW, furthest neighbor, 0.05 distance threshold) led to an overestimate of the number of species in our sample, resulting in 56 OTUs. We found similar OTU variability for 10 additional simulated 16S datasets (see below). This result highlights a fundamental

limitation of hierarchical clustering strategies for 16S rRNA analysis — only 42 of the 49 species present in our sample corresponded to a homogeneous sub-tree within the best hierarchical clustering of our data. The remaining 7 species cannot be correctly clustered irrespective of the similarity threshold chosen.

The results presented above highlight a wide variation in the OTU structure as we explore the parameters of the analysis process. To determine whether such variation is also present in the methodologies used in practice, we compared three analysis methodologies that performed well in our combinatorial search to several methodologies reported in published literature. The results shown in Table 5 indicate that the published methodologies can overestimate the diversity of the simulated environment, sometimes by more than 3-fold. The fragmentation of the resulting OTUs is particularly striking among the most abundant phylotypes (Figure 7), where sequences belonging to the same species are distributed among multiple OTUs. In contrast, the methodologies chosen by our combinatorial search produce few mistakes.



**Figure 6.** (a) The number of OTUs found versus the VI distance from the true species clustering for 749 OTU sets. Generally, smaller clustering distances lead to many OTUs while larger clustering distances result in very few OTUs, both of which poorly approximate the species-level structure in the sample. Near 49 OTUs, the true number of species in the sample, the OTU sets are relatively closer to the true species-level structure. Detail of the lower-left corner of (a) re-colored by (b) MSA, (c) distance threshold, and (d) clustering algorithm.



**Figure 7.** Comparison of OTU sets to true species clusters. The innermost rings represents the 20 most abundant species in the sample. Each species shown has  $\geq 40$  sequences in the dataset (total observations shown next to name of each species). The

middle ring displays OTUs of the methodology using the parameters that resulted in the closest approximation of the species structure. The outer ring is an OTU set generated from methodologies used to study microbial communities of **(a)** soil [14] and **(b)** termite hindguts [63]. We see that the published methodologies partition most species into several OTUs, resulting in a poor approximation of the species-level structure of the environment. Note that OTU sets from the middle and outer rings of (a) grouped the *B. cepacia* sequences with a less abundant species not shown (*B. pseudomallei*). Though the outer ring of (b) did not make this mistake, it heavily partitioned the *B. cepacia* sequences into seven OTUs. This demonstrates the potential variability of OTUs defined using different methodologies.

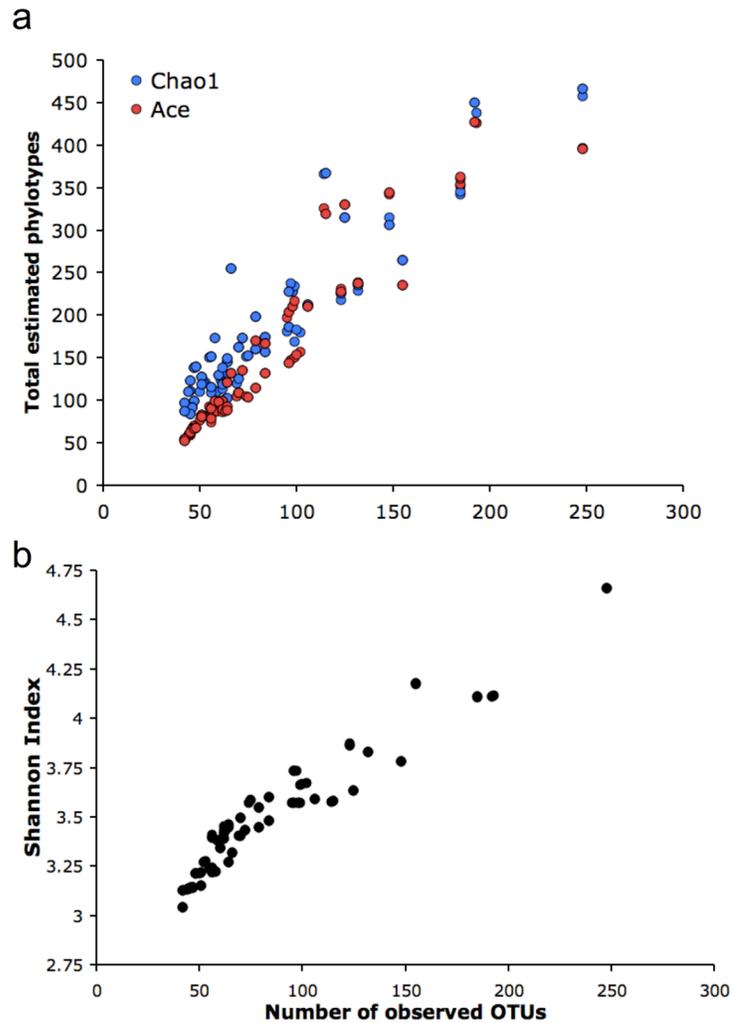
	Correction	MSA	Clustering	Distance	OTUs	Ace	Chao1	Shannon	VI
<u>Optimal</u>	F84	ClustalW	fn	0.05	56	79	116	3.39	0.044
	F84	NAST	fn	0.06	56	78	176	3.39	0.054
	JC	MUSCLE	fn	0.06	54	69	132	3.37	0.068
<i>Drosophila</i> (host) [64]	JC	ClustalW	fn	0.03	70	109	162	3.49	0.087
Marine sponge [65]	F84	ClustalW	fn	0.03	70	109	162	3.49	0.087
Soil [14]	JC	NAST	fn	0.03	99	150	169	3.66	0.157
Deep sea biosphere [66, 67]	JC	MUSCLE	fn	0.03	96	396	466	4.66	0.190
Termite hindgut [63]	JC	NAST	fn	0.01	185	360	351	4.11	0.320

**Table 5.** OTU sets closest to the true species clustering for each multiple sequence alignment. The “VI” column indicates the VI distance of each clustering from the true species clustering. Optimal methods are contrasted with five published methodologies. The “Correction” column corresponds to the evolutionary distance correction. Note that for the optimal methods using ClustalW and NAST alignments, the F84 and K2P

corrections produced identical OTU sets because the distance matrices were very similar, though not identical. All methods in this table used furthest neighbor (fn) clustering. The Ace, Chao1, and Shannon diversity estimators are also provided.

### **Nonparametric estimators of richness and diversity**

The large variability in the OTU estimates produced by different methodologies had a significant effect on the inferred ecological parameters of the environment being studied. The Chao1 [68] and ACE [69] richness estimators and the Shannon diversity index [70] are measures commonly used to estimate the level of diversity present in an environment. These measures were highly sensitive to differences in OTU structure (Figure 8) even when distance thresholds were restricted within the range 0.01-0.05. Under the true species clustering,  $S_{Ace} = 57$ ,  $S_{Chao1} = 67$ , and  $H = 3.41$ .  $S_{Ace}$  and  $S_{Chao1}$  estimates for the computed OTU clusterings ranged from 52 to 427 and 84 to 466 phylotypes, respectively, while Shannon diversity indices ( $H$ ) ranged from 3.04 to 4.66. Accurate estimates of the diversity of an environment are particularly important when planning metagenomics sequencing projects, and a particular environment might not be studied if the diversity is incorrectly perceived to be too high to effectively sample.



**Figure 8.** Variability in nonparametric estimators and diversity indices using a clustering distances 0.01-0.05. Plots of **(a)** Ace and Chao1, and **(b)** Shannon measures reveal significant sensitivity to OTU sets. Each plotted methodology used either the MUSCLE, ClustalW, or NAST MSA; they also used either furthest, nearest, or average neighbor clustering, and one of the following evolutionary distance corrections: JC, K2P, or F84. The observed variability does not include the traditional confidence interval estimation of each statistic, which will add to the uncertainty in the estimators.

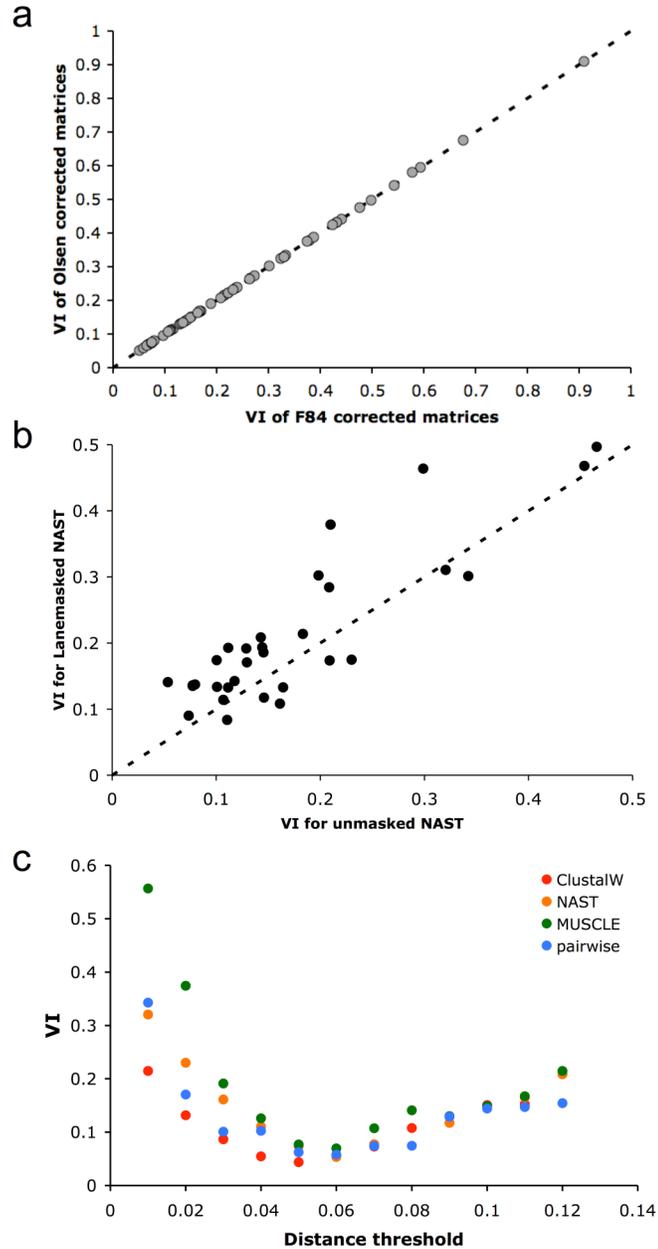
### **Partial masking of MSAs**

To improve phylogenetic analyses, researchers often remove hyper-variable segments of MSAs either manually or using a filter such as LaneMask [71, 72]. We explored the impact of this approach on OTU clustering. Specifically, we used the GreenGenes LaneMask filter, which reduces a NAST alignment to 1287 highly conserved columns. The results are surprising — in our data, LaneMask resulted in a worse approximation of the true species composition than the unmasked alignment (see Figure 9b). This suggests that the use of a generic mask should be critically evaluated in the context of OTU and phylogenetic analyses.

### **Pairwise versus multiple sequence alignments**

Comparison of the OTU clustering to a known standard using the information-theoretic VI distance is a general tool that can be used to evaluate other parameters of 16S rRNA analysis. In particular, we evaluated whether a multiple sequence alignment is needed prior to clustering the data. An MSA is necessary for phylogenetic analyses in order to ensure that the pairwise distances between the sequences are consistent with their evolutionary history. Constructing an MSA, however, is computationally expensive, requiring time proportional to the cube of the number of sequences being analyzed, making this approach impractical for large numbers of sequences (pyrosequencing experiments, for example, often generate hundreds of thousands of sequences). For the purpose of clustering, however, the direct computation of pairwise distances (performed in just quadratic time) appears to be sufficient. For distance thresholds in the range 0.02 - 0.05, distances computed from pairwise alignments resulted in a more accurate OTU

structure than those obtained from MUSCLE or NAST MSAs, but performed slightly worse than those computed from ClustalW MSA (Figure 9c). This indicates that, at least for large datasets, multiple alignments can be replaced by direct computation of distances from pairwise alignments.



**Figure 9. (a)** OTU structures are not highly sensitive to varying distance corrections. Each point plotted uses identical MSA and clustering method (distance thresholds  $D = 0.01 - 0.19$ ) varying only the distance correction. The y-axis represents the VI distance from the true species using Olsen-corrected matrices, while the x-axis is the analogous distance using F84-corrected matrices ( $r^2 = 0.9999$ ). The dashed line is the function  $y = x$ . **(b)** Isolating the affect of using LaneMask on OTU quality. We applied LaneMask to a

NAST alignment provided by the GreenGenes website to check for improved OTU accuracy. Surprisingly, when using distance cutoffs of 0.00 to 0.10, the masked alignment provided a poorer approximation of the true species structure on average than the unmasked alignment. The dashed line is the function  $y = x$ . **(c)** Comparison of pairwise distance methods vs. multiple sequence alignments. The y-axis is the distance from the true species clustering. Pairwise distances produced OTUs with quality comparable to methods employing MSA programs.

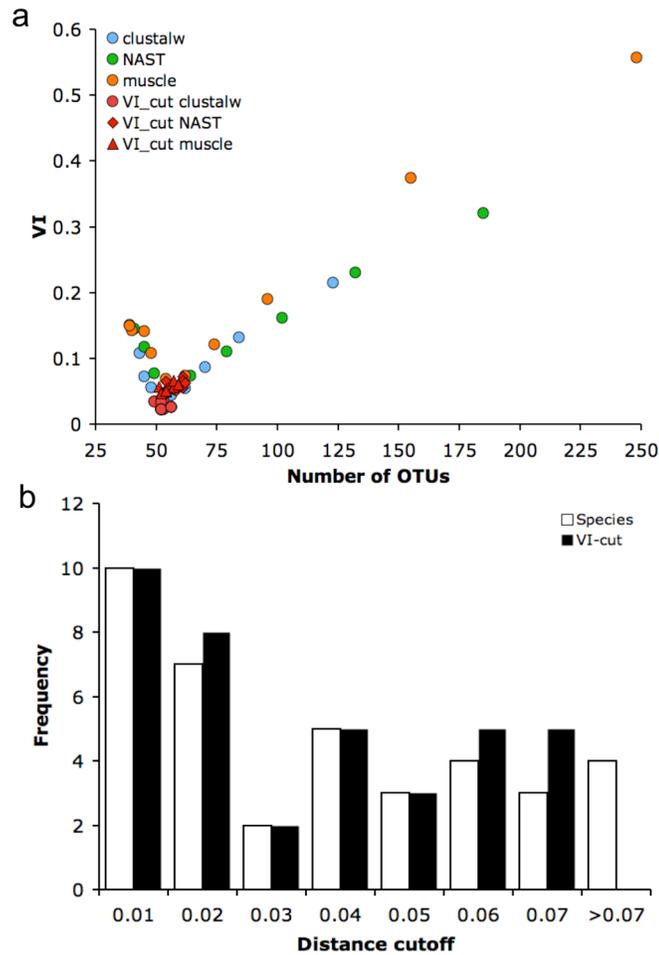
### **Supervised clustering alternatives**

Our analysis has so far made the assumption that one of the primary goals of a 16S analysis pipeline is to estimate the composition of an environment at a pre-specified taxonomic level (e.g. species). As demonstrated by our results, the OTU methodologies proposed in the literature fail to achieve this goal, generally overestimating the number of species. Even by systematically evaluating various settings for the parameters of the analysis process, we could not obtain perfect concordance between the OTU structure and the species composition of the environment. This is in part due to the fact that the concept of “species” is born out of gross morphological and phenotypic traits of microorganisms, and therefore cannot be precisely mapped to fine-scale molecular measurements. Furthermore, the rate of evolution varies across the tree of life, making it unrealistic to rely on a single distance threshold.

As an alternative, we investigated the use of a semi-supervised clustering method to adaptively select a set of local distance thresholds that lead to OTUs that better fit the species composition of the environment. Specifically, we employed *VI-cut* [73], a

clustering approach that identifies a cut within a hierarchical clustering tree that maximizes the fit with a labeled subset of the sequences. In the case of 16S analysis, VI-cut constructs a set of OTUs that optimally matches (in terms of VI distance) the species structure of an environment as inferred from a small subset of sequences that have known taxonomic assignments (for more details see Materials & Methods).

We applied VI-cut to our data by simulating partial taxonomic knowledge of the dataset. For each MSA and the optimal distance correction (shown in Table 5), we randomly selected 10% of the sequences and provided VI-cut with their true labels. To assess the variability in the algorithm's results, we repeated this procedure 20 times. As seen in Figure 10a, VI-cut outperforms methodologies that employ a single distance threshold, irrespective of the MSA employed or the random selection of labeled sequences. The need for an adaptive threshold (such as that provided by the VI-cut approach) is highlighted in Figure 10b — the diameter of clusters corresponding to a single species in our data varies considerably among our sequences (from 0.01 to 0.07) and the semi-supervised learning algorithm implemented in VI-cut is able to closely approximate the true distribution of distance thresholds. Note that perfect concordance between OTUs and species cannot be achieved even with the best hierarchical clustering tree constructed from our data – this suggests that there may be better techniques for analyzing 16S rRNA data than common hierarchical clustering.



**Figure 10.** Results of VI-cut compared to standard methodologies. **(a)** Standard methodologies using a specific MSA with furthest neighbor clustering to find OTUs. Furthest neighbor clustering was used for all standard methodologies plotted. VI-cut was employed using the same MSA and distance correction in each plot. For each VI-cut trial, 10% of the sequences were randomly selected and given labels. Over 20 trials, OTUs determined by VI-cut are stable and more accurate than the standard methodologies. **(b)** Distribution of true species distances. “Species” and distance cutoffs inferred by VI-cut to generate OTUs for one trial shown in (a). Singletons not shown. There is considerable variation ( $D = 0.01-0.07$ ) in the optimal distance threshold among species. While

standard methodologies cut the tree at a single constant threshold, VI-cut allows for variable cutting distances, providing more flexibility for defining OTUs.

### **Consistency of methods across multiple datasets**

To investigate the consistent improvement of the VI-cut methodology over other methods, we created ten additional 16S environmental samples – each sample containing 500 randomly selected sequences from the original dataset. We repeated our comparison of VI-cut to other methods for these 10 simulated samples. Examining the results across each MSA, we found that VI-cut consistently produced the best species-level approximation compared to standard methodologies (Table 6).

Finally, it is important to observe that clustering 16S rRNA sequences into a set of OTUs is a valuable analysis tool even if the resulting OTUs do not correlate with pre-defined taxonomic entities. The *ad hoc* choice of analysis parameters, however, complicates cross-study comparisons. Our results highlight the need for standardizing 16S rRNA metagenomic analysis methods, or in the very least, reporting results obtained with multiple distance thresholds or clustering algorithms. The data used in this study have been deposited in the FAMeS online database (<http://fames.jgi-psf.org>) — a repository for simulated metagenomic data [58].

MSA	Clustering	Distance	mean VI
MUSCLE	VI-cut	adaptive	0.0589
ClustalW	VI-cut	adaptive	0.0595
ClustalW	fn	0.03	0.0688
MUSCLE	fn	0.04	0.0691
ClustalW	fn	0.04	0.0697
MUSCLE	fn	0.05	0.0748
NAST	VI-cut	adaptive	0.0762
ClustalW	fn	0.02	0.0838
NAST	fn	0.05	0.0845
MUSCLE	fn	0.03	0.0860
NAST	fn	0.06	0.0872
ClustalW	fn	0.05	0.0942
NAST	fn	0.04	0.0992
MUSCLE	fn	0.06	0.1025
ClustalW	fn	0.06	0.1176
NAST	fn	0.03	0.1222
MUSCLE	fn	0.02	0.1370
ClustalW	fn	0.01	0.1505
NAST	fn	0.02	0.1633
NAST	fn	0.01	0.2362
MUSCLE	fn	0.01	0.2629

**Table 6.** Top-performing methodologies and performance of VI-cut, ranked by their mean VI-distance over 10 simulated datasets. We constrained the results to commonly accepted methods using furthest neighbor clustering and distance thresholds less than 0.07. A distance threshold of 0.01 is consistently among the worst performing methodologies. VI-cut consistently results in the best clustering for each MSA.

## Materials & Methods

### **Creation of simulated datasets**

The RDP database (release 9.57) [22] was downloaded and reduced to 16S sequences only containing full taxonomic identification. The total number of each species was calculated and 1860 sequences from the 50 most abundant species in the database were

selected. These sequences were required to be at least 800 base pairs. NAST [59] was run with default parameters, successfully aligning 1677 sequences (sequences with less than 75% identity to one of the profile alignment sequences were removed). To screen for false annotations, these sequences were then reclassified down to the genus level using the RDP Naïve Bayesian classifier [52] and GreenGenes SimRank [23]. The RDP classifier assigned all sequences to their correct genus with  $\geq 95\%$  confidence. SimRank also classified all sequences to the correct genus. Finally, we ran BLASTN [74] with a word size of 20 on all 1677 sequences against the reduced RDP database with full taxonomic information. Every sequence in the simulated sample had at least one hit to a different sequence in the database with the same species annotation with an E-value  $< 1e-50$  and a bitscore  $> 1000$ . These three independent methods of validation strongly suggested that there are no spurious annotations in the simulated sample.

### **Multiple sequence alignment**

All 1677 were aligned using MUSCLE, ClustalW, and NAST [59-61] using default parameters. ClustalW was run with the “Fast” option for pairwise alignments. In the NAST alignment, all columns containing only gaps were removed, and each MSA was trimmed so that every sequence spanned the entire alignment.

### **Distance corrections and clustering methods**

Distance matrices with Jukes-Cantor, Kimura 2-parameter, and Felsenstein84 corrections were computed using DNADIST with default parameters from the PHYLIP package [53].

Olsen and F84 distance-corrected matrices were also generated using the ARB package [75] for additional validation. All distance matrices served as input to DOTUR [62] which uses nearest neighbor, average neighbor, and furthest neighbor clustering to create OTUs. DOTUR additionally creates OTUs by varying a constant distance threshold  $D$  which is used as a criterion for merging two clusters in one. Distance thresholds ranged from 0, 0.01, 0.02, ... 0.45, resulting in a total of 749 OTU sets created by different methodologies.

### **Measures of similarity for clusterings**

We employed two measures of similarity between clusterings: the Rand index [76] and the Variation of Information (VI) metric [56]. Examining the values of all clusterings according to the Rand index and the VI, we found identical rankings between the two metrics. Because the Rand index tends to concentrate near 1 given more clusters, we use the VI as the measure of comparison between clusterings. In order to provide a reference set of VI distances for known clusters, we measured the VI between the true species clustering and the true phylum, class, order, family, and genus clusterings (Table 7).

True clustering	VI
Phyla	0.171
Classes	0.109
Orders	0.058
Families	0.026
Genera	0.008
Species	0

**Table 7.** Variation of information (VI) distances of true taxonomic clusterings to the known species clustering.

### Computation of the Rand index and variation of information for clusterings

The variation of information criterion is a measure of similarity between two partitions (or clusterings) of a given set [56]. For this study, the set is the 1677 16S sequences selected for the artificial environmental sample. Mathematically, a given clustering  $C$ , is a partition of a set  $S$  into disjoint subsets (clusters) where:

$$C = \{C_1, C_2, \dots, C_M\}, C_i \cap C_j = \emptyset, \text{ and } \bigcup_{i=1}^M C_i = S.$$

If there are  $m$  elements in set  $S$ , and we let  $m_i$  be the number of elements in cluster  $C_i$ ,

then  $m = \sum_{i=1}^M m_i$ . Given two clusterings,  $C$  and  $D$ , we can examine all pairs of points in  $S$

and see whether  $C$  and  $D$  agree on whether or not they should be in the same cluster. Any pair of points will fall exclusively into one of the four following categories:

*11 – The point pair is in the same cluster for both  $C$  and  $D$ .*

*00 – The point pair is in different clusters for both  $C$  and  $D$ .*

*10 – The point pair is in the same cluster for  $C$ , but not for  $D$ .*

*01 – The point pair is in the same cluster for  $D$ , but not for  $C$ .*

Accordingly, the total number of point pairs falling into each category is  $N_{11}$ ,  $N_{00}$ ,  $N_{10}$ , and  $N_{01}$ . Given these values, the Rand index is computed as:

$$R(C, D) = \frac{N_{11} + N_{00}}{N_{11} + N_{00} + N_{10} + N_{01}}.$$

To compute the Variation of information between two clusterings, we first find the probability that a randomly selected sequence is in a particular cluster, that is,

$P(i) = \frac{m_i}{m}$ . Given this discrete probability distribution, the uncertainty of the random variable  $i$ , is the entropy associated with clustering  $C$ , defined as:

$$H(C) = - \sum_{i=1}^M P(i) \cdot \log P(i).$$

Now, suppose we have two clusterings  $C = \{C_1, C_2, \dots, C_M\}$ , and  $D = \{D_1, D_2, \dots, D_{M'}\}$ . Then

we calculate the joint distribution  $P(i, j) = \frac{|C_i \cap D_j|}{m}$  describing the similarity of all pairs of clusters between  $C$  and  $D$ . The *mutual information* between the clusterings  $C$  and  $D$  is then defined to be

$$I(C, D) = \sum_{i=1}^M \sum_{j=1}^{M'} P(i, j) \log \frac{P(i, j)}{P(i)P(j)},$$

and finally, the variation of information between  $C$  and  $D$  is defined as the sum of the individual clustering entropies less 2 times the mutual information:

$$VI(C, D) = H(C) + H(D) - 2I(C, D).$$

If  $C$  and  $D$  are identical clusterings, then  $H(C)=H(D)=I(C, D)$ , and the  $VI = 0$ . The  $VI$  distance is a true metric, satisfying symmetry, non-negativity, and the triangle inequality.

### **VI-cut method for defining OTUs**

VI-cut is a procedure that finds a clustering from a hierarchical tree decomposition  $T$  that optimally matches a partial set of known labels, as defined by the variation of information metric [73]. A clustering is defined by choosing a set of nodes in  $T$ . Each

chosen node  $c$  corresponds to a single cluster consisting of all the leaves (sequences) in the subtree rooted at  $c$ . The chosen nodes represent a node-cut in the tree such that each leaf belongs to exactly one cluster.

Let  $D$  represent the partial clustering of annotated sequences such that sequences with the same label are grouped together. The set of chosen nodes corresponds to a node-cut  $K$ , which induces a clustering  $A_K$ . The VI-cut algorithm finds the  $A_K$  that minimizes the VI distance to  $D$ :

$$\min_K VI(A_K, D)$$

Although there are exponentially many number of possible node-cuts in  $T$ , VI-cut finds the optimal one efficiently using dynamic programming. For this study, we modified the VI-cut algorithm by incorporating *forbidden nodes*, i.e. nodes in  $T$  that VI-cut is not allowed to choose. Specifically, any node  $n$  with a corresponding distance  $\geq 0.07$  was forbidden. This means that if the cluster induced by  $n$  contains a pair of sequences which have a pairwise distance  $\geq 0.07$  then  $n$  is not allowed to be chosen.

To incorporate forbidden nodes into the VI-cut algorithm, we first ran the standard VI-cut algorithm. If the clustering returned contained a forbidden node  $n$ , we moved down the tree and replaced  $n$  with its closest unforbidden descendants such that each sequence is still placed in only one cluster. This modification forces the method to cut the tree at distances  $< 0.07$ , which helps to cluster large subtrees with multiple species that may not have any known labels.

### **ANOVA of methodology components**

In order to isolate the individual impact of each component in an OTU methodology, we

examined the 200 methodologies resulting in the lowest VI distance from the true species clustering, and performed a multi-way analysis of variance (ANOVA) considering four factors: multiple sequence alignment, evolutionary distance correction, clustering algorithm, and distance threshold. Using a linear model with no interactions, we found that the distance threshold alone explains 56% of the total variance in VI (Table 8). This impact was followed by the MSA, the clustering algorithm, and finally the distance correction, which explained 33%, 7%, <0.01% of the total variance, respectively. This model explains 97% of the total variance, indicating that component interactions are negligible for our purposes. An  $F$  test did not detect any statistically significant differences between distance corrections ( $F = 0.002$ ,  $P = 0.998$ ). We extended this comparison to include the Olsen distance correction in ARB [75], which we found produced OTUs virtually identical to those created using the F84 correction (Figure 6a).

Parameter	Sum of Squares	Degrees of freedom	Mean Sq.	$F$	Prob > $F$
Distance threshold	0.4411	11	0.0401	23.0160	< 0.0001
MSA	0.0480	2	0.0240	13.7843	< 0.0001
Clustering	0.0099	2	0.0050	2.8503	0.0604
Distance correction	< 0.0001	2	< 0.0001	0.0020	0.9980
Error	0.3171	182	0.0017		
Total	0.7910	199	0.0708		

**Table 8.** Multi-way ANOVA table assessing components used in OTU methodologies.

The factor with the largest effect on the quality of the OTUs was the distance threshold, followed by the MSA, and then the clustering algorithm. The distance correction explained < 0.01% of the variance and no statistically significant difference could be detected between the corrections ( $P = 0.998$ ).

## Chapter 4: Statistical methods for detecting differentially abundant features in clinical metagenomic samples

### Background

Broad sequencing of bacterial populations allows us a first glimpse at the many microbes that cannot be analyzed through traditional means (only ~1% of all bacteria can be isolated and independently cultured with current methods [77]). Studies of environmental samples initially focused on targeted sequencing of individual genes, in particular the 16S subunit of ribosomal RNA [67, 78-80], though more recent studies take advantage of high-throughput shotgun sequencing methods to assess not only the taxonomic composition, but also the functional capacity of a microbial community [18, 30, 81].

Several software tools have been developed in recent years for comparing different environments on the basis of sequence data. DOTUR [62], Libshuff [82], *f*-libshuff [83], SONs [84], MEGAN [85], UniFrac [86], and TreeClimber [87] all focus on different aspects of such an analysis. DOTUR clusters sequences into operational taxonomic units (OTUs) and provides estimates of the diversity of a microbial population thereby providing a coarse measure for comparing different communities. SONs extends DOTUR with a statistic for estimating the similarity between two environments, specifically, the fraction of OTUs shared between two communities. Libshuff and *f*-libshuff provide a hypothesis test (Cramer von Mises statistics) for deciding whether two communities are different, and TreeClimber and UniFrac frame this question in a phylogenetic context. Note that these methods aim to assess **whether**, rather than **how** two communities differ. The latter question is particularly important as we begin to

analyze the contribution of the microbiome to human health. Metagenomic analysis in clinical trials will require information at individual taxonomic levels to guide future experiments and treatments. For example, we would like to identify bacteria whose presence or absence contributes to human disease and develop antibiotic or probiotic treatments. This question was first addressed by Rodriguez-Brito *et al.* [88], who use bootstrapping to estimate the p-value associated with differences between the abundance of biological subsystems. More recently, the software MEGAN of Huson *et al.* [85] provides a graphical interface that allows users to compare the taxonomic composition of different environments. Note that MEGAN is the only one among the programs mentioned above that can be applied to data other than that obtained from 16S rRNA surveys.

These tools share one common limitation — they are all designed for comparing exactly two samples — therefore have limited applicability in a clinical setting where the goal is to compare two (or more) treatment populations each comprising multiple samples. In this paper, we describe a rigorous statistical approach for detecting differentially abundant features (taxa, pathways, subsystems, etc.) between clinical metagenomic datasets. This method is applicable to both high-throughput metagenomic data and to 16S rRNA surveys. Our approach extends statistical methods originally developed for microarray analysis. Specifically, we adapt these methods to discrete count data and correct for sparse counts. Our research was motivated by the increasing focus of metagenomic projects on clinical applications (e.g. Human Microbiome Project [36]).

Note that a similar problem has been addressed in the context of digital gene expression studies (e.g. SAGE [89]). Lu *et al.* [90] employ an overdispersed log-linear

model and Robinson and Smyth [91] use a negative binomial distribution in the analysis of multiple SAGE libraries. Both approaches can be applied to metagenomic datasets. We compare our tool to these prior methodologies through comprehensive simulations, and demonstrate the performance of our approach by analyzing publicly available datasets, including 16S surveys of human gut microbiota and random sequencing-based functional surveys of infant and mature gut microbiomes and microbial and viral metagenomes. The methods described in this paper have been implemented as a web server and are also available as free source-code (in R) from <http://metastats.cbcb.umd.edu>.

### *Materials & Methods*

Our approach relies on the following assumptions: (i) we are given data corresponding to two treatment populations (e.g. sick and healthy human gut communities, or individuals exposed to different treatments) each consisting of multiple individuals (or samples); (ii) for each sample we are provided with count data representing the relative abundance of specific *features* within each sample, e.g. number of 16S rRNA clones assigned to a specific taxon, or number of shotgun reads mapped to a specific biological pathway or subsystem (see below how such information can be generated using currently available software packages). Our goal is to identify individual features in such datasets that distinguish between the two populations, i.e. features whose abundance in the two populations is different. Furthermore, we develop a statistical measure of confidence in the observed differences.

The input to our method can be represented as a *Feature Abundance Matrix* whose rows correspond to specific features, and whose columns correspond to individual

metagenomic samples. The cell in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column is the total number of observations of feature  $i$  in sample  $j$  (Figure 11). Every distinct observation is represented only once in the matrix, i.e. overlapping features are not allowed (the rows correspond to a partition of the set of sequences).

	<b>S1</b>	<b>S2</b>	. . . . .	<b>S(N-1)</b>	<b>SN</b>
<b>T1</b>	$c(1,1)$	$c(1,2)$	. . . . .	$c(1,N-1)$	$c(1,N)$
<b>T2</b>	$c(2,1)$	$c(2,2)$			.
.	.				.
.	.				.
<b>T(M-1)</b>	$c(M-1,1)$				
<b>TM</b>	$c(M,1)$		. . . . .		$c(M,N)$

**Figure 11.** Format of the feature abundance matrix. Each row represents a specific taxon, while each column represents a subject or replicate. The frequency of the  $i^{\text{th}}$  feature in the  $j^{\text{th}}$  subject ( $c(i,j)$ ) is recorded in the corresponding cell of the matrix. If there are  $g$  subjects in the first population, they are represented by the first  $g$  columns of the matrix, while the remaining columns represent subjects from the second population.

**Data normalization**

To account for different levels of sampling across multiple individuals, we convert the raw abundance measure to a fraction representing the relative contribution of each feature to each of the individuals. This results in a normalized version of the matrix described

above, where the cell in the  $i^{\text{th}}$  row and the  $j^{\text{th}}$  column (which we shall denote  $f_{ij}$ ) is the proportion of taxon  $i$  observed in individual  $j$ . We chose this simple normalization procedure because it provides a natural representation of the count data as a relative abundance measure, however other normalization approaches can be used to ensure observed counts are comparable across samples, and we are currently evaluating several such approaches.

### Analysis of differential abundance

For each feature  $i$ , we compare its abundance across the two treatment populations by computing a two-sample  $t$  statistic. Specifically, we calculate the mean proportion  $\bar{x}_{it}$ , and variance  $s_{it}^2$  of each treatment  $t$  from which  $n_t$  subjects (columns in the matrix) were sampled:

$$\bar{x}_{it} = \frac{1}{n_t} \sum_{j \in \text{treatment } t} f_{ij}$$

$$s_{it}^2 = \frac{1}{n_t - 1} \sum_{j \in \text{treatment } t} (f_{ij} - \bar{x}_{it})^2$$

We then compute the two-sample  $t$  statistic:

$$t_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}$$

Features whose  $t$  statistics exceeds a specified threshold can be inferred to be differentially abundant across the two treatments (two-sided  $t$ -test).

### Assessing significance

The threshold for the  $t$  statistic is chosen such as to minimize the number of false positives (features incorrectly determined to be differentially abundant). Specifically, we try to control the p-value—the likelihood of observing a given  $t$  statistic by chance. Traditional analyses compute the p-value using the  $t$  distribution with an appropriate number of degrees of freedom. However, an implicit assumption of this procedure is that the underlying distribution is normal. We do not make this assumption, but rather estimate the null distribution of  $t_i$  non-parametrically using a permutation method as described in Storey and Tibshirani [92]. This procedure, also known as the nonparametric  $t$ -test has been shown to provide accurate estimates of significance when the underlying distributions are non-normal [93, 94]. Specifically, we randomly permute the treatment labels of the columns of the abundance matrix and recalculate the  $t$  statistics. Note that the permutation maintains that there are  $n_1$  replicates for treatment 1 and  $n_2$  replicates for treatment 2. Repeating this procedure for  $B$  trials, we obtain  $B$  sets of  $t$  statistics:  $t_1^{ob}, \dots, t_M^{ob}$ ,  $b = 1, \dots, B$ , where  $M$  is the number of rows in the matrix. For each row (feature), the p-value associated with the observed  $t$  statistic is calculated as the fraction of permuted tests with a  $t$  statistic greater than or equal to the observed  $t_i$ :

$$p_i = \frac{\#\{ |t_i^{ob}| \geq |t_i|, b = 1, \dots, B \}}{B}.$$

This approach is inadequate for small sample sizes in which there are a limited number of possible permutations of all columns. As a heuristic, if less than 8 subjects are used in

either treatment, we pool all permuted  $t$  statistics together into one null distribution and estimate p-values as:

$$p_i = \frac{1}{BM} \sum_{b=1}^B \#\{j : |t_j^{ob}| \geq |t_i|, j = 1, \dots, M\}.$$

Note that the choice of 8 for the cutoff is simply heuristic based on experiments during the implementation of our method. Our approach is specifically targeted at datasets comprising multiple subjects — for small data-sets approaches such as that proposed by Rodriguez-Brito et. al. [88] might be more appropriate.

Unless explicitly stated, all experiments described below used 1000 permutations. In general, the number of permutations should be chosen as a function of the significance threshold used in the experiment. Specifically, a permutation test with  $B$  permutations can only estimate p-values as low as  $1/B$  (in our case  $10^{-3}$ ). In datasets containing many features, larger numbers of permutations are necessary to account for multiple hypothesis testing issues (further corrections for this case are discussed below). Precision of the p-value calculations is obviously improved by increasing the number of permutations used to approximate the null distribution, at a cost, however, of increased computational time. For certain distributions, small p-values can be efficiently estimated using a technique called importance sampling. Specifically, the permutation test is targeted to the tail of the distribution being estimated, leading to a reduction in the number of permutations necessary of up to 95% [95, 96]. We intend to implement such an approach in future versions of our software.

## Multiple hypothesis testing correction

For complex environments (many features/taxa/subsystems), the direct application of the  $t$  statistic as described can lead to large numbers of false positives. For example, choosing a p-value threshold of 0.05 would result in 50 false positives in a dataset comprising 1000 organisms. An intuitive correction involves decreasing the p-value cutoff proportional to the number of tests performed (a Bonferroni correction), thereby reducing the number of false positives. This approach, however, can be too conservative when a large number of tests are performed [21].

An alternative approach aims to control the false discovery rate (FDR), which is defined as the proportion of false positives within the set of predictions [97], in contrast to the false positive rate defined as the proportion of false positives within the entire set of tests. In this context, the significance of a test is measured by a q-value, an individual measure of the FDR for each test.

We compute the q-values using the following algorithm, based on Storey and Tibshirani [92]. This method assumes that the p-values of truly null tests are uniformly distributed, assumption that holds for the methods used in Metastats. Given an ordered list of p-values,  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ , (where  $m$  is the total number of features), and a range of values  $\lambda = 0, 0.01, 0.02, \dots, 0.90$ , we compute

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_j > \lambda\}}{m(1 - \lambda)}.$$

Next, we fit  $\hat{\pi}_0(\lambda)$  with a cubic spline with 3 degrees of freedom, which we denote  $\hat{f}$ , and let  $\hat{\pi}_0 = \hat{f}(1)$ . Finally, we estimate the q-value corresponding to each ordered p-value. First,  $\hat{q}(p_{(m)}) = \min(p_{(m)} \times \hat{\pi}_0, 1)$ . Then for  $i = m-1, m-2, \dots, 1$ ,

$$\hat{q}(p_{(i)}) = \min \left( \frac{\hat{\pi}_0 \times m \times p_{(i)}}{i}, \hat{q}(p_{(i+1)}) \right).$$

Thus, the hypothesis test with p-value  $p_{(i)}$  has a corresponding q-value of  $\hat{q}(p_{(i)})$ . Note that this method yields conservative estimates of the true q-values, i.e.  $\hat{q}(p_{(i)}) \geq q(p_{(i)})$ . Our software provides users with the option to use either p-value or q-value thresholds, irrespective of the complexity of the data.

### **Handling sparse counts**

For low frequency features, e.g. low abundance taxa, the nonparametric  $t$ -test described above is not accurate [98]. We performed several simulations (data not shown) to determine the limitations of the nonparametric  $t$ -test for sparsely-sampled features. Correspondingly, our software only applies the test if the total number of observations of a feature in either population is greater than the total number of subjects in the population (i.e. the average across subjects of the number of observations for a given feature is greater than one). We compare the differential abundance of sparsely-sampled (rare) features using Fisher's exact test. Fisher's exact test models the sampling process according to a hypergeometric distribution (sampling without replacement). The frequencies of sparse features within the abundance matrix are pooled to create a 2x2 contingency table (Figure 12), which acts as input for a two-tailed test. Using the notation from Figure 12, the null hypergeometric probability of observing a 2x2 contingency table is:

$$p = \frac{\binom{R_1}{f_{11}} \binom{R_2}{f_{21}}}{\binom{n}{C_1}}, \text{ where } \begin{aligned} R_1 &= f_{11} + f_{12}, \\ R_2 &= f_{21} + f_{22}, \\ C_1 &= f_{11} + f_{21}, \\ n &= f_{11} + f_{12} + f_{21} + f_{22}. \end{aligned}$$

By calculating this probability for a given table, and all tables more extreme than that observed, one can calculate the exact probability of obtaining the original table by chance assuming that the null hypothesis (i.e. no differential abundance) is true [98].

Note that an alternative approach to handling sparse features is proposed in microarray literature. The Significance Analysis of Microarrays (SAM) method [99] addresses low levels of expression using a modified  $t$  statistic. We chose to use Fisher's exact test due to the discrete nature of our data, and because prior studies performed in the context of digital gene expression indicate Fisher's test to be effective for detection of differential abundance [100].

	treatment 1	treatment 2
taxon $i$	$f_{11}$	$f_{12}$
not taxon $i$	$f_{21}$	$f_{22}$

**Figure 12.** Detecting differential abundance for sparse features. A 2x2 contingency table is used in Fisher's exact test for differential abundance between rare features.  $f_{11}$  is the number of observations of feature  $i$  in all individuals from treatment 1.  $f_{21}$  is the number of observations that are not feature  $i$  in all individuals from treatment 1.  $f_{12}$  and  $f_{22}$  are similarly defined for treatment 2.

## **Creating the Feature Abundance Matrix**

The input to our method, the Feature Abundance Matrix, can be easily constructed from both 16S rRNA and random shotgun data using available software packages. Specifically for 16S taxonomic analysis, tools such as the RDP Bayesian classifier [52] and Greengenes SimRank [23] output easily-parseable information regarding the abundance of each taxonomic unit present in a sample. As a complementary, unsupervised approach, 16S sequences can be clustered with DOTUR [62] into operational taxonomic units (OTUs). Abundance data can be easily extracted from the “\*.list” file detailing which sequences are members of the same OTU. Shotgun data can be functionally or taxonomically classified using MEGAN [85], CARMA [101], or MG-RAST [24]. MEGAN and CARMA are both capable of outputting lists of sequences assigned to a taxonomy or functional group. MG-RAST provides similar information for metabolic subsystems that can be downloaded as a tab-delimited file.

All data-types described above can be easily converted into a Feature Abundance Matrix suitable as input to our method. In the future we also plan to provide converters for data generated by commonly-used analysis tools.

## **Data used in this paper**

Human gut 16S rRNA sequences were prepared as described in Eckburg *et al.* and Ley *et al.* (2006) and are available in GenBank, accession numbers: DQ793220-DQ802819, DQ803048, DQ803139-DQ810181, DQ823640-DQ825343, AY974810-AY986384. In our experiments we assigned all 16S sequences to taxa using a naïve Bayesian classifier currently employed by the Ribosomal Database Project II (RDP) [52]. COG profiles of

13 human gut microbiomes were obtained from the supplementary material of Kurokawa *et al.* [102]. We acquired metabolic functional profiles of 85 metagenomes from the online supplementary materials of Dinsdale *et al.* (2008) (<http://www.theseed.org/DinsdaleSupplementalMaterial/>).

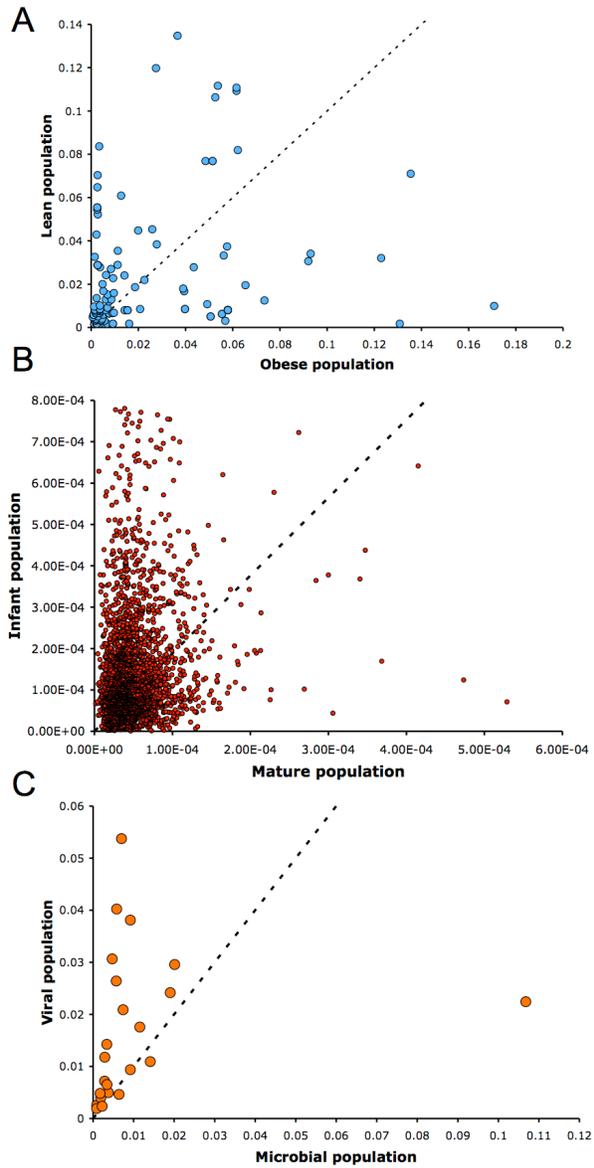
## Results

### **Comparison with other statistical methods**

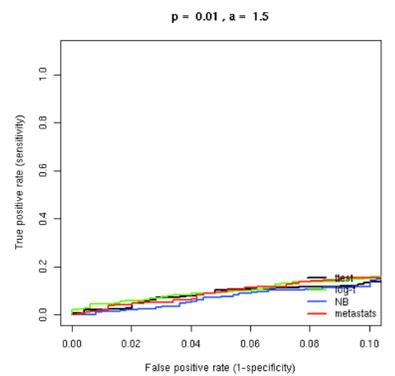
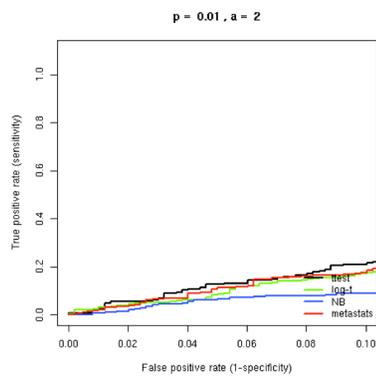
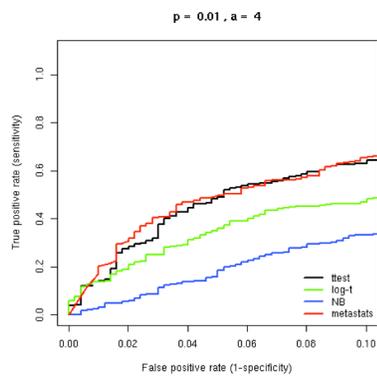
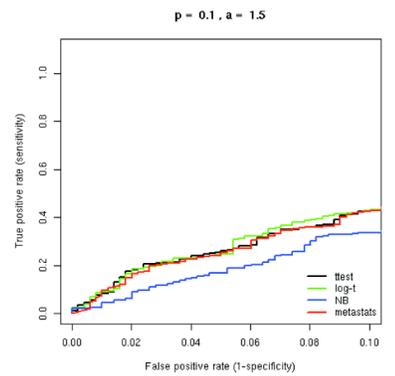
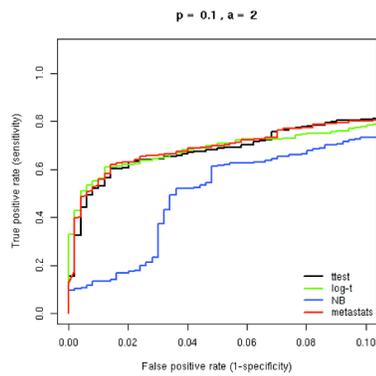
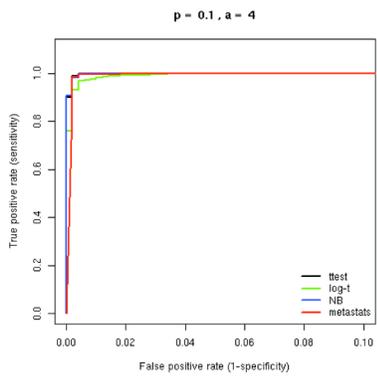
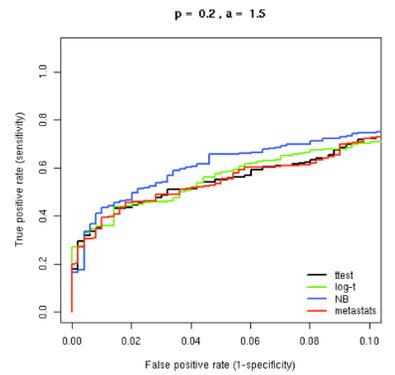
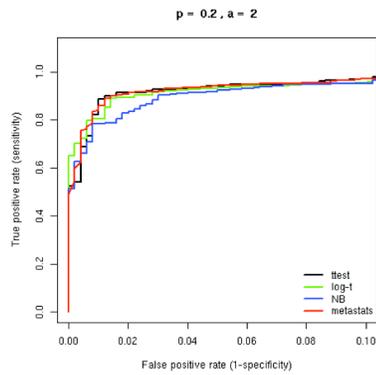
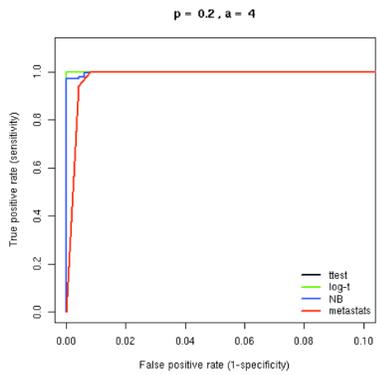
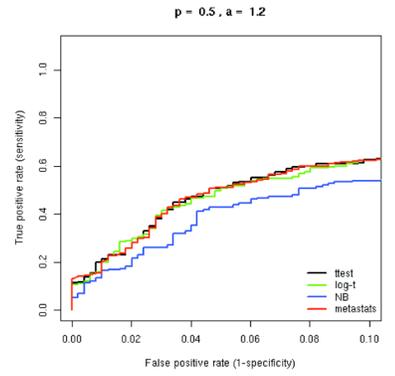
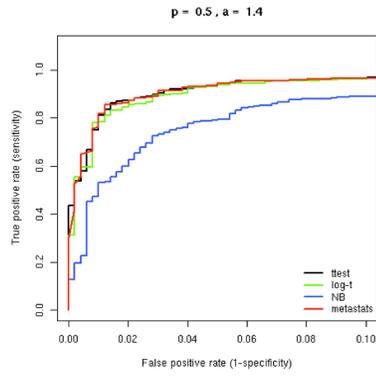
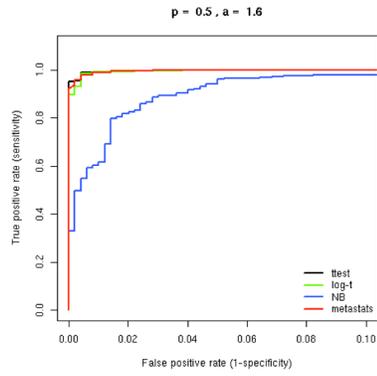
As outlined in the introduction, statistical packages developed for the analysis of SAGE data are also applicable to metagenomic datasets. In order to validate our method, we first designed simulations and compared the results of Metastats to Student's *t*-test (with pooled variances) and two methods used for SAGE data: a log-linear model (Log-t) by Lu *et al.* [90], and a negative binomial (NB) model developed by Robinson and Smyth [91].

We designed a metagenomic simulation study in which ten subjects are drawn from two groups - the sampling depth of each subject was determined by random sampling from a uniform distribution between 200 and 1000 (these depths are reasonable for metagenomic studies). Given a population mean proportion  $p$  and a dispersion value  $\phi$ , we sample sequences from a beta-binomial distribution  $B(\alpha, \beta)$ , where  $\alpha = p(1/\phi - 1)$  and  $\beta = (1-p)(1/\phi - 1)$ . Note that data from this sampling procedure fits the assumptions for Lu *et al.* as well as Robinson and Smyth and therefore we expect them to do well under these conditions. Lu *et al.* designed a similar study for SAGE data, however, for each simulation, a fixed dispersion was used for both populations and the dispersion

estimates were remarkably small ( $\phi = 0, 8e-06, 2e-05, 4.3e-05$ ). Though these values may be reasonable for SAGE data, we found that they do not accurately model metagenomic data. Figure 13 displays estimated dispersions within each population for all features of the metagenomic datasets examined below. Dispersion estimates range from  $1e-07$  to  $0.17$ , and rarely do the two populations share a common dispersion. Thus we designed our simulation so that  $\phi$  is chosen for each population randomly from a uniform distribution between  $1e-08$  and  $0.05$ , allowing for potential significant differences between population distributions. For each set of parameters, we simulated 1000 feature counts, 500 of which are generated under  $p_1 = p_2$ , the remainder are differentially abundant where  $a * p_1 = p_2$ , and compared the performance of each method using receiver-operating-characteristic (ROC) curves. Figure 14 displays the ROC results for a range of values for  $p$  and  $a$ . For each set of parameters, Metastats was run using 5000 permutations to compute p-values. Metastats performs as well as other methods, and in some cases is preferable. We also found that in most cases our method was more sensitive than the negative binomial model, which performed poorly for high abundance features.



**Figure 13.** Dispersion estimates ( $\phi$ ) for three metagenomic datasets used in this study. These plots compare dispersion values between (A) obese and lean human gut taxonomic data, (B) infant and mature human gut COG assignments, and (C) microbial and viral subsystem annotations. We find a wide range of possible dispersions in this data and significant differences in dispersions between two populations.



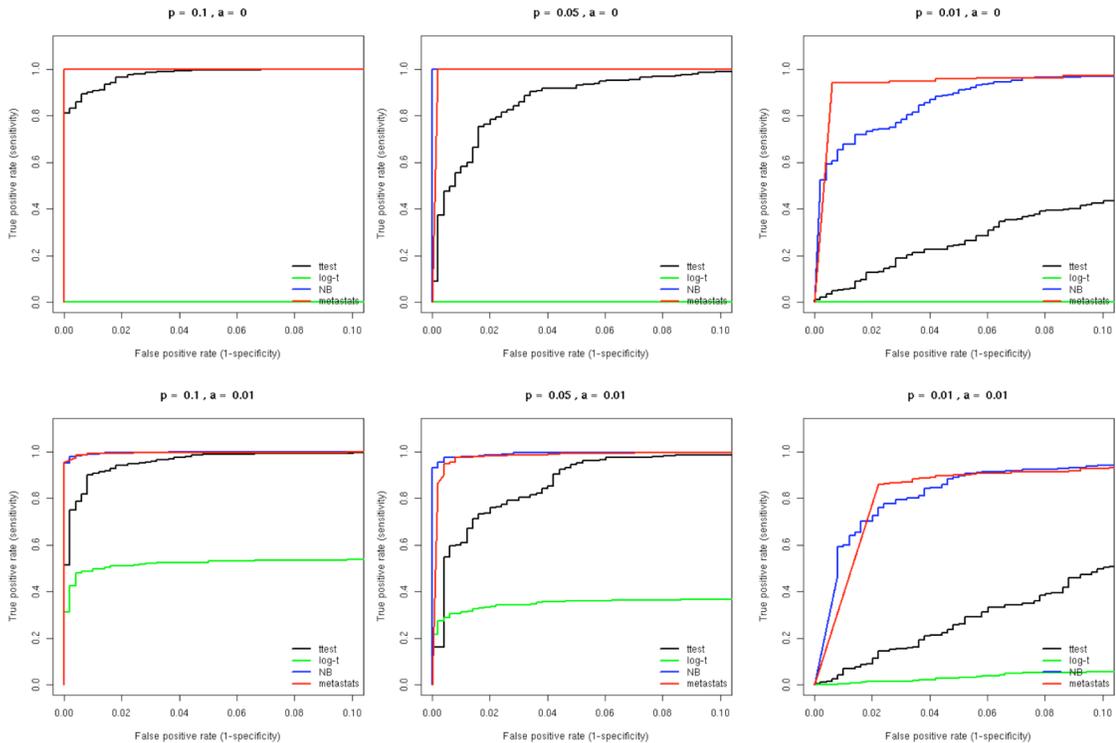
**Figure 14.** ROC curves comparing statistical methods in a simulation study. Sequences were selected from a beta-binomial distribution with variable dispersions and group mean proportions  $p_1$  and  $p_2$ . For each set of parameters, we simulated 1000 trials, 500 of which are generated under the null hypothesis ( $p_1 = p_2$ ), and the remainder are differentially abundant where  $a * p_1 = p_2$ . For example,  $p=0.2$  and  $a=2$  indicates features comprising 20% of the population that differ two-fold in abundance between two populations of interest. Parameter values for  $p_1$  and  $a$  are shown above each plot.

Our next simulation sought to examine the accuracy of each method under extreme sparse sampling. As shown in the datasets below, it is often the case that a feature may not have any observations in one population, and so it is essential to employ a statistical method that can address this frequent characteristic of metagenomic data. Under the same assumptions as the simulation above, we tested  $\alpha = 0$  and 0.01, thereby significantly reducing observations of a feature in one of the populations. The ROC curves presented in Figure 15 reveal that Metastats outperforms other statistical methods in the face of extreme sparseness. Holding the false positive rate (x-axis) constant, Metastats shows increased sensitivity over all other methods. The poor performance of Log-t is noteworthy given it is designed for SAGE data that is also potentially sparse. Further investigation revealed that the Log-t method results in a highly inflated dispersion value if there are no observations in one population, thereby reducing the estimated significance of the test.

Finally, we selected a subset of the Dinsdale *et al.* [18] metagenomic subsystem data (described below), and randomly assigned each subject to one of two populations (20 subjects per population). All subjects were actually from the same population (microbial metagenomes), thus the null hypothesis is true for each feature tested (no feature is differentially abundant). We ran each methodology on this data, recording computed p-values for each feature. Repeating this procedure 200 times, we simulated tests of 5200 null features. Table 9 displays the number of false positives incurred by each methodology given different p-value thresholds. The results indicate that the negative binomial model results in an exceptionally high number of false positives

relative to the other methodologies. Student's  $t$ -test and Metastats perform equally well in estimating the significance of these null features, while Log-t performs slightly better.

These studies show that Metastats consistently performs as well as all other applicable methodologies for deeply-sampled features, and outperforms these methodologies on sparse data. Below we further evaluate the performance of Metastats on several real metagenomic datasets.



**Figure 15.** ROC curves comparing statistical methods in a simulation study for extreme sparse sampling. Sequences were selected from a beta-binomial distribution with variable dispersions and group mean proportions  $p_1$  and  $p_2$ . For each set of parameters, we simulated 1000 trials, 500 of which are generated under the null hypothesis ( $p_1 = p_2$ ), and the remainder are differentially abundant where  $a \cdot p_1 = p_2$ . For example,  $p=0.2$  and  $a=2$

indicates features comprising 20% of the population that differ two-fold in abundance between two populations of interest. Parameter values for  $p_I$  and  $a$  are shown above each plot.

<b>P ≤</b>	<b>Number of False Positives</b>			
	<b>Metastats</b>	<b>Student-t</b>	<b>Log-t</b>	<b>NB</b>
0.001	7	4	4	109
0.005	25	25	24	121
0.01	51	52	43	133

**Table 9.** Comparison of false positives found by different methodologies. Using real metagenomic data, we simulated features with no differential abundance by randomly dividing subjects from a single population into two subpopulations. We found that for a stringent p-value threshold of 0.001, the negative binomial model (NB) resulted in a false positive rate 20 times higher than the other methodologies. The Log-t of Lu *et al.* resulted in the lowest false positive rate among the methods tested while Student’s test and Metastats performed equally well.

### **Taxa associated with human obesity**

In a recent study, Ley *et al.* [20] identified gut microbes associated with obesity in humans and concluded that obesity has a microbial element, specifically that Firmicutes and Bacteroidetes are bacterial divisions differentially abundant between lean and obese humans. Obese subjects had a significantly higher relative abundance of Firmicutes and a lower relative abundance of Bacteroidetes than the lean subjects. Furthermore, obese subjects were placed on a calorie-restricted diet for one year, after which the subjects' gut microbiota more closely resembled that of the lean individuals.

We obtained the 20,609 16S rRNA genes sequenced in Ley *et al.* and assigned them to taxa at different levels of resolution (note that 2,261 of the 16S sequences came from a previous study [19]). We initially sought to re-establish the primary result from this paper using our methodology. Table 10 illustrates that our method agreed with the results of the original study: Firmicutes are significantly more abundant in obese subjects ( $P = 0.003$ ) and Bacteroidetes are significantly more abundant in the lean population ( $P < 0.001$ ). Furthermore, our method also detected Actinobacteria to be differentially abundant, a result not reported by the original study. Approximately 5% of the sample was composed of Actinobacteria in obese subjects and was significantly less frequent in lean subjects ( $P = 0.004$ ). *Collinsella* and *Eggerthella* were the most prevalent Actinobacterial genera observed, both of which were overabundant in obese subjects. These organisms are known to ferment sugars into various fatty acids [103], further strengthening a possible connection to obesity. Note that the original study used Student's *t*-test, leading to a *p*-value for the observed difference within Actinobacteria of 0.037, 9

times larger than our calculation. This highlights the sensitivity of our method and explains why this difference was not originally detected.

To explore whether we could refine the broad conclusions of the initial study, we re-analyzed the data at more detailed taxonomic levels. We identified three classes of organisms that were differentially abundant: Clostridia ( $P = 0.005$ ), Bacteroidetes ( $P < 0.001$ ), and Actinobacteria ( $P = 0.003$ ). These three were the dominant members of the corresponding phyla (Firmicutes, Bacteroides, Actinobacteria, respectively) and followed the same distribution as observed at a coarser level. Metastats also detected nine differentially abundant genera accounting for more than 25% of the 16S sequences sampled in both populations ( $P \leq 0.01$ ). *Syntrophococcus*, *Ruminococcus*, and *Collinsella* were all enriched in obese subjects, while *Bacteroides* on average were eight times more abundant in lean subjects.

For taxa with several observations in each subject, we found good concordance between our results (p-value estimates) and those obtained with most of the other methods (Table 10). Surprisingly, we found that the negative binomial model of Robinson and Smyth failed to detect several strongly differentially abundant features in these datasets (e.g. the hypothesis test for Firmicutes results in a p-value of 0.87). This may be due in part to difficulties in estimating the parameters of their model for our datasets and further strengthens the case for the design of methods specifically tuned to the characteristics of metagenomic data. For cases where a particular taxon had no observations in one population (e.g. *Terasakiella*), the methods proposed for SAGE data seem to perform poorly.

Taxon	Obese	Lean	P values			
			Metastats	Student-t	Log-t	NB
<b>Phyla</b>						
Bacteroidetes	2.902 ±1.067	25.652±4.576	0.0002	0.0000	0.0004	0.0000
Firmicutes	89.318±2.223	72.833±4.812	0.0028	0.0025	0.0030	0.8701
Actinobacteria	4.490±1.345	0.447±0.179	0.0037	0.0371	0.0004	0.0773
<b>Classes</b>						
Bacteroidetes (class)	2.722±1.065	25.652±4.576	0.0001	0.0000	0.0005	0.0001
Actinobacteria (class)	4.490±1.345	0.447±0.179	0.0024	0.0371	0.0004	0.1858
Clostridia	84.633±2.388	66.907±5.799	0.0036	0.0042	0.0052	0.9797
<b>Genera</b>						
<i>Syntrophococcus</i>	2.380±0.383	0.666±0.337	0.0014	0.0077	0.0067	0.4860
<i>Terasakiella</i>	0.000±0	0.115±0.115	0.0016	0.1986	0.9963	0.0166
<i>Ruminococcus</i>	26.276±4.454	10.707±2.094	0.0023	0.0207	0.0039	0.6639
<i>Marinilabilia</i>	0.010±0.010	0.138±0.138	0.0024	0.2353	0.0467	0.0011
<i>Collinsella</i>	3.565±1.187	0.154±0.154	0.0052	0.0451	0.0046	0.6545
<i>Bacteroides</i>	1.841±0.963	14.623±4.444	0.0056	0.0023	0.0105	0.0012
<i>Paludibacter</i>	0.000±0	0.093±0.069	0.0059	0.0896	0.9963	0.0000
<i>Bryantella</i>	0.461±0.051	0.151±0.102	0.0065	0.0072	0.0304	0.0487
<i>Desulfovibrio</i>	0.031±0.031	0.145±0.145	0.0073	0.3390	0.2315	0.0156

**Table 10.** Differentially abundant taxa between lean and obese human gut microflora.

For the phylum, class, and genus levels (mean percentage ± s.e., p-value ≤ 0.01) we successfully re-established the major result of Ley *et al.*, and uncovered a new difference within Actinobacteria. Both Firmicutes and Actinobacteria have significantly higher relative abundances in obese people, while Bacteroidetes make up a higher proportion of the gut microbiota in the lean population. Results reveal Clostridia as the primary component of the differential abundance observed within Firmicutes, and Bacteroidetes and Actinobacteria classes explain the differential abundance observed within the

corresponding phyla. Using this p-value threshold, we expect less than one false positive among these results. The last four columns display the computed p-values for different statistical methods, including Metastats and the overdispersion methods of Lu *et al.* (Log-t) and Robinson and Smyth (NB). These results reveal NB and Student's *t*-test to be overly-conservative.

### **Differentially abundant COGs between mature and infant human gut microbiomes**

Targeted sequencing of the 16S rRNA can only provide an overview of the diversity within a microbial community but cannot provide any information about the functional roles of members of this community. Random shotgun sequencing of environments can provide a glimpse at the functional complexity encoded in the genes of organisms within the environment. One method for defining the functional capacity of an environment is to map shotgun sequences to homologous sequences with known function. This strategy was used by Kurokawa *et al.* [102] to identify clusters of orthologous groups (COGs) in the gut microbiomes of 13 individuals, including four unweaned infants. We examined the COGs determined by this study across all subjects and used Metastats to discover differentially abundant COGs between infants and mature (> 1 year old) gut microbiomes. This is the first direct comparison of these two populations as the original study only compared each population to a reference database to find enriched gene sets. Due to the high number of features (3868 COGs) tested for this dataset and the limited number of infant subjects available, our method used the pooling option to compute p-values (we chose 100 permutations), and subsequently computed q-values for each feature. Using a threshold of  $Q \leq 0.05$  (controlling the false discovery rate to 5%), we

detected 192 COGs that were differentially abundant between these two populations (Table A1). See Table 11 for most abundant detected COGs and others discussed below.

The most abundant enriched COGs in mature subjects included signal transduction histidine kinase (COG0642), outer membrane receptor proteins, such as Fe transport (COG1629), and Beta-galactosidase/beta-glucuronidase (COG3250). These COGs were also quite abundant in infants, but depleted relative to mature subjects. Infants maintained enriched COGs related to sugar transport systems (COG1129) and transcriptional regulation (COG1475). This over-abundance of sugar transport functions was also found in the original study, strengthening the hypothesis that the unweaned infant gut microbiome is specifically designed for the digestion of simple sugars found in breast milk. Similarly, the depletion of Fe transport proteins in infants may be associated with the low concentration of iron in breast milk relative to cow's milk [104]. Despite this low concentration, infant absorption of iron from breast milk is remarkably high, and becomes poorer when infants are weaned, indicating an alternative mechanism for uptake of this mineral. The potential for a different mechanism is supported by the detection of a Ferredoxin-like protein (COG2440) that was 11 times more abundant in infants than in mature subjects, while Ferredoxin (COG1145) was significantly enriched in mature subjects.

COG id	Description	Mature		Infants		Metastat qvalue
		mean	stderr	mean	stderr	
COG0205	6-phosphofructokinase	0.0017	0.0001	0.0006	0.0002	0.0313
COG0358	DNA primase (bacterial type)	0.0024	0.0001	0.0008	0.0001	0.0072
COG0507	ATP-dependent exoDNase (exonuclease V), alpha subunit - helicase superfamily I member	0.0016	0.0001	0.0008	0.0001	0.0349
COG0526	Thiol-disulfide isomerase and thioredoxins	0.0028	0.0002	0.0014	0.0002	0.0371
COG0621	2-methylthioadenine synthetase	0.0017	0.0001	0.0008	0.0002	0.045
COG0642	Signal transduction histidine kinase Predicted	0.0132	0.0009	0.007	0.0004	0.027
COG0667	oxidoreductases (related to aryl- alcohol dehydrogenases)	0.0012	0.0001	0.0021	0.0001	0.0282
COG0739	Membrane proteins related to metalloendopeptidases	0.0024	0.0001	0.0006	0.0001	0.0072
COG0745	Response regulators consisting of a CheY- like receiver domain and a winged-helix DNA-binding domain	0.0076	0.0003	0.0051	0.0004	0.0352
COG0747	ABC-type dipeptide transport system, periplasmic component	0.0011	0.0001	0.0027	0.0003	0.0352

COG1113	Gamma-aminobutyrate permease and related permeases	0.0002	0.0001	0.0018	0.0003	0.0349
COG1129	ABC-type sugar transport system, ATPase component	0.0013	0.0001	0.0028	0.0003	0.0492
COG1145	Ferredoxin	0.0017	0.0001	0.0005	0.0002	0.0217
COG1196	Chromosome segregation ATPases	0.0017	0.0001	0.0007	0.0001	0.0108
COG1249	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) component, and related enzymes	0.0006	0.0001	0.0011	0.0001	0.0349
COG1263	Phosphotransferase system IIC components, glucose/maltose/N-acetylglucosamine-specific Predicted transcriptional regulators	0.0012	0.0001	0.0031	0.0003	0.0313
COG1475		0.0025	0.0002	0.0014	0.0002	0.0435
COG1595	DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog	0.0053	0.0004	0.0013	0.0003	0.0206
COG1609	Transcriptional regulators	0.003	0.0002	0.0092	0.0013	0.0424
COG1629	Outer membrane receptor proteins, mostly Fe transport	0.012	0.0016	0.0013	0.0007	0.0313
COG1762	Phosphotransferase system mannitol/fructose-specific IIA domain (Ntr-type)	0.0004	0.0001	0.0017	0.0002	0.0293

COG1961	Site-specific recombinases, DNA invertase Pin homologs	0.0059	0.0004	0.0018	0.0006	0.0345
COG2204	Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains	0.0019	0.0002	0.0005	0.0002	0.0421
COG2244	Membrane protein involved in the export of O-antigen and teichoic acid	0.0019	0.0001	0.0009	0.0001	0.0229
COG2376	Dihydroxyacetone kinase	0.0002	0	0.0009	0.0001	0.0278
COG2440	Ferredoxin-like protein	0	0	0.0002	0	0.0394
COG2893	Phosphotransferase system, mannose/fructose-specific component IIA	0.0003	0.0001	0.0011	0.0001	0.0352
COG3250	Beta-galactosidase/beta-glucuronidase	0.0056	0.0004	0.0023	0.0006	0.0435
COG3451	Type IV secretory pathway, VirB4 components	0.0033	0.0001	0.0009	0.0003	0.0157
COG3505	Type IV secretory pathway, VirD4 components	0.0029	0.0001	0.001	0.0003	0.0278
COG3525	N-acetyl-beta-hexosaminidase	0.0016	0.0002	0.0004	0.0001	0.0352
COG3537	Putative alpha-1,2-mannosidase	0.002	0.0003	0.0002	0.0002	0.0352
COG3711	Transcriptional antiterminator	0.0004	0.0001	0.002	0.0003	0.0339
COG3712	Fe <sup>2+</sup> -dicitrate sensor, membrane component	0.0023	0.0003	0	0	0.028
COG4206	Outer membrane cobalamin receptor protein	0.0021	0.0003	0.0003	0.0001	0.0313

COG4771	Outer membrane receptor for ferrienterochelin and colicins	0.0039	0.0005	0.0006	0.0003	0.0366
---------	---	--------	--------	--------	--------	--------

**Table 11.** Differentially abundant COGs between infant and mature human gut microbiomes using a q-value threshold of 0.05. Of the 192 differentially abundant COGs detected, this table displays the most abundant 25 COGs in either mature or infant gut microbiomes. Using this threshold we expect less than 10 false positives in this dataset.

### **Differentially abundant metabolic subsystems in microbial and viral metagenomes**

A recent study by Dinsdale *et al.* profiled 87 different metagenomic shotgun samples (~15 million sequences) using the SEED platform (<http://www.theseed.org>) [18] to see if biogeochemical conditions correlate with metagenome characteristics. We obtained functional profiles from 45 microbial and 40 viral metagenomes analyzed in this study. Within the 26 subsystems (abstract functional roles) analyzed in the Dinsdale *et al.* study, we found 13 to be significantly different ( $P \leq 0.05$ ) between the microbial and viral samples (Table 12). Subsystems for RNA and DNA metabolism were significantly more abundant in viral metagenomes, while nitrogen metabolism, membrane transport, and carbohydrates were all enriched in microbial communities. The high levels of RNA and DNA metabolism in viral metagenomes illustrate their need for a self-sufficient source of nucleotides. Though the differences described by the original study did not include estimates of significance, our results largely agreed with the authors' qualitative conclusions. However, due to the continuously updated annotations in the SEED database since the initial publication, we found several differences between our results and those originally reported. In particular we found virulence subsystems to be less abundant overall than previously reported, and could not find any significant differences in their abundance between the microbial and viral metagenomes.

Subsystem	microbial	viral	Metastats p value
Carbohydrates	17.01 ± 0.77	12.87 ± 0.82	0.001
Amino Acids and Derivatives	9.29 ± 0.46	7.58 ± 0.55	0.019
Respiration	8.24 ± 1.34	3.89 ± 0.46	0.001
Photosynthesis	7.13 ± 2.38	1.16 ± 0.36	0.017
Cofactors, Vitamins, and Pigments	5.54 ± 0.27	6.44 ± 0.26	0.022
Experimental Subsystems	4.88 ± 0.31	5.80 ± 0.36	0.050
DNA Metabolism	3.99 ± 0.24	9.18 ± 1.06	0.001
Cell Wall and Capsule	3.73 ± 0.27	5.64 ± 0.71	0.009
RNA Metabolism	3.65 ± 0.21	5.23 ± 0.71	0.033
Nucleosides and Nucleotides	3.38 ± 0.18	7.72 ± 0.74	0.001
Membrane Transport	2.04 ± 0.11	1.30 ± 0.15	0.001
Nitrogen Metabolism	1.47 ± 0.08	0.93 ± 0.10	0.001
Fatty Acids and Lipids	1.46 ± 0.07	1.05 ± 0.11	0.004

**Table 12.** Differentially abundant metabolic subsystems between microbial and viral metagenomes (mean percentage ± s.e., p-values ≤ 0.05). Using this threshold we expect less than one false positive in the dataset. We find that viral metagenomes are significantly enriched for nucleotides and nucleosides and DNA metabolism, consistent with the viruses' need for self-sufficiency. Processes for respiration, photosynthesis, and carbohydrates are overrepresented in microbial metagenomes.

### Discussion

We have presented a statistical method for handling frequency data to detect differentially abundant features between two populations. This method can be applied to the analysis of any count data generated through molecular methods, including random shotgun sequencing of environmental samples, targeted sequencing of specific genes in a metagenomic sample, digital gene expression surveys (e.g. SAGE [100]), or even whole-genome shotgun data (e.g. comparing the depth of sequencing coverage across assembled

genes). Comparisons on both simulated and real dataset indicate that the performance of our software is comparable to other statistical approaches when applied to well-sampled datasets, and outperforms these methods on sparse data.

Our method can also be generalized to experiments with more than two populations by substituting the  $t$ -test with a one-way ANOVA test. Furthermore, if only a single sample from each treatment is available, a chi-squared test can be used instead of the  $t$ -test. [98].

In the coming years metagenomic studies will increasingly be applied in a clinical setting, requiring new algorithms and software tools to be developed that can exploit data from hundreds to thousands of patients. The methods described above represent an initial step in this direction by providing a robust and rigorous statistical method for identifying organisms and other features whose differential abundance correlates with disease. These methods, associated source code, and a web interface to our tools are freely available at <http://metastats.cbcb.umd.edu>.

## Chapter 5: Inferring microbial interaction webs from time-series metagenomic data

*Note in this chapter, my contributions include designing the interaction interference methodology and performing all computational experiments. Peter Turnbaugh and Jeff Gordon performed the actual humanized mouse experiments in a previously published study. I apply my methodology to their data as described below.*

### Background

In the newly established field of metagenomics, high-throughput DNA sequencing technologies enable researchers to examine the taxonomic composition and functional capabilities of complex microbial environments. Most recent metagenomic studies have focused on samples from a single time-point, however evidence is mounting that microbial communities are often not at equilibrium, rather are constantly shifting state and even oscillating [105-108]. Consequently, there is an immediate need for studies examining the temporal variation in microbial populations [26, 109].

Only a limited number of metagenomics studies have investigated the spatial and temporal dynamics of microbial communities. Eckburg *et al.* performed 16S rRNA analyses on mucosal samples along the human endogenous intestinal tract (as well as fecal samples), revealing not only extensive bacterial diversity, but also remarkable variation throughout the major sections of the colon [19]. Ley *et al.* analyzed the temporal changes in obese human gut microbiota over the course of a diet [20]. This study found that in obese subjects placed on a diet, the gut microflora shifted towards a state similar to that of their lean counterparts. Using 16S-based oligonucleotide arrays to characterize taxonomic diversity, Palmer and colleagues followed the development of gut microbiota in infants from birth through the first year of life, and found that although the

same groups of microbes dominate gut microflora, the entire community is highly variable during the first year of life for each individual newborn [80]. Additional metagenomic studies have explored the dynamic change of microbial communities within ocean water and sediment at different depths [110, 111], on apple surfaces during crop cycles [15], and inside the human gut throughout the course of antibiotic treatments [48].

Longitudinal studies will not only describe a new dimension of bacterial populations for scientists, they will also aid in modeling these systems. Computational models, supplemented by longitudinal data, will provide an opportunity to realistically model community dynamics and validate predictions. In this context, mathematical models can be used to study the underlying interactions between microbes, evaluate the effects of environmental factors, and ultimately, forecast the reaction of a microbial population to perturbation.

In this study, we employ the generalized Lotka-Volterra (gLV) model to predict microbe-microbe interactions from time-series metagenomic data. This model has been widely applied in studies of microbial and macro-scale (i.e. animal) ecology to characterize trophic and non-trophic interactions between organisms [112-114]. A variety of dynamic regimes can be captured using the gLV model including equilibrium convergence, periodicity, and chaos (i.e. unpredictable behavior beyond some time window). Furthermore, in contrast to other modeling approaches (such as generalized additive models [115, 116]), the gLV model formulation allows an intuitive interpretation of its parameters as natural ecological measures and interactions between members of a community. For the purposes of this work, we focus on (i) the prediction of interaction

*direction* (e.g. taxon  $i$  inhibits/enhances the growth rate of taxon  $j$ ) and (ii) the ancillary estimation of interaction strengths in the approximated web.

We present a reliable prediction methodology that computes confidence scores for interaction direction based on the distributions of estimated parameters in the gLV model. We further validate our approach on several simulated microbial populations. Applying our method to a metagenomic dataset following “humanized” mouse gut microbiota over a period of eight weeks, we identify several compelling interactions between dominant members of the intestinal tract.

## Materials & Methods

### **Modeling microbial communities**

Recent studies attempting to model interacting microbial populations have typically used one of two approaches: generalized additive models (GAMs) and generalized Lotka-Volterra (gLV) models, which we now compare below.

GAMs are a general statistical regression approach that incorporate smoothing splines to describe nonlinear relationships between the predictor and response variables [117]. In the context of organismal communities, the change in abundance (or logarithmic abundance) of the  $i^{\text{th}}$  organism,  $\Delta N_i$ , is modeled as a sum of nonparametric smooth functions:

$$\Delta N_i = b_i + \sum_j f_{ij}(N_j) + \varepsilon_i. \quad (1)$$

Here,  $N_j$  is the abundance (or log abundance) of taxon  $j$ ,  $b_i$  is an intercept term,  $f_{ij}(N_j)$  is a smooth function specifying the effects of taxon  $j$  on taxon  $i$ , and  $\varepsilon_i$  is a noise term.

Trosvik and colleagues have used this approach to model both artificially constructed and natural bacterial communities [116, 118].

In contrast, generalized Lotka-Volterra models are deterministic models developed specifically for analyzing communities comprised of interacting individuals (originally in the context of predator-prey relationships). These models have been employed for decades in ecological studies of natural interacting populations at macro- and micro-scales [112-114]. A discrete version of the gLV model is formulated as the following system of first-order difference equations:

$$N_i(t+1) = N_i(t) \exp \left\{ r_i \left( 1 + \frac{1}{K_i} \sum_j \alpha_{ij} N_j(t) \right) \right\} \quad (2)$$

where  $N_i(t)$  is the density of taxon  $i$  at time  $t$ ,  $r_i$  is the reproductive rate of taxon  $i$ , and  $K_i$  is its carrying capacity within the environment (i.e. the theoretical equilibrium of taxon  $i$  in the absence of all other taxa). Each coefficient  $\alpha_{ij}$  is a measure of the overall effect of taxon  $j$  on taxon  $i$ .

The general interaction web can be represented as a matrix:

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} \end{bmatrix},$$

in which any symmetric pair  $(\alpha_{ij}, \alpha_{ji})$ , defines the relationship between taxa  $i$  and  $j$  (see Table 13). Organisms within the same taxon are assumed to be competing, thus  $\alpha_{ii} = -1$  for all  $i$ . This model represents the discrete version of the generalized Lotka-Volterra ODE model:

$$\dot{N}_i = r_i N_i + \frac{r_i}{K_i} \sum_{j=1}^n \alpha_{ij} N_i N_j, \quad (3)$$

where the rate of change of the  $i^{\text{th}}$  taxon  $\dot{N}_i$ , depends on its current abundance  $N_i$  as well as the abundances of all other taxa  $N_j$ . An excellent description of the Lotka-Volterra model, as well as other ecological models, is found in [112].

There are several notable differences between the GAM and gLV approaches. GLV models presume the widely accepted *law of mass action*, that is, two populations interact at a rate proportional to the product of their abundances [119, 120], while in a GAM setting such effects are difficult to model (the influence of taxon  $j$  on taxon  $i$  in equation (1) does not depend on the abundance of taxon  $i$ ). Both models require estimation of  $(m^2+m)$  parameters corresponding to  $m^2$  interactions between organisms and  $m$  additional species-specific parameters (intercept terms in GAM, and carrying capacities and growth rates in gLV). In gLV, the interactions are defined by constant parameters, while in GAM each interaction is a smooth function whose parameters (number of knots and their positions) must also be estimated. Thus, GAMs require a larger number of parameters to be estimated, making it difficult to accurately learn these models from the limited data available. In conclusion, the GAM and gLV approaches are complementary — GAMs provide more flexibility in modeling the density dependence of the interaction between organisms, while gLVs better approximate the physical process of interacting populations and are easier to interpret. Here we provide a first application of the gLV ecological models to metagenomic time-series data.

An attractive property of the gLV model is the ability to generalize over aggregates of organisms, e.g. by modeling an environment at a higher taxonomic level.

Suppose we have a community with three interacting taxa, then the equation describing the abundance of taxon 1, is:

$$N_1(t+1) = N_1(t) \exp \left\{ r_1 \left( 1 + \frac{-N_1(t)}{K_1} + \frac{\alpha_{12} N_2(t)}{K_1} + \frac{\alpha_{13} N_3(t)}{K_1} \right) \right\},$$

which reduces to

$$N_1(t+1) = N_1(t) \exp \{ \beta_0 + \beta_1 N_1(t) + \beta_2 N_2(t) + \beta_3 N_3(t) \}.$$

Assuming that taxa 2 and 3 are members of the same higher-level taxon, which we denote as 2', they can be grouped:  $N_{2'}(t) \equiv N_2(t) + N_3(t)$ . Therefore,

$$N_1(t+1) = N_1(t) \exp \{ \beta_0 + \beta_1 N_1(t) + \beta_{2'} N_{2'}(t) \}$$

where

$$\beta_{2'} = \frac{\beta_2 N_2(t) + \beta_3 N_3(t)}{N_{2'}(t)}.$$

Now, we see that the coefficient  $\beta_{2'}$  is dependent on the abundances of its associated members, implying it is changing over time. We must then require that the relative abundances of taxa 2 and 3 do not change with respect to each other, i.e. for some positive rational value  $p$ ,  $N_2(t) = p \cdot N_3(t)$ . Under this criterion,  $\beta_{2'} = \frac{\beta_2 p + \beta_3}{(p+1)}$  and is

constant. Thus the effect of the higher-level taxon (2') is a linear aggregate of its members if the relative abundances of the members do not change over time.

Additionally, we must also show how this aggregation modifies the equations defining taxa 2 and 3 in order to formulate an equation for the higher-level taxon (2'). We need to reconcile the following equations:

$$\begin{aligned}
N_2(t+1) &= N_2(t) \exp\{\xi_0 + \xi_1 N_1(t) + \xi_2 N_2(t) + \xi_3 N_3(t)\} \\
N_3(t+1) &= N_3(t) \exp\{\gamma_0 + \gamma_1 N_1(t) + \gamma_2 N_2(t) + \gamma_3 N_3(t)\} \\
N_{2'}(t+1) &= N_{2'}(t) \exp\{\varphi_0 + \varphi_1 N_1(t) + \varphi_2 N_{2'}(t)\}
\end{aligned}$$

where  $\xi$ 's,  $\gamma$ 's, and  $\varphi$ 's are all real numbers. Using the same requirement relative abundance criterion as above,  $N_2(t) = p \cdot N_3(t)$ , we have:

$$\begin{aligned}
N_2(t+1) &= N_2(t) \exp\{\xi_0 + \xi_1 N_1(t) + \xi_2 N_{2'}(t)\} \\
N_3(t+1) &= N_3(t) \exp\{\gamma_0 + \gamma_1 N_1(t) + \gamma_2 N_{2'}(t)\}
\end{aligned}$$

and summed,

$$\begin{aligned}
N_2(t+1) + N_3(t+1) &= N_2(t) \exp\{\xi_0 + \xi_1 N_1(t) + \xi_2 N_{2'}(t)\} \\
&\quad + N_3(t) \exp\{\gamma_0 + \gamma_1 N_1(t) + \gamma_2 N_{2'}(t)\}. \quad (1)
\end{aligned}$$

Because  $N_{2'}(t) \equiv N_2(t) + N_3(t)$ ,

$$\begin{aligned}
N_{2'}(t+1) &= (N_2(t) + N_3(t)) \exp\{\varphi_0 + \varphi_1 N_1(t) + \varphi_2 N_{2'}(t)\} \\
&= N_3(t)(p+1) \exp\{\varphi_0 + \varphi_1 N_1(t) + \varphi_2 N_{2'}(t)\}. \quad (2)
\end{aligned}$$

From equations (1) and (2), we see that if we require the effects of each taxon on taxa 2 and 3 to be equivalent, (i.e.  $\xi_0 = \gamma_0$ ,  $\xi_1 = \gamma_1$ ,  $\xi_2 = \gamma_2$ ), then we can reconcile the equations:

$$\begin{aligned}
N_2(t+1) + N_3(t+1) &= (N_2(t) + N_3(t)) \exp\{\xi_0 + \xi_1 N_1(t) + \xi_2 N_{2'}(t)\} \\
&= N_{2'}(t+1).
\end{aligned}$$

Therefore, if we assume that the relative abundances of a set of taxa do not change with respect to each other, and further, that other members of the population have identical individual effects on these taxa, then the set may be aggregated into one higher taxon in

the gLV model. This aggregate property generalizes to any number of members in a higher-level taxon.

		$\alpha_{ij}$		
		-	0	+
$\alpha_{ji}$	-	competition	ammensalism	parasitism
	0	ammensalism	neutrality	commensalism
	+	predation	commensalism	mutualism

**Table 13.** Signs of interaction coefficients associated with major population interactions.  $\alpha_{ij}$  and  $\alpha_{ji}$  are symmetric components in the interaction matrix. In this notation, the predation and parasitism relationships imply that taxon  $i$  is the prey and the parasite relative to taxon  $j$ . If  $\alpha_{ij}$  is zero, it implies that taxon  $j$  has no effect on the growth rate of taxon  $i$ . In our model formulation, we assume members of the same taxon are competing, hence,  $\alpha_{ii} = -1$  for all  $i$ .

### Learning a model from the data

Usually the parameters of the gLV model are determined empirically through controlled laboratory experiments. In the context of metagenomic studies, such experiments are impractical or even impossible. Thus, we explored whether these parameters can be directly learned from time-series data. Specifically, we attempt to find a set of parameters that minimize the difference (usually expressed as a least-squares criterion) between the observed data and the model predictions. We evaluated several regression methods: dynamic regression – an analytical method traditionally used in the context of Lotka Volterra models [112-114]; nonlinear least squares – a gradient-descent approach; and

two gradient-free methods: Nelder-Mead and pattern search. These methods are briefly described below:

Dynamic regression converts the Lotka-Volterra model into a linear formulation according to the equation:

$$\begin{aligned} \ln\left(\frac{N_i(t+1)}{N_i(t)}\right) &= r_i + \frac{r_i}{K_i} \sum_j \alpha_{ij} N_j(t) \\ &= \beta_{i0} + \sum_j \beta_{ij} N_j(t) \end{aligned}$$

allowing model parameters to be estimated through linear regression.

The nonlinear least squares method [121] employs a gradient descent algorithm to estimate model parameters. The basic assumption of this approach is that surface of the objective function being minimized (in our case the least squares difference between model predictions and time-series data) is smooth. The Nelder-Mead [122] and pattern search [123] methods do not make this assumption, rather they explore the search space in a systematic fashion while attempting to minimize the objective function. Nelder-Mead explores the space through a series of operations performed on simplices within the search space, while pattern search follows the direction determined by the values of the objective function along the vectors of a positive basis of the search space. These methods were found to be more robust when optimizing potentially non-smooth functions or in high-dimensional spaces where gradient computations are expensive.

Nonlinear least squares, Nelder-Mead, and pattern search optimizations were performed using implemented MATLAB routines. All three methods employed the same constraints on the number of steps allowed ('MaxIter' = 1000) and the maximum number of function evaluations ('MaxFunEvals' = 1e6), and stopping criteria.

To avoid getting trapped in local minima, we restart the NLS, NM, and PS searches from a collection of random starting points. Specifically, when evaluating the performance of different regression methods we relied on simulated communities comprising 2,3, and 4 taxa (as described below), and sampled 10,000 random points within the search space, selected the best 10 in terms of fit between model and data, and used these as the starting point for the optimization procedure. All three methods were run on the **same** set of initial starting points. For the analysis of the actual metagenomic data we selected the best 10,000 starting points for the NLS procedure from among 500,000 random samples of the search space. All computational analyses were performed in MATLAB v. R2009a (The Mathworks Inc. Natick, MA).

### **An inference methodology with confidence values**

Recall that given a time-series metagenomic dataset our goals are (i) to predict the qualitative interaction directions (+/-) for each pair of taxa, and (ii) to accurately estimate the parameters of the full gLV model. The inference methods described in the previous section answer these goals however do not provide any measure of confidence in the predicted parameters given the fact that the underlying data are noisy. In other words, if small changes in the underlying data or the inference algorithm lead to large differences in estimated parameters (either in sign or magnitude) the resulting model cannot be trusted. To evaluate the stability of the fitting procedure we developed a stochastic extension of the NLS technique. As described before, NLS minimizes a least-squares objective function  $O(x)$  starting from some initial point in the parameter space  $x_0$ . The resulting minimizer,  $x_0^*$ , is a set of parameters in the gLV model that minimizes the

difference between the observed longitudinal data and the model predictions. However,  $x_0^*$  is likely only a *local minimizer*, that is, there exists a different set of parameters  $x^*$  such that  $O(x^*) < O(x_0^*)$ . We cannot exhaustively search the parameter space for a global minimum, so instead we randomly select a set of initial points  $x_1, \dots, x_m$  and perform NLS starting from each one, resulting in minimizers  $x_1^*, \dots, x_m^*$ , respectively. We then examine the distribution of each parameter's estimates across these minimizers to compute a confidence value for each predicted interaction.

Given a set of minimizers  $x_1^*, \dots, x_m^*$ , we first sort these in order of goodness fit (that is, by corresponding values of  $O(x_1^*), \dots, O(x_m^*)$  from smallest to largest). To reduce the effect of outliers we focus on just subset of the estimated models by excluding the bottom  $\varphi$  fraction of the models (with respect to goodness of fit). The sign of each interaction coefficient is set to the majority vote within the remaining models. For a particular interaction coefficient,  $\alpha_{ij}$ , the confidence in our prediction is the proportion of selected models that agree with the majority-vote interaction (+ or -) for that coefficient. For example, if  $\varphi = 0.05$ , we compute confidences after discarding the bottom 5% of models. Similarly, the magnitude of each parameter in the gLV model is found by computing its average (and standard error) over all selected models.

### **Small interaction network simulation design**

We simulated 11 five-taxon communities by randomly selecting model parameters according to the following equations:

$$\begin{aligned}
K_i &\sim Unif(500, 10^5) \\
r_i &\sim Unif(0.8r_*, 1.2r_*), \text{ where } r_* = 0.5 \\
\alpha_{ij} &\sim s \cdot Unif(0.5, 4), \text{ where} \\
s &= \begin{cases} 1 & \text{with probability 0.25} \\ -1 & \text{otherwise} \end{cases} \\
N_i(0) &\sim |K_i + K_i \cdot G(0, 0.25)|
\end{aligned}$$

where  $G(0, 0.25)$  is a Gaussian distribution with mean 0 and standard deviation 0.25. All datasets represented fully connected networks, i.e. every taxon influences every other taxon in some way. From these initial abundances, we simulated 20 consecutive time points, and required all taxon abundances to remain within the range of  $[10, 1e6]$  (prior to introducing error), thus preventing extinction or explosion of any taxon.

Once a satisfactory model was generated, we added noise to the data according to an error parameter  $\gamma$  such that

$$N_i(t) = N_i(t) + \gamma \cdot \mu_i \cdot G(0, 1),$$

where  $\mu_i$  is the mean abundance of taxon  $i$  throughout the time-series. The validation set of 11 five-taxon communities used  $\gamma = 0.03$ .

### **Humanized gnotobiotic mouse gut dataset**

In brief, purified adult human fecal microbiota were first transplanted via gavage into germ-free C57BL/6J mice. After initial colonization, mice remained on a low-fat mouse chow diet for four weeks. Subsequently, half of the mice were switched to a model Western diet high in fat and sugar, and followed over the course of two months. Weekly fecal samples were collected from each mouse and prepared for deep 16S rRNA 454 FLX pyrosequencing. Details of experimental protocols including mouse humanization, gut

microbial community DNA preparation, diet treatments, and 16S rRNA environmental pyrosequencing (and assignment) are described in Turnbaugh *et al.* [124]. Relative abundance measurements in each sample were calculated from corresponding 16S sequence taxonomic assignments.

Our analysis screened out rarely observed taxonomic classes (< 1% of the population on average) due to poor measurement of relative abundance. We normalized relative abundances for each sample by multiplying by  $10^5$  (to approximate 16S copies/nl). This roughly corresponds to the  $\sim 10^{11}$  cells/ml observed in mouse and human faecal samples [125].

To compensate for the fact that only weekly time points available per subject, we fit the average abundances of each class using a shape-preserving piecewise cubic Hermite interpolation. Daily abundance numbers were extracted from the interpolated curve in order to ensure a smooth model fit. 10,000 stochastic NLS iterations were run for each individual mouse time-series dataset. Constraints on model parameters during the fitting procedure required: (1) interaction coefficients to remain within (-10, 10), (2) the universal growth rate between 0 and 2, and (3) the carrying capacity of each taxon to remain within its minimum observed value and 10 times its maximum observed value (maximum observed values ranged from  $2.4 \times 10^3$  to  $7.2 \times 10^5$  16S/nl across all taxa). Constraints were required to fit the model to realistic parameters in reasonable computational time, and fitted parameter sets typically did not approach the limits of the constraints.

## Results

### **Prediction of small interaction webs**

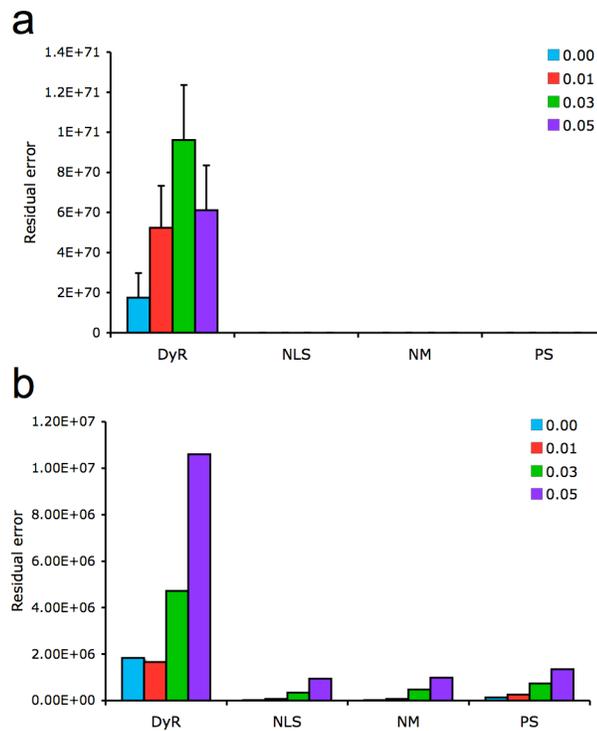
We first designed a simulation study to evaluate the quality of different methods for predicting microbial interaction networks in environments with few taxa. Using the discrete-time gLV model, we generated time-series datasets describing the dynamics of systems with up to four interacting taxa, and then attempted to re-discover the structure of the interaction network, as well as evaluate the quality of the fitted parameters.

The interaction network in each simulation is fully connected (i.e. no interaction coefficients are 0), and allows for mutualistic (+,+), competitive (-,-) and antagonistic (+,-) relationships. We assessed several techniques for data fitting: dynamic regression (DyR), nonlinear least squares (NLS), Nelder-Mead (NM), and pattern search (PS) (see Methods for details). The different procedures are compared through the *false interpretation rate* (FIR), defined as the proportion of interspecific interaction coefficients with incorrect assignments (i.e. the sign of the estimated coefficient is wrong).

Table 14 displays the accuracy of the predictions found in our simulations. In simulations involving two taxa, all methods performed well, frequently resulting in FIRs <1%. In general, the dynamic regression approach handled data with no error very well (FIRs < 4%), but had decreased performance for datasets with high error-rates. Model fits from NLS, NM, and PS methods typically outperformed dynamic regression (Figure 16). On average, the NLS method produced better results than the other methods for realistic data (error rates between 3% and 5%).

		<b>False Interpretation Rate</b>			
		Dynamic regression	NLS	NM	PS
<b>2 taxa</b>	<u>Error rate</u> ( $\gamma$ )				
	0%	0	0	0.015	0
	3%	0.020	0.001	0.005	0
	5%	0.035	0.005	0.005	0.01
<b>3 taxa</b>	<u>Error rate</u> ( $\gamma$ )				
	0%	0.018	0.071	0.168	0.175
	3%	0.125	0.061	0.180	0.170
	5%	0.175	0.070	0.155	0.155
<b>4 taxa</b>	<u>Error rate</u> ( $\gamma$ )				
	0%	0.0367	0.0783	0.2508	0.2742
	3%	0.2158	0.0767	0.2525	0.2575
	5%	0.2825	0.0992	0.2442	0.2442

**Table 14. Structure accuracy results for small networks.** For each error rate, we simulated the dynamics of 100 microbial environments with known interaction webs. Inference methods were run on the same time-series datasets in each trial. False interpretation rates are defined as the proportion of incorrectly inferred interactions across each corresponding set of 100 simulated datasets. See Methods for details of simulated communities.



**Figure 16.** Model accuracy (**a.** mean  $\pm$  s.e.m, **b.** median) in 400 simulated two-taxa systems. One hundred time-series datasets (each with a unique gLV model parameters) were simulated for each error rate (0, 1, 3, and 5%, shown in legend). The NLS, NM, and PS methods resulted in more accurate model fits than dynamic regression in 96%, 99%, and 86% of the trials, respectively. We found in these simulations that dynamic regression often resulted in very poor model fits, preventing further predictive modeling and simulation of microbial systems. In trials where the residual error of the DyR method was beyond floating-point representation (25 of the 400 trials), we reassigned the error to the largest computed residual error found across all DyR trials. Residual error is defined

as the sum-of-squared differences between the observed data and model predictions (see Methods).

We also tested the use of simple Pearson (i.e. linear) correlations of growth rates (and absolute abundance) between taxa as a way to detect interactions. However, strong linear correlations cannot translate to interaction direction between organisms, so one could not infer a competitive, mutualistic, or antagonistic relationship, but simply an interaction of some type. Simulating 1000 4-taxa communities (with an error rate of 3%), we examined what proportion of interactions would be missed if we required growth rates (or absolute abundances) to have a correlation coefficient of at least 0.15 to indicate an interaction. In this case, linear correlations of growth rates and abundances failed to detect 30% and 45% of the true interactions, respectively. Due to their poor detection ability and vague interpretation, we advise against usage of linear correlations to approximate microbial interactions in this context. It is clear from these trials that estimating interactions even for small food webs can be a formidable task.

### **Validation of regression approach**

The methods employed above are limited in that they do not provide a measure of the significance of each inferred interaction coefficient. The challenge in resolving an interaction web requires a more reliable approach where the confidence in each interaction prediction is measured, and predictions can therefore be ranked in terms of

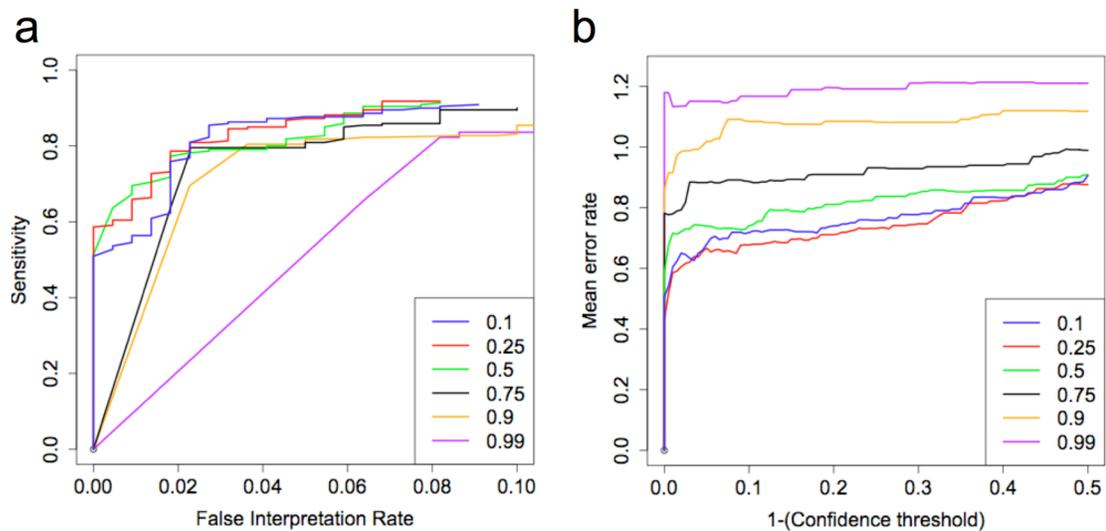
confidence. We propose a stochastic fitting methodology that generates a consensus interaction matrix, in which the sign of each interaction coefficient is given a confidence level based on the distribution of optimized fits.

Briefly, our method first runs the NLS fitting procedure for a predetermined number of iterations. We then sort the resulting parameters sets in order of goodness of fit, and discard a proportion of inferred models (designated by a parameter  $\varphi$  between 0 and 1) before generating consensus confidence values. For example, if  $\varphi = 0.05$ , we compute confidences using the top 95% of optimized fits (i.e. we ignore model fits in the bottom 5%). Given a particular interaction coefficient,  $\alpha_{ij}$ , the confidence value is the proportion of selected fits that agree with the majority-vote interaction (+ or -) for that coefficient. (See Methods for details.)

We validated our methodology using 11 five-taxa simulated datasets, each with a unique set of model parameters. These five gLV models correspond to 220 interspecific interaction coefficients, which form our test set. We ran the stochastic NLS procedure for 10,000 iterations per dataset, and subsequently generated consensus confidence scores using a series of values for  $\varphi$ . Figure 17a displays the computed ROC curves (sensitivity vs. FIR) for the test set. We discovered that culling a large proportion of optimal fits (e.g.  $\varphi = 0.95$  or  $0.99$ ) produced higher FIRs than trials utilizing smaller values of  $\varphi$ . Additionally, we found that applying a stringent confidence threshold allowed for reliable prediction of the majority of interactions with a negligible FIR. As an example, using  $\varphi = 0.5$  with a confidence cutoff of 0.98 (i.e. 98% among the models ranked in the top half, according to goodness of fit to the simulated time-series data), our method correctly predicts 69.5% of the interactions, and achieves an FIR of less than 1%.

The ecology research community has made extensive efforts to describe the “strength” of interactions between members of a food web [126]. Different goals and driving questions among researchers have led to conflicting definitions of interaction strength, but the purposes of this study, we define strength as the quantitative value of the interaction coefficients in the gLV model; these values provide a normalized measure of the per capita effect of members of the population on each taxon. In line with aim (ii), we sought to assess the accuracy of the estimated interaction coefficients in our methodology, considering the same range of values for  $\varphi$  and confidence cutoffs used to generate the ROC curves in Figure 17a. For each confidence cutoff and  $\varphi$  value, we computed the average error rate of the predicted interaction coefficients (Figure 17b), and observed that larger values of  $\varphi$  resulted in more accurate approximations of interaction strength. Indeed, using a confidence threshold of 0.95, a  $\varphi$  value of 0.5 resulted in a 36% decrease in the average error rate over a more stringent  $\varphi = 0.99$ . For each particular value of  $\varphi$ , a decrease in the confidence threshold tended to increase the average error rate.

For  $\varphi$  values  $\leq 0.5$  and confidence thresholds  $\geq 0.75$ , the average error rate remained below 0.8. In our simulations, the magnitude of an interaction coefficient was on average 2.25. This high relative error rate suggests that despite our success in predicting the general interaction (a sign of + or -), there is considerable room for improvement in estimating interaction strength. Considering the empirical performance of the parameters in our validation study, we let  $\varphi = 0.25$  and employ a confidence threshold of 0.85 in all analyses described below.



**Figure 17. Sensitivity analysis on validation data. (a)** Sensitivity vs. FIR for 11 simulated five-taxa communities. Here we define sensitivity as the proportion of interactions that are correctly predicted. The legend displays values of  $\varphi$  used for each ROC curve. By considering a large number of putative model fits, we can infer the majority true interactions between taxa with a reasonably low FIR. Observing the ‘0.99’ curve (in which we only use the top 1% of fits), we see worse performance in predicting interactions between taxa. **(b)** Corresponding mean error rates of estimated interaction coefficients. As  $\varphi$  decreases, the mean error rate of the interaction coefficients decreases significantly, indicating that considering more putative model fits results in better overall estimation of interaction strength.

### **Microbial dynamics of mice on a Western diet**

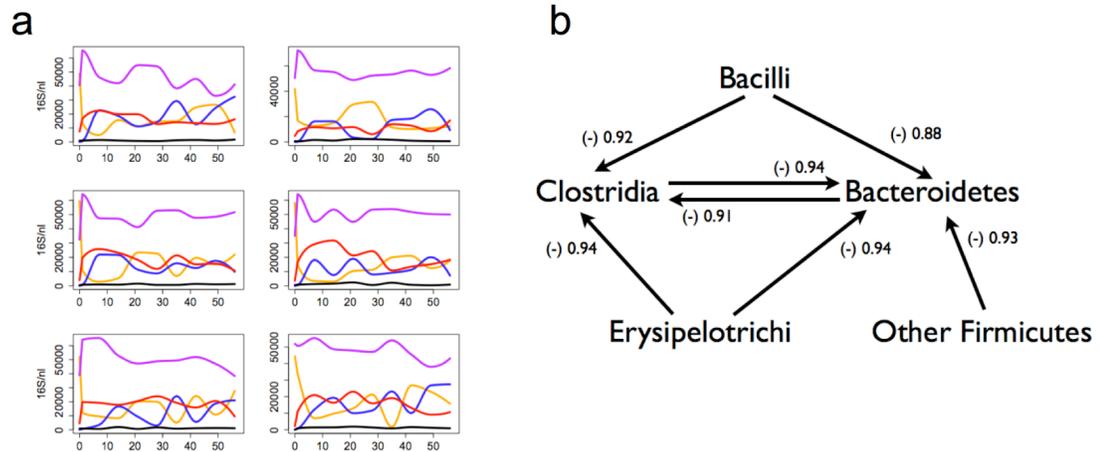
We applied our methodology to data from a metagenomic study investigating the dynamics of humanized mouse gut microflora. Twelve gnotobiotic mice were augmented with human gut microbiota and fed a mouse chow diet for four weeks. Subsequently, six of the mice continued on a mouse chow diet, while the remaining six mice were switched to a representative Western diet high in fat and sugar. For each mouse, deep pyrosequencing of the 16S rRNA V2 hypervariable region was performed on bacterial communities isolated from fecal samples over the course of eight weeks. Sequences were assigned to a taxonomy creating a taxonomic profile for each sample. Though the average gut microflora of the chow-fed mice remained relatively stable, the microbe populations in mice on the Western diet shifted dramatically throughout the study (Figure 18a).

To assess the potential microbial interactions in these models of the human gut, each time-series profile of mice switched to the Western diet was evaluated using our methodology. We assume that the microbial interaction web between any two mice in the study is the same, so our goal was to see if predicted microbial interactions were conserved across the different mice.

**Model consistency.** We separately learned the parameters of the gLV model for each individual mouse and found a high level of concordance between the individual models. Despite the fact that some time-series profiles exhibited remarkably different dynamics over the course of the study (Figure 18a), computed confidence values for each interaction coefficient were strongly correlated across all mice (Table 15). Furthermore,

estimated carrying capacities were highly correlated (mean pairwise Pearson's  $r^2$  value = 0.948). This lends support to the robustness of our methods to variable temporal patterns.

**Inferred growth rates.** Our implementation of the gLV model employs a universal net growth parameter for all taxa. Averaged across all humanized mice, the inferred growth rate was approximately 0.44 (with a standard deviation of 0.01), implying the bacterial population in the distal gut has a very slow turnover rate ( $\sim 1.6$  days). This stagnant state has also been observed previously in humans [20, 127] and has been attributed to host-associated factors including the immune response and the neutral pH levels of the colon [127, 128].



**Figure 18. Humanized mouse gut microbiota analysis. (a)** Time-series 16S profiles of six humanized mice fed a high fat Western diet. Each plot represents a different mouse. The y-axes represent normalized 16S gene copies per nanoliter of fecal material (see Methods for normalization details). Taxa are shown in the corresponding colors: Bacteroidetes (orange), Bacilli (blue), Clostridia (purple), Erysipelotrichi (red), and other Firmicutes (black). **(b)** Predicted interactions with high confidence between bacterial members of the humanized mouse gut community. Indicated interactions maintained confidence values greater than 0.85 for all studied mice. The remaining 13 possible interactions had relatively low levels of confidence. Displayed with each arrow is the general effect in parentheses (+ or -) along with the average confidence value across all mice. In this case, all arrows suggest an overall inhibitory effect (-) of one organism on another. No taxon was found to significantly enhance the growth of any other organism.

	WM1	WM2	WM3	WM4	WM5	WM6
WM1	1	0.941	0.930	0.923	0.977	0.920
WM2		1	0.819	0.829	0.887	0.875
WM3			1	0.981	0.972	0.972
WM4				1	0.975	0.876
WM5					1	0.911
WM6						1

**Table 15.** Pairwise correlation coefficients of confidence values for predicted interactions. Each cell displays the Pearson’s  $r^2$  value of confidence scores between humanized germ-free mice. We observe that computed confidence scores are highly correlated across each time-series dataset, indicating the microbial interactions with the greatest confidence are conserved across mice. Note grey cells are redundant, i.e. the  $r^2$  value of (WM1,WM2) is equal to the  $r^2$  of (WM2,WM1).

	Bacteroid.	Bacilli	Clost.	Erysip.	Other Firm.
Bacteroidetes		0.78 (0.04)	1.15 (0.05)	0.96 (0.05)	1.39 (0.12)
Bacilli	0.03 (0.06)		0.32 (0.02)	0.36 (0.02)	0.12 (0.06)
Clostridia	1.33 (0.12)	1.18 (0.10)		1.28 (0.10)	0.94 (0.13)
Erysipelotrichi	0.19 (0.01)	0.61 (0.02)	0.45 (0.01)		0.20 (0.06)
Other Firmicutes	0.43 (0.02)	0.43 (0.03)	0.51 (0.02)	0.51 (0.01)	

**Table 16.** Average (std. err) of interaction coefficient magnitude estimates across six Western-fed humanized mice.

**Predicted interactions.** A full diagram of the interactions predicted by our model is shown in Figure 18b. Interaction coefficients with the greatest confidence typically involved the Bacteroidetes or Clostridia populations. Our model predicts that Bacilli, Clostridia, Erysipelotrichi, and the subpopulation of remaining Firmicutes all inhibit the growth of Bacteroidetes with confidence values greater than 0.85 (for all individual mice). Similarly, Bacteroidetes, Bacilli, and Erysipelotrichi all inhibit the growth of Clostridia with corresponding confidence values greater than 0.90. No taxa were predicted to enhance the growth of any other group in our results. Table 16 displays the range of estimated magnitudes of all interaction coefficients.

Several of the interactions inferred from the data are supported by prior studies. For example, our model implies that Clostridia and Bacteroidetes are strongly competitive, a result also found in microarray-based studies of infant gut microflora [116]. We have observed this same predicted interaction from preliminary modeling analysis following the gut microbiota of obese humans on a low-calorie fat-restricted or carbohydrate-restricted diet for 1 year (data not shown). Furthermore, a recent genomic study reported on transcription profile modification in members of these two classes when introduced simultaneously in the guts of gnotobiotic mice. When co-colonized with *Eubacterium rectale*, *Bacteroidetes thetaiotaomicron* adapts by up-regulating a subset of genes for degrading glycans that *E. rectale* is unable to metabolize. In turn, *E. rectale* down-regulates a large number of genes encoding for glycoside hydrolases and specializes in the breakdown of simple sugars such as cellobiose and lactose when co-colonized with *B. thetaiotaomicron* [129]. The alteration of each species toward differing metabolisms and the limited effects of co-colonization on gene transcription rates for cell

replication supports the notion of a general competitive interaction between Clostridia and Bacteroidetes as predicted by our methods.

**Carrying capacities.** Our methodology consistently predicted that the Clostridia population had the highest carrying capacity in the environment, followed by Bacteroidetes; Erysipelotrichi and Bacilli had relatively similar carrying capacities. It is crucial to understand these carrying capacities are measured in 16S gene copies per nanoliter, rather than cells/nl, due to the multicopy nature of the 16S rRNA gene (see Discussion below).

Examining predicted interactions that influence the growth rate of Bacteroidetes, the effects of Clostridia and Other Firmicutes were significantly greater than that of Bacilli and Erysipelotrichi (paired T-test,  $P < 0.008$  for all concomitant tests). There was no significant difference between the interaction strengths of Clostridia and Other Firmicutes on Bacteroidetes. Additionally, the interaction strength of Erysipelotrichi on Clostridia was significantly greater than that of Bacilli (paired T-test,  $P < 0.003$ ).

### Discussion

We have presented a systematic methodology for predicting microbial interactions in time-series metagenomic datasets. Our methods were validated using simulated temporal dynamics of interacting communities and we further applied this framework to a series of metagenomic datasets describing mouse gut microbial dynamics during a prototypic Western diet. The key to our approach is a measurement of confidence for each predicted interaction coefficient based on the distribution of parameter estimates in the gLV model.

Our method is linked to the assumptions of the generalized Lotka-Volterra model, which may be violated in some studies of microbial communities. For example, the model also assumes that its parameters (carrying capacities, growth rates, and the interaction coefficients) do not depend on the abundance of the individual members of the community. These assumptions are clearly an over-simplification (e.g. quorum sensing mechanisms are density-dependent), and future work is necessary to evaluate how these simplifications affect the overall results of our analysis.

In the application of our methods to the humanized mouse gut data, our observations are based on the abundances of taxonomic classes, each of which is a combination of multiple species. A related property of the gLV model is the ability to generalize over aggregates of organisms, e.g. by modeling an environment at a higher taxonomic level. However, to reasonably merge a set of taxa  $S$  into a higher taxon, strong assumptions are required (e.g. the relative abundances of the taxa within  $S$  do not change over time). See Methods for a mathematical discussion of these assumptions. Note strict assumptions based-on taxonomic aggregation play a role in other approaches (e.g. generalized additive models [116]), and this issue remains an open problem.

Several inherent limitations exist when studying spatiotemporal dynamics using metagenomic sequence data. Though the 16S rRNA gene is an excellent candidate for amplification with universal PCR primers, the number of known copies per genome ranges from one to 15 (e.g. *Clostridium paradoxum*) [130], suggesting that taxonomic abundance profiles of microbial communities are highly skewed. Additionally, phylogenetic marker studies (and environmental genome shotgun projects in general) often lack the important measure of *cell density* - defined as the number of cells per unit

volume (e.g. cells/mL) (note however for the Western diet mouse gut study described above, previous studies have found stable microbial densities in mouse cecal samples [131]). Several experimental approaches hold significant promise in mitigating these uncertainties, including: microarrays, qPCR [111], flow cytometry [132] and other microfluidics devices [133]. The methods we present in this paper will remain applicable even as metagenomic data improves in accuracy.

While our methods predict qualitative interactions quite well, we have also shown that measurement of interaction strength is significantly more difficult. Because no model can capture all aspects of these communities, methodologies for predictive modeling will require comprehensive datasets for training as well as rigorous experimental evaluation. Nonetheless, in the spirit of optimism and hope, we conclude this discussion with a speculative application in which we use gLV models to study the dynamics of gut microbiota in dieting obese human subjects. Our goal is simple to state but incredibly challenging to reach: determine which taxonomic groups require manipulation in order to shift the obese gut microbiome structure to a lean-like state. Although the following results should be taken with a grain of salt, the approach we take illustrates how we could hypothetically utilize mathematical modeling to forecast the effects of an environmental perturbation to achieve a desired alteration. We have selected this application because of the available available and obesity's poignant impact on our lives – everyone in the United States knows someone who is obese – and this is why you shall humor the pages below.

Recent studies have revealed that human obesity is correlated with a detectable shift in the phylum-level microbiota of the distal gut [20, 21, 134]. Specifically,

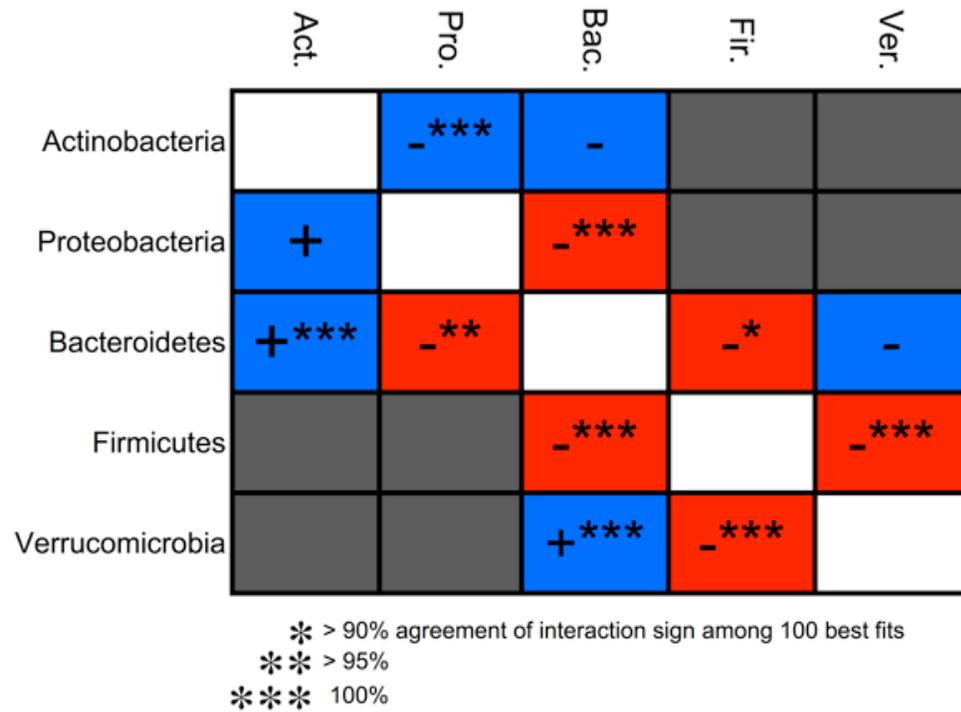
Firmicutes and Actinobacteria are enriched in obese individuals, while lean subjects maintain relatively higher abundances of Bacteroidetes. Ley *et al.* further report that obese subjects following a one-year low-calorie fat-restricted or carbohydrate-restricted diet showed significant changes in their overall gut microbial populations. As the subjects lost weight, their gut microbe levels more closely resembled that of their lean counterparts. Similar studies of germ-free mouse models have re-enforced the correlation of host adiposity not only with taxonomic composition, but also the microbial capacity for energy harvest [30, 135].

We analyzed data from ten obese subjects placed on a low-calorie diet over the course of one year [20]. Each subject provided faecal samples throughout the year; the taxonomic composition of each sample was approximated using targeted 16S rRNA gene sequencing. On average, the relative abundance of Bacteroidetes increased during dieting, while Actinobacteria and Firmicutes populations were depleted. These community shifts are inclined towards a host ecology similar to the lean human population [20, 21].

After fitting our gLV model to the data, we discovered several parameter sets resulted in both a quantitatively and qualitatively sufficient fit. Examining the top 100 model fits, we found parameter sets largely agreed on several phylum interactions (see Figure 19). There exists strong evidence for three competitive interactions: Bacteroidetes-Proteobacteria, Bacteroidetes-Firmicutes, and Firmicutes-Verrucomicrobia. Actinobacteria appear to enhance the growth of Bacteroidetes, and Bacteroidetes in turn promote Verrucomicrobia. In contrast, the overall effect of Bacteroidetes on

Actinobacteria and Verrucomicrobia on Bacteroidetes could not be established due to conflicting interaction signs in the parameter sets.

Despite the consensus on interactions indicated by estimated parameters, the gLV model is very sensitive to minor changes and is capable of exhibiting nonlinear behavior including chaos. We manually manipulated one of the best fitting parameter sets and discovered that altering a single variable can dramatically affect the overall dynamics of the community. Figure 20 displays a bifurcation diagram created by varying only one interaction coefficient, (the effect of Firmicutes on Bacteroidetes). Depending on this parameter, the community may converge to a single equilibrium, a periodic oscillation, or transition to chaotic behavior.



**Figure 19 Phylum-level interaction matrix.** Analyzing parameter estimates of the top 100 model fits, we find strong evidence for several antagonistic and competitive relationships. Each cell in the matrix displays the sign of the estimate, along with a level of confidence based on the top best 100 fits. Red cells indicate competitive interactions between species, while blue cells indicate an asymmetric antagonistic relationship. Gray cells are interactions disregarded by preliminary analysis of the data.

The past decade has seen several advancements in prebiotic and probiotic therapies for maintaining a healthy gut microbial population and treating population imbalances that result in diseases such as ulcerative colitis, inflammatory bowel disease, Crohn's disease, and pouchitis. Prebiotics are non-digestible food elements (e.g. inulin, fructo-oligosaccharides, or galacto-oligosaccharides) aimed to stimulate growth of specific bacterial groups, whereas probiotics contain live bacterial cultures to be ingested in moderation to benefit the host. Regardless of the approach, these treatments seek to encourage a healthy ecosystem through microbial manipulation. Though knowledge of the specific mechanisms of action for these treatments is not known, their effects on the general health of the intestinal tract are well studied [136, 137].

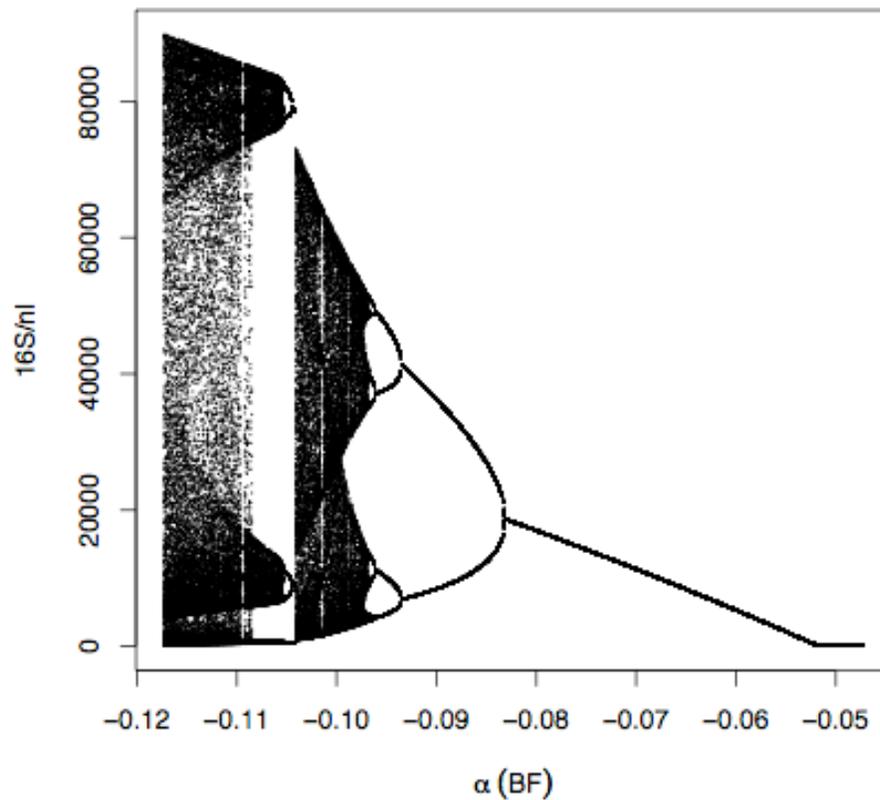
For the average obese subject, our best-fitting model predicts that only the Bacteroidetes and Firmicutes converge to within 5% of their equilibrium in the first year of dieting. The remaining three dominant phyla require at least an additional six months to reach comparable levels. To investigate whether we could increase the overall convergence rate, we simulated the microbial responses to probiotic/prebiotic therapies by altering the initial abundances of each phylum. As most treatments are only known to affect a subset of the microbial population in the gut, we limited our experimentation to manipulating each phylum individually. Each treatment has an associated microbial *load effect* – i.e. the proportional difference between a phylum and its equilibrium after treatment. For example, a load effect of 0.7 means the abundance of the phylum is 30% less than the equilibrium; 1.2 means that abundance is 20% greater than equilibrium. We examined a range of load effects (0.7 – 2.0), as practical treatments will not produce identical results in all patients.

We discovered that boosting the Bacteroidetes levels toward the lean equilibrium increases the convergence rate of all bacterial phyla. Table 17 displays the time each phylum takes to converge given a particular therapy designed to increase the abundance of Bacteroidetes. If a treatment with a load effect of 2.0 exists, our model predicts that all phyla would converge (within 5% of equilibrium) in six months or less; this cuts the overall convergence time roughly in half. If we could achieve such a convergence rate, then the altered microbiota would hypothetically extract less energy from the patient's nutrients and potentially accelerate the fat-depletion and weight loss effects of dieting. Manipulation of other populations did not produce a successful convergence rate increase for all phyla. This result suggests that the Bacteroidetes population may be an ideal target for obesity-related probiotic/prebiotic therapies, which is intriguing given that most probiotic therapies for the gastrointestinal tract utilize members of the Firmicutes or Actinobacteria [128, 138, 139].

To step back for perspective, we have gone from 16S longitudinal data in a clinical study – to modeling the dynamics of the obese gut microbiome – to forecasting which taxonomic groups are targets for shifting this community type toward a lean-like state. The Bacteroidetes hypothesis is most likely incorrect (I hope to eventually check it), but the general approach demonstrates how we can go from data to a dynamic hypothesis with a clear follow-up experiment. It is a beautiful example of the circular relationship of scientific experiments and quantitative analysis.

There is tremendous promise for the field of metagenomics, particularly in its translation to biotechnology and medicine. However, crucial prerequisites for this translation include not only comprehensive temporal datasets, but also novel quantitative

approaches including mathematical modeling to forecast how these complex systems react to treatments and environmental changes. The results we present here are just a first step in this direction.



**Figure 20. Microbial community transition to chaos.** This bifurcation diagram displays the long term dynamics of the Bacteroidetes population versus the value of a single parameter in our model. By varying  $\alpha$ (BF), (the interaction coefficient of Firmicutes and Bacteroidetes) we observe dramatic shifts in the stability of the community. For example, when  $\alpha$ (BF) equals -0.06, the Bacteroidetes population achieves a long-term equilibrium. However, if we set  $\alpha$ (BF) to -0.09, the Bacteroidetes population converges to a period-two steady state oscillation. Further decreasing of this parameter leads to a period doubling cascade and eventual transition into chaos, where the long-term dynamics of the population are highly sensitive to this one parameter.

		No treatment	Probiotic/prebiotic load effect					
			0.70	0.90	1.00	1.50	1.80	2.00
Time to convergence (weeks)	Actinobacteria	79	29	22	21	15	13	12
	Proteobacteria	166	117	102	93	18	6	5
	Bacteroidetes	91	42	2	3	4	4	4
	Firmicutes	28	21	20	20	19	19	19
	Verrucomicrobia	39	27	26	26	24	24	24

**Table 17. Potential population impacts of probiotic/prebiotic therapies on Bacteroidetes.** We define the treatment ‘load effect’ as the relative abundance of Bacteroidetes after treatment compared to its equilibrium abundance predicted by our model (e.g. 0.7 implies the treatment increased the level of Bacteroidetes to within 30% of the true equilibrium, 1.0 implies the treatment increased the Bacteroidetes abundance to the exact equilibrium value.) Time to convergence represents the number of weeks needed to maintain within 5% of the predicted equilibrium. Without treatment, 4 out of 5 phyla converge within two years. Our model predicts that stimulating Bacteroidetes population growth decreases the time to convergence for all observed phyla. We find that overloading the Bacteroidetes levels to twice the equilibrium dramatically increases convergence rates such that all phyla converge within six months. Similar manipulation of other phyla did not produce the same level of success, suggesting the Bacteroidetes should be the focus of future probiotic/prebiotic obesity therapies.

## Chapter 6: Conclusions and further study

The primary goal of my graduate research has been the development of improved methods for metagenomic analysis in order to advance our understanding of the human microbiome and other microbial populations. The ideas presented here represent novel contributions spanning elements of preprocessing, processing, and post-processing of metagenomic sequence data.

Figaro, a novel vector-trimming algorithm, can rapidly detect and remove vector sequence from multiple metagenomic sequence libraries without prior knowledge of the vector sequences themselves, thereby assisting researchers in many aspects of metagenomics including assembly, gene-finding and annotation. Since its publication in 2008, this open-source software has over 950 downloads at SourceForge.net.

In the direct processing of environmental 16S rRNA sequences, we performed a comprehensive analysis of OTU clustering methodologies that have been employed to estimate the diversity of microbial communities in landmark studies for the last decade. We have found that the choice of parameters in these methodologies is extremely important for accurate clusters, and that most studies have used parameters that are too stringent, resulting in inflated estimates of microbial diversity. While this observation has been slow to catch on, many leaders of the HMP are now aware and will hopefully require further validation of OTU-based analysis. As most HMP studies now utilize 454 pyrosequencing technology (and potentially Illumina in the future), there is an immediate need for rigorous evaluation of OTUs created from reads much shorter than the Sanger-based sequences used in our study. Pyrosequencing reads currently cannot span multiple hypervariable regions of the 16S gene, and thus there is less phylogenetic information to

classify and compare sequences. Additionally, 454 technology has been reported to produce unique artifacts in metagenomic data such as perfect and near-perfect replicates, which can severely skew relative abundance estimates [140].

In the context of 16S rRNA surveys, laboratory preparation, PCR primer bias, and chimeric sequences can also dramatically affect results. To understand the extent of each of these effects, a validation study must be performed in which a bacterial community of known composition (e.g. identifiable species, relative abundance information) is sampled and surveyed using standard techniques. Thus, a 16S taxonomic profile could be compared to an approximate truth, and the sequence dataset may be analyzed for sequencing errors, chimeras, unobserved species, and biased relative abundance measurements. This approach would give the microbial ecology community important insight into how well these protocols describe the true microbial population.

For post-processing annotated metagenomic and 16S rRNA datasets, we presented Metastats, a statistical methodology for detecting differentially abundant metagenomic features between two populations in large-scale clinical studies. Implemented as a fully automated webserver, to date Metastats has received over 700 jobs by 80 unique users. In future work, this methodology could be extended to include comparisons of three or more populations using nonparametric ANOVA with the F-statistic, or perhaps multiway-ANOVA to find interactions between multiple factors. Often software packages risk feature overload, thereby alienating the user; Metastats has been designed to be rigorous but streamlined, and additional extensions will need to conform to this standard.

Finally, moving from post-processing into modeling, we described a methodology for inferring microbial interaction webs from time-series 16S datasets. While there are many technical and experimental issues that require further validation, this project represents a step toward the holy grail of metagenomics: to model the dynamics of a microbial community and accurately forecast how a perturbation could attain a desired result. This achievement would dramatically impact many fields of science including medical microbiology, environmental sustainability, bioenergy generation, waste disposal, and industrial crop management.

As we continue to move toward this dream, we already know many of our challenges – specific technological and experimental innovations must be realized including precise estimation of microbial cell density, improvements in DNA sequencing technology, and unbiased taxonomic profiling protocols. These innovations will happen, and they will take us ever closer toward our ultimate goal. The future of metagenomics is not in the hands of microbiologists alone. There is room for many areas of expertise including mathematics, computer science, engineering, medicine, chemistry, geology, and oceanography – and all are vital.

## Appendices

### *Appendix 1: Differentially abundant COGs in comparison of infant and adult gut microbiomes*

<b>COG id</b>	<b>Description</b>	<b>mature mean</b>	<b>infant mean</b>	<b>Metastat qvalue</b>
<i>COG0249</i>	Mismatch repair ATPase (MutS family)	0.001601	0.000527	0.00722
<i>COG0358</i>	DNA primase (bacterial type)	0.002438	0.000766	0.00722
<i>COG0427</i>	Acetyl-CoA hydrolase	0.000542	0.000131	0.00722
<i>COG0482</i>	Predicted tRNA(5- methylaminomethyl-2- thiouridylate) methyltransferase, contains the PP-loop ATPase domain	0.000917	0.000270	0.00722
<i>COG0574</i>	Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	0.001271	0.000407	0.00722
<i>COG0739</i>	Membrane proteins related to metalloendopeptidases	0.002441	0.000608	0.00722
<i>COG0793</i>	Periplasmic protease	0.001465	0.000333	0.00722
<i>COG1808</i>	Predicted membrane protein	0.000168	0.000000	0.00722
<i>COG3152</i>	Predicted membrane protein	0.000086	0.000424	0.00722
<i>COG3956</i>	Protein containing tetrapyrrole methyltransferase domain and MazG-like (predicted pyrophosphatase) domain	0.000301	0.000022	0.00722

<i>COG4277</i>	Predicted DNA-binding protein with the Helix-hairpin-helix motif	0.000319	0.000032	0.00722
<i>COG5000</i>	Signal transduction histidine kinase involved in nitrogen fixation and metabolism regulation	0.000239	0.000003	0.00722
<i>COG5545</i>	Predicted P-loop ATPase and inactivated derivatives	0.001442	0.000185	0.00722
<i>COG0543</i>	2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases	0.001034	0.000581	0.00894
<i>COG0037</i>	Predicted ATPase of the PP-loop superfamily implicated in cell cycle control	0.001276	0.000497	0.01084
<i>COG0332</i>	3-oxoacyl-[acyl-carrier-protein] synthase III	0.000951	0.000243	0.01084
<i>COG0612</i>	Predicted Zn-dependent peptidases	0.001507	0.000451	0.01084
<i>COG1013</i>	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta subunit	0.000794	0.000279	0.01084
<i>COG1014</i>	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, gamma subunit	0.001017	0.000207	0.01084
<i>COG1074</i>	ATP-dependent exoDNase (exonuclease V) beta subunit (contains helicase and exonuclease domains)	0.000933	0.000398	0.01084

<i>COG1112</i>	Superfamily I DNA and RNA helicases and helicase subunits	0.000872	0.000092	0.01084
<i>COG1196</i>	Chromosome segregation ATPases	0.001676	0.000651	0.01084
<i>COG1449</i>	Alpha-amylase/alpha-mannosidase	0.000181	0.000000	0.01084
<i>COG1636</i>	Uncharacterized protein conserved in bacteria	0.000355	0.000034	0.01084
<i>COG2385</i>	Sporulation protein and related proteins	0.000646	0.000112	0.01084
<i>COG4880</i>	Secreted protein containing C-terminal beta-propeller domain distantly related to WD-40 repeats	0.000069	0.000000	0.01084
<i>COG1774</i>	Uncharacterized homolog of PSP1	0.000523	0.000097	0.01275
<i>COG0208</i>	Ribonucleotide reductase, beta subunit	0.000215	0.000675	0.01565
<i>COG0445</i>	NAD/FAD-utilizing enzyme apparently involved in cell division	0.001017	0.000295	0.01565
<i>COG1086</i>	Predicted nucleoside-diphosphate sugar epimerases	0.000869	0.000198	0.01565
<i>COG3451</i>	Type IV secretory pathway, VirB4 components	0.003289	0.000942	0.01565
<i>COG3775</i>	Phosphotransferase system, galactitol-specific IIC component	0.000182	0.000498	0.01565
<i>COG3488</i>	Predicted thiol oxidoreductase	0.000121	0.000000	0.01708
<i>COG1797</i>	Cobyrinic acid a,c-diamide synthase	0.000454	0.000092	0.01934
<i>COG1595</i>	DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog	0.005279	0.001297	0.02057

<i>COG0192</i>	S-adenosylmethionine synthetase	0.000799	0.000479	0.02167
<i>COG0465</i>	ATP-dependent Zn proteases	0.001276	0.000704	0.02167
<i>COG1145</i>	Ferredoxin	0.001656	0.000496	0.02167
<i>COG2059</i>	Chromate transport protein ChrA	0.000818	0.000186	0.02167
<i>COG0514</i>	Superfamily II DNA helicase	0.001192	0.000560	0.02270
<i>COG2244</i>	Membrane protein involved in the export of O-antigen and teichoic acid	0.001940	0.000897	0.02291
<i>COG0466</i>	ATP-dependent Lon protease, bacterial type	0.000871	0.000320	0.02330
<i>COG3884</i>	Acyl-ACP thioesterase	0.000300	0.000000	0.02330
<i>COG1881</i>	Phospholipid-binding protein	0.000052	0.000359	0.02348
<i>COG0674</i>	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha subunit	0.000832	0.000256	0.02435
<i>COG1748</i>	Saccharopine dehydrogenase and related proteins	0.000493	0.000063	0.02518
<i>COG0323</i>	DNA mismatch repair enzyme (predicted ATPase)	0.000773	0.000366	0.02701
<i>COG0642</i>	Signal transduction histidine kinase	0.013205	0.007023	0.02701
<i>COG0653</i>	Preprotein translocase subunit SecA (ATPase, RNA helicase)	0.001037	0.000452	0.02701
<i>COG2878</i>	Predicted NADH:ubiquinone oxidoreductase, subunit RnfB	0.000526	0.000121	0.02701
<i>COG4864</i>	Uncharacterized protein conserved in bacteria	0.000190	0.000017	0.02701
<i>COG1162</i>	Predicted GTPases	0.000691	0.000241	0.02783
<i>COG2376</i>	Dihydroxyacetone kinase	0.000187	0.000919	0.02783

<i>COG3505</i>	Type IV secretory pathway, VirD4 components	0.002877	0.000955	0.02783
<i>COG1493</i>	Serine kinase of the HPr protein, regulates carbohydrate metabolism	0.000463	0.000110	0.02795
<i>COG3712</i>	Fe <sup>2+</sup> -dicitrate sensor, membrane component	0.002259	0.000031	0.02795
<i>COG0667</i>	Predicted oxidoreductases (related to aryl-alcohol dehydrogenases)	0.001171	0.002134	0.02818
<i>COG1409</i>	Predicted phosphohydrolases	0.001283	0.000361	0.02818
<i>COG3294</i>	Uncharacterized conserved protein	0.000116	0.000000	0.02818
<i>COG5368</i>	Uncharacterized protein conserved in bacteria	0.000226	0.000000	0.02818
<i>COG1864</i>	DNA/RNA endonuclease G, NUC1	0.000215	0.000020	0.02823
<i>COG1762</i>	Phosphotransferase system mannitol/fructose-specific IIA domain (Ntr-type)	0.000436	0.001694	0.02929
<i>COG4877</i>	Uncharacterized protein conserved in bacteria	0.000182	0.000000	0.02932
<i>COG1083</i>	CMP-N-acetylneuraminic acid synthetase	0.000191	0.000025	0.03131
<i>COG1629</i>	Outer membrane receptor proteins, mostly Fe transport	0.011997	0.001276	0.03131
<i>COG2344</i>	AT-rich DNA-binding protein	0.000424	0.000061	0.03131
<i>COG3385</i>	FOG: Transposase and inactivated derivatives	0.000276	0.000033	0.03131
<i>COG0205</i>	6-phosphofructokinase	0.001712	0.000558	0.03131

<i>COG0602</i>	Organic radical activating enzymes	0.000839	0.000437	0.03131
<i>COG0731</i>	Fe-S oxidoreductases	0.000191	0.000000	0.03131
<i>COG1035</i>	Coenzyme F420-reducing hydrogenase, beta subunit	0.000351	0.000063	0.03131
<i>COG1072</i>	Panthothenate kinase	0.000007	0.000230	0.03131
<i>COG1263</i>	Phosphotransferase system IIC components, glucose/maltose/N-acetylglucosamine-specific	0.001155	0.003075	0.03131
<i>COG1350</i>	Predicted alternative tryptophan synthase beta-subunit (paralog of TrpB)	0.000213	0.000002	0.03131
<i>COG1351</i>	Predicted alternative thymidylate synthase	0.000225	0.000000	0.03131
<i>COG1541</i>	Coenzyme F390 synthetase	0.000647	0.000180	0.03131
<i>COG1757</i>	Na <sup>+</sup> /H <sup>+</sup> antiporter	0.001080	0.000395	0.03131
<i>COG2152</i>	Predicted glycosylase	0.000518	0.000046	0.03131
<i>COG3426</i>	Butyrate kinase	0.000291	0.000054	0.03131
<i>COG3635</i>	Predicted phosphoglycerate mutase, AP superfamily	0.000230	0.000000	0.03131
<i>COG3943</i>	Virulence protein	0.000900	0.000197	0.03131
<i>COG4206</i>	Outer membrane cobalamin receptor protein	0.002057	0.000252	0.03131
<i>COG4658</i>	Predicted NADH:ubiquinone oxidoreductase, subunit RnfD	0.000552	0.000203	0.03131
<i>COG3414</i>	Phosphotransferase system, galactitol-specific IIB component	0.000062	0.000556	0.03388
<i>COG3711</i>	Transcriptional antiterminator	0.000408	0.001954	0.03388

<i>COG1961</i>	Site-specific recombinases, DNA invertase Pin homologs	0.005869	0.001827	0.03454
<i>COG2365</i>	Protein tyrosine/serine phosphatase	0.000250	0.000040	0.03454
<i>COG0033</i>	Phosphoglucomutase	0.000050	0.000381	0.03492
<i>COG0507</i>	ATP-dependent exoDNase (exonuclease V), alpha subunit - helicase superfamily I member	0.001603	0.000850	0.03492
<i>COG0708</i>	Exonuclease III	0.000799	0.000381	0.03492
<i>COG0790</i>	FOG: TPR repeat, SEL1 subfamily	0.000726	0.000169	0.03492
<i>COG1113</i>	Gamma-aminobutyrate permease and related permeases	0.000178	0.001834	0.03492
<i>COG1160</i>	Predicted GTPases	0.000899	0.000499	0.03492
<i>COG1188</i>	Ribosome-associated heat shock protein implicated in the recycling of the 50S subunit (S4 paralog)	0.000388	0.000155	0.03492
<i>COG1249</i>	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) component, and related enzymes	0.000635	0.001138	0.03492
<i>COG1440</i>	Phosphotransferase system cellobiose-specific component IIB	0.000100	0.000672	0.03492
<i>COG2195</i>	Di- and tripeptidases	0.001438	0.000580	0.03492
<i>COG2509</i>	Uncharacterized FAD-dependent dehydrogenases	0.000871	0.000205	0.03492
<i>COG2887</i>	RecB family exonuclease	0.000219	0.000035	0.03492

<i>COG3487</i>	Uncharacterized iron-regulated protein	0.000084	0.000000	0.03492
<i>COG3560</i>	Predicted oxidoreductase related to nitroreductase	0.000182	0.000015	0.03492
<i>COG3935</i>	Putative primosome component and related proteins	0.000472	0.000189	0.03492
<i>COG3968</i>	Uncharacterized protein related to glutamine synthetase	0.001022	0.000193	0.03492
<i>COG4912</i>	Predicted DNA alkylation repair enzyme	0.000322	0.000068	0.03492
<i>COG0337</i>	3-dehydroquinate synthetase	0.000588	0.000290	0.03515
<i>COG0367</i>	Asparagine synthase (glutamine-hydrolyzing)	0.000776	0.000280	0.03515
<i>COG0549</i>	Carbamate kinase	0.000230	0.000763	0.03515
<i>COG0686</i>	Alanine dehydrogenase	0.000259	0.000039	0.03515
<i>COG0724</i>	RNA-binding proteins (RRM domain)	0.000296	0.000000	0.03515
<i>COG0745</i>	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	0.007579	0.005100	0.03515
<i>COG0747</i>	ABC-type dipeptide transport system, periplasmic component	0.001052	0.002682	0.03515
<i>COG1592</i>	Rubryerythrin	0.000845	0.000235	0.03515
<i>COG1875</i>	Predicted ATPase related to phosphate starvation-inducible protein PhoH	0.000200	0.000003	0.03515
<i>COG2239</i>	Mg/Co/Ni transporter MgtE (contains CBS domain)	0.000604	0.000184	0.03515
<i>COG2374</i>	Predicted extracellular nuclease	0.000185	0.000012	0.03515

<i>COG2893</i>	Phosphotransferase system, mannose/fructose-specific component IIA	0.000255	0.001083	0.03515
<i>COG3525</i>	N-acetyl-beta-hexosaminidase	0.001608	0.000385	0.03515
<i>COG3537</i>	Putative alpha-1,2-mannosidase	0.001956	0.000221	0.03515
<i>COG3950</i>	Predicted ATP-binding protein involved in virulence	0.000174	0.000026	0.03515
<i>COG4422</i>	Bacteriophage protein gp37	0.000472	0.000085	0.03515
<i>COG4657</i>	Predicted NADH:ubiquinone oxidoreductase, subunit RnfA	0.000468	0.000113	0.03515
<i>COG5495</i>	Uncharacterized conserved protein	0.000248	0.000037	0.03515
<i>COG1125</i>	ABC-type proline/glycine betaine transport systems, ATPase components	0.000075	0.000381	0.03560
<i>COG3292</i>	Predicted periplasmic ligand-binding sensor domain	0.000649	0.000000	0.03560
<i>COG2002</i>	Regulators of stationary/sporulation gene expression	0.000478	0.000078	0.03627
<i>COG3774</i>	Mannosyltransferase OCH1 and related enzymes	0.000245	0.000026	0.03627
<i>COG0019</i>	Diaminopimelate decarboxylase	0.001153	0.000618	0.03664
<i>COG0137</i>	Argininosuccinate synthase	0.000502	0.000220	0.03664
<i>COG0164</i>	Ribonuclease HII	0.000532	0.000271	0.03664
<i>COG0540</i>	Aspartate carbamoyltransferase, catalytic chain	0.000610	0.000339	0.03664
<i>COG0618</i>	Exopolyphosphatase-related proteins	0.000544	0.000108	0.03664
<i>COG0833</i>	Amino acid transporters	0.000077	0.000493	0.03664

<i>COG1390</i>	Archaeal/vacuolar-type H <sup>+</sup> -ATPase subunit E	0.000176	0.000040	0.03664
<i>COG3882</i>	Predicted enzyme involved in methoxymalonyl-ACP biosynthesis	0.000080	0.000000	0.03664
<i>COG4771</i>	Outer membrane receptor for ferrienterochelin and colicins	0.003898	0.000551	0.03664
<i>COG0526</i>	Thiol-disulfide isomerase and thioredoxins	0.002811	0.001441	0.03706
<i>COG0572</i>	Uridine kinase	0.000665	0.000211	0.03770
<i>COG0459</i>	Chaperonin GroEL (HSP60 family)	0.000655	0.000433	0.03826
<i>COG0532</i>	Translation initiation factor 2 (IF-2; GTPase)	0.000875	0.000453	0.03826
<i>COG0632</i>	Holliday junction resolvasome, DNA-binding subunit	0.000477	0.000223	0.03826
<i>COG0646</i>	Methionine synthase I (cobalamin-dependent), methyltransferase domain	0.000453	0.000155	0.03826
<i>COG0785</i>	Cytochrome c biogenesis protein	0.000040	0.000118	0.03826
<i>COG1077</i>	Actin-like ATPase involved in cell morphogenesis	0.000730	0.000281	0.03826
<i>COG1089</i>	GDP-D-mannose dehydratase	0.000324	0.000051	0.03826
<i>COG1262</i>	Uncharacterized conserved protein	0.000177	0.000014	0.03826
<i>COG1362</i>	Aspartyl aminopeptidase	0.000764	0.000195	0.03826
<i>COG1579</i>	Zn-ribbon protein, possibly nucleic acid-binding	0.000139	0.000000	0.03826
<i>COG2234</i>	Predicted aminopeptidases	0.000437	0.000096	0.03826
<i>COG2264</i>	Ribosomal protein L11 methylase	0.000436	0.000190	0.03826
<i>COG3174</i>	Predicted membrane protein	0.000092	0.000000	0.03826

<i>COG3643</i>	Glutamate formiminotransferase	0.000194	0.000014	0.03826
<i>COG4360</i>	ATP adenylyltransferase (5',5''-P-1,P-4- tetraphosphate phosphorylase II)	0.000049	0.000000	0.03826
<i>COG5015</i>	Uncharacterized conserved protein	0.000163	0.000012	0.03826
<i>COG0124</i>	Histidyl-tRNA synthetase	0.000618	0.000350	0.03908
<i>COG2859</i>	Uncharacterized protein conserved in bacteria	0.000113	0.000000	0.03908
<i>COG3176</i>	Putative hemolysin	0.000335	0.000053	0.03908
<i>COG4833</i>	Predicted glycosyl hydrolase	0.000248	0.000017	0.03908
<i>COG3206</i>	Uncharacterized protein involved in exopolysaccharide biosynthesis	0.000553	0.000073	0.03923
<i>COG2440</i>	Ferredoxin-like protein	0.000016	0.000182	0.03938
<i>COG2273</i>	Beta-glucanase/Beta- glucan synthetase	0.000311	0.000036	0.03991
<i>COG4804</i>	Uncharacterized conserved protein	0.000768	0.000139	0.04142
<i>COG0326</i>	Molecular chaperone, HSP90 family	0.000763	0.000233	0.04210
<i>COG0536</i>	Predicted GTPase	0.000688	0.000357	0.04210
<i>COG0676</i>	Uncharacterized enzymes related to aldose 1-epimerase	0.000058	0.000253	0.04210
<i>COG0781</i>	Transcription termination factor	0.000416	0.000252	0.04210
<i>COG1589</i>	Cell division septal protein	0.000081	0.000344	0.04210
<i>COG1643</i>	HrpA-like helicases	0.000149	0.000649	0.04210
<i>COG1696</i>	Predicted membrane protein involved in D- alanine export	0.000768	0.000253	0.04210
<i>COG1819</i>	Glycosyl transferases, related to UDP- glucuronosyltransferase	0.000174	0.000049	0.04210

<i>COG2081</i>	Predicted flavoproteins	0.000734	0.000238	0.04210
<i>COG2204</i>	Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains	0.001946	0.000471	0.04210
<i>COG2755</i>	Lysophospholipase L1 and related esterases	0.001422	0.000516	0.04210
<i>COG3408</i>	Glycogen debranching enzyme	0.000492	0.000092	0.04210
<i>COG4856</i>	Uncharacterized protein conserved in bacteria	0.000217	0.000080	0.04210
<i>COG1609</i>	Transcriptional regulators	0.003002	0.009242	0.04240
<i>COG1052</i>	Lactate dehydrogenase and related dehydrogenases	0.000930	0.000367	0.04298
<i>COG4123</i>	Predicted O-methyltransferase	0.000501	0.000181	0.04298
<i>COG1115</i>	Na <sup>+</sup> /alanine symporter	0.001166	0.000400	0.04313
<i>COG2768</i>	Uncharacterized Fe-S center protein	0.000548	0.000122	0.04313
<i>COG4775</i>	Outer membrane protein/protective antigen OMA87	0.001002	0.000177	0.04313
<i>COG4092</i>	Predicted glycosyltransferase involved in capsule biosynthesis	0.000040	0.000000	0.04341
<i>COG0083</i>	Homoserine kinase	0.000105	0.000369	0.04346
<i>COG1198</i>	Primosomal protein N' (replication factor Y) - superfamily II helicase	0.000829	0.000412	0.04346
<i>COG1475</i>	Predicted transcriptional regulators	0.002502	0.001376	0.04346
<i>COG1649</i>	Uncharacterized protein conserved in bacteria	0.000652	0.000107	0.04346
<i>COG2966</i>	Uncharacterized conserved protein	0.000510	0.000305	0.04346

<i>COG3250</i>	Beta-galactosidase/beta-glucuronidase	0.005604	0.002297	0.04346
<i>COG4231</i>	Indolepyruvate ferredoxin oxidoreductase, alpha and beta subunits	0.000375	0.000064	0.04346
<i>COG0621</i>	2-methylthioadenine synthetase	0.001714	0.000781	0.04498
<i>COG0707</i>	UDP-N-acetylglucosamine:LPS N-acetylglucosamine transferase	0.000795	0.000474	0.04498
<i>COG4099</i>	Predicted peptidase	0.000171	0.000015	0.04589
<i>COG1129</i>	ABC-type sugar transport system, ATPase component	0.001257	0.002756	0.04924

**Table 18 Differentially abundant COGs in comparison of infant and adult gut microbiomes.**

## Bibliography

1. Snyder, L. and W. Champness, *Molecular genetics of bacteria*. 3rd ed. 2007, Washington, D.C.: ASM Press. xvii, 735 p.
2. Wheelis, M., *Principles of Modern Microbiology* 2008, Sudbury, MA: Jones and Bartlette Publishers. 11-25.
3. American Society for Microbiology., D.H. Bergey, and R.S. Breed, *Bergey's manual of determinative bacteriology : a key for the identification of organisms of the class Schizomycetes*. 1923, Baltimore: Williams & Wilkins company. xi,442p.
4. Razumov, A.S., *The direct method of calculation of bacteria in water. Comparison with the Koch method*. Mikrobiologiya, 1932. **1**(2): p. 131-146.
5. Staley, J.T. and A. Konopka, *Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats*. Annu Rev Microbiol, 1985. **39**: p. 321-46.
6. Torsvik, V., J. Goksoyr, and F.L. Daae, *High diversity in DNA of soil bacteria*. Appl Environ Microbiol, 1990. **56**(3): p. 782-7.
7. Woese, C.R., J. Maniloff, and L.B. Zablen, *Phylogenetic analysis of the mycoplasmas*. Proc Natl Acad Sci U S A, 1980. **77**(1): p. 494-8.
8. Woese, C.R., et al., *The phylogenetic relationships of three sulfur dependent archaebacteria*. Syst Appl Microbiol, 1984. **5**: p. 97-105.
9. Woese, C., et al., *A comparison of the 16S ribosomal RNAs from mesophilic and thermophilic bacilli: some modifications in the Sanger method for RNA sequencing*. J Mol Evol, 1976. **7**(3): p. 197-213.
10. Fox, G.E., et al., *Classification of methanogenic bacteria by 16S ribosomal RNA characterization*. Proc Natl Acad Sci U S A, 1977. **74**(10): p. 4537-4541.
11. Lane, D.J., et al., *Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses*. Proc Natl Acad Sci U S A, 1985. **82**(20): p. 6955-9.
12. Pace, N.R., *A molecular view of microbial diversity and the biosphere*. Science, 1997. **276**(5313): p. 734-40.
13. Rusch, D.B., et al., *The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific*. PLoS Biol, 2007. **5**(3): p. e77.
14. Schloss, P.D. and J. Handelsman, *Toward a census of bacteria in soil*. PLoS Comput Biol, 2006. **2**(7): p. e92.
15. Ottesen, A.R., et al., *Impact of organic and conventional management on the phyllosphere microbial ecology of an apple crop*. J Food Prot, 2009. **72**(11): p. 2321-5.
16. Delbes, C., L. Ali-Mandjee, and M.C. Montel, *Monitoring bacterial communities in raw milk and cheese by culture-dependent and -independent 16S rRNA gene-based analyses*. Appl Environ Microbiol, 2007. **73**(6): p. 1882-91.
17. Yamane, K., et al., *Diversity and similarity of microbial communities in petroleum crude oils produced in Asia*. Biosci Biotechnol Biochem, 2008. **72**(11): p. 2831-9.
18. Dinsdale, E.A., et al., *Functional metagenomic profiling of nine biomes*. Nature, 2008. **452**(7187): p. 629-32.

19. Eckburg, P.B., et al., *Diversity of the human intestinal microbial flora*. Science, 2005. **308**(5728): p. 1635-8.
20. Ley, R.E., et al., *Microbial ecology: human gut microbes associated with obesity*. Nature, 2006. **444**(7122): p. 1022-3.
21. Turnbaugh, P.J., et al., *A core gut microbiome in obese and lean twins*. Nature, 2009. **457**(7228): p. 480-4.
22. Cole, J.R., et al., *The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis*. Nucleic Acids Res, 2005. **33**(Database issue): p. D294-6.
23. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB*. Appl Environ Microbiol, 2006. **72**(7): p. 5069-72.
24. Meyer, F., et al., *The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes*. BMC Bioinformatics, 2008. **9**(1): p. 386.
25. Covacci, A., et al., *From microbial genomics to meta-genomics*. Drug Development Research, 1997. **41**(3-4): p. 180-192.
26. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*, N.R.C.o.t.N.A. Committee on Metagenomics: Challenges and Functional Applications, Editor. 2007, The National Academies Press.
27. Tyson, G.W., et al., *Community structure and metabolism through reconstruction of microbial genomes from the environment*. Nature, 2004. **428**(6978): p. 37-43.
28. Venter, J.C., et al., *Environmental genome shotgun sequencing of the Sargasso Sea*. Science, 2004. **304**(5667): p. 66-74.
29. DeLong, E.F., *Microbial community genomics in the ocean*. Nat Rev Microbiol, 2005. **3**(6): p. 459-69.
30. Turnbaugh, P.J., et al., *An obesity-associated gut microbiome with increased capacity for energy harvest*. Nature, 2006. **444**(7122): p. 1027-31.
31. Venter, J.C., et al., *The sequence of the human genome*. Science, 2001. **291**(5507): p. 1304-51.
32. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
33. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome*. Nature, 2002. **420**(6915): p. 520-62.
34. Adams, M.D., et al., *The genome sequence of Drosophila melanogaster*. Science, 2000. **287**(5461): p. 2185-95.
35. *Human Microbiome Project - Overview*, P. Division of Program Coordination, and Strategic Initiatives, Editor. 2009.
36. Turnbaugh, P.J., et al., *The human microbiome project*. Nature, 2007. **449**(7164): p. 804-10.
37. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors*. Nature, 2005. **437**(7057): p. 376-80.
38. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*. Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
39. Andersson, B., et al., *A "double adaptor" method for improved shotgun library construction*. Anal Biochem, 1996. **236**(1): p. 107-13.

40. Chou, H.H. and M.H. Holmes, *DNA sequence quality trimming and vector removal*. *Bioinformatics*, 2001. **17**(12): p. 1093-104.
41. Read, T.D., et al., *Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae*. *Nucleic Acids Res*, 2003. **31**(8): p. 2134-47.
42. Richards, S., et al., *Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution*. *Genome Res*, 2005. **15**(1): p. 1-18.
43. Delcher, A.L., et al., *Fast algorithms for large-scale genome alignment and comparison*. *Nucleic Acids Res*, 2002. **30**(11): p. 2478-83.
44. Kurtz, S., et al., *Versatile and open software for comparing large genomes*. *Genome Biol*, 2004. **5**(2): p. R12.
45. Rabinowicz, P.D. and J.L. Bennetzen, *The maize genome as a model for efficient sequence analysis of large plant genomes*. *Curr Opin Plant Biol*, 2006. **9**(2): p. 149-56.
46. Myers, E.W., et al., *A whole-genome assembly of Drosophila*. *Science*, 2000. **287**(5461): p. 2196-204.
47. Seshadri, R., et al., *Complete genome sequence of the Q-fever pathogen Coxiella burnetii*. *Proc Natl Acad Sci U S A*, 2003. **100**(9): p. 5455-60.
48. Dethlefsen, L., et al., *The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing*. *PLoS Biol*, 2008. **6**(11): p. e280.
49. Grice, E.A., et al., *A diversity profile of the human skin microbiota*. *Genome Res*, 2008. **18**(7): p. 1043-50.
50. Huse, S.M., et al., *Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing*. *PLoS Genet*, 2008. **4**(11): p. e1000255.
51. Chen, K. and L. Pachter, *Bioinformatics for whole-genome shotgun sequencing of microbial communities*. *PLoS Comput Biol*, 2005. **1**(2): p. 106-12.
52. Wang, Q., et al., *Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy*. *Appl Environ Microbiol*, 2007. **73**(16): p. 5261-7.
53. Felsenstein, J., *PHYMLIP - phylogeny inference package (Version 3.2)*. 1989, *Cladistics* 5.
54. Hugenholtz, P., B.M. Goebel, and N.R. Pace, *Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity*. *J Bacteriol*, 1998. **180**(18): p. 4765-74.
55. Sait, M., P. Hugenholtz, and P.H. Janssen, *Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys*. *Environ Microbiol*, 2002. **4**(11): p. 654-66.
56. Meila, M., *Comparing clusterings - an information based distance*. *Journal of Multivariate Analysis*, 2007. **98**(5): p. 873-895.
57. Lambais, M.R., et al., *Bacterial diversity in tree canopies of the Atlantic forest*. *Science*, 2006. **312**(5782): p. 1917.
58. Mavromatis, K., et al., *Use of simulated data sets to evaluate the fidelity of metagenomic processing methods*. *Nat Methods*, 2007. **4**(6): p. 495-500.

59. DeSantis, T.Z., Jr., et al., *NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W394-9.
60. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity*. BMC Bioinformatics, 2004. **5**: p. 113.
61. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
62. Schloss, P.D. and J. Handelsman, *Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness*. Appl Environ Microbiol, 2005. **71**(3): p. 1501-6.
63. Warnecke, F., et al., *Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite*. Nature, 2007. **450**(7169): p. 560-5.
64. Corby-Harris, V., et al., *Geographical distribution and diversity of bacteria associated with natural populations of Drosophila melanogaster*. Appl Environ Microbiol, 2007. **73**(11): p. 3470-9.
65. Kennedy, J., et al., *Diversity of microbes associated with the marine sponge, Haliclona simulans, isolated from Irish waters and identification of polyketide synthase genes from the sponge metagenome*. Environ Microbiol, 2008. **10**(7): p. 1888-902.
66. Huber, J.A., et al., *Microbial population structures in the deep marine biosphere*. Science, 2007. **318**(5847): p. 97-100.
67. Sogin, M.L., et al., *Microbial diversity in the deep sea and the underexplored "rare biosphere"*. Proc Natl Acad Sci U S A, 2006. **103**(32): p. 12115-20.
68. Chao, A., *Non-parametric estimation of the number of classes in a population*. Scand. J. Stat., 1984. **11**: p. 265-270.
69. Chao, A. and S.M. Lee, *Estimating the Number of Classes Via Sample Coverage*. Journal of the American Statistical Association, 1992. **87**(417): p. 210-217.
70. Shannon, C.E., *A Mathematical Theory of Communication*. Bell System Technical Journal, 1948. **27**(4): p. 623-656.
71. Hugenholtz, P., *Exploring prokaryotic diversity in the genomic era*. Genome Biol, 2002. **3**(2): p. REVIEWS0003.
72. Lane, D.J., *16S/23S rRNA sequencing*, in *Nucleic Acid Techniques in Bacterial Systematics*. 1991, Wiley: New York. p. 115-175.
73. Navlakha, S., et al., *Finding Biologically Accurate Clusterings in Hierarchical Decompositions Using the Variation of Information*. 2008.
74. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
75. Ludwig, W., et al., *ARB: a software environment for sequence data*. Nucleic Acids Res, 2004. **32**(4): p. 1363-71.
76. Rand, W.M., *Objective Criteria for Evaluation of Clustering Methods*. Journal of the American Statistical Association, 1971. **66**(336): p. 846-&.
77. Schloss, P.D. and J. Handelsman, *Metagenomics for studying unculturable microorganisms: cutting the Gordian knot*. Genome Biol, 2005. **6**(8): p. 229.

78. Bik, E.M., et al., *Molecular analysis of the bacterial microbiota in the human stomach*. Proc Natl Acad Sci U S A, 2006. **103**(3): p. 732-7.
79. Batzoglou, S., et al., *ARACHNE: a whole-genome shotgun assembler*. Genome Res, 2002. **12**(1): p. 177-89.
80. Palmer, C., et al., *Development of the Human Infant Intestinal Microbiota*. PLoS Biol, 2007. **5**(7): p. e177.
81. Gill, S.R., et al., *Metagenomic analysis of the human distal gut microbiome*. Science, 2006. **312**(5778): p. 1355-9.
82. Singleton, D.R., et al., *Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples*. Appl Environ Microbiol, 2001. **67**(9): p. 4374-6.
83. Schloss, P.D., B.R. Larget, and J. Handelsman, *Integration of microbial ecology and statistics: a test to compare gene libraries*. Appl Environ Microbiol, 2004. **70**(9): p. 5485-92.
84. Schloss, P.D. and J. Handelsman, *Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures*. Appl Environ Microbiol, 2006. **72**(10): p. 6773-9.
85. Huson, D.H., et al., *MEGAN analysis of metagenomic data*. Genome Res, 2007. **17**(3): p. 377-86.
86. Lozupone, C. and R. Knight, *UniFrac: a new phylogenetic method for comparing microbial communities*. Appl Environ Microbiol, 2005. **71**(12): p. 8228-35.
87. Schloss, P.D. and J. Handelsman, *Introducing TreeClimber, a test to compare microbial community structures*. Appl Environ Microbiol, 2006. **72**(4): p. 2379-84.
88. Rodriguez-Brito, B., F. Rohwer, and R.A. Edwards, *An application of statistics to comparative metagenomics*. BMC Bioinformatics, 2006. **7**: p. 162.
89. Velculescu, V.E., et al., *Serial analysis of gene expression*. Science, 1995. **270**(5235): p. 484-7.
90. Lu, J., J.K. Tomfohr, and T.B. Kepler, *Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach*. BMC Bioinformatics, 2005. **6**: p. 165.
91. Robinson, M.D. and G.K. Smyth, *Moderated statistical tests for assessing differences in tag abundance*. Bioinformatics, 2007. **23**(21): p. 2881-7.
92. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies*. Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.
93. Troyanskaya, O.G., et al., *Nonparametric methods for identifying differentially expressed genes in microarray data*. Bioinformatics, 2002. **18**(11): p. 1454-61.
94. Efron, B. and R. Tibshirani, *An introduction to the bootstrap*. 1993, New York: Chapman & Hall. xvi, 436.
95. Hesterberg, T., *Control variates and importance sampling for efficient bootstrap simulations*. Statistics and Computing, 1996. **6**(2): p. 147-157.
96. Johns, M.V., *Importance Sampling for Bootstrap Confidence-Intervals*. Journal of the American Statistical Association, 1988. **83**(403): p. 709-714.
97. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing*. Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.

98. Zar, J.H., *Biostatistical analysis*. 4th ed. 1999, Upper Saddle River, N.J.: Prentice Hall. 1 v. (various pagings).
99. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response*. Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
100. Ruijter, J.M., A.H. Van Kampen, and F. Baas, *Statistical evaluation of SAGE libraries: consequences for experimental design*. Physiol Genomics, 2002. **11**(2): p. 37-44.
101. Krause, L., et al., *Phylogenetic classification of short environmental DNA fragments*. Nucleic Acids Res, 2008. **36**(7): p. 2230-9.
102. Kurokawa, K., et al., *Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes*. DNA Res, 2007. **14**(4): p. 169-81.
103. Tuomanen, E., *Microbial inhabitants of humans - Their ecology and role in health and disease*. Science, 2005. **308**(5722): p. 635-635.
104. Picciano, M.F. and R.H. Deering, *The influence of feeding regimens on iron status during infancy*. Am J Clin Nutr, 1980. **33**(4): p. 746-53.
105. van der Gast, C.J., A.S. Whiteley, and I.P. Thompson, *Temporal dynamics and degradation activity of a bacterial inoculum for treating waste metal-working fluid*. Environ Microbiol, 2004. **6**(3): p. 254-63.
106. Becks, L., et al., *Experimental demonstration of chaos in a microbial food web*. Nature, 2005. **435**(7046): p. 1226-1229.
107. Pringault, O., et al., *Temporal variations of microbial activity and diversity in marine tropical sediments (New Caledonia lagoon)*. Microb Ecol, 2008. **55**(2): p. 247-58.
108. Cook, K.L., et al., *Spatial and temporal changes in the microbial community in an anaerobic swine waste treatment lagoon*. Anaerobe, 2009.
109. Raes, J. and P. Bork, *Molecular eco-systems biology: towards an understanding of community function*. Nat Rev Microbiol, 2008. **6**(9): p. 693-9.
110. DeLong, E.F., et al., *Community genomics among stratified microbial assemblages in the ocean's interior*. Science, 2006. **311**(5760): p. 496-503.
111. Biddle, J.F., et al., *Metagenomic signatures of the Peru Margin seafloor biosphere show a genetically distinct environment*. Proc Natl Acad Sci U S A, 2008. **105**(30): p. 10583-8.
112. McCallum, H., *Population parameters : estimation for ecological models*. Methods in ecology. 2000, Oxford ; Malden, Mass.: Blackwell Science. x, 348 p.
113. Mounier, J., et al., *Microbial interactions within a cheese microbial community*. Applied and Environmental Microbiology, 2008. **74**(1): p. 172-181.
114. Pfister, C.A., *Estimation Competition Coefficients from Census Data: A Test with Field Manipulations of Tidepool Fishes*. The American Naturalist, 1995. **146**(2): p. 271-291.
115. Hastie, T. and R. Tibshirani, *Generalized additive models*. 1st ed. Monographs on statistics and applied probability ; 43. 1990, London ; New York: Chapman and Hall. xv, 335 p.
116. Trosvik, P., N.C. Stenseth, and K. Rudi, *Convergent temporal dynamics of the human infant gut microbiota*. Isme J, 2009.

117. Wood, S.N., *Generalized additive models : an introduction with R*. Texts in statistical science. 2006, Boca Raton, FL: Chapman & Hall/CRC. xvii, 391 p.
118. Trosvik, P., et al., *Characterizing mixed microbial population dynamics using time-series analysis*. *Isme J*, 2008. **2**(7): p. 707-15.
119. Eskola, H.T.M. and S.A.H. Geritz, *On the mechanistic derivation of various discrete-time population models*. *Bulletin of Mathematical Biology*, 2007. **69**(1): p. 329-346.
120. Berryman, A.A., *The Orgins and Evolution of Predator-Prey Theory*. *Ecology*, 1992. **73**(5): p. 1530-1535.
121. Coleman, T.F. and Y.Y. Li, *An interior trust region approach for nonlinear minimization subject to bounds*. *Siam Journal on Optimization*, 1996. **6**(2): p. 418-445.
122. Lagarias, J.C., et al., *Convergence properties of the Nelder-Mead simplex method in low dimensions*. *Siam Journal on Optimization*, 1998. **9**(1): p. 112-147.
123. Lewis, R.M. and V. Torczon, *Pattern search methods or linearly constrained minimization*. *Siam Journal on Optimization*, 2000. **10**(3): p. 917-941.
124. Turnbaugh, P., et al., *The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice*. *Sci Transl Med*, 2009. **1**(6): p. 6ra14.
125. Ley, R.E., D.A. Peterson, and J.I. Gordon, *Ecological and evolutionary forces shaping microbial diversity in the human intestine*. *Cell*, 2006. **124**(4): p. 837-48.
126. Berlow, E.L., et al., *Interaction strengths in food webs: issues and opportunities*. *Journal of Animal Ecology*, 2004. **73**: p. 585-598.
127. Macfarlane, G.T., S. Macfarlane, and G.R. Gibson, *Validation of a Three-Stage Compound Continuous Culture System for Investigating the Effect of Retention Time on the Ecology and Metabolism of Bacteria in the Human Colon*. *Microb Ecol*, 1998. **35**(2): p. 180-7.
128. Guarner, F. and J.R. Malagelada, *Gut flora in health and disease*. *Lancet*, 2003. **361**(9356): p. 512-9.
129. Mahowald, M.A., et al., *Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla*. *Proc Natl Acad Sci U S A*, 2009. **106**(14): p. 5859-64.
130. Klappenbach, J.A., et al., *rrndb: the Ribosomal RNA Operon Copy Number Database*. *Nucleic Acids Res*, 2001. **29**(1): p. 181-4.
131. Sonnenburg, J.L., C.T. Chen, and J.I. Gordon, *Genomic and metabolic studies of the impact of probiotics on a model gut symbiont and host*. *PLoS Biol*, 2006. **4**(12): p. e413.
132. Faulwetter, J.L., et al., *Microbial processes influencing performance of treatment wetlands: A review*. *Ecological Engineering*, 2009. **35**(6): p. 987-1004.
133. Moon, S., et al. *Escherichia coli counting using lens-free imaging for sepsis diagnosis*. in *Proc. SPIE*. 2009. Berlin, Germany.
134. White, J.R., N. Nagarajan, and M. Pop, *Statistical methods for detecting differentially abundant features in clinical metagenomic samples*. *PLoS Comput Biol*, 2009.
135. Ley, R.E., et al., *Obesity alters gut microbial ecology*. *Proc Natl Acad Sci U S A*, 2005. **102**(31): p. 11070-5.

136. Macfarlane, S., G.T. Macfarlane, and J.H. Cummings, *Review article: prebiotics in the gastrointestinal tract*. *Aliment Pharmacol Ther*, 2006. **24**(5): p. 701-14.
137. Rastall, R.A., et al., *Modulation of the microbial ecology of the human colon by probiotics, prebiotics and synbiotics to enhance human health: an overview of enabling science and potential applications*. *FEMS Microbiol Ecol*, 2005. **52**(2): p. 145-52.
138. Camilleri, M., *Probiotics and irritable bowel syndrome: rationale, mechanisms, and efficacy*. *J Clin Gastroenterol*, 2008. **42 Suppl 3 Pt 1**: p. S123-5.
139. O'Sullivan, G.C., et al., *Probiotics: an emerging therapy*. *Curr Pharm Des*, 2005. **11**(1): p. 3-10.
140. Gomez-Alvarez, V., T.K. Teal, and T.M. Schmidt, *Systematic artifacts in metagenomes from complex microbial communities*. *Isme J*, 2009. **3**(11): p. 1314-7.