ABSTRACT

Title of Document: NORMALIZATION AND DIFFERENTIAL ABUNDANCE ANALYSIS OF METAGENOMIC BIOMARKER-GENE SURVEYS

Joseph Nathaniel Paulson
Doctor of Philosophy
2015

Directed By: Professor Mihai Pop
Department of Computer Science
Professor Héctor Corrada Bravo
Department of Computer Science

High-throughput technologies such as whole targeted sequencing of marker-genes and whole metagenomic shotgun (WMS) sequencing have provided unprecedented insight into microbial communities and the interactions between their members. Statistical inference is a challenging task in analyzing these communities while accounting for a far too common limitation of metagenomic datasets: under-sampling. In this dissertation I present novel and robust methods for normalization and differential abundance testing of marker-gene surveys and whole metagenomic shotgun sequencing experiments. Using these methods I analyze one particular microbial community of interest, gut microbiota associated with diarrhea.

One central problem in almost any metagenomic analysis is under-sampling of

the microbial community. The analysis and interpretation of both marker-gene surveys and WMS sequencing data can bias mean and variance estimates due to the misinterpretation of zero valued counts. Even in very deep sequencing surveys, the nature of the "counting experiment" that is a metagenomic analysis can skew representative population estimates for community members. To address this issue, I characterize the biases that sparsity has on association testing of various metagenomic experiments. I developed sparsity-aware methods to 1) control for the variability in sequencing depth with a novel normalization algorithm and 2) associate gene abundance with host phenotypes. The central idea in testing associations is to weight zero values of a gene or taxa according to the posterior probability of not being observed due to under-sampling. These methods have broad general applicability in the analysis of large, relatively sparse data sets, they will provide better insight into the biological properties of complex microbial communities and their potential roles in various environmental niches.

In applying these methods to ecosystems previously unexplored I was able to obtain novel insights in the microbial community of healthy and diseased children from low-income countries. I analyzed 992 children under five years of age from low-income countries, including, The Gambia, Mali, Bangladesh, and Kenya. Approximately half of the samples were from children diagnosed with moderate-to-severe diarrhea. In applying the methods developed we recovered known diarrhea-causing pathogens, including *Escherichia/Shigella* and *Campylobacter* species. We also detected previously unknown associations with disease for several bacteria including *Granulicatella* species and *Streptococcus mitis/pneumonia* groups.

NORMALIZATION AND DIFFERENTIAL ABUNDANCE ANALYSIS OF METAGENOMIC BIOMARKER-GENE SURVEYS

By

Joseph Nathaniel Paulson

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
2015

Advisory Committee:
Professor Mihai Pop, Chair/Advisor
Professor Héctor Corrada Bravo, Chair/Advisor
Professor Doron Levy
Professor O. Colin Stine
Professor Steve Mount, Dean's Representative

Preface

The methods, tools and results in this dissertation have either been published in peer-reviewed academic journals, are in submission, or are under preparation for submission. At the time of submission, Chapters 2 and 3 have been published. Chapters 4 will be submitted soon and Chapter 5 is in preparation for submission. My contribution in the third chapter consisted of analyzing the resulting sequence data and did not involve any data preparation or sequencing.

My co-authors, advisors and mentors for each of the projects have contributed to my improved research and understanding of computational, statistical and biological perspectives and challenges. This dissertation is a combination of my own individual research as well as the contribution from my co-authors and collaborators. I am so fortunate to have been able to surround myself with experts without whom none of this work would be attainable.

A list of the papers that constitute my dissertation and a selection of other papers I have contributed to thus far.

Chapter 2:

- Paulson, JN., Stine, OC., Bravo, HC., Pop, M. (2013). Differential abundance analysis for marker-gene surveys. Nature Methods, Volume 10, pg. 1200-1202.
- Paulson, JN., Bravo, HC., Pop, M. (2014). Reply to: "a fair comparison". Nature Methods, Volume 11, pg. 359-360.

Chapter 3:

- Pop, M.*, Walker, AW.*, Paulson, JN.*, Lindsay, B.*, Antonio, M.*, Hossain MA.*, Oundo J.*, Tamboura B.*, Mai V.*, Astrovskaya I., Bravo, HC., Rance R., Stares M., Levine MM., Panchalingam S., Kotloff K., Ikumapayi UN., Ebruke C., Adeyemi M., Ahmed D., Ahmed F., Alam MT., Amin R., Siddiqui S., Ochieng JB., Ouma E., Juma J., Mailu E., Omore R., Morris JG., Breiman RF., Saha D., Parkhill J., Nataro JP., Stine, OC. (2014). Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. Genome Biology, Volume 15, pg. R76.

Chapter 4:

- Paulson, JN., Hill, CM., Almeida, M., Bravo, HC., Pop, M. Differential abundance analysis for whole metagenomic shotgun sequencing. In preparation.

Chapter 5:

- Paulson, JN.*, Dinalankara, W.*, Mendelowitz, L., Kross, S., Bravo, HC, Pop, M. (2015). The gut microbiome strongly predicts phenotype X. In preparation.

A selection of papers that I have contributed to (being a co-author):

- Talukder, H.*, **Paulson, JN.***, Stine, OC., Pop, M., Bravo, HC. (2015). Longitudinal differential abundance analysis of microbial marker-gene surveys. In preparation.
- Astrovskaya, I., **Paulson, JN.**, Chakraborty, S., Lindsay BR., Li, S., Pop, M., Bravo, HC., Harro, C., Walker, AW., Parkhill, J., Walker, RI., Rance, R., Sack, DA., Stine, OC. (2015). Reduction in intestinal microbiota diversity during diarrhea and treatment with ciprofloxacin in volunteers. In preparation.
- Chelaru, F., **Paulson JN.**, Pop, M., Bravo, HC. Metaviz: interactive visualization of metagenomic data. In preparation.

* Indicates co-first authors

Dedication

To my family.

Acknowledgements

First and foremost I would like to thank my advisors Professor Mihai Pop and Professor Héctor Corrada-Bravo. I owe much of my academic growth and achievements to their mentorship. Mihai revealed to me a universe I knew nothing of before and gave me opportunities to work on problems I find interesting and important. Héctor's continuous guidance and support has saved me numerous times. Their lessons in statistics, computational biology, and life will serve me well into the next chapters of life. Thank you, I am grateful.

I would like to thank each of my committee members, Professor O. Colin Stine, Professor Doron Levy, and Professor Steve Mount for providing their advice and insight. Colin in particular has been a constant inspiration, cheering me on and providing me with opportunities to work on meaningful data and explore the world!

Many thanks to my friends and colleagues in the Math department and at CBCB without whom the road would certainly have been much more difficult. Dave Darmon and Stefan Doboszaczak started this journey with me and have been wonderful friends and roommates. Rob Patro, Darya Filippova, Geet Duggal, and Hao Wang helped fill my visits to the submarine with lots of `cha`. My lab mates Chris Hill, Dave Kelley, Bo Liu, Mohammad Ghodsi, Sean Kross, Serge Koren, and Todd Treangan contributed to many insightful discussions. Justin Wagner, Florin Chelaru, Wikum Dinalankara, Senthil Muthiah – you all are awesome.

I want to send a special thank you to Lee Mendelowitz, Kwame Okrah, and Hisham Talukder – you three have been great housemates, lab mates and revelers of the night. Joyce Hsiao thank you being the supportive older sister in times of need.

Vivey Chen thank you for being there every step of the way. I'll miss and cherish the time with each of you and am excited for the new memories we'll make together.

This dissertation would not have been possible without the support of my family. Mamajon, Babajon, Mehrdad, Haleh, Sheila, Geoff, Layla, Sorab, Akila, Sasha, Maxx, Matthew, Maddy, Josh, Gaga, Papa Jerry, Mom and Dad. Each one of you has given me so much love and support.

Thank you.

Table of Contents

List of Figures

1 Metagenomics has flourished in charting uncharacterized ecosystems

The democratization of high-throughput DNA sequencers has allowed exploring deeply and simultaneously the entire microbial DNA present in an ecosystem, also known as the metagenome [1]. Due to the massive amount of data generated by a single run of these high-throughput sequencers a new discipline emerged, called metagenomics, with the aim to understand the microbes present in each environmental niche. Metagenomic data includes the DNA of the many organisms present, often hundreds of thousands of organisms. High-throughput technologies such as targeted sequencing of marker-genes and whole community metagenomic shotgun (WMS) sequencing have provided unprecedented insight into microbial communities and the interactions between their members. While metagenomic studies originally focused on exploratory and validation projects, they now are rapidly being applied to clinical settings. In these settings, researchers are interested in finding characteristics of the microbiome that correlate with the health or disease status of patients.

Researchers have focused, for example, on the study of microbial communities associated with diarrhea [2], periodontitis [3], bacterial vaginosis [4], diabetes [5, 6], obesity [7, 8], liver disease [9] and eczema [10]. In these settings, the identification of potentially pathogenic or probiotic bacteria revealing significant differences in their abundance in a disease population is critical. While methods for whole-scale community comparisons are commonly used [11, 12] there is a need for

tools that discern taxon-specific disease associations in both marker-gene surveys and WMS studies. Global community comparisons allow for an ecological understanding of the source of multiple samples, but are limited in detailing the specific bacteria that differentiate phenotypes.

1.1 Metagenomic data generation and preparation

There are two commonly used high-throughput techniques to generate metagenomic data. The cheaper method in characterizing microbial communities has been the targeted sequencing of the 16S ribosomal RNA gene from selected samples. 'Universal' primers, strands of nucleic acids that act as the starting points for DNA synthesis, amplify specific hyper-variable regions within the 16S rRNA gene, and the corresponding segments are sequenced using high-throughput long-read technology. The output of the sequencer is converted through a process of base calling to strings of letters representing the four bases comprising the DNA - AGCT (adenine, guanine, cytosine, thymine). The strings 'read' by the sequencer are referred to as sequencing reads. Sequence reads are first clustered into operational taxonomic units (OTUs) [13] and representative sequences from each cluster are then annotated with their taxonomic origin by searching against a database of 16S rDNA reference sequences – a commonly used database is the Ribosomal Database Project (RDP) [14]. The number of reads assigned to a particular organism is assumed to represent an approximation of the abundance of that organism within the community.

The 16S ribosomal RNA gene is commonly used due to its ubiquity in prokaryotes and the species-specific signatures in hypervariable regions along the gene. However, targeted sequencing studies are limited by the marker-gene chosen

and how conserved a region is for a given species or at other taxonomic levels. Additionally, obtaining the exact biological functions performed by the organisms containing a particular 16S rRNA gene is limited by potential horizontal gene transfer and to the annotation of genomes sequenced [15].

An alternative approach involves sequencing the entire DNA of a community – a process called whole metagenomic sequencing (WMS). WMS provides information about the entire genomic content of organisms, thereby providing a better understanding of the biological functions performed by these organisms. One strategy for analyzing WMS data is to estimate the abundance of microbial genes by aligning the raw reads to a non-redundant microbial gene catalogue [16]. Alignments can be summarized into count matrices representing how many times each gene is seen in each sample, information used for all downstream analyses. While both methods generate count matrices differently, many of the general characteristics confounding inference are similar.

1.2 Normalization

Metagenomic data in its rawest form can be thought of as a sparse positive integer matrix. Given m features and n samples, the elements in a count matrix (m, n), c_{ij}, are the number of reads annotated for a particular feature i (whether it be OTU or gene cluster) in sample j.

$$\mathrm{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & & c_{1n} \\ c_{21} & c_{22} & \cdots & & c_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & & c_{mn} \end{bmatrix}$$

Variation in depth of coverage - the total number of sequences produced for

each sample - yields incomparable read count measurements. Samples with few sequencing reactions generate fewer reads than those with more sequencing reactions. As a result each sample's output is variable solely due to the technology. To make samples comparable despite variable output we preprocess the count matrix through a process known as normalization.

Every normalization method makes a biological assumption about the invariant properties of the data that relate to sampling rate. By either scaling or transforming the count matrix the count distributions of two distinct samples are now presumed to be comparable. Normalization schemes often fall into either scaling methods, e.g. average of house-keeping genes, spike-in measurements, assuming equal biomass (leading to scaling by sequencing depth), or transformation methods, e.g. assumption of equal count distributions (leading to quantile normalization). Each count c_{ij} is either scaled or transformed depending on the normalization scheme chosen. Normalization is an important pre-processing procedure that can alter downstream analyses and can bias results if performed incorrectly.

By far, the most commonly used approach in metagenomics normalizes the data by dividing feature read counts by the total number of reads (denoted $S_j = \sum_{i=1}^{m} c_{ij}$) in each sample, i.e., converting feature counts to proportions. This normalization procedure which we refer to as total-sum normalization has been shown to incorrectly bias differential abundance estimates in data derived through high-throughput technologies [17, 18] and potentially biases metagenomic data. Furthermore, statistics derived by this normalization may induce spurious correlations between features, a fact known since 1897 in the work of Pearson [19].

Normalization has a long history in high-throughput experiments measuring gene expression with microarray technology [20–22] and sequencing-based assays such as RNA-seq [17, 23, 24]. However, there are key differences between RNA-seq data and metagenomic data. Methods developed for RNA-seq data do not take into account the largest technical artifact in metagenomic data, sparsity due to under-sampling.

In the next section we highlight a few of the more popular normalization methods from the microarray and RNA-seq literature and the assumptions that led to their development.

1.2.1 Transformation methods

Quantile normalization is a transformation method that equates count distributions to eliminate inter-sample sampling depth differences. Quantile normalization assumes count distributions are the same across samples. However, the community makeup of each microbial community and shape of the distribution is dependent on the environmental niche and numerous factors leading to high intra-sample variability.

This method was first proposed for microarray data, but has been generalized and applied to other high-throughput sequencing data types as well. The distributions of samples $1, \dots, n$ are the same if they have the same quantiles. Denote k quantiles of sample j as q_{kj}, $k = 1, \dots, p$ and $j = 1,2, \dots, N$, the k^{th} quantile vector across N samples as $\mathrm{q}_k = (q_{k1}, q_{k2}, \dots, q_{kN})'$, and the set of all quantile vectors as $\mathrm{q} = (\mathrm{q}_1, \mathrm{q}_2, \dots, \mathrm{q}_p)$. The objective is a projection of each of the quantile vectors onto an orthonormal vector in the N-dimensional vector space. Consider a N-dimenstional

unit vector $\mathrm{d} = (\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}})$. Using the Gram-Schmidt process, we can find a projection of each of the quantile vectors as follows:

$$Proj_{\mathrm{d}}(\mathrm{q}_k) = \frac{\mathrm{q}_k \cdot \mathrm{d}}{\mathrm{d} \cdot \mathrm{d}} \mathrm{d}$$

$$= \frac{\frac{1}{\sqrt{N}} \sum_{j=1}^{N} q_{kj}}{1} \times \frac{1}{\sqrt{N}} 1$$

$$= \frac{1}{N} \sum_{j=1}^{N} q_{kj} \times 1$$

where 1 is a N-dimensional column vector of 1's.

Therefore, we can make the samples to have the same distributions by substituting the quantiles of each sample with the corresponding average quantile across samples. For every feature i, sample j, if the abundance value c_{ij} is the k^{th} quantile in sample j, then replace c_{ij} with y_{ij}, such that: $y_{ij} = \frac{1}{N} \sum_{j=1}^{N} q_{kj}$. The quantile normalized abundance estimates are then subsequently log transformed. This is a special case of a transformation using the empirical distribution for each sample and the empirical distribution of the averaged sample quantiles.

1.2.2 Scaling methods

Scaling methods take advantage of the invariant properties of the data to estimate a sample's sampling rate. The estimates are used to scale counts in correcting for technical artifacts. As an example, upper quantile normalization makes the assumption that the positive count distributions are similar at the low to mid quantiles and only begin to deviate from each other at the 75$^{\text{th}}$ quantile. The normalization factor for this method would be a sample's 75$^{\text{th}}$ quantile and the

normalized count is y_{ij}, such that: $y_{ij} = c_{ij}/q_{75j}$ [17].

In DESeq, count values are expected to follow closely the geometric mean of gene expression values across samples with the assumption that read count frequencies increase exponentially with sequencing depth. As such the geometric mean is the choice of central tendency measure [23]. A normalizing factor is calculated for each sample as the median of the ratio between feature read counts and the geometric mean of read counts across all samples. For samples $j = 1, \dots, N$, adjusted scaling factor S'_j is calculated as $S'_j = \underset{i}{median} \frac{c_{ij}}{(\prod_{j=1}^{N} c_{ij})^{\frac{1}{N}}}$. As such, the normalized count y_{ij} is $y_{ij} = \frac{c_{ij}}{S'_j}$

TMM makes the biological assumption that the majority of genes are not differentially expressed. TMM, instead of using all samples, makes use of a single sample as reference. The method attempts to estimate the sampling rate by calculating the average of log-fold-changes after removing outliers (defined as the outer 30%). Using notation from [24]: $\log_2(TMM_j^{(r)}) = \frac{\sum_{i \in G*} w_{ij}^r M_{ij}^r}{\sum_{i \in G*} w_{ij}^r}$ where $M_{ij}^r = \frac{\log_2(\frac{c_{ij}}{S_j})}{\log_2(\frac{c_{ir}}{S_r})}$ and $w_{ij}^r = \frac{S_j - c_{ij}}{S_j c_{ij}} + \frac{S_r - c_{ir}}{S_r c_{ir}}$ where $G*$ is the set of features post-trimming. The weights, w_{ij}^r, are the inverse of the approximate asymptotic variances for M_i.

There are many other methods for normalization that have been considered in the literature and we refer the reader for a greater overview and in depth analysis of the normalization schemes that attempt to force samples to have equal biomass, similar specific quantiles, and overview the concept of housekeeping genes that certain genes are invariant across condition and samples [18].

1.3 Differential abundance analysis for biomarker discovery

The goal in many metagenomic studies is to test for organisms associated with a phenotype. We can formalize the problem as testing the null hypothesis that the average abundance of feature (OTU or gene) i is not different between conditions:

$$H_0: \mu_{i1} = \mu_{i2}$$

$$H_a: \mu_{i1} \neq \mu_{i2}$$

To test for features that are differentially abundant, features that are significantly different between groups, it is crucial to take into account technically confounding characteristics. A number of issues in analyzing high-throughput sequencing data include accounting for preferential sequencing, potentially confounding factors (sub-group information), variable sequencing depth, and multiple testing (on often thousands of organisms). Each of the issues can be addressed in a statistical manner by including covariates in the design, normalization of the count matrix, or performing multiple testing correction (e.g. False Discovery Rate (FDR)). Specialized methods in RNA-seq were developed specifically to account for over-dispersion (technical variability) observed in RNA-seq data. However, one of the major confounding technical artifacts of metagenomic data is sparsity due to under-sampling, which metagenomic specific methods did not account for [25–27].

1.3.1 Methods developed specifically for metagenomic data

Xipe was one of the first statistical applications to metagenomic data. The method compared two sample's count distributions in a non-parametric manner. Xipe tests if two samples are statistically significantly different from each other by sampling independently from each sample population and calculating the difference

between samples and reporting the median value on 10,000 permutations. Separately, the two samples are mixed and a null distribution of difference ranges is generated to compare the median value against. One severe limitation in the methodology included the limitation that only two samples could be compared at a time and did not remark on any particular OTU's significance.

Metastats was the first methodology designed for the analysis of groups of metagenomic samples. Metastats was the first statistical method developed specifically to manipulate the clinical metagenomic datasets of large sample sizes. Metastats tests the hypothesis of equal means between treatment groups by calculating a t-statistic for features with sufficient presence and sample sizes. In a non-parametric fashion Metastats calculates a null distribution of bootstrapped t-statistics to test for a feature's significance.

Define the number of samples in treatment 1, n_1 and the number of samples in treatment 2, n_2. The two-sample t-statistic is computed using sample means and variances from the two groups. Using the t-statistic they assess significance by permuting treatment labels for the samples and estimated a null distribution of t-statistics. These calculated t-statistics, $t_1^{0b}, \ldots, t_M^{0b}$, for B permutations generate the null distribution of the desired statistic. Finally p-values are calculated:

$$p_i = \frac{\{\# \mid t_i^{0b} \mid \geq \mid t_i \mid, b = 1, \ldots, B\}}{B}$$

For multiple hypothesis testing correction White et al. aims to control type I error by estimating the FDR defined as the expected value of the proportion of false positives within a set of predictions.

Given a hypothesis truth table:

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

The expected value of the proportion of false positive features to the number of features called significant is the FDR [28].

$$FDR = \mathrm{E}\left[\frac{F}{F+T}\right] = \mathrm{E}\left[\frac{F}{S}\right]$$

To estimate the FDR they follow the algorithm described in [28] and compute the proportion of truly null (with $\lambda = 0,0.01,0.02,\ldots,0.90$)

$$\hat{\pi}_0(\lambda) = \frac{\{\#p_i > \lambda\}}{M(1-\lambda)}$$

and fit $\hat{\pi}_0(\lambda)$ with a cubic spline, f and set $\pi_0 = f(1)$. Subsequently, for each ordered p_i they calculate a q-value given $\hat{q}(p_M) = \min(p_M \times \pi_0, 1)$

$$\hat{q}(p_i) = \min\left(\frac{\pi_0 \times M \times p_i}{i}, \hat{q}(p_{i+1})\right)$$

LefSe was another methodology for biomarker discovery. LefSe's goal is similar to Metastats, to specify the biologically relevant organisms that are differentially abundant between two or more groups. LefSe first uses non-parametric ranking statistics to determine which of the features are differentially abundant. Following significant feature detection, LefSe uses sub-group metadata to do pairwise testing in order to avoid a high false positive rate and report only biologically relevant features. These subsets of features should have abundances compatible to, as they describe, "algorithmically encoded biological hypothes[es]". For visualization, Segata et al. subsequently calculate effect sizes on significant features using linear discriminant analysis.

LefSe estimates the impact of features through the non-parametric factorial Kruskal-Wallis sum-rank test. For features that reject the hypothesis at a given nominal value, α, pairwise tests among sub-groups are performed with the Wilcoxon rank-sum test. The two tests are non-parametric methods where Kruskal-Wallis tests the mean ranks of multiple groups are the same and the Wilcoxon statistic measures the median difference between two groups is zero. LefSe determines each feature to be non-significant if at least one comparison between subtreatments has $p > \alpha$ or is significant in the opposite direction. The assumptions for these tests are that the shape of the count distributions to be identical.

1.3.2 Methods developed specifically for RNA-seq data

DESeq (Differential Expression Analysis for Sequence Count Data) and edgeR [23, 29] are two of the most popular RNA-seq analysis methods. They propose that RNA-seq is count data with additional technical variability – over-dispersion. The two methods decompose the variance of a feature as a combination of technical variability and biological variability. Technical variability in the expression or abundance measurement due to experimental errors or other factors not measured in the experiment. Biological variability contributes to differential expression between genes. In their models, the amount of biological variability is directly proportional to the amount of over-dispersion. In both methods, the assumption is that a negative binomial distribution is more appropriate than a Poisson distribution in modeling the data and accounting for this over-dispersion.

DESeq and edgeR assume that the normalized gene expression values $y_{ij}, i = 1, \dots, m, j = 1, \dots, n$ follows a negative binomial distribution: $y_{ij} \sim$

$NB(\mu_{ij}, \sigma_{ij}^2)$ where $y_{ij} = 0,1, \ldots$ μ_{ij} is the expected count for gene i in sample j (averaged normalized count value) and σ_{ij}^2 is the variance of the count values of gene g and is modeled as the sum of two components: $\sigma_{ij}^2 = S_j\ \mu_{ij} + \alpha_j S_j^2 \mu_{ij}^2$ where S_j is the sequencing depth of sample j and α is the dispersion parameter for gene j. The first component $S_j\ \mu_{ij}$ is termed the technical variability, and the second component is the biological variability of interest. To model the over-dispersion parameter DESeq and edgeR describe the relationship between over-dispersion and mean in a generalized linear model.

A number of other methods have been developed for the analysis of microarray and RNA-seq data. We refer the reader to the papers on Myrna, limma, and voom [30–32] for details on normality assumed methods. However, none of the methods developed for RNA-seq data account for sparsity induced by under-sampling, the largest confounder of metagenomic sequencing data. In the next section we highlight our approach to account for sparsity due to under-sampling in better estimating robust fold-changes in detecting differentially abundant species and genes.

2 Differential abundance analysis for marker-gene surveys

2.1 Background

2.1.1 Context

The majority of the chapter work was published in Nature Methods [33]: Paulson, JN., Stine, OC., Bravo, HC., Pop, M. (2013). Differential abundance analysis for marker-gene surveys. Nature Methods, 10, pg. 1200-1202. Questions from researchers in the field lead to further greater characterization of the effect under-sampling has on resulting variability following log transformation and total sum normalization with alternative pseudo-counts for which we published [34]: Paulson, JN., Bravo, HC., Pop, M. (2014). Reply to: "a fair comparison". Nature Methods, Volume 11, pg. 359-360.

Prior to the methods and analysis techniques developed in the following chapter, the state of the art was to make use of either LefSe or Metastats. The method proposed below goes beyond either one in introducing a novel normalization technique and a method to account for a ubiquitous issue in marker-gene microbial community abundance estimates, under-sampling. We compared our novel methodology to LefSe, Metastats and a number of other RNA-seq methods developed.

That said, there are still questions currently unaddressed by these methods. Like many high-throughput sequencing assays, profiling the microbial community is complicated by sequencing errors. Even small sequencing errors results in clustering sequences from the same taxonomic unit into different OTUs. Future potential work would account for the highly-correlated abundance profiles across samples of distinct

OTUs corresponding to the same taxonomic entity. Not accounting for this correlation will produce biased larger variance estimates degrading performance.

At the time of this writing, metagenomeSeq is a software package available through Bioconductor that is still in active development and support. The software is available at http://tinyurl.com/metagenomeSeq and has had over 3300 distinct downloads in the last year.

2.1.2 Abstract

We introduce a novel methodology for differential abundance analysis of sparse high-throughput data from large-scale surveys of marker genes for microbial communities. Our approach relies on cumulative sum scaling (CSS) normalization - a novel count data normalization technique - and the zero-inflated Gaussian (ZIG) model as a statistical method for detecting differential abundance of taxonomic features. We show that our CSS normalization technique significantly improves sample clustering by phenotypic similarity compared to existing normalization methods for count data. We also show that our ZIG method significantly improves the robustness of differential abundance inferences compared to existing methods for count data by accounting for bias introduced by the under-sampling of microbial communities commonly found in large-scale marker gene studies. We have implemented these methods in the publicly available Bioconductor package metagenomeSeq (http://cbcb.umd.edu/software/metagenomeSeq).

We demonstrate our methods by applying metagenomeSeq to simulated data, as well as several published microbiota datasets including cross-sectional oral samples, and a time-series study on the effect of diet in gnotobiotic mice. We also

provide a thorough comparison of our new method to existing approaches for differential abundance analysis used in microbiota and/or sequence-based transcriptome analysis.

2.2 Overview of recent health related applications

Studies of marker genes to survey communities are recently being applied to clinical settings with the intent of understanding the structure and function of healthy microbial communities and the association of the microbiota with disease including Crohn's disease [35], bacterial vaginosis [4], diabetes [5, 6], eczema [10], obesity [36] and periodontal disease [3]. In this setting, the identification of potentially pathogenic or probiotic bacteria revealing significant differences in their abundance in a disease population is critical. While methods for whole-scale community comparisons are commonly used [11, 12] there is a need for tools that discern taxon-specific disease associations in marker-gene surveys.

2.2.1 Data preparation for marker-gene targeted sequencing

A common approach in marker-gene surveys is the targeted sequencing of the 16S ribosomal RNA gene from selected samples. ‘Universal’ primers amplify specific hyper-variable regions within the 16S rRNA gene, and the corresponding segments are sequenced using high-throughput long-read technology (often from the 454 platform). Sequence reads are first clustered into operational taxonomic units (OTUs) [13] and representative sequences from each cluster are then annotated against a database of 16S rDNA reference sequences including Ribosomal Database Project (RDP) [14].

Data preprocessing and differential abundance analysis has a long history in high-throughput experiments measuring gene expression with microarray technology [20–22] and sequencing-based assays such as RNA-seq [17, 23, 24]. However, there are key differences between RNA-seq data and marker-gene data. In marker-gene surveys, clustering and annotation of read sequences is neither fixed nor known *a priori* over which differential abundance is tested. Furthermore, most taxonomic features in marker-gene studies are rare (absent from a large number of samples) in contrast to RNA-seq studies where a much more complete representation of features is encountered. These significant differences require the development of specific methods for determining differential abundance in marker-gene surveys. However, existing methods for differential abundance analysis of marker-gene data do not address issues stemming from the preprocessing of count measurements inherent to these assays [25–27].

2.2.2 Motivation for normalization

Variation in depth of coverage - the total number of sequences produced for each sample yields incomparable read count measurements. Data normalization is an initial step in most differential abundance analysis aimed at making feature counts comparable across samples. By far, the most commonly used approach normalizes the data by dividing feature read counts by the total number of reads in each sample, i.e., converts feature counts to appropriately scaled ratios. This normalization procedure which we refer to as total-sum normalization has been shown to incorrectly bias differential abundance estimates in RNA-seq data derived through high-throughput technologies [17, 18] and potentially biases marker gene survey data. Furthermore,

statistics derived by this normalization may induce spurious correlations between features, a fact known since 1896 in the work of Pearson [19]. Our first contribution is a novel data normalization technique that corrects bias in the assessment of differential abundance introduced by total-sum normalization.

2.2.3 Motivation for differential abundance testing

The number of taxonomic features discovered in a specific sample is strongly related to the sample's sequencing depth. Specifically, samples with low depth may have counts equal to zero in a high proportion of taxonomic features due strictly to under sampling of the microbial community. Under-sampling affects estimates of organism presence and abundance in sample groups and also analyses of differential abundance as a result. Our second contribution is a zero-inflated Gaussian distribution mixture-model that accounts for biases resulting from under-sampling the microbial community.

2.2.4 Summary of results

We show that these methods significantly improve the accuracy of detecting differential abundance using simulated and published data. We also show improved performance in analyzing two datasets: 1) mouse intestinal microbiota differences between diets, and 2) subgingival plaque samples and tongue samples from the same individual. These new methods are implemented in the Bioconductor package metagenomeSeq (http://cbcb.umd.edu/software/metagenomeSeq).

2.3 Results

2.3.1 Data preprocessing

2.3.1.1 Cumulative sum scaling

We introduce cumulative sum scaling (CSS) normalization a novel normalization technique that controls for biases in measurements across taxonomic features. A recent proposal for normalization of RNA-seq data is to scale counts by the 75th percentile of each sample's non-zero count distribution [17]. That scaling method was motivated by the observation that a few measurements (e.g., taxa or genes) are sampled preferentially as sequencing yield increases, and have an undue influence on normalized counts derived by the usual normalization procedure. In that case, the 75th percentile was chosen as it behaved consistently across samples. We observed a similar trend in many 16S rDNA datasets, but there was variability in which percentile captures consistency across samples in each dataset (Figure **1**). Therefore, we developed the cumulative sum scaling normalization method as an extension that is better suited for marker gene survey data. Our method scales counts by the cumulative sum of counts up to a percentile determined using a data-driven method.

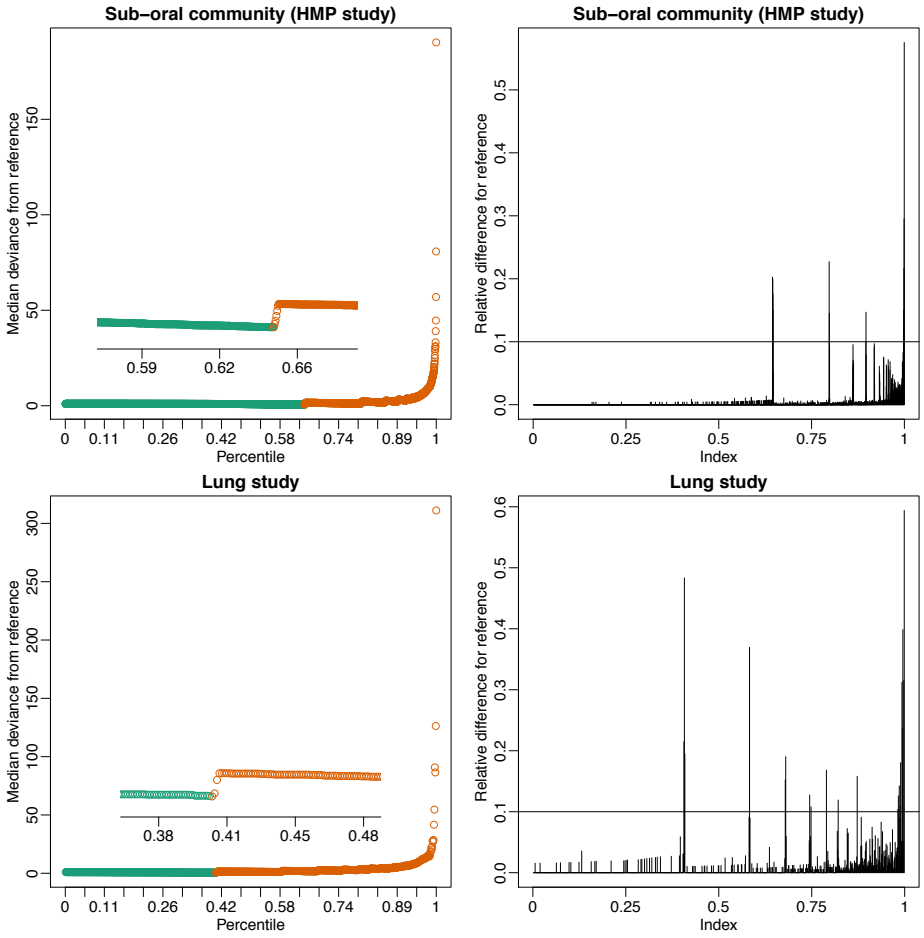

Figure 1: Data-driven adaptive method for selecting normalization scale quantile

Left, we plot d_l (see 2.3.1.1 Cumulative sum scaling) for our oral sub-community and lung microbiome datasets. In the oral sub-community (HMP) dataset we observe sample count distributions differ greatly from the reference at the 65th percentile where relative difference of the median deviation is greater than 10%. For the lung microbiome dataset, this occurs at the 41st percentile. Right, relative difference of the median deviation of sorted sample counts from reference for the oral sub-community and lung microbiome datasets. We observe that sample counts follow similar distributions up to the 65th and 41st percentile respectively. The reference is calculated as the row means of sorted counts.

Assume raw data is given as count matrix $M(m, n)$ where m and n are the number of features and samples, respectively. The raw data in this matrix is represented by counts c_{ij} representing the number of times taxonomic feature i was observed in sample j. Denote the sum of counts for sample i as $s_j = \sum_i c_{ij}$. The usual

normalization procedure for marker gene survey data corresponds to producing normalized counts $\tilde{c}_{ij} = \frac{c_{ij}}{s_j}$. We refer to this procedure as total-sum normalization.

We introduce a new normalization method, cumulative sum scaling normalization (CSS), to remove biases in the count data. The biases come from features that are preferentially amplified in a sample-specific manner. Denote the l^{th} quantile of sample j as q_j^l, that is, in sample j there are l taxonomic features with counts smaller than q_j^l. For $l = \lfloor 0.95 * m \rfloor$, q_j^l corresponds to the 95$^{\text{th}}$ percentile of the count distribution for sample j.

Also denote $s_j^l = \sum_{i|c_{ij} \leq q_j^l} c_{ij}$ as the sum of counts for sample j up to the l^{th} quantile. Using this notation, the total sum $s_j = s_j^m$. Our normalization chooses a value $\hat{l} \leq m$ to define a normalization scaling factor for each sample to produce normalized counts $\tilde{c}_{ij} = \frac{c_{ij}}{s_j^{\hat{l}}} N$ where N is an appropriately chosen normalization constant. We scale all samples using the same constant N so normalized counts have interpretable units. We recommend using the median scaling factor $s_j^{\hat{l}}$ across samples. Counts for samples with scaling factor close to N can be interpreted as reference samples, and counts for other samples are interpreted relative to the reference. In our datasets the median $s_j^{\hat{l}}$ was close to 1000 and thus used this value in our analysis. Note that ratios are also used in this procedure, assuming there is a finite capacity to the size of microbial communities. This is the same assumption that underlies total-sum normalization. However, our method seeks to avoid placing undue influence on features that are preferentially sampled. The relative proportion of the features is

unaffected by the normalization as $s_j = \sum_i c_{ij}$ and $\tilde{s}_j = \frac{\sum_i c_{ij}}{s_j^{\hat{l}}}$, this implies $p_i = \frac{c_{ij}}{s_j} = \frac{s_j^{\hat{l}} * c_{ij}}{s_j^{\hat{l}} * \sum_i c_{ij}} = \frac{\tilde{c}_{ij}}{\tilde{s}_j} = \tilde{p}_i$.

The choice of the appropriate quantile given by $\hat{l}$ above is critical for ensuring that the normalization approach does not introduce normalization-related artifacts in the data. At a high level, the count distribution of samples should all be roughly equivalent and independent of each other up to this quantile under the assumption that, at this range, counts are derived from a common distribution. The specific value for the chosen quantile is project-specific and likely depends on the complete experimental details (including all the sample preparation, sequencing, and subsequent bioinformatics analysis).

We use an adaptive, data-driven, method to determine $\hat{l}$ based on the observation above. We find a value $\hat{l}$ where sample-specific count distributions deviate from an appropriately defined reference distribution. Specifically, denote $\overline{q^l} = med_j\{q_j^l\}$, the median l^{th} quantile across samples, as the l^{th} quantile of the reference distribution. Note that this is exactly the way a reference distribution is defined in the commonly used quantile normalization approach [17]. Denote as $d_l = med_j \left| q_j^l - \overline{q^l} \right|$. This is the median absolute deviation of sample-specific quantiles around the reference. Under the methods assumptions, this quantity d_l is stable for low quantiles and shows high instability in high quantiles. Our method defines $\hat{l}$ as the smallest value where high instability is detected (Figure **1**). We measure instability in this case by using relative first differences. Specifically, we set

$\hat{l}$ to the smallest l that satisfies $d^{l+1} - d^{l} \geq 0.1\, d^{l}$. The value 0.1 is set arbitrarily and may be substituted by another value to determine high instability.

We also found that CSS-normalized sample abundance measurements are well approximated by a log-normal distribution in studies with large number of samples (Figure 2) and therefore also applied a logarithmic transform to the normalized count data. This transformation controls the variability of taxonomic feature measurements across samples (Figure 2).

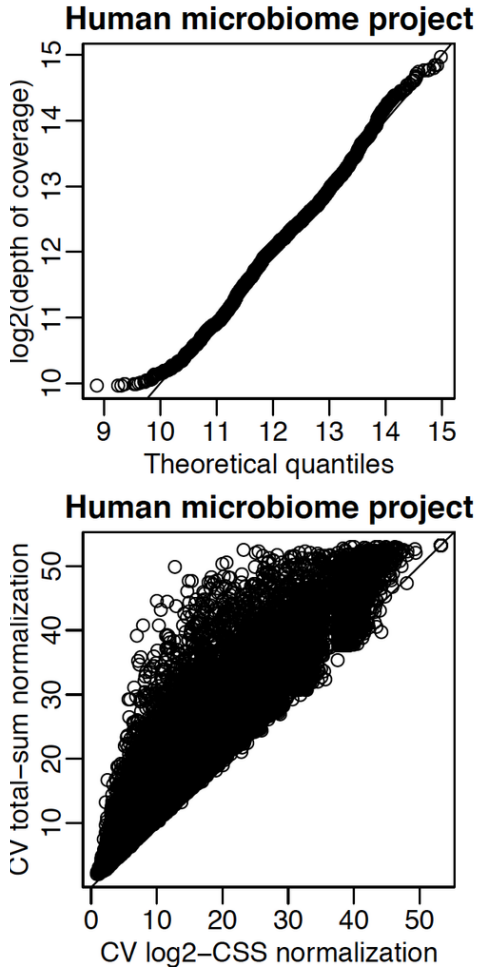

Figure 2: Depth of coverage follows a log-normal distribution and cumulative sum scaling normalization controls dispersion.

Top, quantile-quantile plot of sample sequencing depth and a log-normal distribution. We see that sample depth of coverage closely follows a log-normal distribution.

Bottom, coefficient of variation for total-sum normalized counts versus CSS normalized counts. We observe that dispersion is always greater in total-sum normalized counts.

To illustrate the effects of our data preprocessing method, we use a longitudinal dataset that tracked the gut microbial community of twelve gnotobiotic mice. All of the mice were put on a low-fat, plant-polysaccharide-rich (LF/PP) diet for four weeks and then half of the mice were shifted to a high-fat/high-sugar Western diet. The remaining mice were kept on the original diet for the same time periods. To assess the effect of normalization on distinguishing samples by phenotypic similarity we performed a multi-dimensional scaling analysis of data normalized using our method, DESeq [23] size factors, TMM [24] and total-sum normalization (Figure 3A-D). CSS normalization was able to best separate samples based on diet while controlling within-group variance. We quantified this observation using linear discriminant analysis and observed that CSS normalization performed the best in distinguishing samples by phenotypic similarity (Figure 3E). We observed similar results when comparing CSS normalization to other frequently used normalization methods (Figure 4).

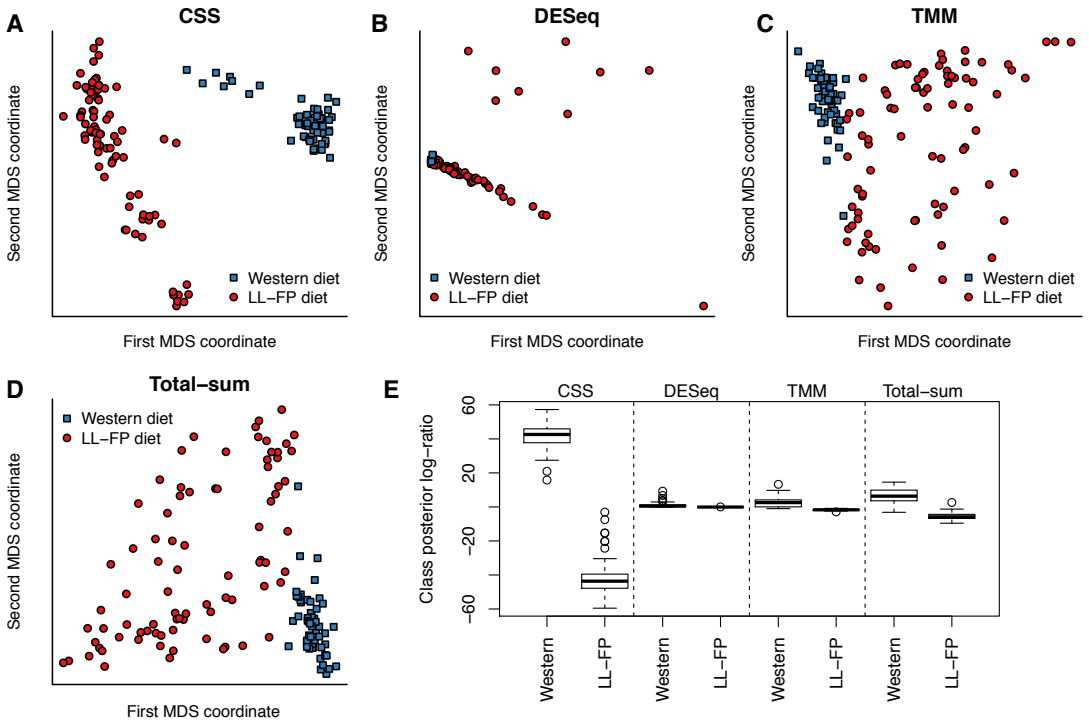

Figure 3: Clustering analysis is improved substantially by CSS normalization.

We plot the first two principal coordinates in a multi-dimensional scaling analysis of mouse stool data normalized by (A) CSS, (B) DESeq size factors, (C) trimmed mean of M-values, and (D) total-sum. Colors indicate clinical phenotype (diet). CSS normalization data successfully separates samples by diet while controlling within-group variability. (E) Class posterior probability log-ratio for Western diet obtained from linear discriminant analysis (LDA). Each box corresponds to the distribution of leave-one-out posterior probability of assignment to the "Western" cluster across normalization methods (whiskers indicate 1.5 times inter-quartile range). Samples were best distinguished by phenotypic similarity using CSS normalization.

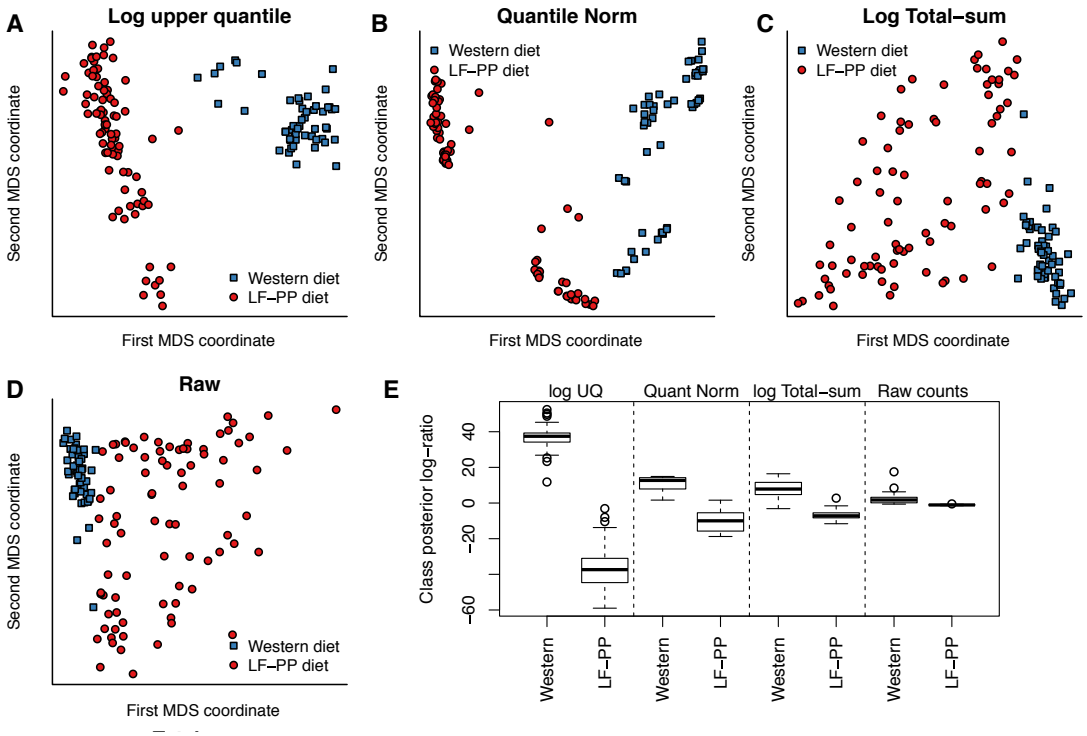

Figure 4: Effect of normalization on clustering analysis.

We plot the first two principal coordinates in a multi-dimensional scaling analysis for count data normalized by (A) logged upper quantile, (B) quantile normalization, (C) logged total-sum scaling, and (D) the raw counts. Colors represent the sample phenotype (diet). (E) Class posterior probability log-ratio for Western diet obtained from linear discriminant analysis (LDA). Each box in the plot corresponds to the distribution of leave-one-out posterior probability of assignment to the "Western" cluster for samples of each type across normalization methods. Clustering analysis is improved significantly by CSS normalization and Logged upper quantile scaling.

2.3.1.2 The devil is in the details

The purpose of the analysis reported in Figure 3 of was to compare the normalization procedure proposed there (CSS + log-transform) to established practices in the field. The prevalent practice in metagenomics is to use total-sum scaling, without transformation. In our analyses, a continuity correction of 1 is used when applying a log transform to preserve the sparsity of metagenomic data since this is an important intrinsic characteristic of the data, as opposed to a mere necessity for

modeling. While conceptual simplicity is a good characteristic of a data analysis tool, its use without further validation and testing is unwarranted. Advocating a log-transform with a carefully chosen pseudo-count is not conceptually simple. To complete an analysis of the effect of pseudo-counts with log transformation, we show the effect of multiple pseudo-counts on clustering analysis [34].

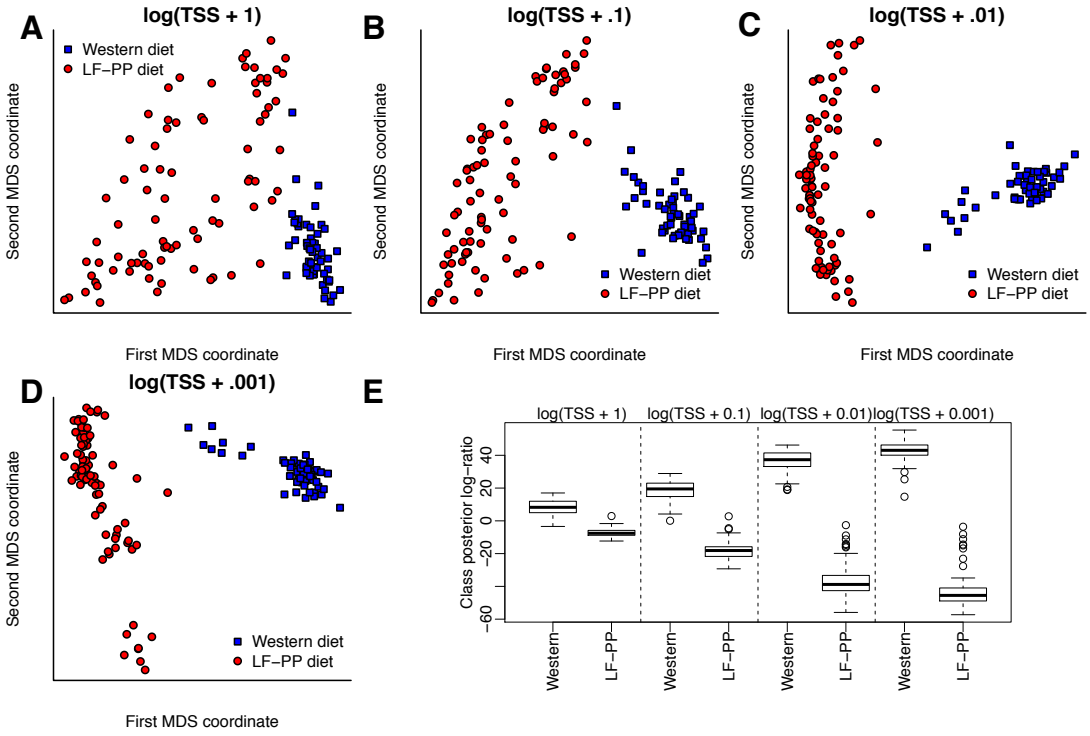

Figure 5: Effect of pseudo-count choice on total sum normalization.

Panels A-D correspond to exponentially decreasing pseudo-counts from 1 to 10^{-3}. Panel E provides the corresponding class posterior log-ratios for each pseudo-count value.

2.3.1.3 Cumulative sum distribution

In 2002, quantile normalization for micro-array data was shown to be the ideal method for normalization [21]. The technique is a method meant to make two distributions identical in statistical properties and remove the variation of non-biological origin. The motivation is coupled by the fact that certain measurements are

sampled preferentially.

Similarly to quantile normalization, the assumption follows that the rate of sampling a particular measurement is similar for those with similar proportions of identified taxa. We too show a technique for making two distributions identical in statistical properties with the additional metagenomic assumption that there is a finite capacity in a metagenomic community. As such, the cumulative summation of a samples' 16s or metagenomic count should follow a similar rate to that of other samples with similar proportions of zeros at an OTU level.

Our algorithm follows (wording similar to [21]):

Cumulative distribution normalization

Step-1: Bin samples into k groups of similar sparsity levels
Step-2: Given n_i samples $\in k_i$ all of length m, form matrix X_i of dim. m x n_i

For each X_i
Step-3a: Sort each column of X_i producing $X_{i,sorted}$
Step-3b: Replace each column of $X_{i,sorted}$ with the cumulative sum of column
Step-3c: Build $X_{i,sorted}$ reference with the row means
Step-3d: Assign i^{th} values within each sample to the reference and take the inverse cumulative sum
Step-3e: Unsort modified $X_{i,sorted}$

Step-4: Scale X_i to the median reference (optional).

2.3.1.4 Assessment of normalization methods

To assess the effect of normalization on distinguishing samples by phenotype we performed a multi-dimensional scaling analysis of count data normalized by using CSS, total-sum scaling, logged total-sum scaling, geometric mean, trimmed mean by M values, quantile scaling, and quantile normalization. We excluded cumulative distribution normalization due to a higher false discovery rate for pairwise correlations between normalized features (not shown).

We calculated the 1000 taxonomic features with largest variance after each normalization method and used subsequent counts in the MDS analysis. We also used linear discriminant analysis (LDA) to distinguish samples by diet. We calculated the log-ratio of class posterior probabilities for each sample x using leave-one-out cross-validation:

$$log \frac{f_w(x)\pi_w}{f_l(x)(1-\pi_w)}$$

where π_w is the proportion of samples on the "Western" diet, and f_w and f_l are normal densities for each of the diets, with a common variance. Parameters in each leave-one-out fold are estimated from the remaining samples. The class posterior probability should be large and positive for "Western" samples and small and negative for samples in the other group. We measure the performance of each normalization method by the difference in the distribution of the class posterior probabilities (Figure 3E, Figure 4E and Figure **5**E).

2.3.2 Differential abundance analysis

2.3.2.1 Under-sampling implications in marker-gene surveys

We found a strong correlation between the number of detected OTUs and sample sequencing depth in 16S rDNA studies using high-throughput sequencing methods (R^2=0.92-0.97, Figure **6**). This suggests that measurements of differential abundance suffer from biases resulting from the possible misinterpretation of zero counts in samples with low coverage as taxonomic features not present in the microbial community as opposed to interpreting their absence as a result of under-sampling the microbial community. The degree of sparsity observed in marker-gene

experiments is much higher than usually seen in other abundance assays such as transcriptome profiling from single genomes (Figure **7**). An analysis of a collection of RNA-seq datasets [37] reveals that the proportion of features found in all samples (15-85%) is much higher than that proportion in the 16S rDNA studies we surveyed (1-3%).

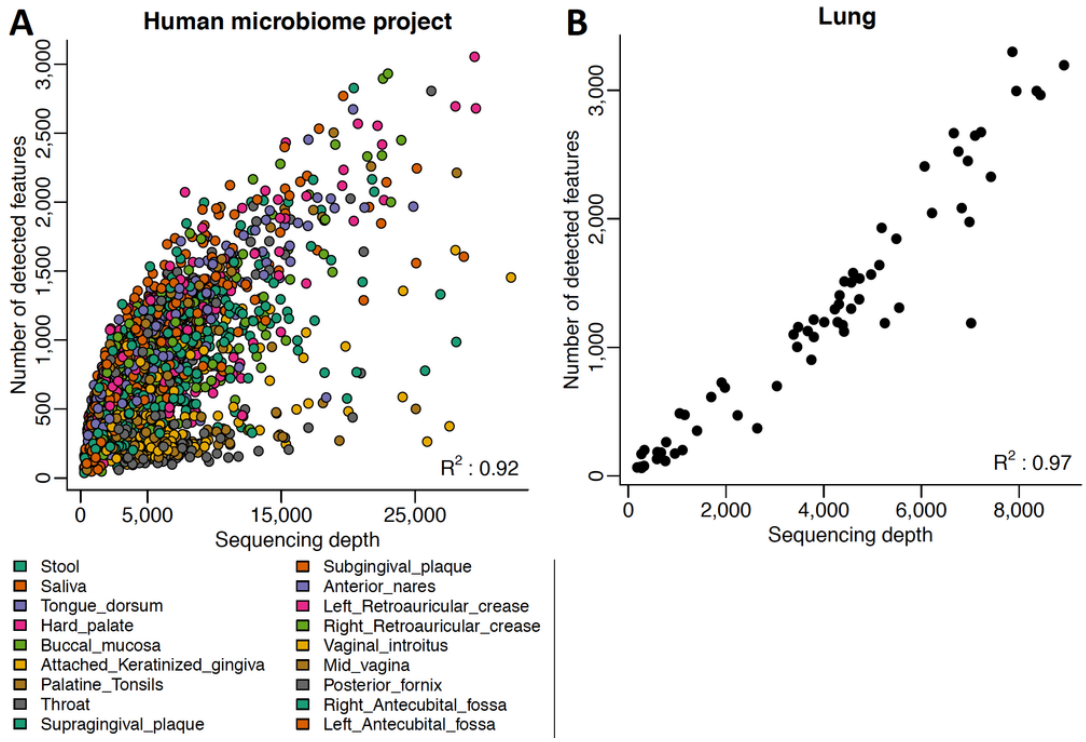

Figure 6: The number of OTUs detected in a sample depends on sequencing depth and phenotypic characteristics.

We plot the number of detected OTUs in a sample as a function of sequencing depth for the Human Microbiome Project [38] (A) and the lung microbiota study [39] (B). There is a strong dependency between sequencing depth and number of detected OTUs. Neither dataset shows that the number of OTUs stabilizes for samples with high depth, indicating that in both cases sequencing depth may not be sufficient for comprehensive profiling of the microbial community. We found that a significant proportion of variability in the number of OTUs detected is explained by sequencing depth. We found that including clinical covariates, e.g., body site sampled, we obtained significantly improved fits with higher adjusted R^2. The same dependency is observed in all studies analyzed for this paper.

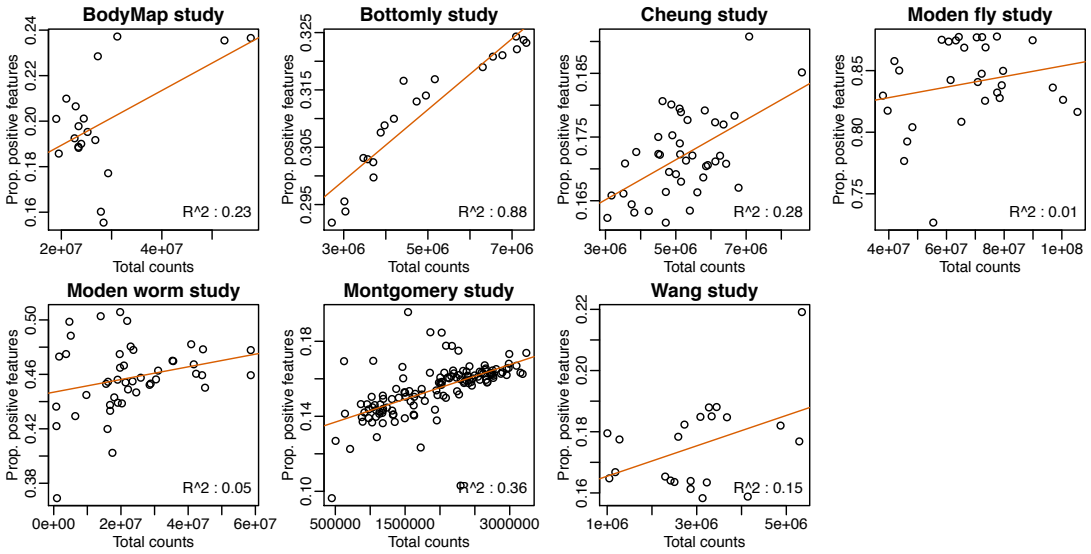

Figure 7: Effect of sequencing depth on the number of genes detected in RNA-seq.

We plot the number of genes detected in a RNA-seq sample as a function of sequencing depth. In the lower corner of each plot is the adjusted R^2 value representing how much of the variation in each sample's detected genes is described by the depth of coverage. The proportion of genes detected in any particular sample is much higher in RNA-seq datasets (15-85%) compared to marker gene survey data samples (1-3%). Depths of coverage are also much larger in RNA-seq. Datasets obtained from ReCount: http://bowtie-bio.sourceforge.net/recount/

To explicitly account for under-sampling, we developed a novel statistical model of differential abundance as a mixture-model that implements a zero-inflated Gaussian distribution of mean group abundance of each taxonomic feature (see below and Figure **8**). The components of the mixture model correspond to normally distributed log-abundances in each group of interest, e.g., case or control (represented as the count distribution in Figure **8**) and a spike-mass at zero indicating absence of the feature due to under-sampling (represented as the detection distribution in Figure **8**). Our model seeks to directly estimate the probability that an observed zero is generated from the detection distribution due to under-sampling or from the count distribution (absence of the taxonomic feature in the microbial community). We estimate the expected value of latent component indicators based on sample

sequencing depth of coverage using an expectation maximization algorithm (see 2.3.2.3 Expectation Maximization Algorithm).

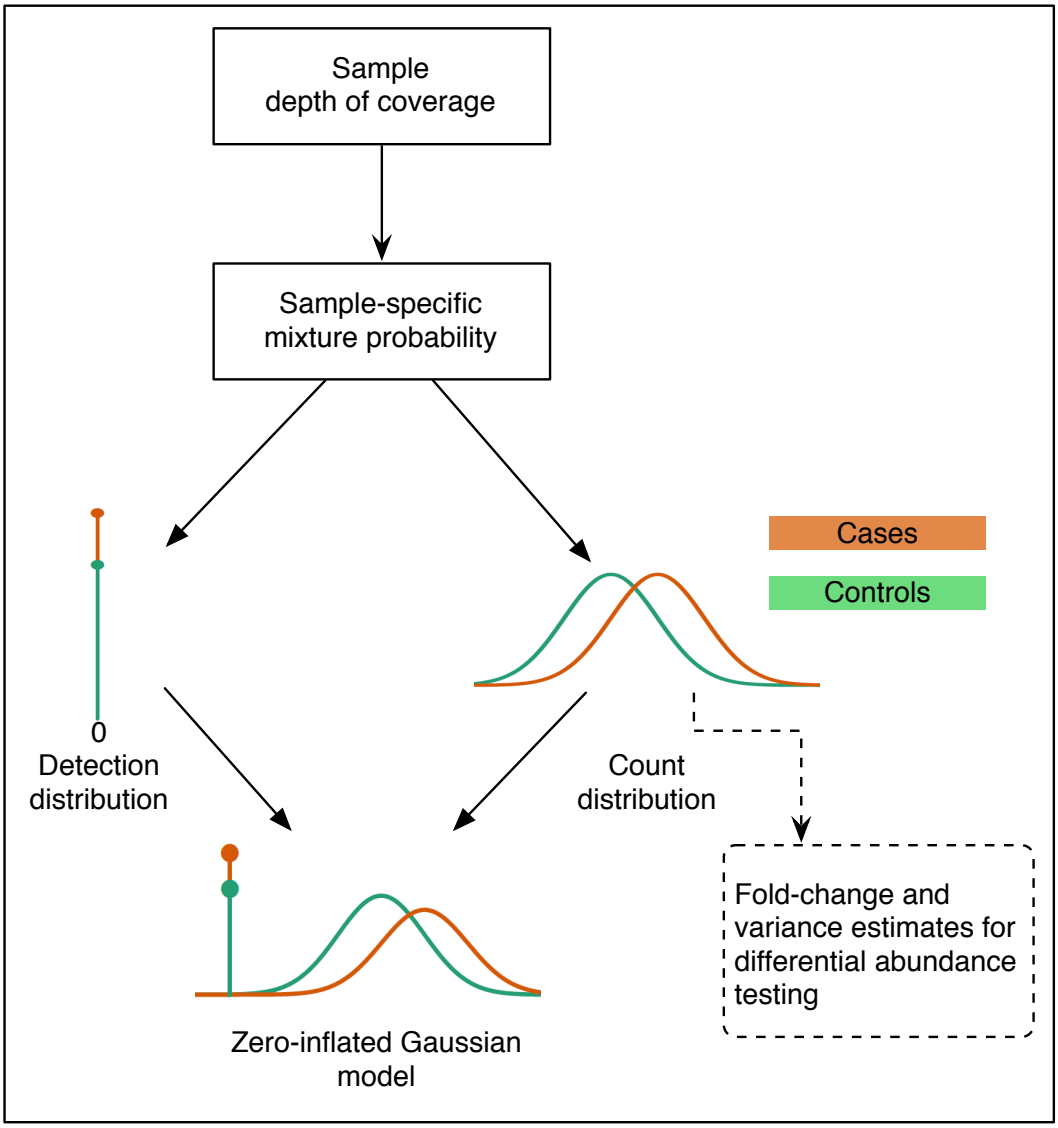

Figure 8: A graphical representation of the zero-inflated Gaussian mixture model.

Counts are modeled with a mixture distribution with components corresponding to normally-distributed log-abundances in each group, e.g. cases (orange) and controls (green), and a spike-mass at zero corresponding to a detection distribution that depends on each sample's sequencing depth. We use fold-change and variance estimates derived from the count component of the mixture to test for differential abundance between groups.

Given that zero counts are modeled as generated from either a spike mass at

zero, $f_0(y)$, or a Gaussian distribution, $f_{count}(y_{ij};\ \mu_i, \sigma_i^2) \sim N(\mu_i, \sigma_i^2)$ our final model for log-counts is:

$$f_{zig}(y_{ij}|\theta) = \pi_j \times f_0(y_{ij}) + (1 - \pi_j) \times f_{\text{count}}(y_{ij}; \mu_i, \sigma_i^2)$$

We use a linear mean model in the count component of the mixture (f_{count} f_{count}) following standard conventions in methods for testing differential abundance in gene expression [31]. Posterior probabilities of zero counts being generated from f_{count} are used as weights when estimating parameters of the linear model in f_{count} from which differential abundance is determined. Other zero-inflated models have been developed with mixtures of Poisson and Binomial distributions and applied to ecological data [40].

We illustrate the effect of the ZIG model on differential abundance analysis in Figure **9** where we plot data for one OTU from the Human Microbiome Project [38]. Using posterior probability estimates that account for community under-sampling as weights to estimate count distribution parameters reduces the estimated fold-change between the two groups under study. Furthermore, counts after accounting for under-sampling are better fit by a log-normal distribution (Shapiro-Wilks test P=0.78) than normalized counts (Shapiro-Wilks test P=0.08).

2.3.2.2 Zero-inflated Gaussian mixture model

Our zero-inflated Gaussian (ZIG) mixture model is motivated by the observed relationship between depth of coverage and the number of OTUs detected (Figure **6**). Below we provide full details for our method.

Count data is modeled from two populations, each with n_A and n_B samples and with m features (OTUs). The raw count for sample j and feature i is denoted by

c_{ij}. The class indicator function is defined as $k(j) = I\{j \in \text{group } A\}$.

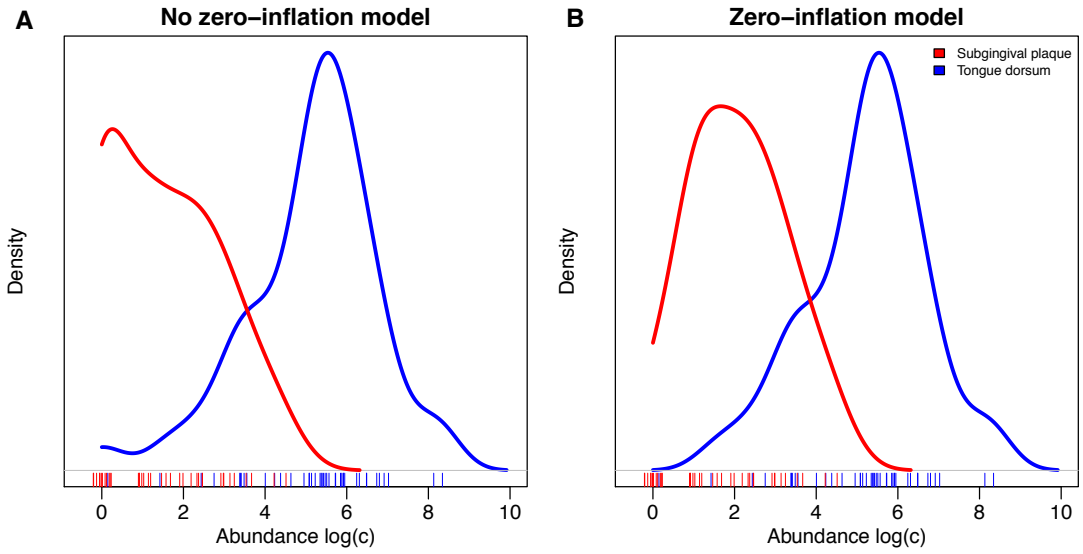

Figure 9: Illustration of the effect of zero-inflated Gaussian mixture model on differential abundance.

(A) Kernel density estimates of log-counts for an *Granulicatella para-adiacens* OTU in samples from the Human Microbiome Project. We see that subgingival plaque samples have a large number of zero counts that result in large differential abundance estimate when a model without zero-inflation is used. (B) Weighted kernel density estimates for the same samples. Weights are obtained from the posterior probabilities for zero counts due to under-sampling of the microbial community. After accounting for zero-inflation, the differential abundance estimate is moderated for this feature. This plot also suggests that the log-normal distributional assumption used in this paper is supported.

The zero-inflated model is defined for the continuity-corrected $\log_2$ of the raw count data:

$$y_{ij} = \log_2(c_{ij} + 1)$$

as a mixture of a point mass at zero $I_{\{0\}}(y)$ and a count distribution $f_{count}(y; \mu, \sigma^2) \sim N(\mu, \sigma^2)$. Given mixture parameters π_j, we have that the density of the zero-inflated Gaussian distribution for feature i, in sample j with s_j total counts is:

$$f_{zig}(y_{ij}; s_j, \beta, \mu_i, \sigma_i^2) = \pi_j(s_j) \cdot I_{\{0\}}(y_{ij}) + (1 - \pi_j(s_j)) \cdot f_{count}(y_{ij}; \mu_i, \sigma_i^2)$$

We specify the mean model as:

$$E(y_{ij} \mid k(j)) = \pi_j \cdot 0 + (1 - \pi_j) \cdot \left(b_{i0} + \eta_i \log_2 (\frac{s_j^{\hat{l}} + 1}{N}) + b_{i1} k(j) \right).$$

In this case, parameter b_{i1} is an estimate of fold-change in mean normalized counts between the two populations. The term including the logged normalization factor $\log_2 \left(\frac{s_j^{\hat{l}}}{N} \right)$ captures OTU-specific normalization factors through parameter η_i. This can capture feature specific biases, for instance in PCR amplification efficiency. The model can also be specified without OTU-specific normalization, in which case the term including the normalization factor is treated as an offset in the linear model. This is equivalent to defining a model on logged normalized count data without including the normalization offset term in the linear model.

For large marker gene survey studies in clinical and epidemiological settings, it is essential to include possible sources of confounding error when testing the association between the abundance of taxonomic features and a clinical phenotype of interest (disease, for instance). Our linear model methodology can easily incorporate these confounding covariates in a straightforward manner. Other zero-inflated models have been developed mixing the Poisson and Binomial distributions. These models have had applications to ecological count data. Based on the observation that the number of zero-valued features of a sample depends on its total number of counts, we model the mixture parameters $\pi_j(s_j)$ as a binomial process:

$$\log \frac{\pi_j}{1 - \pi_j} = \beta_0 + \beta_1 \cdot \log(s_j)$$

2.3.2.3 Expectation Maximization Algorithm

Denote the full set of estimates as $\theta_{ij} = \{\beta_0, \beta_1, b_{i0}, \eta_i, b_{i1}\}$. Maximum-likelihood estimates are approximated using the EM algorithm where we treat mixture membership $\Delta_{ij} = 1$ if y_{ij} is generated from the zero point mass as latent indicator variables. The log-likelihood in this extended model is then:

$$l(\theta_{ij}; y_{ij}, s_j) = (1 - \Delta_{ij}) \log f_{count}(y; \mu_i, \sigma_i^2) + \Delta_{ij} \log \pi_j(s_j) + (1 - \Delta_{ij}) \log\{1 - \pi_j(s_j)\}$$

Estimates responsibilities, $z_{ij} = Pr(\Delta_{ij} = 1)$, given current estimates $\hat{\theta}_{ij}$ as

$$\hat{z}_{ij} = \frac{\hat{\pi}_j \cdot I_{\{0\}}(y_{ij})}{\hat{\pi}_j \cdot I_{\{0\}}(y_{ij}) + (1 - \hat{\pi}_j) f_{count}(y_{ij}; \hat{\theta}_{ij})}$$

Notice $\hat{z}_{ij} = 0 \; \forall \; y_{ij} > 0$.

Estimate parameters θ_{ij} given current estimates $\hat{z}_{ij}$:

To compute b, we use weighted least squares, with weights $1 - \hat{z}_{ij}$. Note that only samples with $y_{ij} = 0$ potentially have weights < 1. Estimates of standard error are also obtained using $1 - \hat{z}_{ij}$ as weights. The mixture parameter is estimated as $\pi_j = \sum_{i=1}^{G} \hat{z}_{ij} / G$, from which we estimate β, using least squares on the logit model as: $\log \frac{\hat{\pi}_j}{1 - \hat{\pi}_j} = \beta_0 + \beta_1 \log(s_j)$.

From the estimated fold-change (b_{1i}) and its standard error, we construct a moderated t-statistic by Empirical Bayes [31] and use a parametric t-distribution to obtain p-values for the test $b_{1i} = 0$. Notice that this only incorporates the count component of the zero-inflated mixture model. We interpret this test as the expected difference in abundance between groups conditioned on feature detection.

The moderated t-statistic is defined as $t_i = \frac{b_{1i}}{(\tilde{s}_i^2 / \sum_j (1-z_{ij}))^{1/2}}$, where $\tilde{s}_i^2$ is obtained by pooling all features' variances as described in [31], $\tilde{s}_i^2 = \frac{d_0 s_0^2 + d_i s_i^2}{d_0 + d_i}$ where s_i^2 and d_i are respectively the observed feature variance and degrees of freedom and d_0 and s_0^2 are estimated using the method of moments incorporating all feature variances and degrees of freedom. We found that by using a log-normal distribution, the moderated t-test was appropriate. As in the previous Metastats version, we use the q-value method to correct for multiple testing.

We chose to use a log-normal distribution in the count component of the mixture instead of a generalized linear model, negative binomial [23, 24], for both computational and statistical reasons. On the computational side, we would need to estimate a weighted generalized linear model using an iterative method at each maximization step. That is too computationally intense and numerically unstable. On the statistical side, we find that the log-normal distribution is appropriate since the type of marker gene survey study we are targeting tend to have moderate to large sample sizes (Figure 2). This is consistent with recent observations in the literature [32].

2.3.2.4 Ambiguous read assignment to OTUs

Ambiguous read assignment is an important consideration in testing differential abundance in count data, and in particular RNA-seq [41–45]. There are two sources for ambiguous read assignment in RNA-seq data that may apply to marker gene survey data. In isoform ambiguity, where a sequence read could be generated from sequencing one of multiple possible isoforms for a gene. In this case gene-level abundance measurements are convolutions of isoform-level abundances

which may bias differential abundance inferences in the presence of differential abundance at the isoform level. In marker gene survey data, reads are clustered based on sequence similarity and the resulting clusters define features over which differential abundance testing is performed. The type of convolution occurring in RNA-seq data would occur when an OTU defined by clustering contains two distinct functional OTUs. We have chosen a sequence similarity threshold (99%) that was previously shown to be more stringent than similarity at species-level [46] thus reducing the possibility of convolution to occur. We therefore believe that this type of ambiguity does not arise frequently in this setting. On the other hand, less stringent sequence similarity thresholds, which would increase the frequency of convolution, still exhibit high sparsity as previously reported. We believe sparsity does indeed drive the improvement in results we see as methods using non-zero inflated negative binomial models, including Cuffdiff2, are not suitable.

However, there is a second source of ambiguity in RNA-seq data that can occur in marker gene survey data. In RNA-seq analysis there is ambiguity for some reads with multiple potential 'optimal' mappings along the genome, so called 'multi-mapped' reads. This analogously occurs in marker gene survey data in assigning reads to OTU clusters and subsequently the count observed for given OTUs. In this case, reads may be assigned to more than one cluster if it is within the given similarity threshold for more than one cluster representative sequence. The default option in our pipeline, based on DNAClust, does not guarantee that a sequence can be uniquely placed. Reads are assigned to a particular cluster by choosing the best alignment and largest OTU representative center for a given set of clusters and can have more than

one possible placement. However, there an option in DNAClust, 'non-overlapping', that results in less ambiguously assigned reads by restricting reads that are within the radius of two or more clusters to not get assigned. This is similar to a commonly used approach in RNA-seq analysis of discarding multi-mapped reads.

To test the effect of a potentially ambiguous read mapping to a cluster we re-ran DNAClust with the 'non-overlapping' option on the full HMP dataset to compare rarefaction and sparsity results observed earlier. The rarefaction effect (association between depth of coverage and the number of detected features) and sparsity is essentially unchanged (Figure 10). The least sparse sample after filtering OTUs (less than 5 positive samples or reads present) and samples (< 1,000 total counts or > 35,000) in the 'non-overlapping' run is 97.48% non-positive while we observed 97.46% with default DNAclust options.

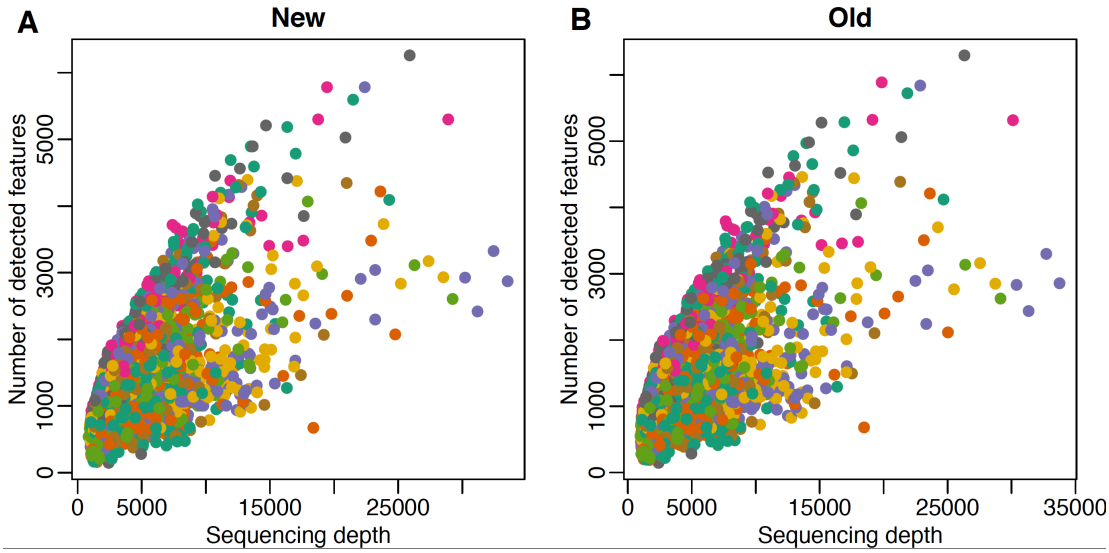

Figure 10: Effect of unambiguously placing reads in OTU centers on rarefaction.

Plot of the same samples for the 99% ‘perfect clusters’ (A) and 99% ‘exact clusters’ (B) run through DNAclust after trimming OTUs to be positive in at least five samples and removing outlying samples. Notice that rarefaction and sparsity is not affected by the DNAclust option.

We subset the data to the same subgingival and tongue samples and trimmed OTUs (less than 5 positive samples) as previously performed for the subgingival and tongue analysis. We observed 23,275 OTUs, a total of 410 fewer OTUs. We reran DESeq and the zero-inflated Gaussian mixture model on this less ambiguous dataset and compared fold-change estimates between DESeq and the zero-Inflated Gaussian mixture (Figure 11A). We observed the same phenomena and show in the figure below that ZIG is adjusting fold-changes on sparse OTUs as previously described. Also, we observed that the over dispersion estimates are similar to our previous run (Figure 11B).

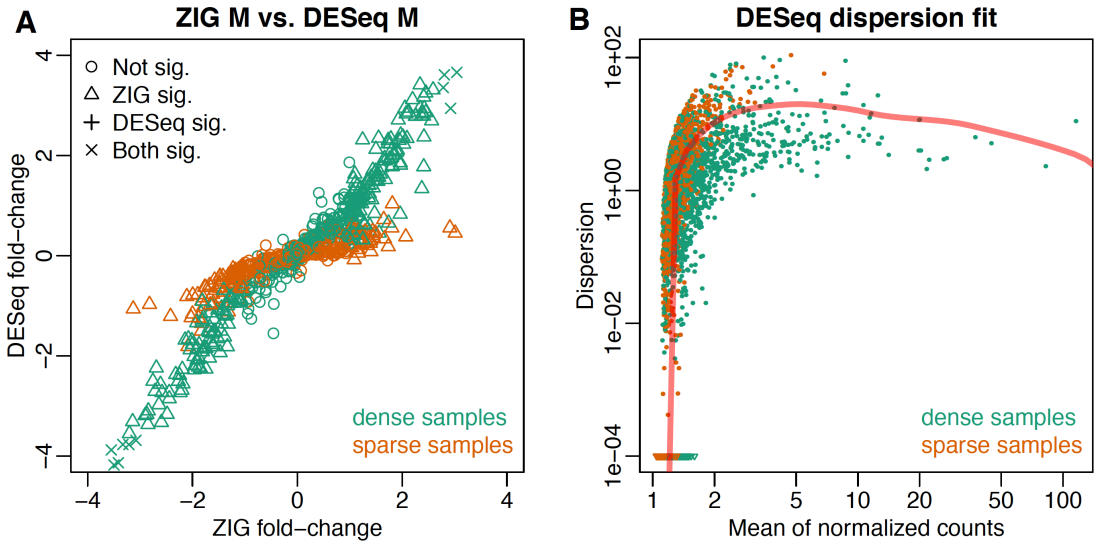

Figure 11: Effect of unambiguously placing reads in OTU centers on differential abundance.

Plots are created using OTUs created running the 'perfect clustering' option in DNAclust. (A) Fold-change estimates for metagenomeSeq and DESeq. We see that metagenomeSeq and DESeq agree in these estimates for dense features (green), while for sparse features (orange) metagenomeSeq adjusts fold-changes weighting zeros according to sample depth of coverage. (B) Estimate of dispersion as a function of mean feature counts by DESeq. Sparse features, indicated in orange, display high dispersion relative to dense features (green). This makes DESeq overestimate dispersion in general resulting in a large number of missed discoveries.

2.3.2.5 Simulation study

We evaluated our mixture model and compared it to currently used methods for differential abundance analysis using a simulation study where count data from two populations was generated producing datasets with varying levels of sparsity and association strength (quantified by fold-change). We compared our approach to existing tools for metagenomic analysis: the original Metastats method [26], Xipe [25], and a Kruskal-Wallis test as used in LefSe [27]. We also compared to commonly-used, representative methods for RNA-seq analysis: a non-zero inflated log-normal model [30], DESeq [23] and edgeR [29]. We report area under receiver operating characteristic curves (AUROC) as a summary of each method's sensitivity and specificity as simulation parameters are varied (Figure **12**).

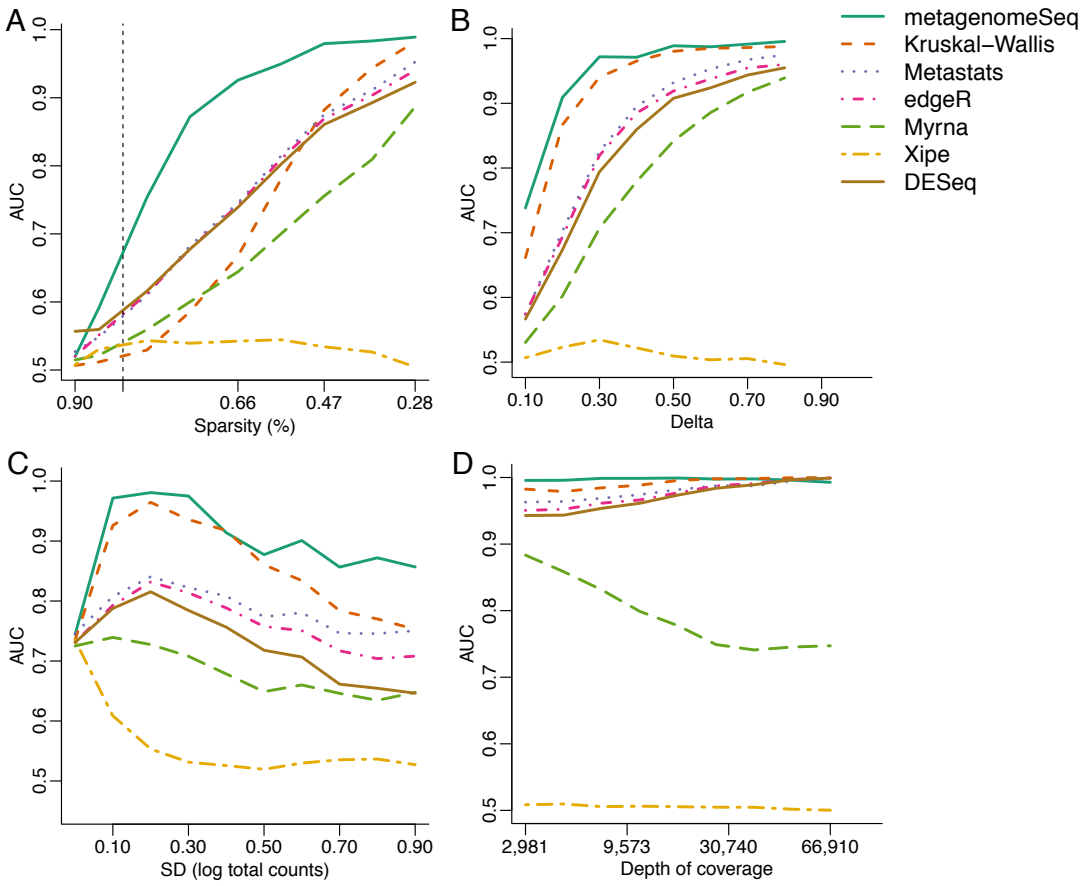

Figure 12: Simulation results indicated that metagenomeSeq has greater sensitivity and specificity in a variety of settings.

We used the area under the receiver operating characteristic curve (AUROC) to compare Metastats [26], Xipe [25], and Kruskal-Wallis test as used in LefSe [27], a nonzero inflated log-normal model (myrna) [30], DESeq [23] and edgeR [29]. (A) AUROC as data set sparsity decreases. metagenomeSeq achieved larger AUROC values than any other method in data sets with high sparsity (vertical dashed line represents the least sparse metagenomic data set). (B) AUROC as the effect size between two conditions (Delta) increases. Both metagenomeSeq and LefSe were better at detecting features with small effect size. (C) AUROC as the variability in depth of sequencing increases. metagenomeSeq and the Kruskal-Wallis test were robust with respect to high variability in sequencing depth. SD, sampling depth s.d. (D) AUROC as average sequencing depth increases. All models (except the nonzero inflated log-normal model and Xipe) performed similarly well at sufficient depth of coverage.

Our approach (metagenomeSeq) and to a lesser degree, the Kruskal-Wallis test consistently produced high area under the curve scores across most simulation

settings. The highest AUROC score is achieved by metagenomeSeq in cases with high measurement sparsity which are consistent with actual metagenomic datasets. The original Metastats, edgeR, Myrna and DESeq have similar performance characteristics, with smaller AUROC scores than metagenomeSeq and Kruskal-Wallis. Xipe performed poorly across most simulation settings, as expected, since this method does not account for population variability.

2.3.2.5.1 Simulation study design

We simulated OTU level datasets with 1000 features. A sample's total count was sampled from a log-normal distribution with μ = 7.5 and a standard deviation of 0.3. These values represent similar total counts to those observed in data. The first 50 features were chosen to be "significant". In one of the populations, for the first 25 significant features, we changed the proportion of the total counts for those features by adding $1x10^{-2} \cdot \delta$ percentage of the particular sample's total counts. For the remaining 25 we subtracted $1x10^{-2} \cdot \delta$ percentage of the sample's total counts. We used a logistic regression model of the proportion of zeros as a function of depth of coverage in a standard marker gene survey to build a plausible simulation model for sparsity. Given a sample's depth of coverage s_j an expected proportion of zero features π_j is obtained from the logistic regression fit. For each feature we randomly drew from a Bernoulli trial with probability π_j to spuriously set the feature to zero. Finally, we assigned randomly to 5% of the data an additional 1.3% (a value obtained from a standard marker gene survey) of the mean of the total counts to introduce extremely abundant features.

2.3.2.5.2 Subgroup simulation study design

We simulated data from two populations where each population consisted of two subpopulations representing a case-control study where cases and controls were collected from multiple sites. We simulated OTU level datasets with 1000 features as in 2.3.2.5.1 Simulation study design. The second subgroup had a relatively larger abundance of the significant features. This represents potential greater feature enrichment in a site's sub population. The trend across populations in either subgroup is to either increase or decrease in cases or controls.

2.3.3 Comparison of methods on subgingival plaque and tongue microbiota

We obtained data of oral microbiota from the Human Microbiome Project, the largest comparative 16S rDNA dataset to date [38], to compare methods for differential abundance analysis (Figure **13**). We used the original Metastats, edgeR, DESeq and metagenomeSeq to identify OTUs whose abundance differs significantly between tongue and subgingival plaque samples. The comparison was restricted to those OTUs with sufficient samples having non-zero counts for reliable estimation of fold-changes. For each method, differentially abundant OTUs were determined at FDR<0.05 where the OTU is at least twice as abundant in one group compared to the other (absolute log fold-change greater than 1). The original Metastats method and edgeR declared the largest number of OTUs to be significant (533 and 524, respectively), while metagenomeSeq (360) and, especially, DESeq (20) declared fewer significant OTUs.

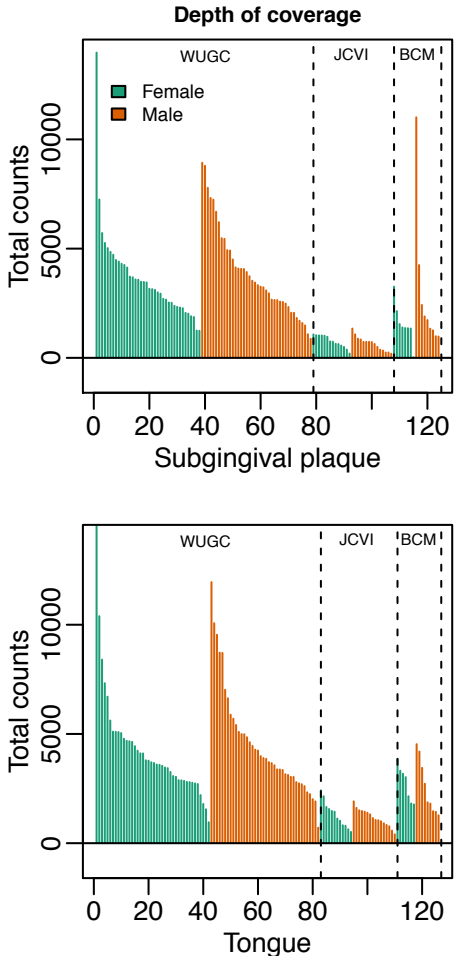

Figure 13: Sample sequencing depth for oral sub-community samples.

Samples are ordered by sequencing center (WUGC, JCVI, BCM), sex (female, male), and raw depth of coverage. The top graphs represent subgingival plaque samples, the bottom represent tongue samples.

2.3.3.1 Comparison with DESeq

Overall, metagenomeSeq and DESeq showed high agreement in fold-change estimates (Figure 14). Specifically, metagenomeSeq and DESeq fold-change estimates were very similar on features exhibiting low sparsity, but DESeq fails to declare as significant the majority of dense features declared significant by metagenomeSeq (only 20 of 244). We found that the mean-dispersion trend estimated by DESeq for this dataset was uncharacteristic of those estimated from RNA-seq data

(Figure 14B) and DESeq consistently overestimated dispersion for dense features resulting in a large number of failed discoveries (e.g., the feature shown in Figure 14C) is not declared as differentially abundant by DESeq). Features with high sparsity drive the poor dispersion estimate in DESeq and are also features where fold-change estimates between metagenomeSeq and DESeq disagreed (e.g. Figure 14D). By controlling for low sequencing depth, metagenomeSeq is able to detect these population differences appropriately.

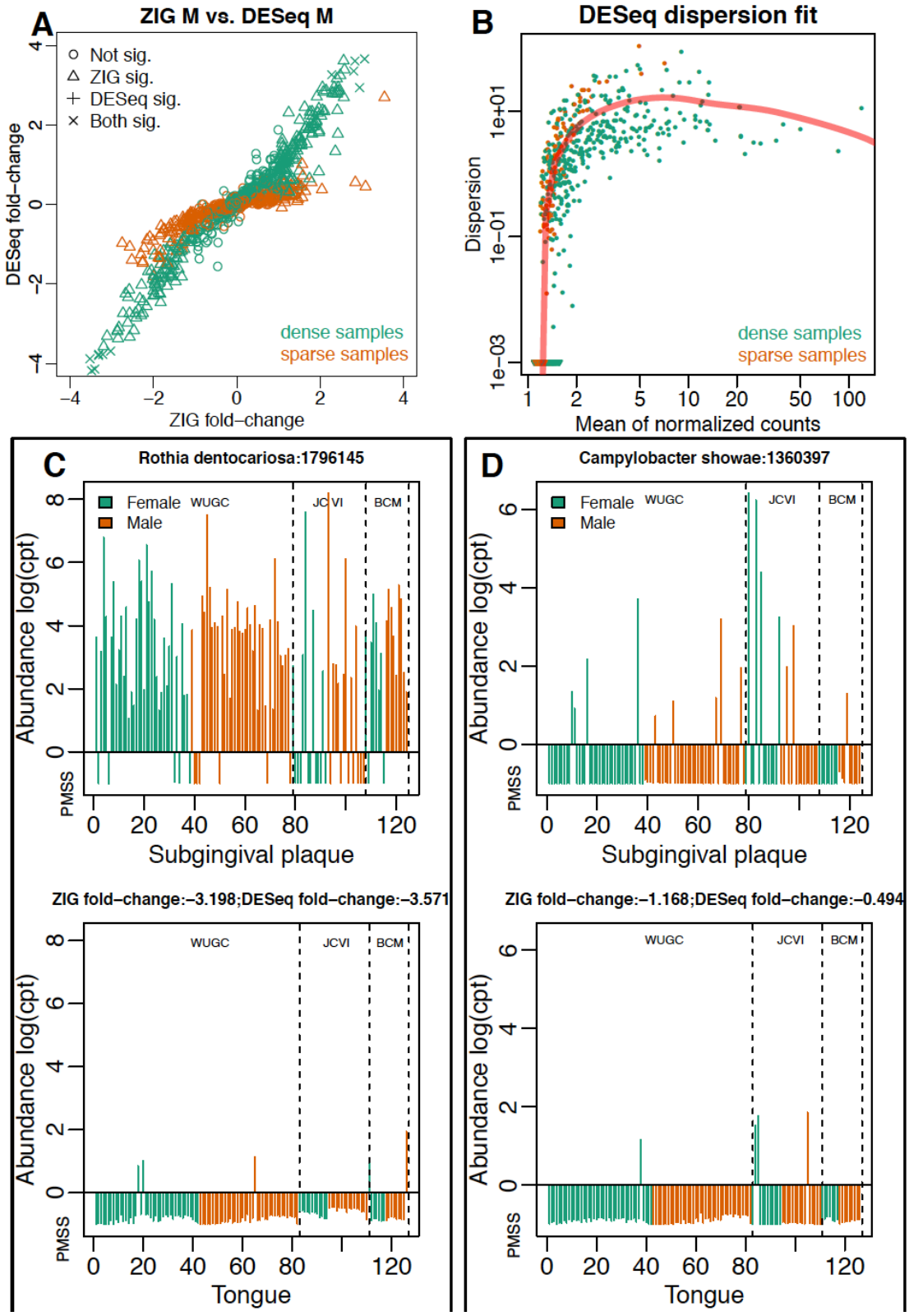

Figure 14: Comparison of metagenomeSeq differential abundance detection to detection by DESeq.

(A) Fold-change estimates for metagenomeSeq and DESeq. We see that metagenomeSeq and DESeq agree in these estimates for dense features (green), while for sparse features (orange) metagenomeSeq adjusts fold-changes weighting zeros according to sample depth of coverage. (B) Estimate of dispersion as a function of mean feature counts by DESeq. Sparse features, indicated in orange, display high dispersion relative to dense features (green). This makes DESeq overestimate dispersion in general resulting in a large number of missed discoveries. (C) A feature where dispersion overestimates by DESeq lead to a false negative call. Each panel plots the CSS normalized counts with samples ordered by sequencing center (WUGC, JCVI, BCM), sex (female, male), and depth of coverage. The top panel shows subgingival plaque samples, the bottom panel shows tongue samples. The bottom strip in each panel indicates the posterior probability estimates for zeros due to community sub-sampling (PMSS). (D) A sparse feature where weighting of zero counts by metagenomeSeq results in a differential abundance call missed by DESeq. Panels follow convention in sub-figure (C).

2.3.3.2 Comparison with edgeR

The package edgeR consistently estimated larger fold-changes in comparison with both metagenomeSeq (Figure **15**A) and DESeq (Figure **15**B). In edgeR total sample counts are included as a term in a generalized linear model, while our model includes the CSS normalization term in the log-normal linear model of the count distribution. Artifacts arising from normalization using total counts lead to many false differential abundance predictions made by edgeR (e.g., Figure **15**C), which are avoided by using our proposed normalization method. The dispersion trend estimate in edgeR is also uncharacteristic for this dataset (Figure **15**D) and the deviation is again driven by feature sparsity.

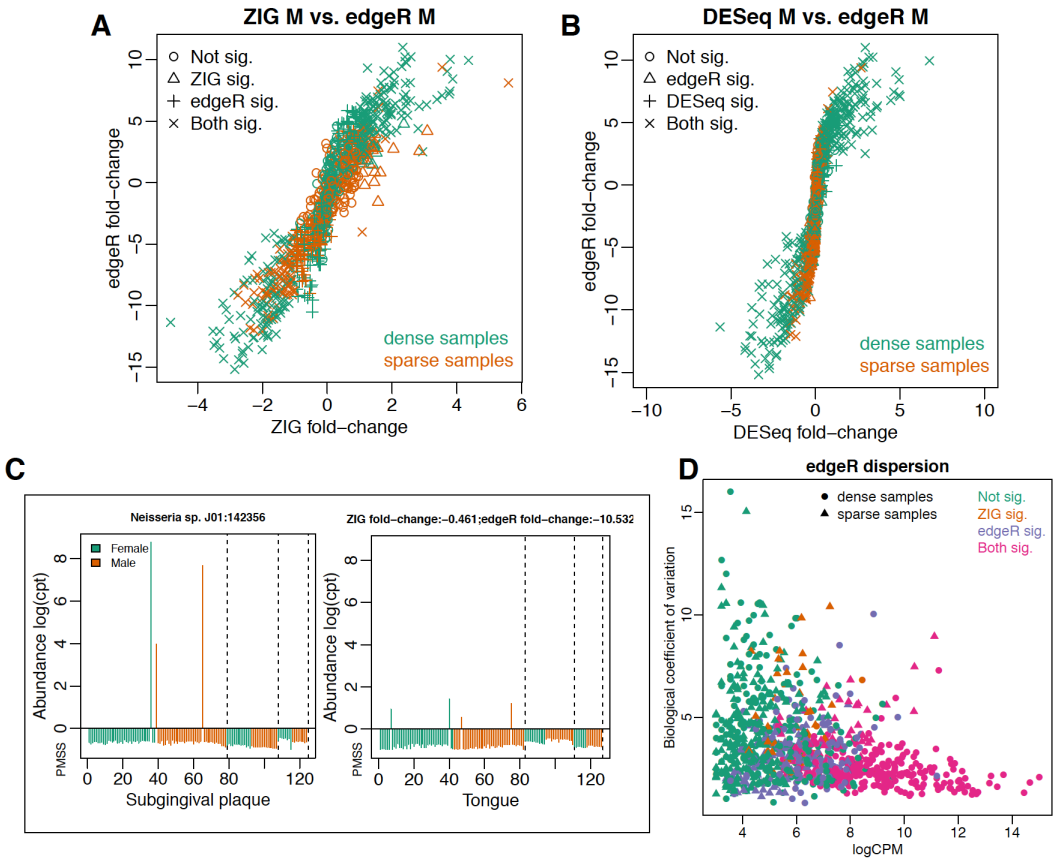

Figure 15: Comparison of edgeR differential abundance detection to detection in metagenomeSeq and DESeq.

(A) Fold-change estimates for metagenomeSeq and edgeR. We see that fold-changes are not consistent across features with edgeR significantly overestimating many fold changes. (B) Fold-change estimates for DESeq and edgeR. Again, edgeR overestimates many fold-changes. (C) Feature not considered significant according to metagenomeSeq, but edgeR normalization using total counts results in a significant call. Panels follow convention in Figure 14. Counts plotted are CSS normalized counts. (D) Dispersion as estimated by edgeR. Dispersion does not follow typical RNA-seq experiments. Features declared significant by metagenomeSeq, but not edgeR (orange) have high variability estimates in edgeR and tend to be sparse features (triangles). metagenomeSeq and edgeR agree on abundant dense features (magenta circles). Features declared significant by edgeR but not metagenomeSeq (purple) have moderate abundance driven by few high-count features resulting from normalization artifacts.

2.3.3.3 Comparison with Metastats

Metastats was more consistent with edgeR (Figure 16A) than metagenomeSeq (Figure 16B) or DESeq (Figure 16C) due to its use of total normalization. Even

though we are testing for overall differential abundance between oral microbiota, regardless of sequencing site, Metastats consistently called features as significant where differences are specific to sequencing site (e.g., samples sequenced in JCVI as shown in Figure 16D). Large survey studies may obtain samples from a variety of locations, over heterogeneous populations. Analysis methods used in these studies require the ability to interpret differential abundance taking into account the heterogeneity of these populations. Both RNA-seq methods and metagenomeSeq use linear models that can include possible confounding sources of variability, in this case sequencing site or gender, to aid interpretation in differential abundance testing. Our software and both RNA-seq methods are able to detect site-specific differential abundance between microbiota using an interaction model. Metastats was not designed to carry out this kind of test.

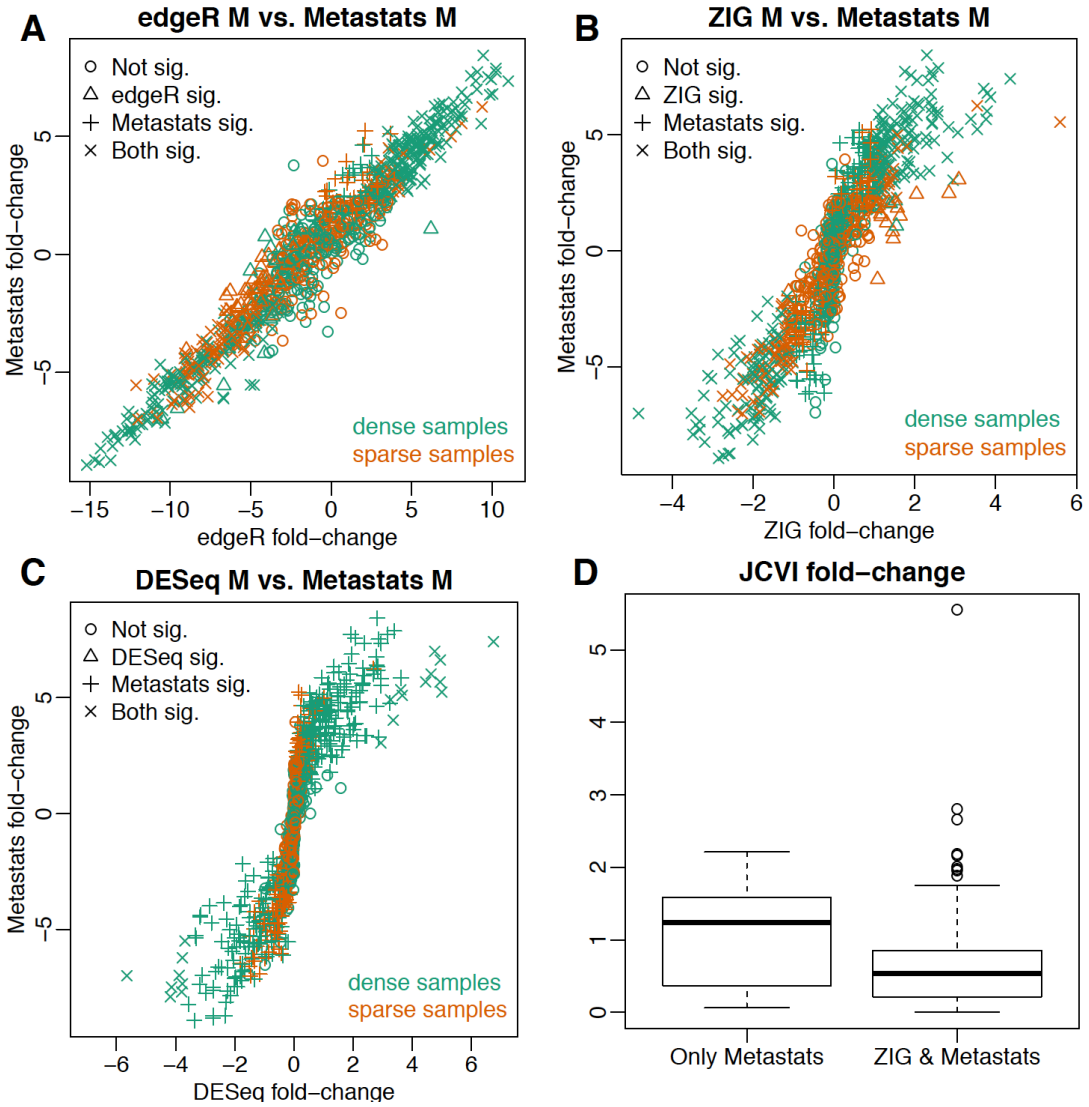

Figure 16: Comparison of Metastats differential abundance detection to detection in metagenomeSeq, DESeq and edgeR.

(A) Fold-change estimates for Metastats and edgeR showing high consistency between the two methods. (B) Fold-change estimates for metagenomeSeq and Metastats where Metastats consistently over-estimates fold-changes especially for sparse features (orange). (C) Fold-change estimates for DESeq and Metastats, where again Metastats overestimates fold-changes for sparse features (orange). (D) Sequencing site fold-change estimates from metagenomeSeq for features detected as differentially abundant by the original Metastats, but not by metagenomeSeq. Many differential abundance discoveries in Metastats are solely due to site specific effects.

2.3.3.1 Comparison with LefSe

As further validation we compare our results with those obtained by Segata, et

al. using the LefSe tool on the same oral dataset [27]. We confirm all their findings and identify three additional differentially abundant species missed by their analysis. Specifically, we find *Atopobium parvulum*, *Lautropia sp.*, and *Desulfotomaculum sp.* to be enriched in subgingival plaque (Figure 17). All of these were fairly abundant in the samples, representing at least 4% the population, and represent previously characterized members of the normal subgingival microbiota [47–49].

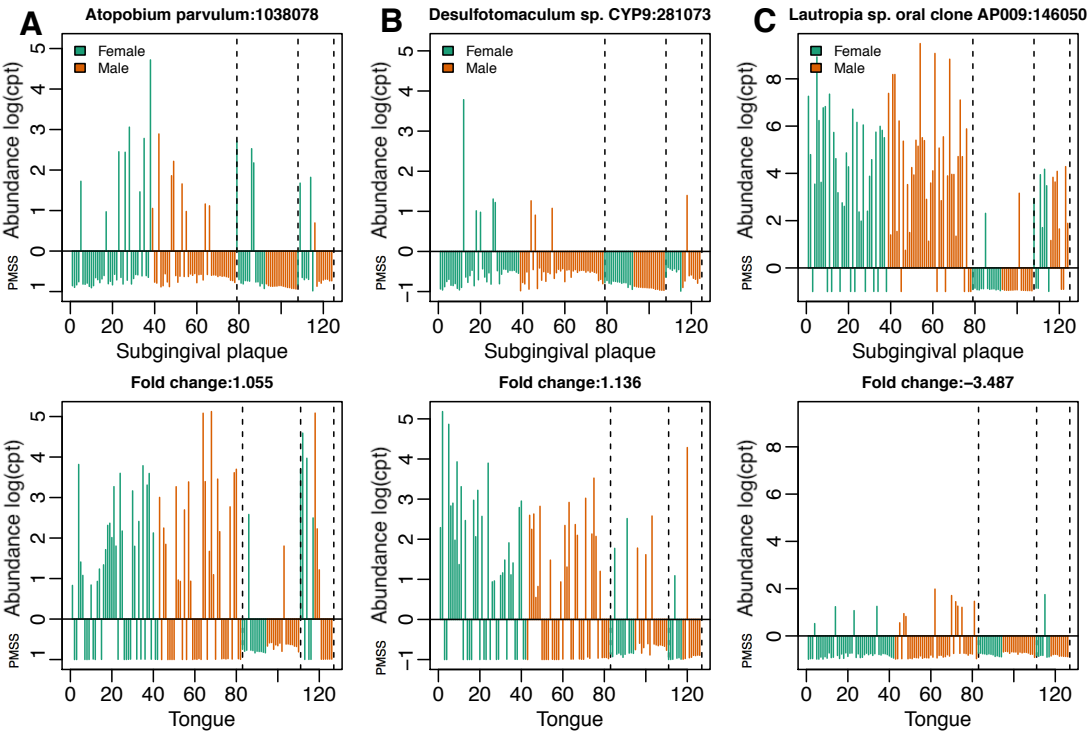

Figure 17: Novel species detected as differentially abundant in tongue and subgingival microbiomes.

Plot of CSS normalized counts for three OTUs. Samples are ordered by sequencing center (WUGC, JCVI, BCM), sex (female, male), and depth of coverage. Underneath each panel are probabilities of missing sequences due to subsampling (PMSS). The top graphs represent subgingival plaque samples, the bottom represent tongue samples.

Our ZIG model uses linear modeling following standard conventions in methods for testing differential abundance in gene expression [31] that control for

confounding factors. In contrast, LefSe uses an *ad hoc* heuristic approach to account for subpopulations in large marker studies that is overly conservative and prone to low sensitivity. We observed by simulation (Figure 18) that metagenomeSeq was more sensitive than LefSe (0.95 vs. 0.01, respectively) and retained high specificity (0.96 vs. 1) when confounding subpopulations were included among tested groups.

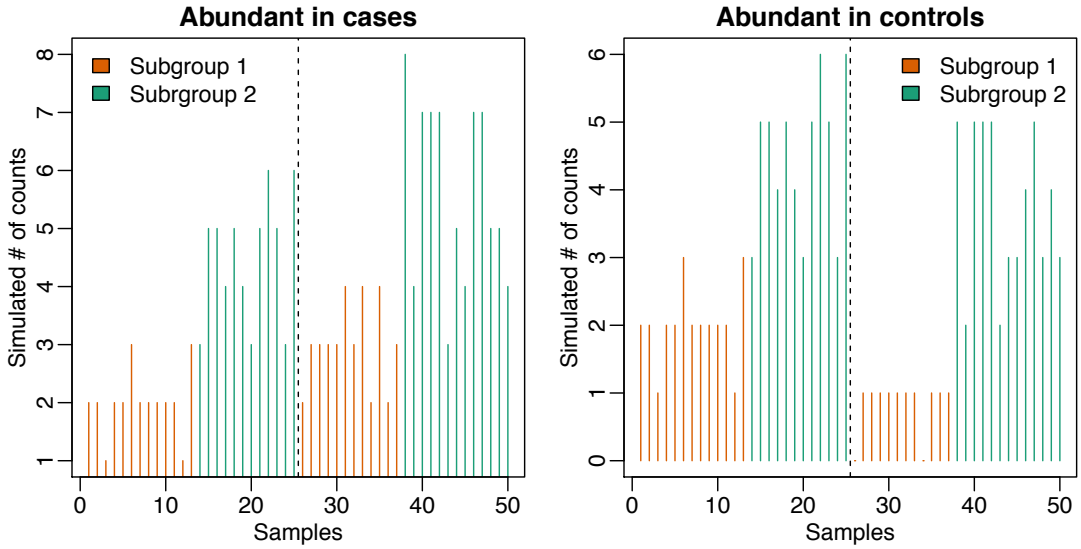

Figure 18: Abundance of two simulated features from the subgroup simulation.

Plot of simulated counts for two OTUs. Samples are ordered by case control and subgroup population. The dotted line separates the two populations, controls and cases and the colors for each line represents the subgroup. Notice the trend from controls to cases is consistent across subgroups, but LefSe is confounded by the cross population subgroups.

2.3.4 Materials

2.3.4.1 Marker-gene survey data

2.3.4.1.1 Humanized gnotobiotic mouse gut

Twelve germ-free adult male C57BL/6J mice were fed a low-fat, plant polysaccharide-rich diet. Each mouse was gavaged with healthy adult human fecal material. Following the fecal transplant, mice remained on the low-fat, plant

polysacchaaride-rich diet for four weeks, following which a subset of 6 were switched to a high-fat and high-sugar diet for eight weeks. Fecal samples for each mouse went through PCR amplification of the bacterial 16S rRNA gene V2 region weekly. Details of experimental protocols and further details of the data can be found in Turnbaugh et al. [50] OTUs were classified by RDP [14] and annotated (minimum confidence level of 0.8). Sequences can be found at: http://gordonlab.wustl.edu/TurnbaughSE_10_09/STM_2009.html.

2.3.4.1.2 Subgingival plaque and tongue dorsum

Subgingival plaque and tongue dorsum samples were a part of the Human Microbiome Project [38] dataset used in this analysis. The samples were part of a larger study aimed at cataloging the healthy human microbiome. Reads were deposited into the Data Analysis and Coordination Center (DACC) at http://www.hmpdacc.org/. In particular, reads and metadata were downloaded from http://www.hmpdacc.org/HMR16S/. Further information on data collection protocol and samples is available at http://www.hmpdacc.org/ and in HMP. Only patients from their earliest visit were considered as were only samples properly annotated. Following OTU propagation (described below), singletons (up to 5 positive samples) were trimmed. To consider solely differential abundance estimates, we report on OTUs present in at least approximately 2% of the population. For each differential abundance method compared, differentially abundant OTUs were determined at FDR<0.05 where the OTU is at least twice as abundant in one group compared to the other (absolute estimated fold-change greater than 1). We used LefSe's default detection method (as no fold-change estimate is provided).

2.3.4.1.3 Human Microbiome Project data

Data used in Supplementary Figure 3 was a part of the Human Microbiome Project[24] dataset used in this analysis. The samples were a catalog of the healthy human microbiome. Reads were organized into OTUs by QIIME[8] and deposited in the Data Analysis and Coordination Center (DACC) at http://www.hmpdacc.org/. In particular, OTUs and metadata was downloaded from http://www.hmpdacc.org/HMQCP/. Further information on data collection protocol and samples is available at http://www.hmpdacc.org/ and in HMP.

2.3.4.1.4 Lung microbiome

The lung microbiome consisted of respiratory flora sampled from six healthy individuals. Three healthy nonsmokers and three healthy smokers. The upper lung tracts were sampled by oral wash and oro-nasopharyngeal swabs. Up to a patients' glottis, samples were taken using two bronchoscopes a serial bronchoalveolar lavage and lower airway protected brushes. More detailed information about the lung Microbiome samples, collection and protocols is available in Charlson et al. [39] Reads and barcodes were provided by Frederic Bushman. Following OTU propagation (described in 2.3.4.2 OTU identification and annotation), OTUs were trimmed if they were not present in approximately 8% of the population.

2.3.4.2 OTU identification and annotation

454 SFF files and barcode dictionaries were downloaded and run through the same pipeline. Conservative Operational Taxonomic Units (OTUs) were constructed by pooling together the sequences from all samples, then clustered using DNAclust [13] with default parameters (99% identity clusters) to ensure that the definition of an

OTU is consistent across all samples. To obtain taxonomic identification, a representative sequence from each OTU was aligned to Ribosomal Database (rdp.cme.msu.edu, release 10.4) using Blastn with long word length (-W 100) in order to only detect nearly-identical sequences. Sequences without a nearly-identical match to RDP were marked as having "no match" and assigned an OTU identifier. The resulting data was organized into a collection of tables at many different taxonomic levels containing each taxonomic group as a row and each sample as a column. These tables formed the substrate for the statistical analyses described. This process was performed for the human microbiome project and the human lung microbiome datasets. After removing OTUs present in less than 5 samples, the HMP dataset consisted of 23,685 OTUs, whereas the human lung microbiome consisted of 2,365 OTUs. We explored the effect of ambiguously assigned reads (sequences that have good matches to two or more OTUs) by running DNAclust in 'non-overlapping' mode – a mode that ensures high separation between clusters and eliminates ambiguous reads. We provide further discussion of the ambiguity of mapping reads to OTUs in (2.3.2.4 Ambiguous read assignment to OTUs).

2.3.4.3 RNA-seq data

RNA sequencing counts were downloaded from ReCount [37], http://bowtie-bio.sourceforge.net/recount/. Only datasets with at least 15 samples were considered.

2.3.4.4 Software

The following software versions were used for analysis on the following platform. DESeq version 1.8.3 [23] and edgeR version 2.6.12 [29] and limma version 3.12.3 [31] were used in the comparisons. Personalized R scripts were written for the

other methods and all analyses were performed on R version 2.15.1 on a Red Hat Enterprise Linux Server release 5.9 (Tikanga) 64-bit platform.

2.4 Discussion

We present a novel normalization and differential abundance analytical methods tailored to the specific characteristics of data from surveys of marker-genes for microbiota. The CSS normalization approach is an extension of methods used in normalization of microarray [20–22] and RNA-seq [17, 23, 24], modified using adaptive methods to determine normalization parameters. This extension determines normalization parameters that better account for the strong dependence of the number of detected features on sampling depth encountered in marker-gene studies.

Rarefaction is a common phenomenon in molecular surveys of bacterial communities [51], where the number of taxonomic features detected in a sample depends on the amount of sequencing performed. This large variation in the number of taxonomic features detected in each sample, contributes to the inherent sparsity of metagenomic data where most features are only found in a few samples, as previously reported [52, 53]. To accurately estimate differential abundance, we explicitly model the effect of under-sampling on the ability to detect a particular feature. Although under-sampling is ubiquitous in marker-gene survey data, to our knowledge, the approach presented here is the first to correct for this phenomenon. While our focus is on data generated in microbial community surveys sparsity may also be an issue in some RNA-seq experiments, and thus our methods may have broader applicability. The evaluation of our methods in that context is, however, beyond the scope of this work and will be addressed in future studies.

We have shown that the use of our methods yields a more precise biological interpretation of the data – in mouse stool data the CSS normalization helps distinguish clinical phenotypes that are confounded by commonly used normalization methods, while in the oral microbiome, the combined differential abundance modeling approach identifies additional associations that were missed by commonly used tools. The additional organisms found enriched in subgingival plaque are fairly abundant well known members of the periodontal microbiome and include sulfate-reducing bacteria, which have been proposed as potential pathogenicity factors in periodontal disease [47].

Recent publications have relied on machine learning techniques, such as random forests, to identify microbiota signatures correlated with phenotypic observations [52, 54]. Our work targets a complementary task — the feature-by-feature assessment of differential abundance based on an appropriately defined linear model that accounts for specific microbiota features and confounding factors. The methods developed here, in particular the ZIG mixture model, can be incorporated into machine-learning based predictive models that seek to identify multiple features for specific phenotypes.

Other types of analyses adversely affected by high sparsity and sampling bias include clustering and co-occurrence network discovery. While the normalization approach presented here can help control biases in analyses based on simple correlation measures, methods developed to specifically discover significant correlations between sparse features in marker-gene survey data are better suited for the task [55, 56].

This work directly addresses some of the main challenges to robust analysis of marker-gene surveys in clinical and epidemiological settings: technical biases in the sequencing libraries, leading to variable depth of coverage across samples and the resulting rarefaction effect; and confounding due to technical and population characteristics. We have demonstrated that our methods outperform approaches that are widely used in the field, and hope that the improved analysis approaches we propose will help biologists achieve the full promise of marker-gene surveys in clinical research. Many of the ideas developed here can also be applied to data derived from metagenomic experiments and we plan to evaluate their application to such data in future work.

3 Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition

3.1 Background

3.1.1 Context

This work was a collaborative effort of many individuals and institutions resulting in one of the largest marker-gene surveys published at the time and the first high-throughput sequencing study of diarrhea culminating in the Genome Biology paper [2]: Pop, M.*, Walker, AW.*, Paulson, JN.*, Lindsay, B.*, Antonio, M.*, Hossain MA.*, Oundo J.*, Tamboura B.*, Mai V.*, Astrovskaya I., Bravo, HC., Rance R., Stares M., Levine MM., Panchalingam S., Kotloff K., Ikumapayi UN., Ebruke C., Adeyemi M., Ahmed D., Ahmed F., Alam MT., Amin R., Siddiqui S., Ochieng JB., Ouma E., Juma J., Mailu E., Omore R., Morris JG., Breiman RF., Saha D., Parkhill J., Nataro JP., Stine, OC. (2014). Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. Genome Biology, Volume 15, pg. R76. My contributions to the work included the application of the methods presented in the previous chapter to this dataset and a number of side-analyses in summarizing the distributional properties of OTUs and annotation, the effects of OTU radii on fold-change estimates, Shannon diversity analysis, MDS analysis, and the application of SparCC. Additionally, I developed an interactive browser of the marker-gene survey data available at http://epiviz.cbcb.umd.edu/shiny/MSD1000 with code available at https://github.com/nosson/MSD1000.

In the paper we characterized the microbial communities of 992 infants from four developing countries, The Gambia, Mali, Bangladesh and Kenya. One major caveat to the study is that these associations cannot be described as causal, rather, they must be attributed to as associations till experimental evidence points otherwise. All of this work would not be possible without the many field workers.

The methods developed in Chapter Differential abundance analysis for marker-gene surveys` were initially designed on the dataset presented below. The method described in the previous chapter was designed due to the lack of suitable methods in the analysis of this significant dataset. Applying the methods developed in Chapter 2 we analyzed the microbial communities and searched for bacteria novelly associated with health, diarrhea, and dysentery.

3.1.2 Abstract

Background: Diarrheal diseases continue to contribute significantly to morbidity and mortality in infants and young children in developing countries. There is an urgent need to better understand the contributions of novel, potentially uncultured, diarrheal pathogens to severe diarrheal disease, as well as distortions in normal gut microbiota composition that might facilitate severe disease.

Results: We use high-throughput 16S rRNA gene sequencing to compare fecal microbiota composition in children under five years of age who have been diagnosed with moderate to severe diarrhea (MSD) with the microbiota from diarrhea-free controls. Our study includes 992 children from four low-income countries in West and East Africa, and Southeast Asia. Known pathogens, as well as bacteria currently not considered as important diarrhea-causing pathogens, are positively associated

with MSD, and these include *Escherichia/Shigella*, and *Granulicatella* species, and *Streptococcus mitis/pneumonia* groups. In both cases and controls, there tend to be distinct negative correlations between facultative anaerobic lineages and obligate anaerobic lineages. Overall genus-level microbiota composition exhibit a shift in controls from low to high levels of *Prevotella* and in MSD cases from high to low levels of *Escherichia/Shigella* in younger versus older children; however, there was significant variation among many genera by both site and age.

Conclusions: Our findings expand the current understanding of microbiota-associated diarrhea pathogenicity in young children from developing countries. Our findings are necessarily based on correlative analyses and must be further validated through epidemiological and molecular techniques.

3.2 Diarrheal disease continues to be a major cause of childhood mortality

Diarrheal diseases continue to be major causes of childhood mortality, ranking among the top four largest contributors to years of life lost in sub-Saharan Africa and South Asia [57]. The proportion of deaths attributed to diarrhea among children aged under 5 years is estimated to be approximately 15% worldwide [58], and as high as approximately 25% in Africa and 31% in South East Asia [59]. More than two dozen enteric pathogens, belonging to diverse branches of the tree of life, are known to cause diarrhea and can be tested for in a clinical setting. However, it is likely that additional pathogens remain to be identified among the enteric microbiota.

In response to important unanswered questions surrounding the burden and etiology of childhood diarrhea in developing countries, the William and Melinda

Gates Foundation commissioned the Global Enterics Multicenter Study (GEMS) [60], which recently reported the pathogens responsible for cases of moderate-to-severe diarrhea (MSD) in seven impoverished countries of sub-Saharan Africa and south Asia. Importantly, for approximately 60% of MSD cases in GEMS, no known pathogen could be implicated by conventional diagnostic methods [61]. These observations highlight the potential presence of previously undiscovered pathogens, and/or possible interactions between pathogens and other members of the intestinal microbiota (both pathogenic and commensal) that may either exacerbate the clinical manifestation or protect the host from disease.

Here we apply molecular techniques to survey the intestinal microbiota in a subset of GEMS cases and controls. Our study comprises 992 children from four under-developed countries in West Africa (The Gambia and Mali), East Africa (Kenya), and South Asia (Bangladesh), representing a subset of the over 25,000 GEMS children enrolled. Our results shed additional light on potential mechanisms underlying MSD in children of developing countries. Prior to presenting these results we would like to stress that our analyses are, by necessity, correlative and the results presented here must be validated through epidemiological and molecular analyses, several of which are already underway.

3.3 Results and discussion

3.3.1 Description of data

Our data comprise roughly equal proportions of cases and controls (0.51 *vs.* 0.49, respectively) from four sites: Bangladesh (N = 206), The Gambia (N = 269), Kenya (N = 305), and Mali (N = 212). Approximately 55% of the subjects were

boys. Of 992 samples, 508 were from patients with MSD. The children ranged in age from newborn to 59 months. We stratified them into five age categories: 0 to 5 months (N = 112), 6 to 11 months (N = 308), 12 to 17 months (N = 173), 18 to 23 months (N = 146), and 24 to 59 months (N = 253). There were no significant differences between the proportion of cases and controls in each country and from each age group. The sequencing of PCR amplified 16S rRNA genes resulted in 3,584,096 reads passing quality checks. Each sample had at least 1,000 reads, and there were an average of 3,613 reads per sample. The reads were clustered using DNAclust [13] into 97,666 operational taxonomic units (OTUs). Of these, 21,247 passed chimera checking, were detected in more than five samples, or represented at least 20 sequences in a single sample, and were included in further analysis. The number of OTUs per sample ranged from 55 to 1252, with a median of 380 and an average of 412. The mean OTU size was 138, ranging from 5 (by definition) to 192,978 (with median OTU size = 15 sequences). Representative sequences from the 21,247 OTUs matched 728 distinct taxa from 161 genera. Among these, 4,730 (22 %) did not have good (>100 bp exact match, >97% identity) matches to isolate sequences from the Ribosomal Database Project (RDP). These were flagged as 'unassigned' in our analysis and are discussed further below. These sequences are not simply an artifact of our stringent alignment criteria as evidenced by the fact that a re-analysis of the 6,879 most abundant OTUs using the reference-based OTU picking algorithm implemented in QIIME [11] failed to classify a similar proportion of sequences (2,162 or 31% of the abundant OTUs).

3.3.2 Microbiota variations by age

The well documented [62–64] succession of the intestinal microbiota during child development is apparent in our non-diarrheal control samples (Figure 19A). During the first year of life, the 'healthy' gut microbiota in our infant cohorts is characterized by comparatively low overall diversity and a relatively high proportion of facultatively anaerobic, and potentially pathogenic, organisms (for example, the *Escherichia/Shigella* group, which cannot be distinguished from each other by 16S rRNA gene sequences), organisms that are believed to play a role in the development of the host immune system [63, 64]. In older ages, the dominance of these organisms is reduced, replaced by a corresponding increase in overall diversity (Figure 19B), accompanied by a particularly pronounced increase in the proportional abundance of the bacterial genus *Prevotella*. These changes are most evident in our non-diarrheal control samples, where the genus *Prevotella* increases from approximately 12% to approximately 48% proportional abundance during the first 5 years of life, while the *Escherichia* genus drops from about 20% proportional abundance in infants under 6 months of age to approximately 1% in 2- to 5-year-olds. Two other genera, *Veillonella* and *Streptococcus* also exhibit significant decreases with increasing age. Our data also show an increase with increasing age in the proportion of a range of organisms (labeled 'unassigned' in Figure 19A) that have no good quality matches to cultured isolates in public databases, and which appear to belong predominantly to obligate anaerobic bacteria (over 60% can be assigned by the RDP classifier to the *Ruminococcaceae* and *Lachnospiraceae* families of the Firmicutes phylum, which are relatively poorly represented in culture

collections [65], as well as the *Bacteroidaceae* family). These previously-uncultured putative obligate anaerobes increase in proportional abundance from approximately 8% in diarrhea-free young children to approximately 23% in the older age group, consistent with increase in diversity within the intestinal microbiota and the known expansion of these groups, which are able to colonize the intestine in greater numbers as the complex polysaccharides they utilize for growth become a greater feature of the host diet [54].

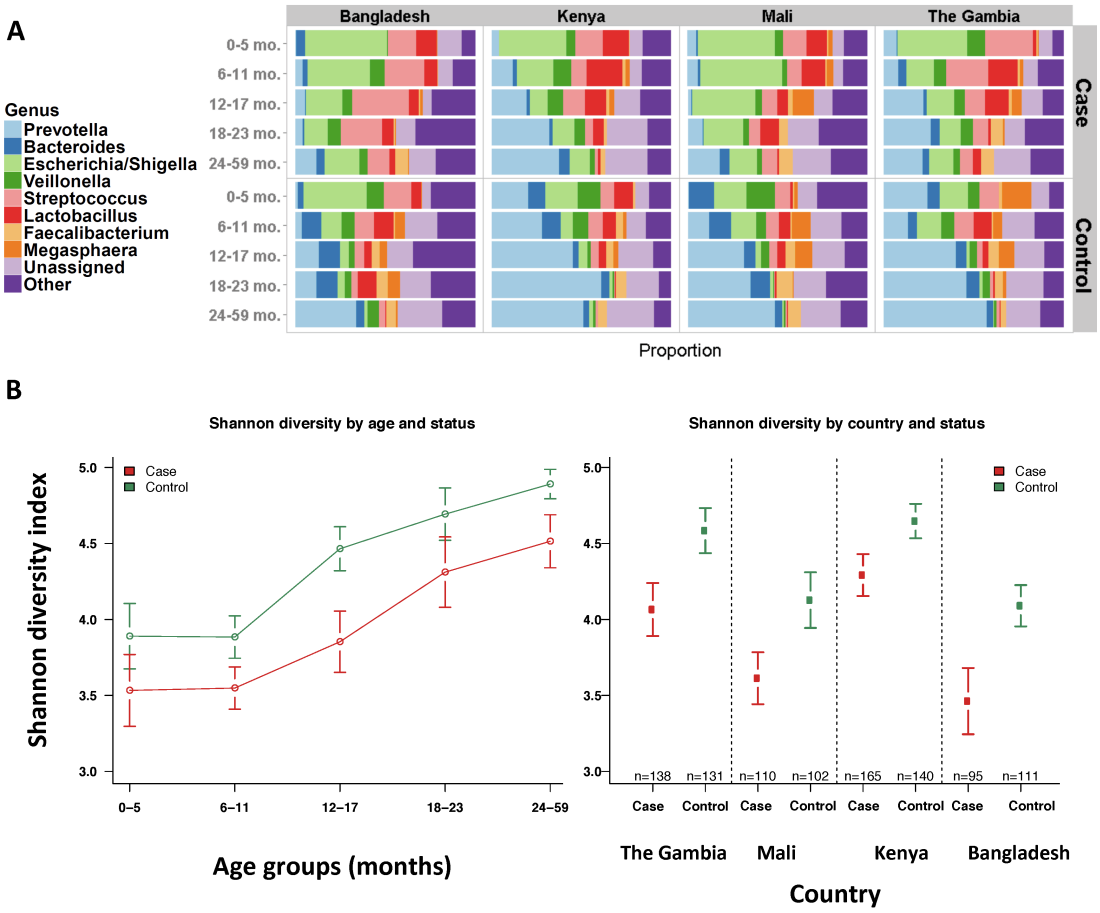

Figure 19: Comparison of diarrheal and non-diarrheal stool.

(A) Proportional abundance of genera in non-diarrheal controls and MSD cases in different age categories. Each color represents a different group. The order and color for each group is the same for controls (patients without MSD) and MSD cases. The eight groups most frequently found in controls (*Prevotella, Bacteroides, Escherichia/Shigella, Veillonella, Streptococcus, Lactobacillus, Faecalibacterium,*

Megasphaera, plus unassigned and other) are depicted. (B) Shannon diversity index across ages and diarrheal status. Average Shannon diversity indices for the five different age strata as well as the corresponding 95% confidence intervals. Both cases and controls exhibited higher mean Shannon diversity index scores at higher age groups compared to lower age groups (P <0.001, one-way ANOVA). The diversity of healthy samples positively correlates with age in the first 2 years of life, as previously reported [64]. The diversity index for cases is significantly less than that for controls within each country (P <0.02, Tukey's t-test corrected for multiple comparisons). Also see Figure **20**.

These observations broadly hold when stratifying by country of origin; however, country-specific effects are also apparent. For example, the samples from Bangladesh are different from the African countries, particularly in the younger age groups, and are characterized by a lower proportion of *Prevotella* sequences and a higher proportion of organisms from the *Escherichia/Shigella* and *Streptococcus* genera (Figure 19A).

The patterns observed within control samples were significantly different from patterns from patients with MSD; however, some overall age-related trends were similar. For example, *Prevotella* abundance correlates with age, albeit reaching a much lower peak, with only 23% abundance in the oldest age group (*vs.* 48% in controls, P <10-^{16}). Other obligate anaerobic microbes have lower proportional abundance among cases compared to controls: *Bacteroides* and the unclassified putative anaerobes are both 5% lower in cases, consistent with previous observations that indicate intestinal dysbiosis is associated with a decrease in the proportional abundance of obligate anaerobes [66]. Among cases, *Escherichia/Shigella* and *Streptococcus* spp. maintain a high proportion across all age groups, though their preponderance drops significantly (41% to 13% and 18.5% to 7.5%, respectively) as children age. Furthermore it appears that *Prevotella* and *Escherichia/Shigella* are

negatively correlated in MSD cases (Spearman rho = -0.55, P <0.0001). The disruption associated with diarrhea is also reflected in lower diversity values in MSD cases in every age group (Figure 19B, Figure **20**A-D).

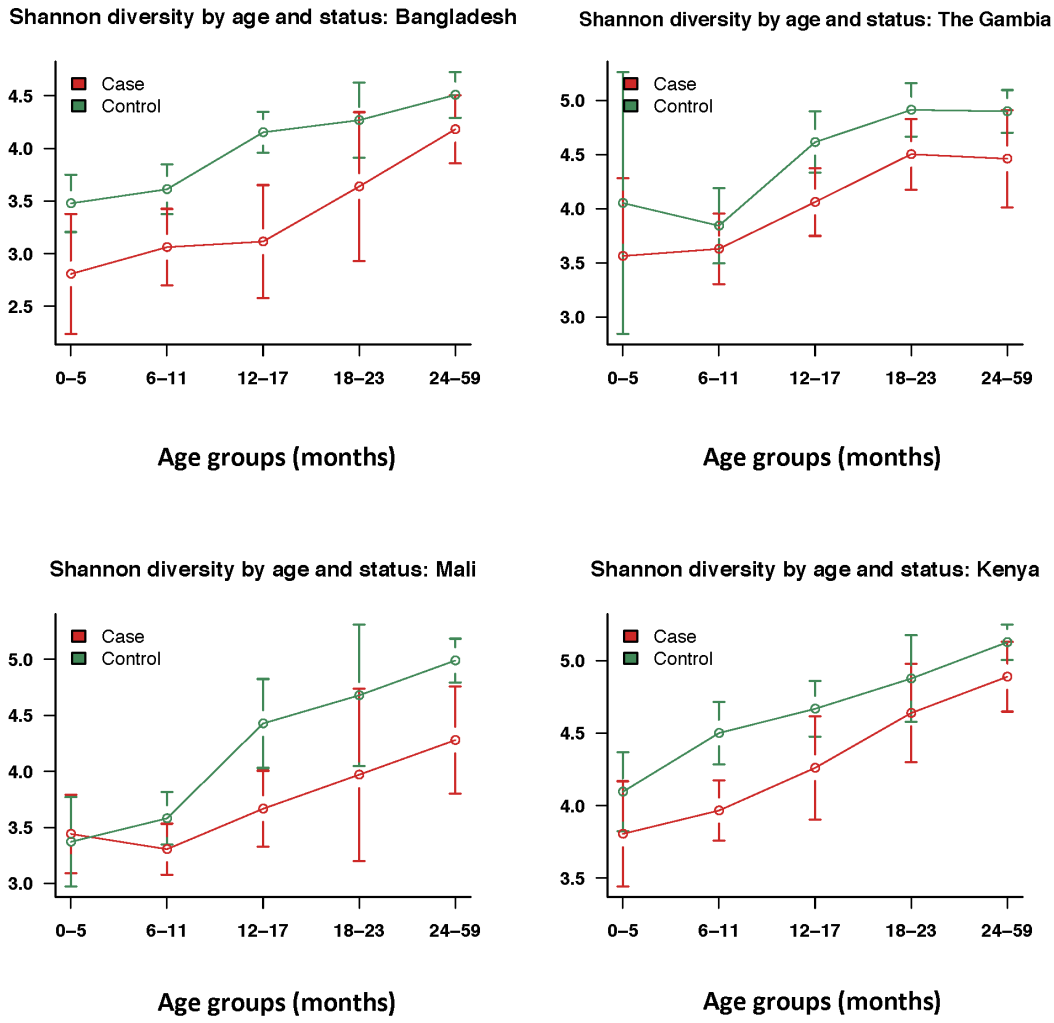

Figure 20: Shannon diversity indices across age and country.

Points represent the average Shannon diversity stratified by country, age, and health status. Each panel consists of data from a different country and colors represent the health status where red represents cases and green, controls. The upper and lower bars correspond to the 95% confidence interval. There is a strong trend for both cases and controls to become more complex with age. Diversity appears to increase greatly following the 6-11 month age group in Mali and The Gambia. Diversity increases throughout age in Kenya and Bangladesh.

Country-specific effects were also observed in diarrheal stool; for instance, in Kenya, diarrhea appeared to have a less marked effect on the microbiota (Figure 19A). *Escherichia/Shigella* spp. were most common in Mali, accounting for 34% of the sequences, next most common in Bangladesh (24%) and least common in The Gambia (15%). *Prevotella* spp. were found in high proportional abundances in The Gambia (18%) and Kenya (19%). The genus *Streptococcus* is found in relatively high abundances in Bangladesh (21%) and The Gambia (13%) with lower abundances in Mali (10%) and Kenya (9%). As expected, the taxonomic diversity (Shannon diversity index) is significantly different between cases and controls in all countries ($P<0.005$, pairwise t-test). Of note, where *Prevotella* is more common (The Gambia and Kenya), the diversity is higher (Figure 19B).

3.3.3 Taxonomic groups statistically increased or decreased in diarrhea

Multidimensional scaling analysis could not separate the diarrhea and diarrhea-free bacterial communities due to high inter-personal variation (Figure **21**). We estimated the association of individual OTUs with disease using statistical tests addressing both presence-absence statistics (Fisher's exact test and logistic regression) and abundance-dependent statistics (using generalized linear models) that account for the number of OTU-specific sequences in each stool, and potential confounders such as sampling depth, age, and country. The former address similar questions to those commonly targeted by the traditional culture-based epidemiological studies, while the latter allow us to assess how pathogen proportional abundance correlates with morbidity.

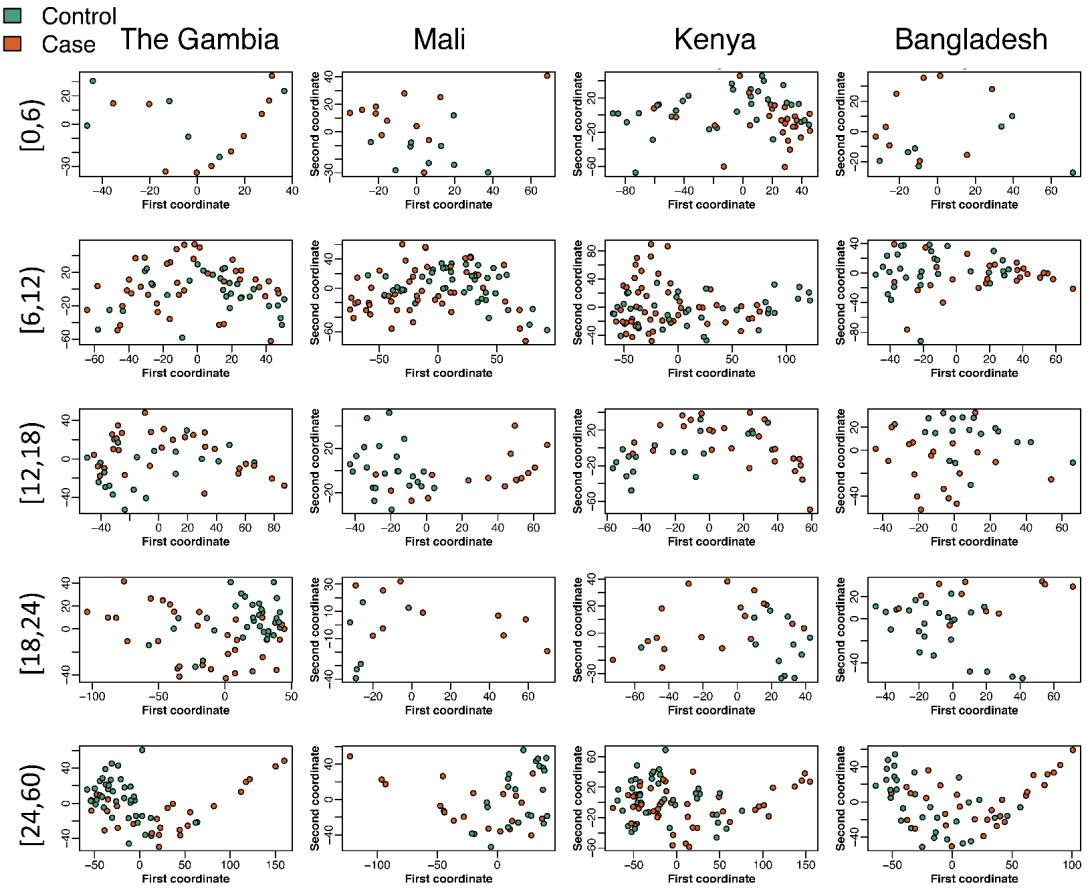

Figure 21: PCA analysis fails to distinguish cases from controls even after stratifying by age and country, likely due to high inter-personal variability.

The plots represent the first two components from a principal component analysis of the OTUs found to be most strongly associated with MSD. The distance used was Euclidean, however other commonly used distances also yield similar plots (data not shown). The columns correspond to country in the order: Gambia, Mali, Kenya and Bangladesh. The rows correspond to age categories from youngest to oldest. Controls are colored green and cases are colored orange.

Ten OTUs were found to be positively associated with diarrhea by all statistical tests. The OTUs associated with MSD have high-similarity matches against database sequences from bacterial taxa in the *Escherichia/Shigella*, *Granulicatella* spp., and *Streptococcus mitis/pneumonia* groups. When only abundance-dependent statistics are used to determine significance, an additional 18 OTUs are found to be highly associated with diarrhea, corresponding to the bacterial

species *Escherichia/Shigella*, *Campylobacter jejuni,* and *Streptococcus pasteurianus*. When only considering presence/absence statistics, 43 additional OTUs are found to be associated with diarrhea, comprising the bacterial groups already discussed above as well as members of the genera *Lactobacillus*, *Neisseria*, *Citrobacter*, *Erwinia,* and *Haemophilus*. It is noteworthy that all of these organisms are either facultatively anaerobic or microaerophilic.

On the other hand, there were no OTUs positively associated with healthy stools by both statistical methods, reflecting the higher degree of inter-individual variation in microbiota content in healthy individuals. Considering only presence/absence statistics, there are 43 OTUs associated with non-diarrheal control samples. The genera associated with these control samples include members of the clostridial families *Peptostreptococcaceae*, *Eubacteriaceae,* and *Erysipelotrichaceae*, and the genera *Clostridium sensu stricto*, *Dialister*, *Enterococcus*, *Prevotella*, *Ruminococcus,* and *Turicibacter*. When considering only abundance statistics, an additional 19 OTUs are significantly associated with non-diarrhea samples and have high quality matches to database sequences corresponding to *Bacteroides fragilis*, *Dialister*, *Megasphaera*, *Mitsuokella/Selenomonas*, *Prevotella* spp., and *Clostridium difficile*. Thus, it can be seen that many obligate anaerobic bacterial lineages correlate with healthy status.

3.3.4 Functional differences between cases and controls

The broad statements made above about oxygen tolerance in the diseased microbiota are supported by PICRUST [67] analyses of our data. Specifically, this showed putative signatures of obligate anaerobic gut lineages to be enriched in the

diarrhea-free samples (for example, glycolysis, $P = 10^{-9}$; pyruvate metabolism, $P = 10^{-7}$; short chain fatty acid biosynthesis, $P = 10^{-3}$; xylene degradation, $P = 10^{-7}$; and so on; all P values by Welch's t-test as computed by STAMP [68]), while oxygen dependent pathways (for example, the TCA cycle, $P < 10^{-15}$) are enriched in diseased samples.

3.3.5 Taxonomic groups correlated with dysentery

We segregated diarrheal stool based on diagnosis of dysentery (presence of blood) and found a total of 30 OTUs that were strongly correlated with dysentery when comparing with non-dysentery diarrheal stool (metagenomeSeq [33], $P < 0.05$). These include several well-known pathogens such as *Enterococcus faecalis*, *Campylobacter jejuni*, *Bacteroides fragilis*, *Clostridium perfringens*, *Enterobacter cancerogenus*, and members of the *Granulicatella*, *Haemophilus*, *Klebsiella,* and *Escherichia/Shigella* genera. Also associated with dysentery were members of the *Streptococcus pasteurianus* and *Streptococcus salivarius* groups. A single OTU, corresponding to *Lactobacillus ruminis*, was found to be negatively associated with dysentery. A genus-level representation of these findings is shown in Figure 22.

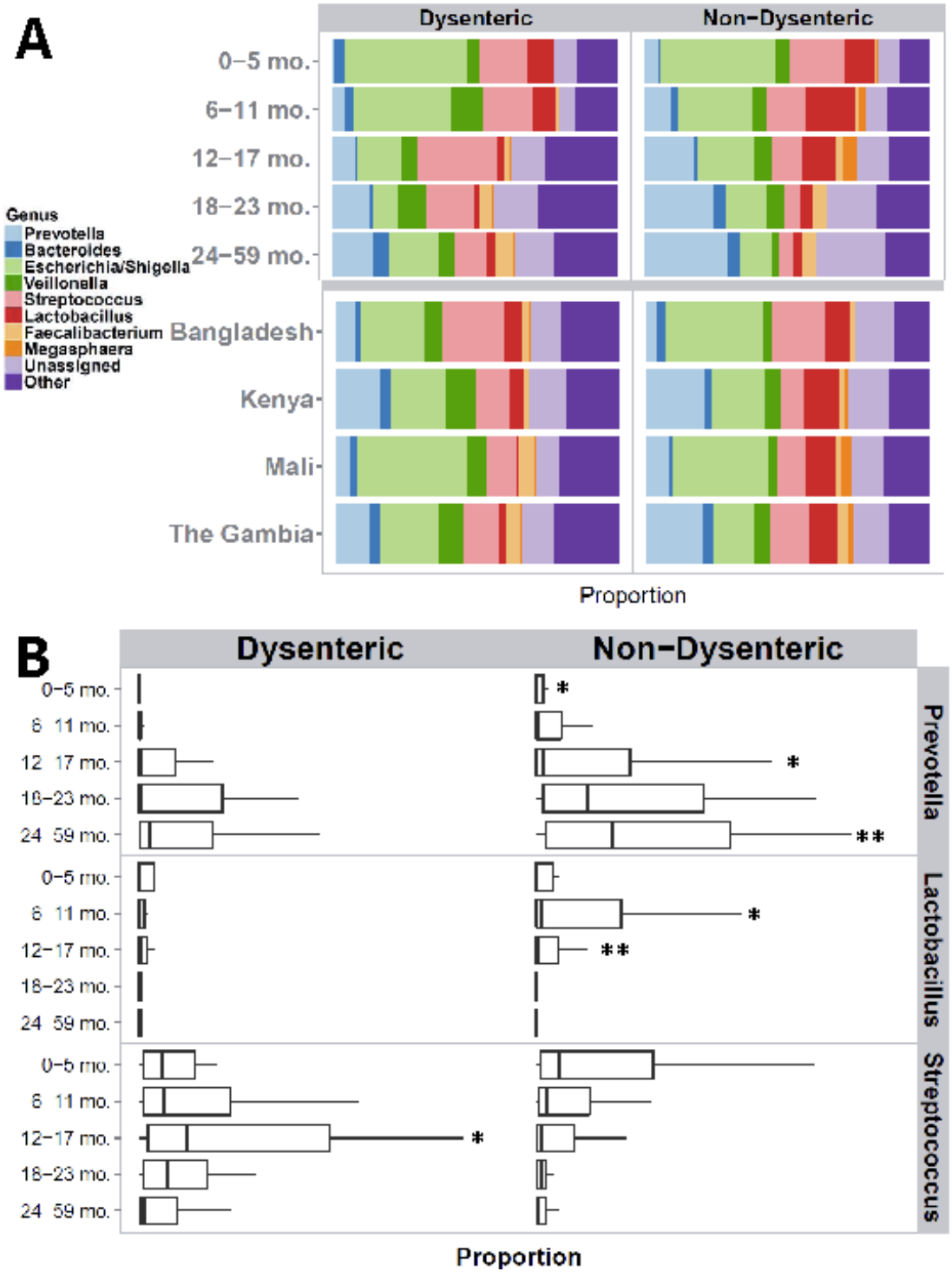

Figure 22: Comparison of dysenteric and non-dysenteric stool.

(A) Genus-level comparison of dysenteric and non-dysenteric diarrheal stool (top) stratified by age; (bottom) stratified by country. (B) Proportional abundance boxplots of *Prevotella, Lactobacillus*, and*Streptococcus* in dysenteric and non-dysenteric

diarrheal stools by age category. The upper whisker extends from the 75th percentile to the highest value that is within 1.5 * IQR of the hinge, where IQR is the inter-quartile range, or distance between the first and third quartiles. The lower whisker extends from the hinge to the lowest value within 1.5 * IQR of the hinge. Data beyond the end of the whiskers are outliers and are not plotted. Asterisks above the whisker indicate a statistically significant difference (by t-test) between dysenteric and non-dysenteric stools placed in the panel with the more abundant mean. A single asterisk indicates $P < 0.05$; double asterisks indicate $P < 0.01$. *Prevotella* is significantly associated with non-dysenteric cases overall ($P = 0.0003$) and in age groups 0 to 6 months ($P = 0.01$), 12 to 17 months ($P = 0.03$), and 24 to 59 months ($P = 0.001$). *Lactobacillus* is significantly associated with non-dysenteric cases overall ($P = 0.0002$) and in children 6 to 11 months (0.02) and 12 to 17 months ($P = 0.003$), while the genus *Streptococcus* is associated with dysentery overall ($P = 0.007$), particularly in children aged 12 to 17 months ($P = 0.01$).

3.3.6 Network view of diarrheal illness

The overall results presented above are also borne out in correlation networks constructed from the data (Figure **23**). At the broad level, in both MSD cases and controls, it can be seen that there tend to be negative correlations between facultative anaerobic lineages and obligate anaerobic lineages. The most obvious example is the negative correlation of the potentially protective *Prevotella* genus with that of potential pathogens such as *Escherichia/Shigella*. Similarly, there are also positive correlations within these two phenotypic subgroupings, such that obligate anaerobic genera such as *Prevotella*, *Roseburia,* and *Dialister* are correlated with each other, while facultative anaerobic or microaerophilic genera such as *Streptococcus*, *Lactobacillus*, *Escherichia/Shigella*, and other Proteobacteria are also correlated with each other. The diarrhea-free network appears to be more tightly connected than the diarrheal network, consistent with ecological theories that equate environment diversity and connectedness with ecosystem stability/health [69, 70]. At the same time, we would like to note that our data do not allow a reliable quantitative assessment of such phenomena due to the large level of inter-personal variation.

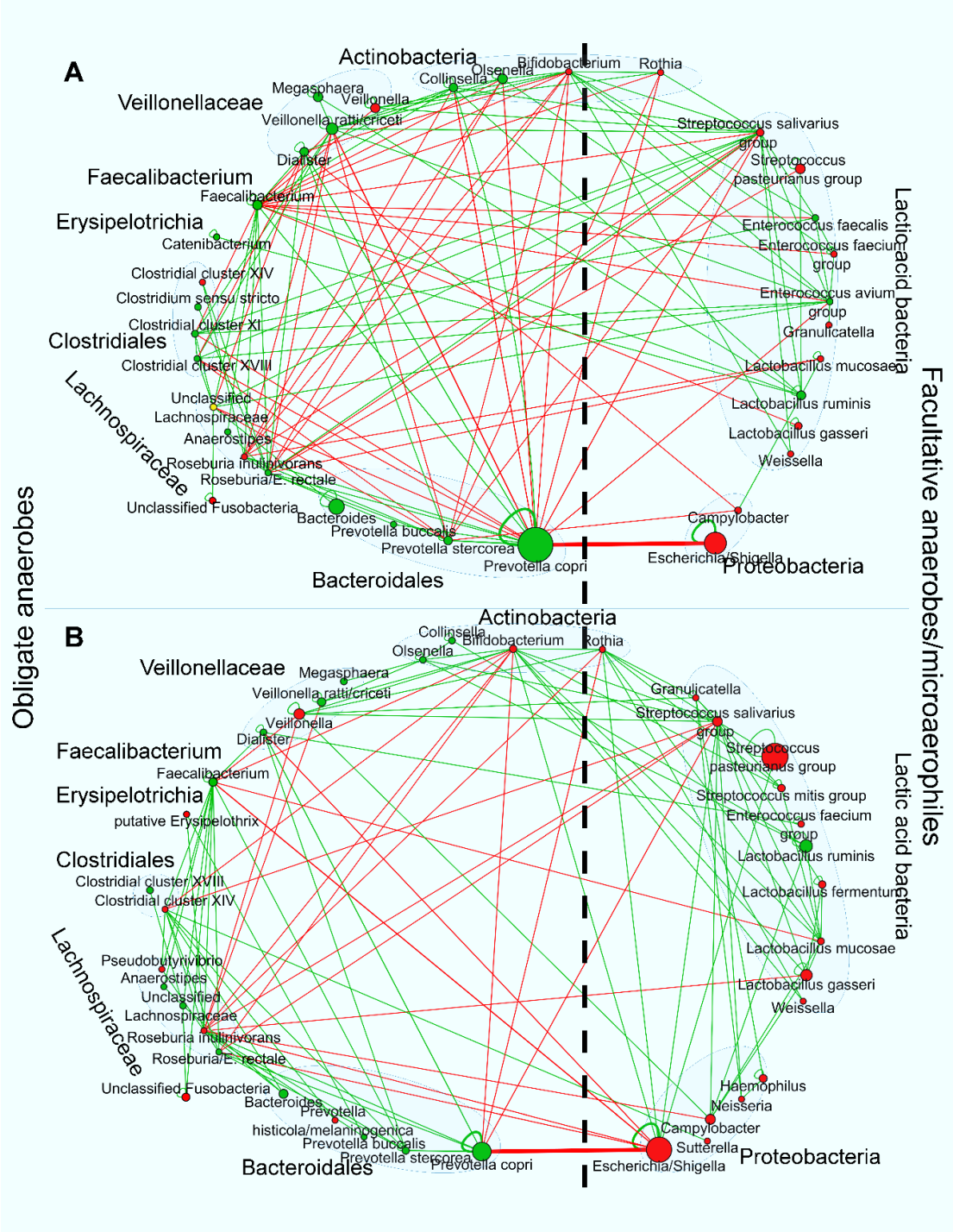

Figure 23: Correlation networks constructed on the controls (A) and on the MSD case (B) samples.

Each node represents a taxon, and each edge represents the presence of significant correlations between OTUs assigned to the corresponding taxa. The diameter of a node is proportional to the normalized abundance of the particular taxon. The edge width is proportional to the number of OTU-OTU correlations linking together the

corresponding taxa. Negative correlations are represented by red edges and positive correlations are represented by green edges. For clarity, only dominant genera and correlations shown (genera with > 500 normalized abundance, median SPARCC [17] correlation > 0.09). Broad taxonomic groups are highlighted through shaded areas. The broken line bisects the figure into obligate anaerobic lineages (left hand side) and facultatively anaerobic/microaerophilic lineages (right hand side).

3.3.7 Discussion

Our analysis of the 16S rRNA gene-based taxonomic profile of diarrheal and control stool samples has demonstrated a strong association between acute diarrheal disease and the overall taxonomic composition of the stool microbiota in young children from the developing world. We have identified statistically significant disease associations with several organisms already implicated in diarrheal disease, such as members of the *Escherichia/Shigella* genus and *C. jejuni*. In addition, we have uncovered an association with diarrheal disease for several organisms not widely believed to cause this disease, such as *Streptococcus* and *Granulicatella*. Streptococcal OTUs associated with disease primarily belong to either the *Streptococcus pneumoniae/mitis* group (indistinguishable within the 16S rRNA gene regions targeted by our study), which contains several important human pathogens, or the *Streptococcus pasteurianus* group. These results merit further exploration as recent studies provide evidence of *Streptococcus*-related diarrheal cases [71, 72]. It is important to stress that pathogenicity is only one of many possible explanations for these findings and the organisms associated with disease status may also either: (1) usually inhabit the upper GI tract and become apparent in diarrheal stool due to dislodging and reduced transit time during disease; (2) thrive in disturbed gut environments; (3) may be better able to persist/resist dislodgement during a diarrheal purge; or (4) a combination of pathogens may cause disease in these

children [73]. Prior evidence certainly suggests that facultative anaerobes (many of which we find associated with diarrhea) tend to flourish in a variety of perturbed gut environments, possibly because the reducing power of the microbiota is affected by the loss of obligate anaerobes following perturbation [66]. Any causality would need to be demonstrated through further experimentation. At the same time, streptococci are also found in our study to be associated with more severe forms of diarrhea (dysentery), thereby strengthening the case for a possible causal connection. Despite uncertainty regarding the causes and effects of microbiota perturbations in the setting of MSD, dissecting the physiologic implications is warranted. For example, an increase in streptococcal or other species in the setting of diarrhea may confer or exacerbate diarrheal effects. *S. mutans* has recently been postulated to have a role in human enteritis. Our work represents an important first step in understanding the complex interaction between microbiota and diarrheal pathogens in developing country settings.

Our study has also revealed a high prevalence of members of the *Prevotella* genus (primarily *Prevotella copri*) in the stool of developing world children, as well as the negative correlation of this genus with disease. These organisms are prevalent in the developing world [54], yet are relatively poorly studied due to fairly low prevalence in the industrialized world [74]. Samples containing high proportions of members of the *Prevotella* genus also have higher overall bacterial diversity, potentially driven by the level of complex polysaccharides/starchy fiber in the diet. Recent evidence suggests that *Prevotella* spp. are particularly abundant in rural African children consuming a high fiber diet [75]. This is in stark contrast to

Western children, who typically have much higher abundances of *Bacteroides* spp., and very little *Prevotella*, a difference that is believed to be linked to diet [76].

Our co-occurrence network analyses (Figure **23**) and proportional abundance analysis suggest potential negative interactions between *Prevotella* and enteric pathogens, such as members of the *Escherichia/Shigella* genus, raising the possibility for the development of novel *Prevotella*-based therapeutic strategies. Another possible probiotic organism identified in our study is *Lactobacillus ruminis*. This organism was found to be associated with non-diarrheal stool and also with less severe forms of diarrhea when comparing diarrheal to dysenteric stool. Although the increase in frequency of these taxa in diarrhea could be due to shortened intestinal transit time, the difference in prevalence of *Lactobacillus* between cases of MSD and dysentery are less likely to represent this effect. *Lactobacillus ruminis* has immunomodulatory properties and has been previously suggested as a potential probiotic [77].

Among OTUs found associated with non-diarrheal stool are sequences classified as *Clostridum difficile*, a surprising finding given that this organism is a common cause of enteric disease, primarily in hospitalized elderly patients. However, although *C. difficile* can be an important pathogen, it is actually carried asymptomatically by around 60% of infants [78]. We also found a conflicting association of OTUs assigned as *Bacteroides fragilis* with both the diarrhea-free status and dysentery, a finding that can perhaps be explained by strain-to-strain variation. Enterotoxigenic *B. fragilis* strains are well characterized diarrheal agents in children [79] whereas, in contrast, non-toxigenic *B. fragilis* has been linked to anti-

inflammatory protective effects in mouse models [80]. It is therefore possible that different strains, which cannot be differentiated through 16S rRNA gene sequencing, might account for these opposing results.

Our study identified many sequences that do not have good matches against cultured organisms in current 16S rRNA gene databases. Many of these sequences only have high-quality matches to other uncultivated and uncharacterized intestinal microbes, highlighting the presence of a large reservoir of uncharacterized microbes in the intestinal tract of children within the developing world, as reported before [81]. Many of the unknown sequences appear to belong to obligate anaerobic lineages of the Firmicutes phylum, which are under-represented in culture collections compared to other intestinal dwelling groups such as *Bacteroides* and bifidobacteria. The prevalence of such 'unknown' sequences is higher in controls and several of these uncharacterized organisms exhibit strong associations with diarrhea-free samples, highlighting their potential role in the maintenance of a healthy gut microbiota, and suggesting the need for a better in-depth characterization of the gut microbiota of children within the developing world, complementing resources recently developed in Europe [16] and the US [38].

Our observations related to the microbial succession in the developing infant gut microbiota carry several caveats. A single sample was collected from each child at a single point in time, and we lack extensive data on prior history of diarrhea. While the data are suggestive of a progression in microbiota structure, monitoring of a birth cohort will be necessary to fully understand the progression of gut microbiota, and assess the impact of diarrhea (including, potentially, multiple episodes of

diarrhea) on this process. At a technical level, we would also note that the primer sets used in this study (targeting the V1-V2 hypervariable regions of the 16S rRNA gene) do not effectively amplify bifidobacteria [82, 83], known to be dominant members of the intestinal microbiota of breast-fed infants, but this bias is likely to be uniform between cases and controls. We purposefully selected a primer set better targeted towards bacterial groups containing known and potential pathogens, such as *Enterobacteriaceae*, to improve our chances of detecting novel pathogens at the cost of obtaining less information about the already well-established early dominance by bifidobacteria.

Our study revealed the limitations of existing molecular and bioinformatics approaches employed in a clinical setting for performing taxonomic surveys of stool samples. The use of the 16S rRNA gene, for example, does not afford a sufficient discrimination within taxonomic groups containing known or putative pathogens (*Escherichia/Shigella*, *Streptococcus*, and so on) indicating the pressing need for the development of new cost-effective and relatively unbiased molecular approaches [84] for increasing the resolution of epidemiological surveys such as ours. Relatedly, the accurate taxonomic assignment of sequences generated in studies such as ours is hampered by numerous errors in public databases and by the use of simplistic 'lowest common ancestor' heuristics by software tools faced with ambiguous taxonomic information. The results presented in this paper were obtained through the careful manual annotation of all the OTUs found to be associated with disease state. Finally, we had to develop a novel statistical method [33] for identifying disease association in order to appropriately address data rarefaction as well as to control for the high

inter-personal variability, a typical feature of the healthy gut microbiota [85], and other confounding factors.

3.3.8 Conclusions

Overall our study demonstrates that the major differences in the microbiota between diarrheal and normal stools are quantitative differences in the proportions of the most prevalent taxa. Such quantitative differences were also observed in our previous qPCR-based study where we found that 80% (1,665/2,072) of controls and 89% (1,307/1,461) of MSD cases had detectable levels of *Shigella*. Quantitative measurements of *Shigella* abundance were critical to assessing attributable risk [84]. Among the known causes of diarrhea (rotavirus, *Shigella*, *Cryptosporidium*, Enterotoxigenic *E. coli*, and so on) the attributable fraction of diarrhea in young children is estimated to be just 43% [61]. Our study provides initial evidence for the existence of novel pathogenic agents. The most likely candidates from our study are members of the *Enterobacteriaceae* and streptococci, taxa which already contain many known human pathogens. Further exploration of these organisms is necessary to better understand their pathogenic potential and the likelihood of their emergence as major pathogens through the acquisition of additional pathogenicity factors. Importantly, our study reveals a possible protective role against diarrhea for the *Prevotella* genus and *Lactobacillus ruminis*. Understanding such effect is important. For example, microbiological [86] or dietary [76] interventions may be possible in the supportive treatment of diarrhea in children similar to approaches used in the management of enteric infections in adults [86–88]. Further genomic and epidemiological studies are necessary to better characterize this genus and to assess

the potential development of diet- or microbiological-based therapeutics.

3.4 Materials

3.4.1 Study design and participants

Stool samples were selected from a large case/control study of moderate-to-severe diarrhea in children aged under 5 years [89]. Cases were enrolled upon presentation to a health clinic reporting MSD. MSD eligibility criteria included sunken eyes, loss of normal skin turgor, a decision to initiate intravenous hydration or to hospitalize the child, or the presence of blood in the stool. Controls were sought following case enrollment, sampled from a demographic surveillance database of the area. Individuals were excluded if they were unable to produce a sufficient amount of stool volume for testing or they were unable or unwilling to consent to involvement in the study. Every participant was consented prior to collection of their stool and their data. Consent was given by the caregiver (usually mother) because the patients are all children aged less than 5 years. All samples were collected between March of 2008 and June of 2009. One sample was collected for each child and no time-series analyses were conducted. The Institutional Review Boards (IRBs) at all cooperating institutions have reviewed and approved the protocol. The IRB Federal Wide Assurance numbers for all the sites are as follows: University of Maryland Baltimore FWA00007145, The Gambia, Medical Research Council Labs FWA 00006873, Kenya Medical Research Institute FWA 00002066, University of Mali Faculty of Medicine Pharmacy and Dentistry FWA 00001769, and International Centre for Diarrhoeal Disease Research, Bangladesh FWA 00001468. Further details on study design are described by Kotloff *et al.* [89].

3.4.2 Microbiology methods

Stool specimens were collected in sterile containers and examined within 24 h. Stools were stored at 2 to 8°C while in transit to the laboratory. Each fresh stool specimen was aliquoted into multiple tubes. All samples were analyzed by traditional microbiological tests for known bacterial, viral, and eukaryotic pathogens. Details of these methods can be found in Panchalingam *et al*. [90] DNA was isolated using a bead beater with 3 mm diameter solid glass beads (sigma Life Science), and subsequently 0.1 mm zirconium beads (BIO-SPEC Inc.) to disrupt cells. The cell slurry was then centrifuged at 16,000 *g* for 1 min, the supernatant removed and processed using the Qiagen QIAamp® DNA stool extraction kit. Extracted DNA was precipitated with 3 M sodium acetate and ethanol and the DNA shipped to the USA.

3.4.3 Amplification and sequencing

DNA was amplified using 'universal' primers targeting the V1-V2 region of the 16S rRNA gene (small subunit of the ribosome) in bacteria (338R (5'-CATGCTGCCTCCCGTAGGAGT-3' and 27 F (5'-AGAGTTTGATCCTGGCTCAG-3'). Both forward and reverse primers had a 5' portion specific for use with 454 FLX sequencing technology and the forward primers contained a barcode between the FLX and gene specific region, so that samples could be pooled to a multiplex level of 96 samples per instrument run.

3.4.4 Data availability

Sequencing data and sample metadata are available at the NCBI archive under project PRJNA234437.

Source code and documentation for the analysis pipeline are available at

GitHub: https://github.com/MihaiPop/GEMS-db.

Abundance table and metadata are available, in BIOM [91] format, at http://www.cbcb.umd.edu/datasets/gems-study-diarrheal-disease.

Additional information on the study as well as links to all resources outlined above are made available at http://www.cbcb.umd.edu/research/projects/GEMS-pathogen-discovery.

3.4.5 Analysis pipeline

The individual reads were filtered for quality using custom in-house scripts that perform the following checks suggested in Huse *et al.* [92]: (1) sequences containing any ambiguity codes (N) are removed; (2) sequences that were shorter than 75 cycles of the 454 instrument were removed (each cycle yields an average of 2.5 bp depending on the sequence composition); (3) sequences for which a barcode could not be identified were removed. These checks are similar to those that can be performed by Mothur [93]. The high quality sequences were separated into 992 sample-specific sets according to the multiplexing barcodes. Conservative OTUs were clustered using DNAclust [13] with parameters (-r 1) (99% identity radius) thus ensuring that the definition of an OTU is consistent across all samples. To obtain taxonomic identification, a representative sequence from each OTU was aligned to Ribosomal Database (RDP) [94] (rdp.cme.msu.edu, release 10.4) using blastn with long word length (-W 100) in order to only detect nearly identical sequences. Sequences without a nearly identical match to RDP (>100 bp perfect match and >97% identity, as defined by BLAST) were marked as being ‘unassigned’ and assigned an OTU identifier. The resulting data were organized into a collection of

tables at several taxonomic levels containing each taxonomic group as a row and each sample as a column.

We note that the clustering criteria we use (<2% divergence, including insertions and deletions) are more conservative than commonly used definitions of 'species-level' OTUs (<2% divergence excluding indels). We used conservative clustering because no universal cutoff applies to all organisms [46] and in order to avoid merging together organisms with potentially different phenotypes (for example, closely-related strains, see Figure **24** for an example in closely-related *Escherichia/Shigella* OTUs). Similar considerations have led to the development of specialized software for the analysis of vaginal 16S rRNA gene survey data [95]. Our approach provides a good tradeoff between mitigating the effect of errors and allowing an unbiased analysis of the data. Furthermore, an exploration of increasingly permissive clustering thresholds reveals that our conservative clustering strategy does not lose statistical power (see Figure **25** and Figure **26**). Chimera checking was performed with Uchime 4.2.40 [96].

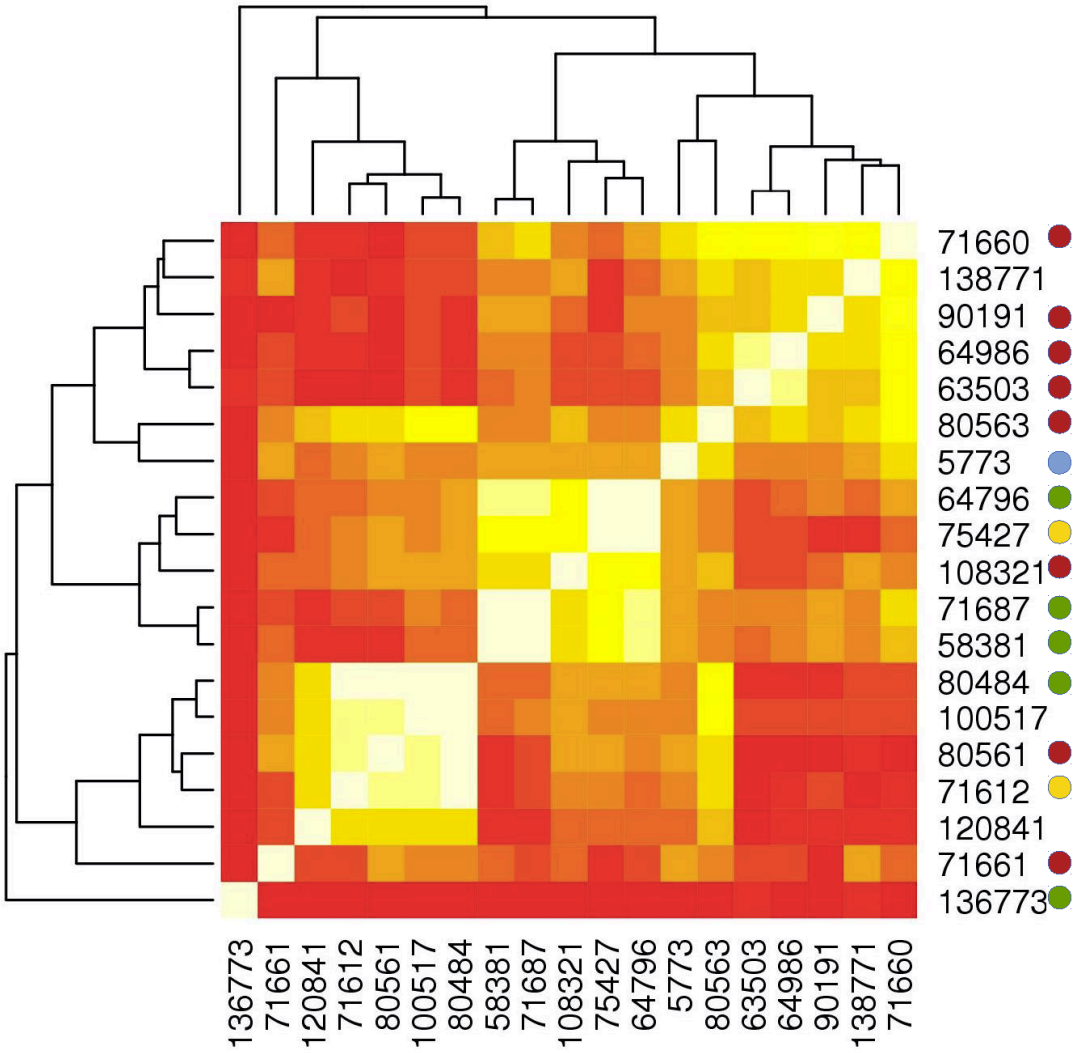

Figure 24: Loose clustering criteria may aggregate phenotypically distinct organisms.

Correlation between the abundance profiles of 19 different OTUs classified as *Escherichia/Shigella*. Despite high sequence similarity between the corresponding sequences, the individual OTUs (defined as DNAclust radius 0.01) have different abundance profiles, falling into roughly 6 distinct clusters, implying distributional or phenotypic differences between the corresponding organisms. The colored circles on the right represent the result of clustering the data at a looser radius of 0.02 (each cluster labeled by a different color). The phenotypically-defined and sequence-defined clusters are not fully concordant highlighting that sequence alone is not sufficient to define OTUs that are biologically relevant.

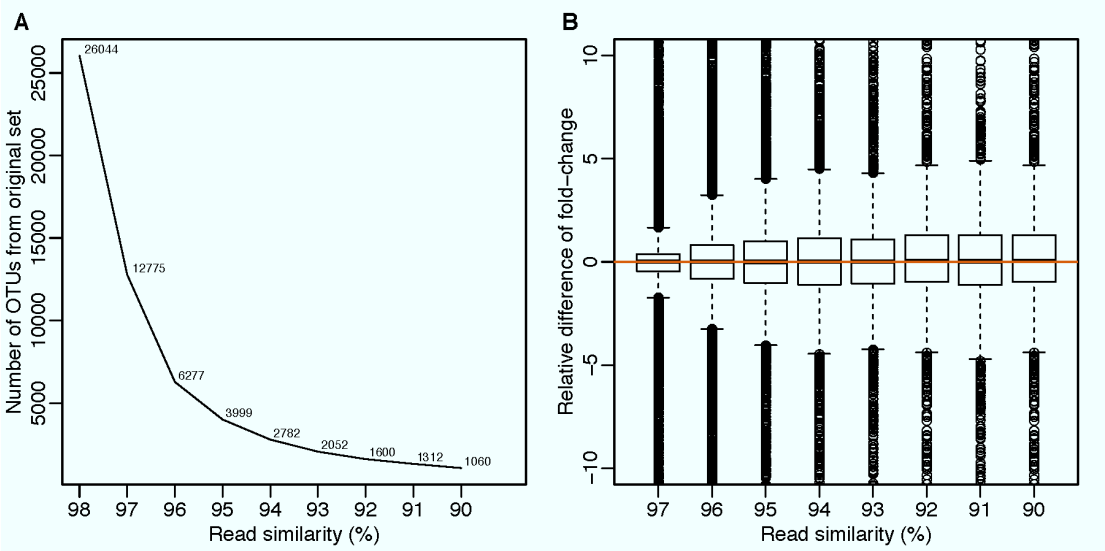

Figure 25: Differential abundance analysis is robust to changes in OTU radius.

(A) Approximate number of OTUs (see below) detected as sequence similarity cutoff is relaxed. As expected, the number of OTUs decreases with decreasing similarity stabilizing at approximately 95% similarity. (B) Relative difference between fold-changes (with respect to 98% identity) as sequence similarity cutoff is relaxed. On average, there is little effect to fold-change with an increase in radius size while there are increased number of outlier features (with bigger fold-change differences starting at 95% similarity). Approximate OTUs (A) were constructed from the set OTUs used in the main manuscript by re-clustering the representative reads for each OTU were with 97-90 % identity and log2 fold-changes for new OTU centers were tracked and compared to the original centers. Relative fold-change difference (B) of re-clustered OTUs is computed as $\frac{|fc_{98} - fc_{xx}|}{fc_{98}}$. Using 99% is necessary when attempting to distinguish species or even strains in a clinical dataset.

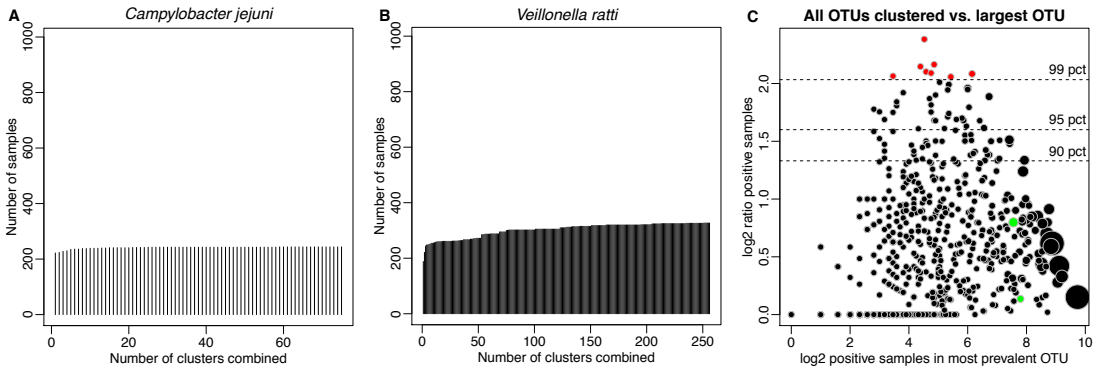

Figure 26: OTU prevalence is robust to OTU radius choice.

(A and B). Bar plot of the number of unique samples positive for *Campylobacter jejuni* (A) and *Veillonella ratti* (B) as OTU clusters annotated to each species are merged in order from most to least prevalent. For *Campylobacter jejuni* (A), known to be well-defined at the species level by 16S rRNA, prevalence at the OTU level (for the most prevalent cluster) is very close to prevalence at the species level. For *Veillonella ratti*, this ratio is smaller. (C) Ratio of OTU-level prevalence (for most prevalent OTU) to species-level prevalence for all species detected. The x-axis represents the $\log_2$ prevalence (number of positive samples) for the most prevalent OTU cluster in each annotated species in our dataset. The y-axis is the $\log_2$ ratio between the total prevalence of a species (sum of all OTUs) and the prevalence of the most abundant OTU. *Campyllobacter jejuni* (A) and *Veillonella ratti* (B) are shown as green points. Red points are those features in the top percentile (with largest ratio) where the species-level prevalence increase by at least 4.09 times. The size of each point corresponds to the number of OTUs with the same species-level annotation. For 82% of species this ratio is below 1, *i.e.*, aggregating over all OTUs labeled with the same species label would at most double the prevalence of a specific species. This result is consistent with the observation that, for the majority of species, prevalence at the OTU level captures species level prevalence as well. The extreme outliers are 8 organisms (red dots) classified as *Enterobacter sp. Y5*, *Prevotella intermedia*, *Prevotella sp. Oral clone BP1-16*, *Ruminococcus albus*, *Salmonella bongori*, *Shigella boydii*, *Streptococcus dysgalactiae*, and *Veillonella sp. MY-P9*.

3.4.6 PICRUST analysis

The most abundant 6879 OTUs were reprocessed using QIIME [11] version 1.8.0-dev as recommended on the PICRUST website (specifically OTUs were constructed with the pick_closed_reference_otus.py script against the latest version (version 13.5) of the Greengenes [97] database) and the resulting information was processed with PICRUST [15] version 1.0.0-dev using the KEGG analysis module

and aggregating the results to level 3. The results were further explored with STAMP [68] version 2.0.2, using the two-group analysis module, focusing on known aerobic and anaerobic pathways.

3.4.7 Data normalization

In order to avoid the bias that may be introduced by preferential amplification or sequencing of specific sequences, we scaled the counts by the 56th percentile of the number of OTUs in each sample. The 56th percentile was empirically determined from the distribution of non-zero counts required to behave consistently across our samples. We normalized with a Cumulative Sum Scaling approach, which scales counts by dividing the sum of each sample's counts up to and including the p^{th} quantile (that is, for all samples j, $S_p = \sum_i (c_{ij} | c_{ij} \leq q_{pj})$, where q_{pj} is the p^{th} quantile of sample *j*). Normalized counts are then given by $\frac{c_{ij}}{s_{pj}} 1000$. This method constrains communities with respect to a total size, but does not place undue influence on features (OTUs) that are preferentially sampled. A full description of the methodology is provided in Paulson *et al*. [33].

3.4.8 Statistical approaches

To test for presence and absence of an organism we performed Fisher's test stratifying by positive and negative samples. Samples were stratified as positive for an organism if the sample had one or more sequences of the organism with a sample being negative if there was absence of sequences. The totals were calculated for each taxa, a minimum of 20 positive samples was required for a statistical test to be attempted. To correct for multiple comparisons we minimized the expected proportion of false positives following Benjamini and Hochberg [98].

Differential abundance was assessed with the package metagenomeSeq [33] - a statistical approach that models confounding such as age and country, and also the effect of under-sampling on the observed counts. Significant findings were reported for OTUs that satisfied the following criteria: (1) OTU was abundant (≥12 normalized counts per sample) in cases or controls; (2) OTU was prevalent (present in ≥10 cases and controls); (3) OTU had fold change or odds ratio exceeding 2 in either cases or controls; and (4) statistical association was significant (P <0.05) after Benjamini-Hochberg correction for multiple testing.

Analyses were performed using the R software package 3.0.1 and packages, Vegan 2.0-7 and metagenomeSeq 1.4.21.

3.4.9 Correlation network construction

Correlation networks were constructed separately on cases and controls to characterize the dependencies between 268 differentially abundant OTUs (Table S3). Each network was built using SparCC [55], a tool specifically developed for assessing the correlation structure within microbial communities. The statistical significance for each OTU-OTU-interaction was obtained with an empirical null distribution using 1,000 bootstrap iterations. The P values were further adjusted for multiple comparisons using the Benjamini and Hochberg [98] correction. All OTU-OTU-interactions with FDR<=0.05, were considered significant and were represented as edges in the network.

For simplicity of visual representation, OTUs were aggregated at genus or lower taxonomic levels using the median normalized abundance of the aggregated OTUs as the abundance of the corresponding taxonomic group. We omitted all

taxonomic groups with median abundance lower than 500 normalized counts, as well as all edges with SparCC correlation lower than 0.09. The plots were drawn in Cytoscape 3.0.1 [99].

4 Differential abundance analysis of WMS sequencing data

4.1 Background

4.1.1 Context

Following the analysis of marker-gene surveys in chapter 2, "Differential abundance analysis for marker-gene surveys", a logical extension is an investigation of the applicability of our methods to whole metagenomic shotgun sequencing (WMS) data. As described in the first chapter, whole metagenomic shotgun sequencing data is generated in a different manner from marker-gene surveys. Understanding the role sparsity and under-sampling plays in WMS datasets is useful for studies investigating genes associated with a phenotype and we characterize that sparsity in a number of WMS datasets. We developed a feature specific zero-inflated model that addresses limitations of the method developed earlier. This work is to be submitted soon for publication. It is joint work with Chris Hill, Mathieu Almeida, Héctor Corrada Bravo, and Mihai Pop.

4.1.2 Abstract

Associating bacteria with disease has predominantly been performed using targeted sequencing of the 16S marker-gene due to cost-effectiveness. However, whole metagenome shotgun sequencing (WMS) can generate entire bacterial gene content for a microbial community. Many of the characteristics, e.g. sparsity, found in marker-gene surveys are similarly present in WMS datasets. However, analyzing bacterial genes results in differential abundance testing of many more features.

We present an investigation of the sparsity in existing metagenomic data and highlight the impact of various counting strategies for generating gene count matrices

on sparsity and resulting gene count distributions. Additionally, we investigate how under-sampling results in fewer detected genes, thereby potentially biasing differential abundance estimates. We perform a novel experiment in analyzing the correlation between normalization factors and raw count abundances and the effect sparsity has on normalization scaling methods as well as commonly used differential abundance statistics. Finally, we demonstrate a new approach by estimating a zero-inflated log-normal mixture model for each gene and highlight the usefulness in both WMS data as well as marker-gene surveys. We term this new statistical analysis approach metagenomeSeq2 and is available within our metagenomeSeq open-source software package through the Bioconductor project.

4.2 Overview of our analysis of whole metagenomic shotgun sequencing

Whole metagenomic shotgun assays generate abundances of a microbial communities' entire bacterial gene content. Associating microbial gene abundances in disease can highlight if particular organisms or potentially expressed functions are enriched or depleted in health or disease. Associating bacteria with disease has predominantly been performed using targeted sequencing of the 16S rRNA marker gene due to cost-effectiveness. However, whole metagenome shotgun sequencing (WMS) can generate entire bacterial gene content for a microbial community. Many of the characteristics, e.g. sparsity, found in marker-gene surveys are similarly present in WMS datasets.

An expansive statistical methodological literature exists on analyzing RNA-seq data and should be evaluated for metagenomic data. A number of key differences

between RNA-seq and metagenomic data necessitate an investigation of the applicability of RNA-seq statistical methods to metagenomic data. For RNA-seq data, genes are pre-defined and well annotated. An important aspect in analyzing RNA-seq data is properly accounting for biological variability in count models, and capturing a variance-mean dependency in the count data. In metagenomics there are many more bacterial genes in a (much larger) reference catalogue, many of which are unique to an individual contributing to greater sparsity levels. A more pronounced concern when performing association testing of metagenomic data is sparsity potentially due to under-sampling.

To demonstrate the extent of sparsity in whole metagenomic datasets we examine multiple mapping strategies and quantify the level of sparsity of each gene in each sample. In determining appropriate scaling normalization methods we perform an experiment based on the distribution of correlations between all raw gene abundances and scaling values in evaluating optimal sampling rate estimates.

To demonstrate the need for accounting for sparsity in association studies, we analyzed a number of whole metagenomic datasets and propose a zero-inflated log-normal parameterization as an improvement to a zero-inflated Gaussian previously proposed for marker-gene surveys [33]. The zero-inflated log-normal parameterization provides a number of advantages. The first advantage is the feature specific model for which we include parameters to capture the effect of under-sampling within the zero-inflated component. The second advantage is that we avoid an iterative expectation-maximization algorithm as proposed in Paulson, et al. [33]. Thirdly, in our results section we highlight a lower false positive rate, compared to

metagenomeSeq, based on previously published simulation settings [100]. Lastly, we compare results of the analysis of a Chinese gut microbiome comparing individuals with and without Type II diabetes to several standard methods used in the field, from marker-gene surveys (metagenomeSeq), metagenomic specific (TSS + Wilcoxon), and RNA-seq (DESeq), demonstrating a zero-inflated log-normal model performs the best.

4.3 Results

4.3.1 Differences in sequencing depth induce systematic variation in gene presence detection sensitivity

We analyzed three WMS studies, a Chinese T2D study, a European female T2D study and oral microbiome data from the HMP study. We observed high levels of sparsity, the number of zero valued samples or genes, in both within-sample and within-gene count distributions (Figure **27**). In all three WMS datasets analyzed, over 63% of the count matrix is zero after filtering (filtering procedure described in Methods). We observed that in the gut microbiome datasets, the median proportion of absent genes in a sample is at least 71%, and the median proportion of samples in which a gene is not detected is at least 79%. In the oral microbiome, these proportions are 61% and 73% respectively.

Less than 0.1% of genes were found in all Chinese or European samples (Figure **28**, Figure **29**) and 0.2% in all HMP oral microbiomes (Figure **30**). For example, in the count abundance matrix produced by Qin et al. only 393 genes out of 1.5 million were found present in all samples (Figure **29**). In these studies the maximum number of genes detected in a sample ranges from 600,000 to 800,000 or

less than 20% of the set of 1.5 million filtered genes.

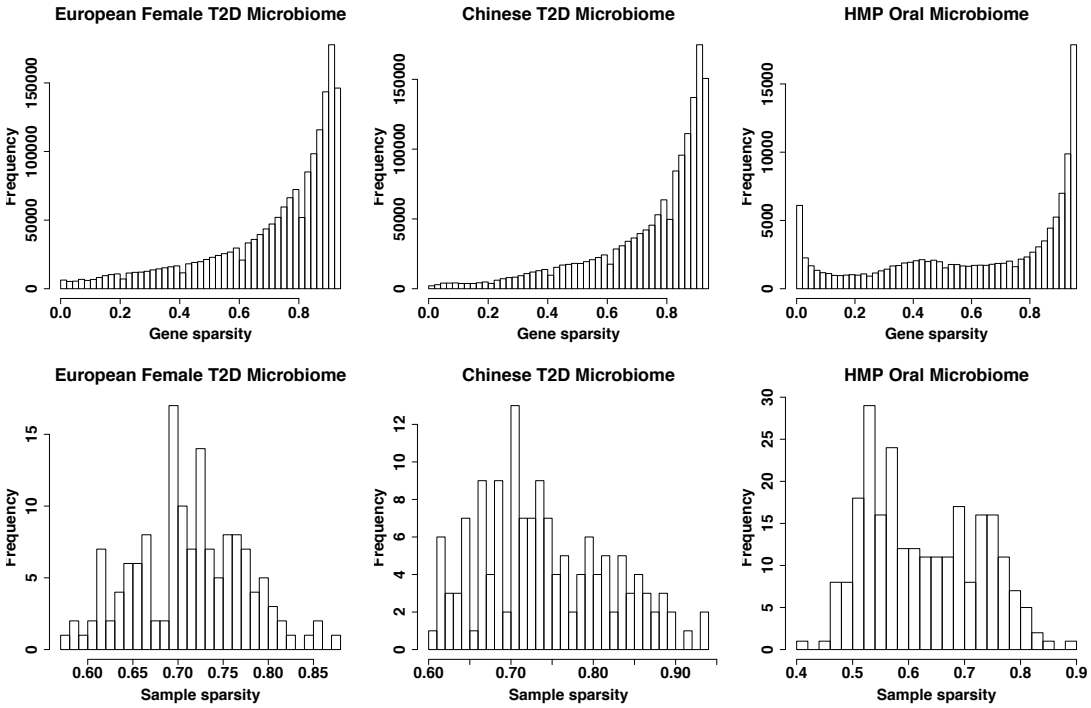

Figure 27: Gene and sample sparsity characteristics in two Type II diabetes microbiome datasets and the HMP oral microbiome.

Top: Histogram of the proportion of samples with zero abundance values for all genes after filtering in the European female T2D, Chinese T2D and HMP oral (supragingival plaque and tongue dorsum) datasets. Bottom: Histogram of the proportion of genes with zero abundance values within the samples after filtering in the European female T2D, Chinese T2D and HMP oral (supragingival plaque and tongue dorsum) datasets.

Despite large output of tens of millions of reads per sample, the observed high sparsity in WMS datasets may be due to under-sampling of these microbial communities. To explore this relationship, we analyzed the association between sample sequencing depth and number of genes detected in these datasets. There was a positive relationship between increasing sequencing depth and the number of detected genes (Figure **28**, Figure **29**, Figure **30**). The relationship was stronger when introducing a threshold for gene count (Figure **28**, Figure **29**). Rarefaction in the gut

microbiome studies shows no signs of saturation as sequencing depth increases. The supragingival plaque oral microbiome displayed a similar relationship between the number of detected genes and sequencing depth (Figure **30**). There was a saturation of the number of genes detected in the tongue dorsum suggesting that rarefaction effects are specific to community structure. In general, these results indicate that the sequencing depth for the samples analyzed is not sufficient for a comprehensive profiling of the microbial gene content.

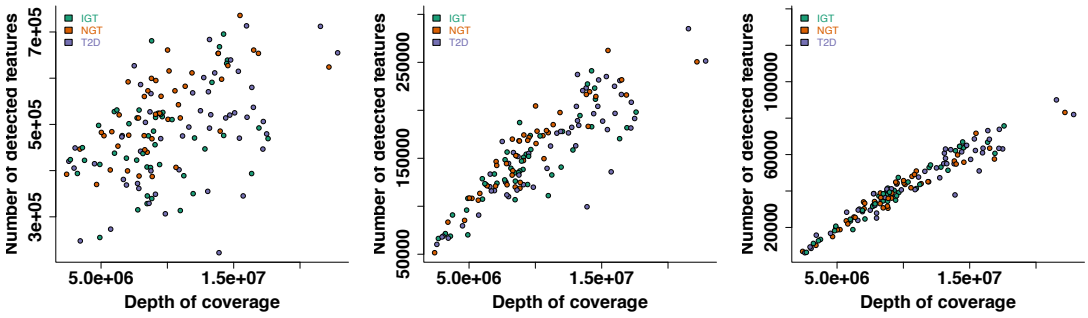

Figure 28: The number of genes detected in a sample depends on sequencing depth and phenotypic characteristics in the European T2D microbiome dataset.

The number of detected genes in a sample as a function of the sequencing depth for the European T2D microbiome dataset. Left) The given dataset whereas the middle and right assume a threshold of 10 and 50 reads mapped to a given gene, respectively. Sample colors follow, green are impaired glucose tolerance (IGT), red are normal glucose tolerance (NGT) and purple are Type II diabetic (T2D).

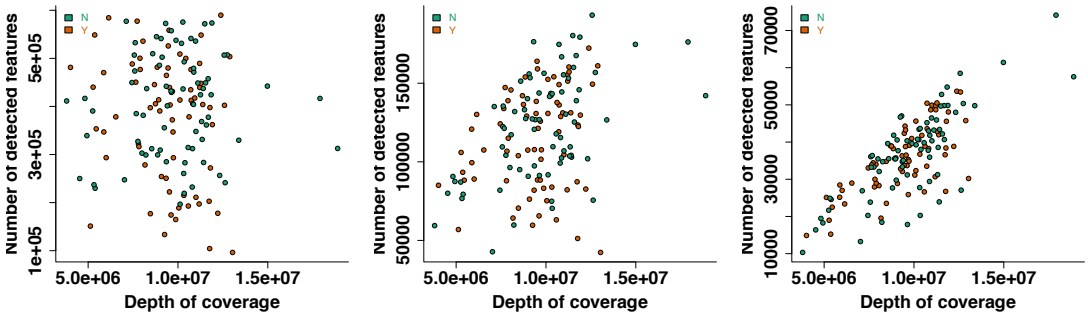

Figure 29: The number of genes detected in a sample depends on sequencing depth and phenotypic characteristics in Chinese microbiome dataset.

The number of detected genes in a sample as a function of the sequencing depth for the European microbiome dataset. Left is the given dataset, whereas the middle and right assumes a threshold of 10 and 50 reads mapped to a given gene, respectively. Color indicates sample status with green representing control and orange T2D sample.

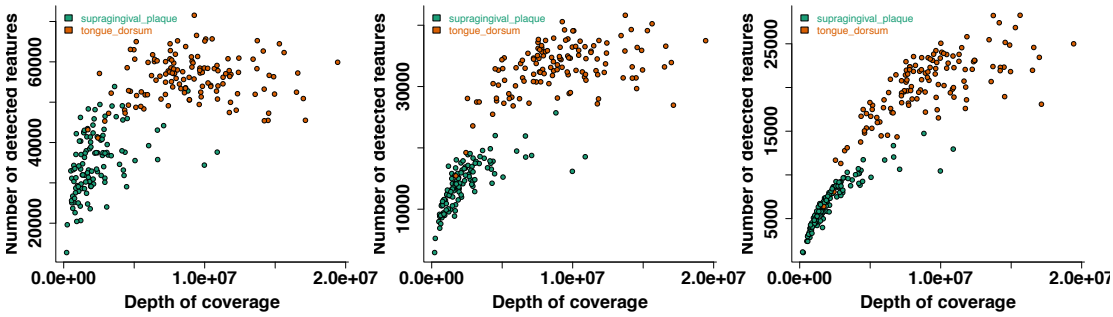

Figure 30: The number of genes detected in a sample depends on sequencing depth and phenotype in the HMP oral (supragingival plaque and tongue dorsum) microbiome.

The number of detected genes in a sample as a function of the sequencing depth for the HMP microbiome dataset. The left is the given dataset, whereas the middle and right assumes a threshold of 10 and 50 reads mapped to a given gene, respectively. Color is indicative of body site, supragingival plaque being green and tongue dorsum orange.

4.3.2 Sparsity persists in non-unique read mapping count procedures

Many WMS studies only consider uniquely mapped reads, ignoring multi-hit reads [5–7, 9, 101]. To ensure sparsity and under-sampling is not an artifact of read mapping, we examined the effect of alternate read mapping strategies. We defined count matrix summaries according to five different methods outlined below.

The first method, termed **unique**, counts reads that uniquely align to a particular gene. The second, **proportional**, assigns reads to a gene based on the proportion of *uniquely* aligned reads relative to other genes. The third method, **all**, assigns a count to every mapping even if ambiguous. The fourth method, **random**, assigns a count to a random gene from among the possible matches. The final method, **fractional**, evenly splits a single count amongst the genes to which it aligns.

For greater detail see Methods. Other potential counting strategies exist, including, counting with an expectation-maximization algorithm [42], but will have an intermediate sparsity between unique and all and were therefore excluded.

Regardless of read mapping method, most genes were absent from the majority of samples. Assigning a fractional or full count value to every single gene mapped by a read generated count datasets with the least sparsity. These methods still produced sparse matrices and resulted in a median sample sparsity (proportion of zero valued genes in a sample) greater than 58% and gene sparsity (proportion of zero valued samples for a gene) greater than 67%.

4.3.3 Sparsity adversely affects scaling normalization methods

Count normalization is an essential part of any statistical analysis of WMS datasets. Normalization is essential in accounting for variability induced by differences in sequencing depth across samples. The majority of metagenomic studies normalize count matrices by scaling counts to proportions and assuming sequencing rates scale all gene abundances uniformly. For example, doubling the number of sequencing reads should give twice the gene abundance for all genes.

To evaluate this assumption, we analyzed the count distributions in the Qin et al. dataset as a function of sequencing depth using quantile-quantile (QQ) plots. Under the previous assumption, a QQ plot of two sample count distributions would follow a straight line, with the slope determined by the ratio of sequencing depths. However, we observed deviations from this assumption primarily at the tails of the distribution when comparing average count distributions of samples with different sequencing depths (Figure **31**). The highest quantiles deviate significantly from the

uniform scaling ratio, implying possible sequencing artifacts. This implies that scaling normalization is not appropriate. Scaling by sequencing depth to convert counts to proportions is particularly problematic as departures from the assumptions behind scaling normalization are pronounced in higher quantiles, implying sequencing depth is a poor estimate of the sampling rate.

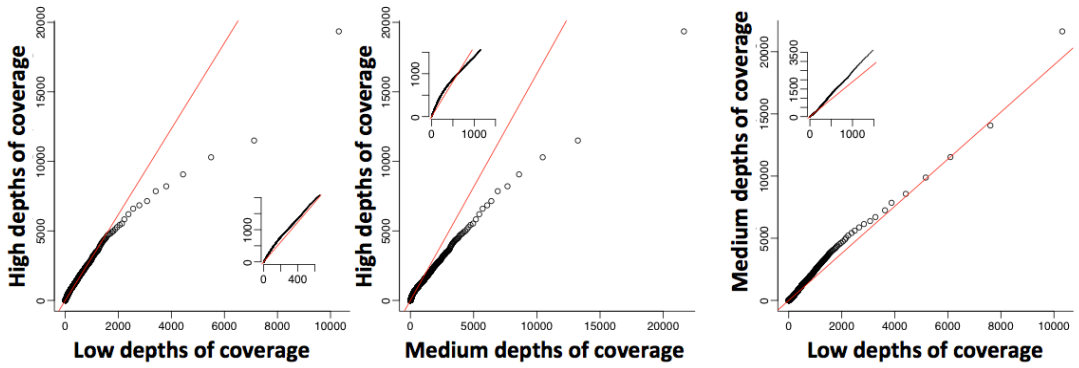

Figure 31: QQ plots of average count distributions.

Each figure is a quantile-quantile plot of the average positive count distribution for various groupings of samples of various depths of coverage. Each subplot consists of the same QQ plot with magnified axes revealing otherwise unobservable elements. A) is the averaged count distribution of samples with large sequencing depths (y-axis) vs. those with low sequencing depths. B) is the averaged count distribution of samples with large sequencing depths (y-axis) vs. those with medium sequencing depths. C) is the averaged count distribution of samples with medium sequencing depths (y-axis) vs. those with low sequencing depths.

A number of scaling normalization techniques have been proposed to address this issue and account for departures from the assumptions: cumulative sum scaling [33], DESeq [23], TMM [24], upper quartile [17]. We next sought to characterize the performance of each of these methods (along with the standard proportion method Total Sum Scaling, TSS). Each normalization value calculated by the various methods is an estimate of the sampling rate used to scale counts and make samples comparable despite variable sampling depths. Under the assumption of scaling

normalization, estimates of the sampling rate would show positive correlation with genes that are not differentially abundant across samples, since differences in abundance in these genes is determined strictly by sampling rate. We computed the correlation between sampling rate estimates derived by each normalization method and gene abundance across samples and analyzed the resulting distributions of correlations (Figure **32**, Figure **33**).

Here we report the results using Spearman correlation, however similar results are obtained for Pearson correlation. The median Spearman correlation for TSS was 0.01, for TMM and upper quartile they were -0.09 and -0.11 respectively. The sampling rate estimates showing highest median correlation were DESeq (0.20) and CSS (0.18, using the median, and 0.16, using a data adaptive quantile). We observed that methods with similar correlations to the gene abundance distributions yield similar estimates of sampling rate despite very different estimation approaches and assumptions as described in the first chapter (Figure **34**).

The sparsity of a gene, whether due to biological condition or under-sampling, presents a challenge in this experiment because of the difficulty in estimating a meaningful correlation due to missing data. However, we can analyze the distribution of correlations at multiple sparsity levels for each gene. We observed that all normalization techniques were strongly biased by the sparsity of a gene (Figure **35**, Figure **36**). The median correlation between sequencing depth (TSS) and gene abundances was centered close to zero for all genes, except a slight increase for Pearson correlations for non-sparse genes (Figure **32**A, Figure **33**A, Figure **35**, Figure **36**). The distributions of the correlation between TMM and the UQ scaling factors

with gene abundances were poorly correlated with the majority of gene abundances with the exception for a small number of highly present genes (Figure **32**BC, Figure **33**BC, Figure **35**, Figure **36**). CSS and DESeq's scaling factors showed consistently higher correlation across sparsity levels, with increasing correlation for highly present genes (Figure **32**DEF, Figure **33**DEF, Figure **35**, Figure **36**). In summary, while scaling normalization methods are not optimal, DESeq and CSS normalization consistently outperform other scaling normalization methods and should be used as part of standard data analysis pipelines.

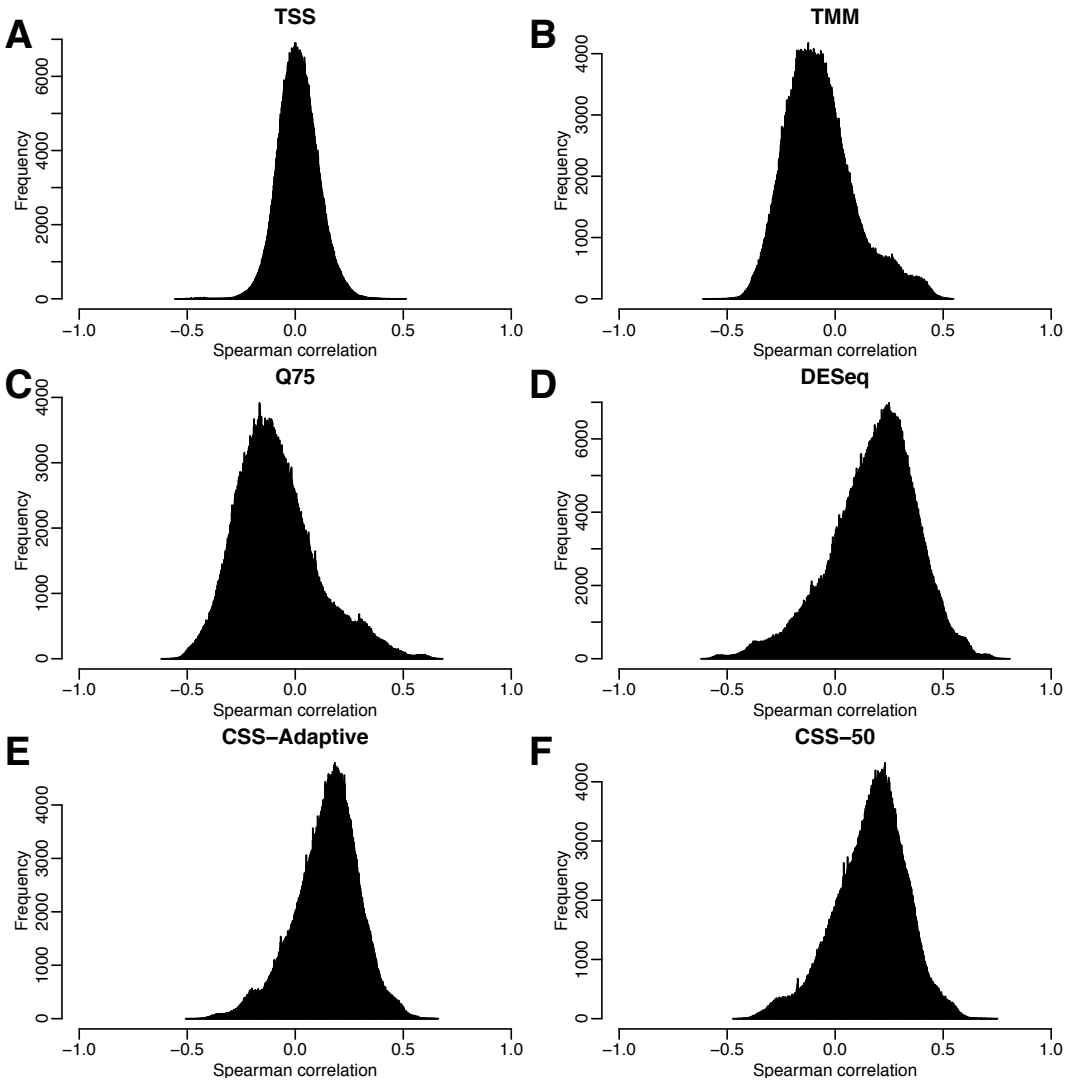

Figure 32: Histograms of Spearman correlations between normalization factors and raw abundances.

Each plot is a histogram of the Spearman correlation of feature abundances with the normalization scaling factor for (top to bottom, left to right) TSS, DESeq, TMM, upper-quartile, CSS adaptive, and CSS 50th quantile.

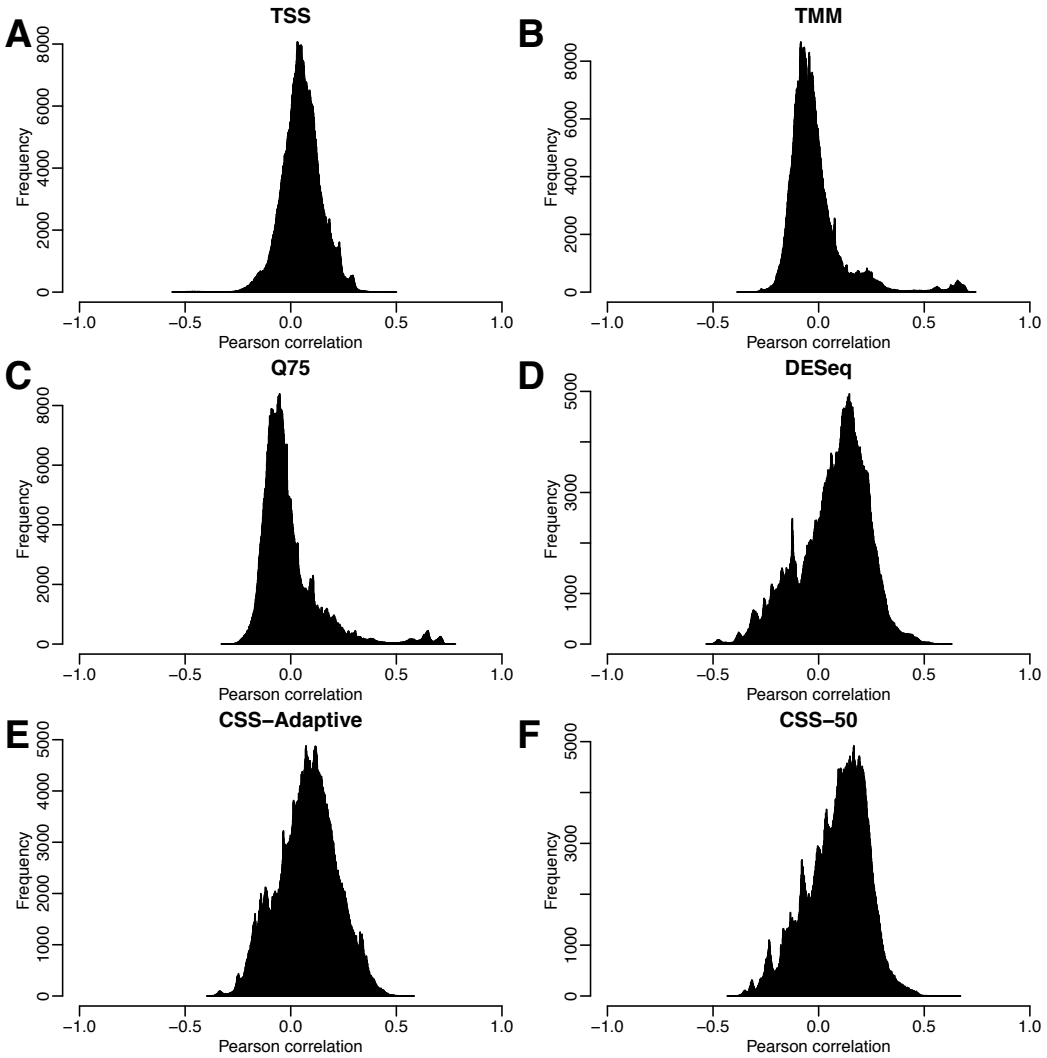

Figure 33: Histogram of Pearson correlations of normalization factors and feature abundances.

Each plot is a histogram of the Pearson correlation of feature abundances with the normalization scaling factor for (top to bottom, left to right) TSS, DESeq, TMM, upper-quartile, CSS adaptive, and CSS 50th quantile.

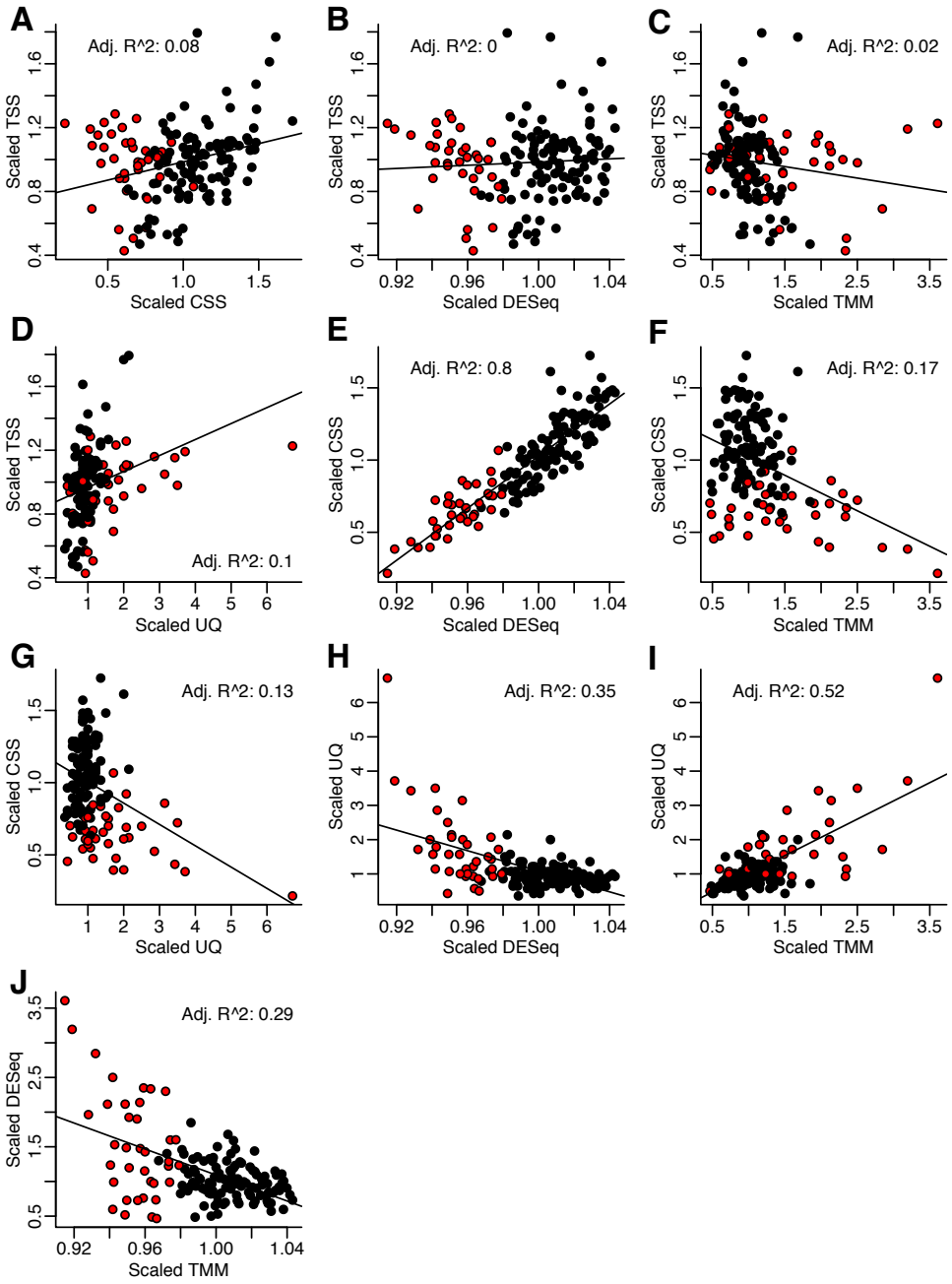

Figure 34: Pairwise comparison of normalization methods.

Scatter plots comparing the normalization factors for each method per sample. For each normalization scheme sample normalization factors were calculated and scaled by the median. Y-axis is the normalization factor described on the left and X-axis is the normalization factor described on the bottom. Red samples are in the top quarter of sample sparsity. Top to bottom, left to right: TSS vs. CSS, DESeq, TMM, UQ; CSS vs. DESeq, TMM, UQ; UQ vs. DESeq, TMM; DESeq vs. TMM.

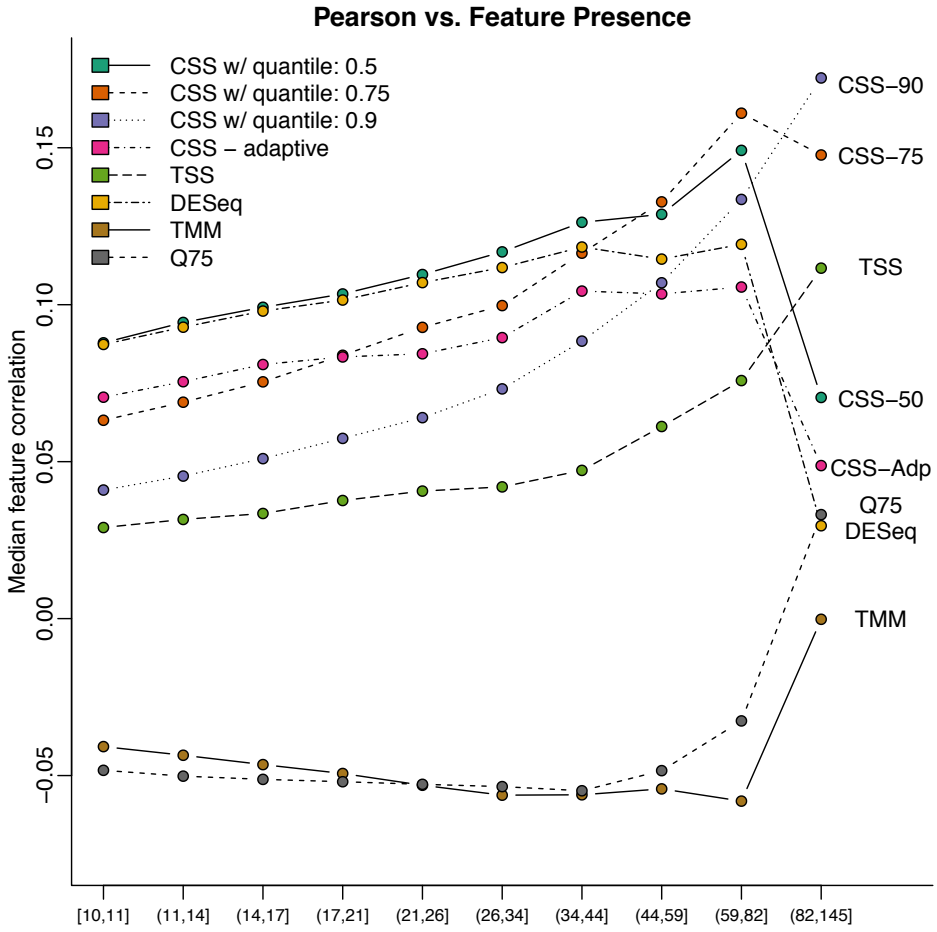

Figure 35: Median Pearson correlations of normalization scaling factors as a function of sparsity.

Median Pearson correlation between gene abundance and sampling rate estimates.

Genes were stratified by sample presence.

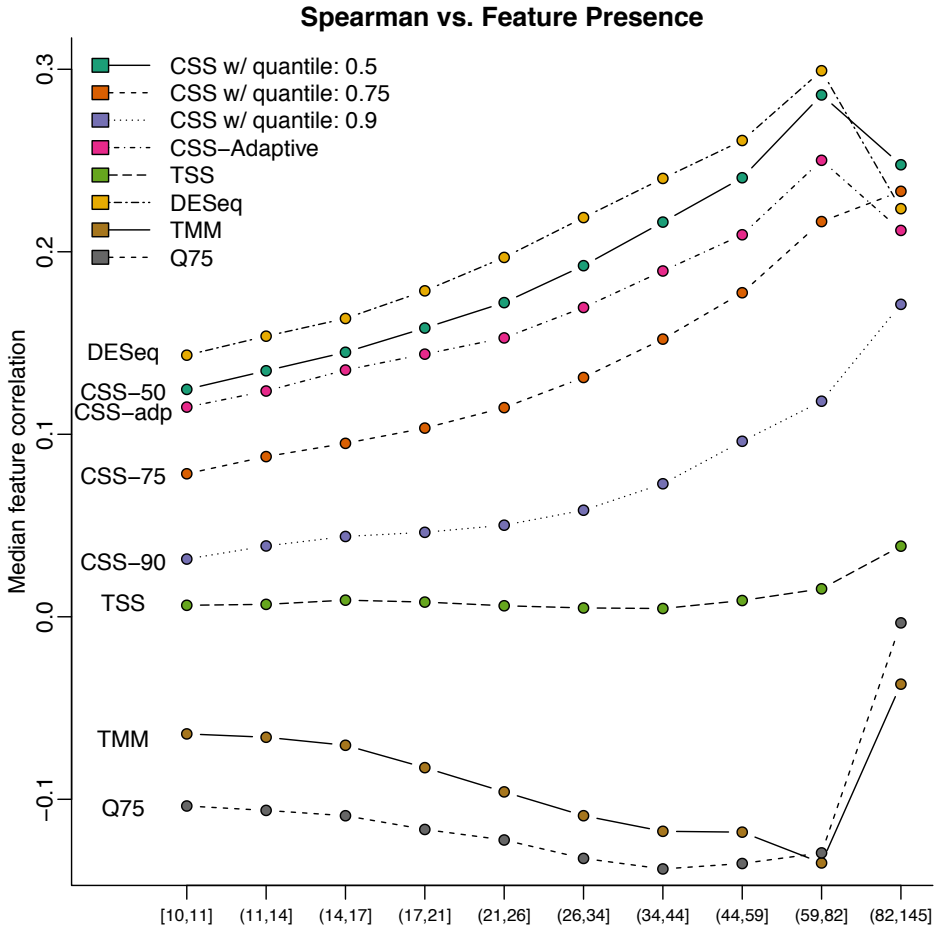

Figure 36: Median Spearman correlation of normalization factors as a function of sparsity.

Median Spearman correlation of evenly binned features by sample presence for each normalization method.

4.4 Feature-specific zero-inflated model for differential abundance testing

To account for the sparsity in metagenomic data we present a zero-inflated log-normal mixture model specifically defined to account for sparsity due to under-sampling. Using the zero-inflated log-normal model we independently fit each gene, allowing for a feature specific zero-inflated model that can estimate the probability a sample has a positive gene count. We compare to other popular methods after describing the design and highlight improved false positive rates compared to other

methods in simulations developed by McMurdie et al. [100].

In describing our model we follow the conventions of Mills [102] and specify the model to include terms accounting for sampling rate. Following estimation, we perform an empirical Bayes shrinkage of parameter estimates to control for high false positive rates observed in our previous model [100]. In our method, fold-change estimates are similar to those produced with a zero-inflated Gaussian model. Additionally, there has been concern regarding the appropriate value for pseudo-counts for continuity correction [34, 103]. Using a zero-inflated log-normal model, the likelihood can be decomposed into a zero and a positive component. The positive component, a log-normal, is maximized on solely positive counts and does not require any continuity correction.

4.4.1 Two-part model

Count data is modeled from multiple populations (for simplicity two groups), each with n_0 and n_1 samples and with m features. We consider each feature independently. We define $k(i)$ as class membership.

A normalized count for sample j is denoted by Y_j. We model the counts as a mixture of two components, a zero-component, including parameters to capture the effect of under-sampling, and a log-normal component, representing a positive count distribution. We represent the vector of covariates for sample j as X_j and in the full set of samples as X. We represent an additional vector of covariates that potentially affect detection in the zero-component for sample j (for example sequencing rate for sample j) as Z_j and in the full set of samples as Z.

We denote the probability of observing a positive count as $p_j = P(Y_j > 0 \mid$

X_j, Z_j). We define two sets of parameters used in modeling each part of the mixture model. Let β represent the parameters used in modeling the probability of positive counts and θ representing parameters of the conditional distribution of positive counts.

Our zero-component is represented with the indicator function, $I(Y_j > 0)$, defined such that if count $Y_j > 0$, then $I(Y_j > 0) = 1$. To model the probability of observing a positive count, p_j, we use the logit link, $\text{logit}(p_j) = [X_j, Z_j]\beta$.

$$\log \frac{p_j}{1 - p_j} = \text{logit}(p_j) = [X_j, Z_j]\beta = \beta_0 + \beta_1 k(j) + \beta_2 \log S_j$$

Our β parameters include an overall mean, β_0, a group effect β_1, and parameter β_2 measuring the effect of each sample's sequencing rate, (S_j), on the probability of observing a positive count. This model is in contrast to the zero-inflated Gaussian mixture where we estimated posterior probabilities a zero belonged to the spike-mass at zero or the Gaussian. In the previous model we did not include case/control status and used a sample's depth of coverage in the zero component of the mixture.

The log-normal component defined on the conditional distribution of normalized positive counts is modeled as $LN_\theta(Y_j \mid Y_j > 0)$ where θ is the mean and variance parameters of a log-normal model, with

$\mu_j = E\big(\, log\big(Y_j\big) \mid Y_j > 0, X_j \,\big) = X_j b = b_0 + b_1 k(j)$ where b_1 is a group effect.

Our mixture probability density function is:

$$f(y_j) = (1 - p_j) I(y_j = 0) + p_j * \frac{1}{y_j \sqrt{2\pi\sigma^2}} exp(-\frac{(ln(y_j) - \mu_j)^2}{2\sigma^2})$$

And full likelihood:

$$L(\beta, b, \sigma^2) = L_1(\beta)L_2(b, \sigma^2)$$

$$= \left\{\prod_{Y_j=0}(1-p_j)\prod_{Y_j>0}p_j\right\}\left\{\prod_{Y_j>0}\frac{1}{y_j\sqrt{2\pi\sigma^2}}exp(-\frac{(ln(y_j)-\mu_j)^2}{2\sigma^2})\right\}$$

The result of the full likelihood being factorable is found in Duan et al. ([104]) and is a result of zero not being included in the support of the log-normal distribution.

4.4.2 Estimation of the zero-inflated log-normal model

In estimating our model we need to maximize the two components of the probability density function (pdf) using the likelihood above. Given the result provided in Duan et al. ([104]) we can separate the full pdf's likelihood $L(\beta, b, \sigma^2)$ into two separate components $L_1(\beta)L_2(b, \sigma^2)$ where the L_1 and L_2 are independent of each other.

The count distribution - the log-normal count distribution is directly computed on normalized counts and is implemented using limma (Smyth 2004). Denote the vector of positive counts as y_c and the design matrix for samples with positive counts as X_c. Maximum likelihood estimates for the log-normal model are given by $\hat{b} = (X_c'X_c)^{-1}X_c'log(y_c)'$. The maximum likelihood estimate of σ^2 is $\hat{\sigma}^2 = \frac{\sum_{Y_j>0}(log(y_j)-X_{cj}\hat{b})^2}{n_+-n_q}$. With $n_q - 1$ covariates included in the model, where n_+ is the number of samples with positive counts.

The zero component of the model is estimated using logistic regression and can be estimated using an optimization algorithm like Newton-Raphson as implemented in the glm R function [105]. In the situation where there are only zero

counts in a group the non-zero component will not have any positive counts to estimate the group effect parameter.

4.4.3 Estimating fold-changes

For the sake of exposition, we first show the derived estimates in the simplified model with no 'nuisance' parameters followed by the inclusion of parameters to capture the effect of under-sampling. In estimating the fold-change we discuss it in terms of a groups' overall mean of the form:

$$M_k = E(Y \mid k(j) = k) = P(Y > 0 \mid k(j) = k)E(Y \mid Y > 0, k(j) = k) = p_k \mu_k$$

given group $k \in \{0,1\}$.

The conditional mean is $\mu_k = E(Y \mid Y > 0, k(j) = k) = e^{b_0 + b_1 k + \frac{\sigma^2}{2}}$. The probability of a positive count is $p_k = \frac{e^{\beta_0 + \beta_1 k}}{1 + e^{\beta_0 + \beta_1 k}}$. For the full model we can rewrite each in sample specific manner: $\mu_j = e^{X_j b + \sigma^2/2}$ and $p_j = \text{logit}(\left[X_j, Z_j\right]\beta)$

As such, the overall mean is:

$$M_k = E(Y \mid k(j) = k) = e^{b_0 + b_1 k + \frac{\sigma^2}{2}} \frac{e^{\beta_0 + \beta_1 k}}{1 + e^{\beta_0 + \beta_1 k}}$$

And the fold-change is defined as the ratio of the means between two groups, in this context is:

$$fc = \frac{M_1}{M_0} = \frac{E(Y \mid k(j) = 1)}{E(Y \mid k(j) = 0)}$$

Where the fold-change estimate simplifies for the two-group model to:

$$fc(\beta, b) = \frac{M_1}{M_0} = \frac{E(Y \mid k(j) = 1)}{E(Y \mid k(j) = 0)} = \frac{e^{b_0 + b_1 + \frac{\sigma^2}{2}} p_1}{e^{b_0 + \frac{\sigma^2}{2}} p_0}$$

We can express the log fold-change as the effect of the positive counts plus an

adjustment factor:

$$\log(fc(\beta, b)) = b_1 + \log(\frac{p_1}{p_0})$$

When we solely include log-transformed sample sequencing rate, $log(S_j)$, in the zero component additional design matrix with parameter estimate β_2 we must calculate the marginal group means due to the changes in p_k. We marginalize over the empirical distribution of the adjustment factor, S_j [102]:

$$\begin{aligned} M_k &= E(Y \mid k(j) = k, S) \\ &= E_S(E(Y \mid k(j) = k)) \\ &= \frac{1}{n}\sum_{j=1}^{n} E\,(Y \mid k(j) = k, S_j) \\ &= \frac{1}{n}\sum_{j=1}^{n} P\,(Y > 0 \mid k(j) = k, S_j) E(Y \mid Y > 0, k(j) = k, S_j) \\ &= \frac{1}{n}\sum_{j=1}^{n} e^{b_0 + b_1 k + \frac{\sigma^2}{2}} \frac{e^{\beta_0 + \beta_1 k + \beta_2 \log(S_j)}}{1 + e^{\beta_0 + \beta_1 k + \beta_2 \log(S_j)}} \end{aligned}$$

As such, we can express the fold-change as:

$$\begin{aligned} fc(\beta, b) = \frac{M_1}{M_0} &= \frac{E(Y \mid k(j) = 1)}{E(Y \mid k(j) = 0)} \\ &= \frac{\frac{1}{n}\sum_{j=1}^{n} e^{b_0 + b_1 + \frac{\sigma^2}{2}} \frac{e^{\beta_0 + \beta_1 + \beta_2 \log(S_j)}}{1 + e^{\beta_0 + \beta_1 + \beta_2 \log(S_j)}}}{\frac{1}{n}\sum_{j=1}^{n} e^{b_0 + \frac{\sigma^2}{2}} \frac{e^{\beta_0 + \beta_2 \log(S_j)}}{1 + e^{\beta_0 + \beta_2 \log(S_j)}}} \\ &= e^{b_1} \frac{\frac{1}{n}\sum_{j=1}^{n} \frac{e^{\beta_0 + \beta_1 + \beta_2 \log(S_j)}}{1 + e^{\beta_0 + \beta_1 + \beta_2 \log(S_j)}}}{\frac{1}{n}\sum_{j=1}^{n} \frac{e^{\beta_0 + \beta_2 \log(S_j)}}{1 + e^{\beta_0 + \beta_2 \log(S_j)}}} \\ &= e^{b_1} \frac{\frac{1}{n}\sum_{j=1}^{n} p_{1j}}{\frac{1}{n}\sum_{j=1}^{n} p_{0j}} \end{aligned}$$

And the log fold-change as:

$$\log(fc(\beta, b, S_j)) = b_1 + \log\left(\frac{1}{n}\sum_{j=1}^{n}\frac{e^{\beta_0+\beta_1+\beta_2\log(S_j)}}{1+e^{\beta_0+\beta_1+\beta_2\log(S_j)}}\right) - \log(\frac{1}{n}\sum_{j=1}^{n}\frac{e^{\beta_0+\beta_2\log(S_j)}}{1+e^{\beta_0+\beta_2\log(S_j)}})$$

$$= b_1 + \log(\frac{1}{n}\sum_{j=1}^{n} p_{1j}) - \log(\frac{1}{n}\sum_{j=1}^{n} p_{0j})$$

$$= b_1 + \log(\frac{\frac{1}{n}\sum_{j=1}^{n} p_{1j}}{\frac{1}{n}\sum_{j=1}^{n} p_{0j}})$$

An estimator for the variance of the two-group model $\log(fc)$ is:

$$\hat{V}ar(\log(fc)) = \sigma^2(\frac{n_0\hat{p}_0 + n_1\hat{p}_1}{n_0\hat{p}_0 n_1\hat{p}_1}) + \frac{1-\hat{p}_0}{n_0 p_0} + \frac{1-\hat{p}_1}{n_1 p_1}$$

The full group model log fold-change has a standard error:

$$\hat{V}ar(\log(fc)) = n\Delta'_{\log(\hat{fc})}\,\Sigma_{\hat{\beta},\hat{b}}\,\Delta_{\log(\hat{fc})}$$

Where $\Delta_{\log(\hat{fc})}$ is the gradient of $\log(fc(\beta, b))$ and $\Sigma_{\hat{\beta},\hat{b}}$ is the Fisher Information matrix (expectation of the Hessian) [102].

4.4.4 Shrinkage of parameter estimates using empirical Bayes

We follow the same procedure as in limma (Smyth 2004) to shrink variance estimates. In this manner we shrink variances using the empirical Bayes method that assumes an inverse Chi-square prior for each variance estimate, σ_i^2 with mean s_0^2 and f_0 degrees of freedom. The posterior variance estimates are:

$$\hat{s}_i = \frac{f_0 s_0^2 + f_i s_i^2}{f_0 + f_i}$$

Using these moderated variances we then shrink the positive component's contribution to the log fold-change, b_1. Consider a normal-normal model with prior mean equal to μ.

$$Y_i = \theta_i + \varepsilon_i$$
$$\varepsilon \sim N(0, \sigma_i^2)$$
$$\theta_i \sim N(0, \tau^2)$$

where σ_i^2 is known and τ^2 is unknown. The posterior distribution $\theta_i \mid y_i \sim N(\lambda_i y_i, \lambda_i \sigma_i^2)$. And our posterior mean estimates are:

$$\hat{\theta}_i = (1 - \hat{\lambda}_i) y_i$$

Where λ_i is a shrinkage parameter leading to either greater shrinkage to zero of fold-change estimates when fold-change estimates are more homogeneous. We utilize the connection between ridge regression, also known as Tikhonov regularization, and the hierarchical model where we assume a 0 mean prior setting $\lambda_i = \sigma_i^2 / n_i \tau^2$ to be the ratio of the within and between variance parameter estimates [106]. In this manner, our shrunken parameter estimates are [107]:

$$\widehat{b_i}^{ridge} = argmin_b \{ \sum_j^N \left(y_{ij} - b_0 - \sum^p x_{ijp} \, b_p \right)^2 + \lambda_i \sum^p b_p^2$$

Estimation of the new ridge parameters becomes, on centered data after removing the intercept: $\widehat{b_i}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y_i$.

In a similar fashion we allow the user to shrink parameter estimates for the group effect in the zero component of the mixture model. The shrinkage adjustment factor, λ, is calculated as $1/(n\tau^2)$. We specifically do not shrink the intercept and coefficient in the zero model corresponding to the sampling rate as we want to allow for the zero-inflated model to fully account for sparsity due to low depths of coverage. Our software allows users to optionally shrink either parameters in the zero, positive or both components and provide a comparison of results on a number of marker-gene surveys below. By default we recommend shrinkage of the positive

component parameter estimates without shrinking the zero component parameter estimates.

4.4.5 Permutation based analysis

For small feature and sample sizes we have also implemented a permutation based method for significance testing. In avoiding the assumption of normality one can build an empirical null distribution of z-scores. In this manner, we can define our p-values as the proportion of permuted z-scores, z_i^{0b}, greater than or equal to the observed z_i: $p_i = \frac{\#\{\left|z_i^{0b}\right| \geq |z_i|, b=1,\ldots,B\}}{B}$. However, this is a computationally intensive feature for whole metagenomic shotgun studies. We provide this feature as a potential alternative for small sets of genes or small marker-gene surveys and plan to evaluate a comparison of the parametric and non-parametric approaches in the future. This procedure was in general too expensive in our previous model employing the EM algorithm.

4.4.6 Effects of fold-change estimation moderation

To investigate the effects of shrinkage we fitted four models on the Qin et al. dataset and a lung marker-gene survey [39] shrinking one of the mixture components, both or none. We observed that shrinkage on all datasets behaved similarly in reducing the difference in log fold-change as a function of OTU or gene presence. In controlling for sparsity, with fewer positive samples, greater shrinkage of the log fold-change estimates occurred in genes or OTUs with the least number of positive samples (Figure **37**).

We highlight the effect of the sampling rate parameters in Figure **38** revealing

a positive relationship between a sample's probability of having a positive count and its' sampling rate. In estimating the parameter that captured the effect of under-sampling, we observed that the majority of features had positive estimates Figure **38**A.

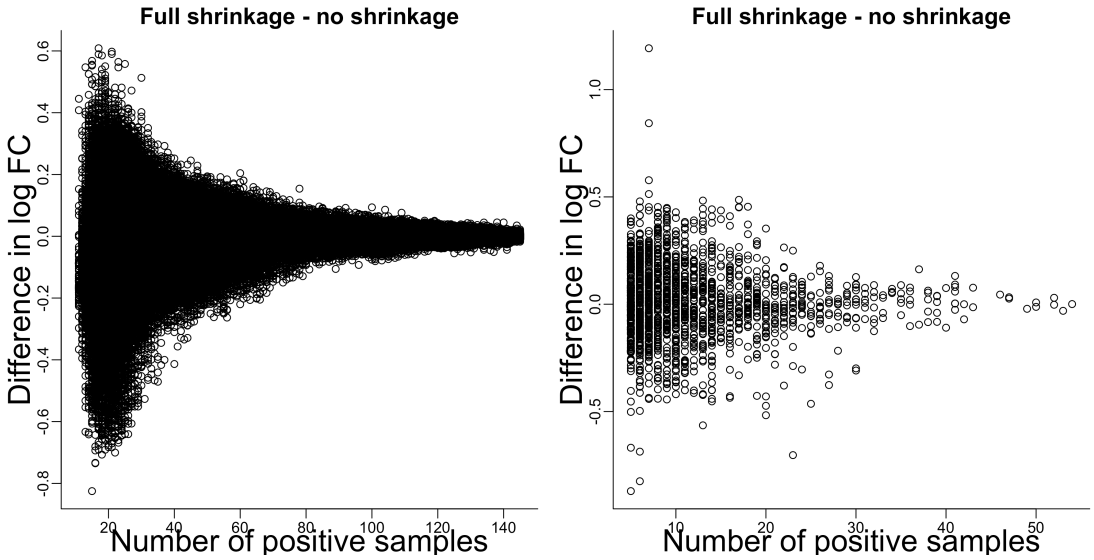

Figure 37: Greater effects of shrinkage on log fold-changes for sparse features.

Scatterplot of the difference in log fold-change estimates between fitting the zero-inflated model with and without shrinkage as a function of feature presence on A) Qin et al. data and B) a lung marker-gene survey. We observe that there is a greater moderation of log fold-change estimates for sparser features. The positive component of the mixture model over-estimates the fold-change and is moderated by shrinking.

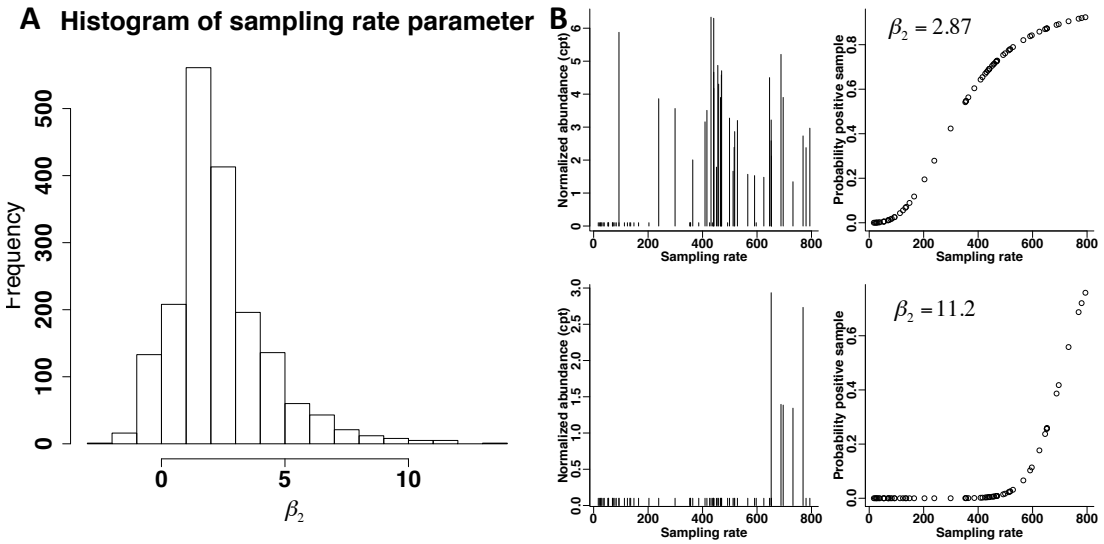

Figure 38: Sample rate parameter estimates have a positive effect on sampling on feature presence probability.

A) a histogram of the sample rate parameter estimates for the lung microbiome marker-gene survey. B) Left, plots of (y-axis) normalized abundance for a specific OTU versus (x-axis) sampling rate estimate using CSS. On the right, a sample's probability of a positive count that increases with sampling rate.

4.5 Zero-inflation models accounting for sequencing depth improve differential abundance testing in WMS datasets.

We investigated the effects of sparsity, as it relates to community under-sampling, on frequently used differential abundance testing methods, including, a zero-inflated log-normal mixture model (metagenomeSeq2), a zero-inflated Gaussian mixture model (metagenomeSeq), DESeq, and Wilcoxon rank-sum.

We first evaluated metagenomeSeq and metagenomeSeq2 on the Chinese T2D microbiome dataset. Utilizing the significant features produced from our new zero-inflated feature model we were able to achieve a Type II diabetes biomarker gene index with much higher AUROC scores compared to Qin et al. (See classification section).

A zero-inflated log-normal model, as described previously accounts for under-sampling and is flexible in including multiple covariates of interest. In comparing to the global model of [33] we observed that fold-change estimates are very similar, except for some shrinkage of larger fold-change estimates in the new model (Figure **39**). Comparing to the new model there were many more genes declared significant at 5% FDR (Figure **40**). In particular, 110,308 genes (26%) were considered significant at 5% FDR, suggesting a potentially higher false positive rate due to the moderated *t*-statistics (not shown). In contrast, the newer feature specific model, metagenomeSeq2, produced 4,418 genes (0.01%). Additionally, z-scores for our

resulting differential abundance test statistic are normally distributed and not biased by sparsity (Figure **41**).

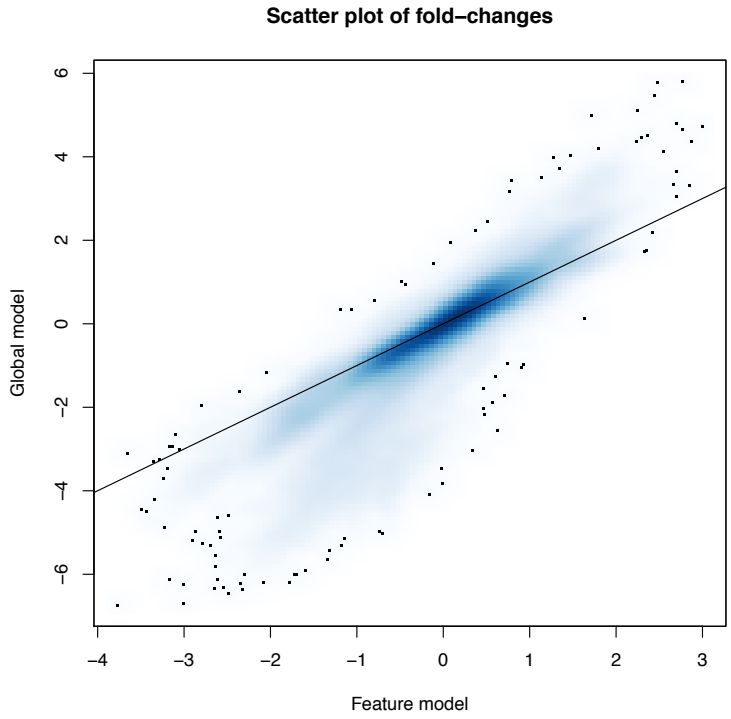

Figure 39: Scatter plot of metagenomeSeq and metagenomeSeq2 fold-change estimates.

The majority of metagenomeSeq and metagenomeSeq2 fold-change estimates are similar.

We next evaluated DESeq on the same dataset. A core challenge in RNA-seq data analysis is accounting for over-dispersion and a variance-mean dependency in count data. Accurate estimation of over-dispersion is critical to the sensitivity of differential expression analysis, especially for lowly expressed genes. However, variability in WMS experiments are atypical compared to RNA-seq experiments, but more similar to that observed in metagenomic marker-gene survey studies [33]. A major difficulty in metagenomic analysis is the sparsity innate to datasets.

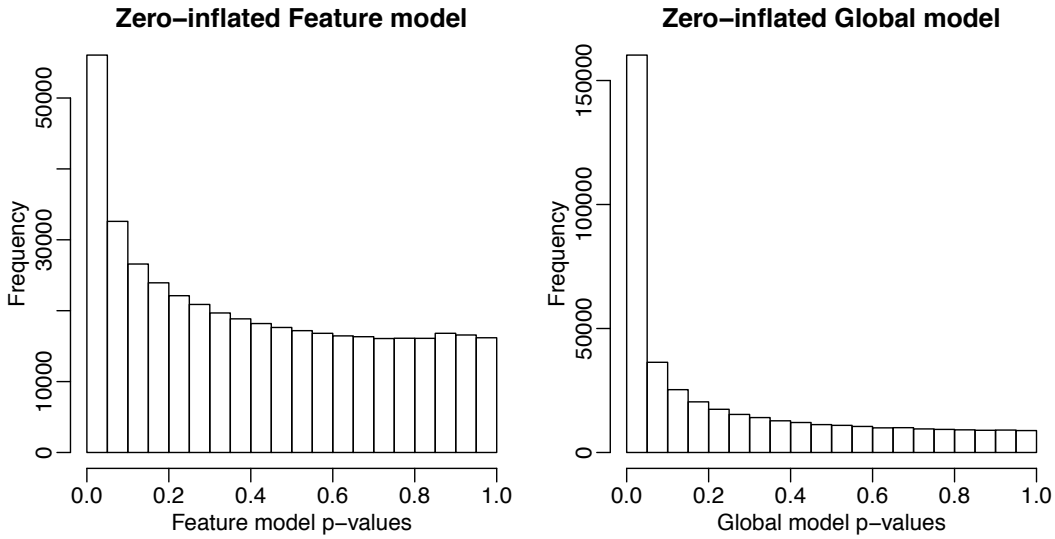

Figure 40: Feature model p-value distributions are more uniformly distributed.

Histograms of the p-values from two zero-inflated models. Left) A zero-inflated feature model and right) global model on the Qin et al. dataset.

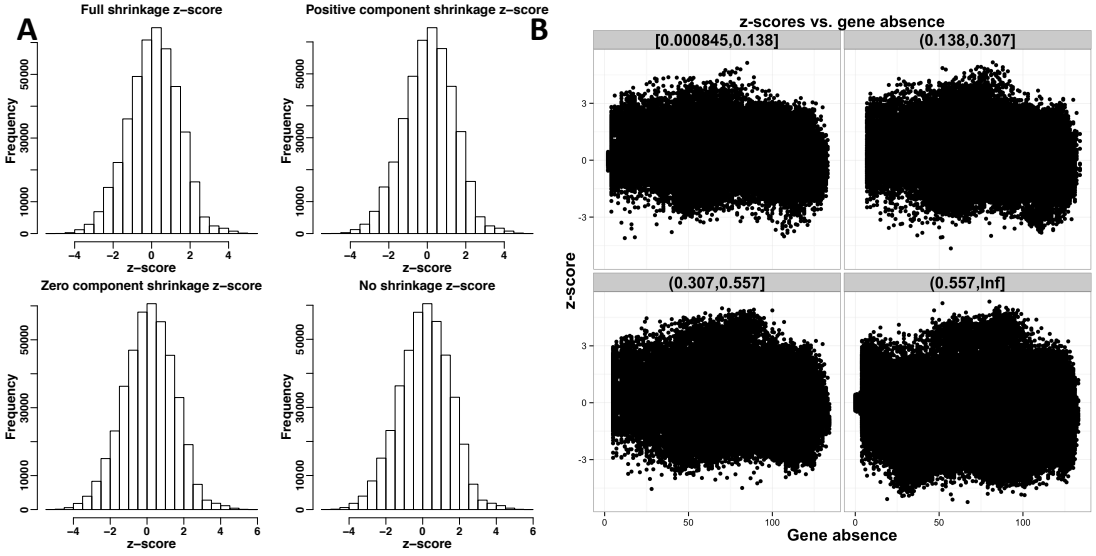

Figure 41: WMS fitted z-scores are normally distributed and not biased by gene presence.

Left) Histograms of the z-scores for various shrinkage methods. Right) Scatter plots of the zero-inflated log-normal mixture model z-scores as a function of feature sparsity faceted by equal bins of absolute log odds ratio.

Both WMS and marker-gene survey studies have high dispersion estimates correlated with increasing sparsity, regardless of normalized count means (Figure **42**).

This is not similar to the typical RNA-seq experiment. Because of the large dispersion estimates approximately 32% of the genes had fold-change estimates that did not converge. Considering only genes for which the negative binomial estimation algorithm converged, no genes were considered differentially abundant with an FDR less than 0.05.

Lastly, we discuss the advantages of the model proposed compared to a non-parametric Wilcoxon rank-sum test as used in Qin et al. [5]. With a zero-inflated log-normal framework it is possible to include potential confounders as covariates. Additionally, Wilcoxon rank-sum loses power with an increase of ties, including zeros, for which work has been developed to mitigate this loss in power [102, 108]. This loss of power is easily observed in the Qin et al. dataset with the strong relationship between the Wilcoxon rank-sum statistic and gene presence. As gene sparsity increases, the W-statistic shrank toward the null in contrast to our model (Figure **43**).

In analyzing the Qin et al. data there were considerable differences in reported genes and their associated p-values including the reporting of certain genes that did not pass our filtering criteria. Of the 52,484 genes reported in Qin et al. 32,847 had fewer than 35 samples with at least one count per million and were therefore filtered by us. Stage I samples were used in the study to report an initial set of genes with p-values less than 0.05 resulting in a reported 278,167 genes; however, reanalyzing the same data resulted in 7,668 fewer significant genes from the Stage I analysis.

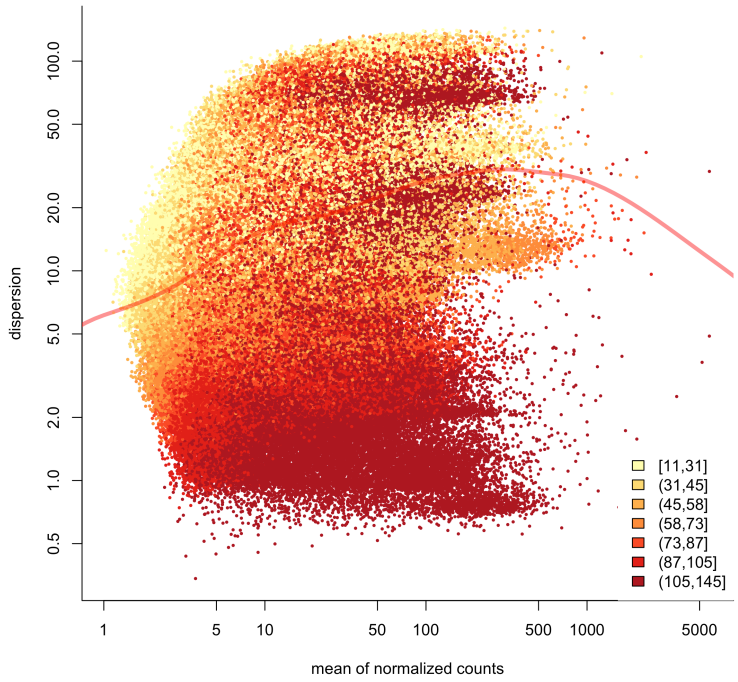

Figure 42: DESeq's dispersion estimates on Chinese T2D metagenomic study is similar to large marker-gene survey studies.

A) Dispersion estimates for filtered features plotted against normalized means. Dispersion estimates are greater than the average RNA-seq experiment and confounded by the sparsity of a feature. Colors are defined by the legend in the bottom right by equal bins of groups of features with similar levels of gene presence.

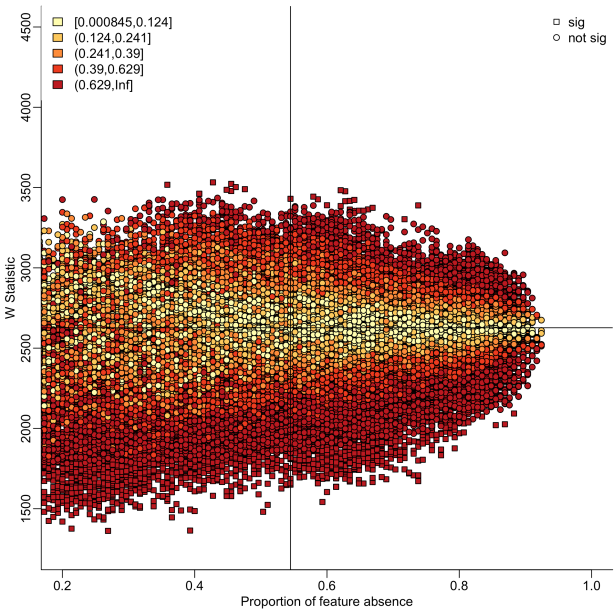

Figure 43: Wilcoxon statistic's relationship to sparsity in the Chinese T2D microbiome dataset.

A) scatterplot of the W statistic (y-axis) vs. proportion of samples absent for a given feature (x-axis) on a subset of the Qin et al. dataset. Colors within circles represent the absolute log odds ratios with red representing features whose presence dominated by a single group. Square features are significant after multiple testing correction and circular features are not. Horizontal line is at 2627, the statistic expected for uniformly distributed rankings given the sample size. The vertical line is at the median gene absence.

4.5.1 Simulations

We performed simulations on a number of statistical methods developed for high-throughput sequencing datasets. In particular, we were interested in comparing the false positive and false negative rates of our zero-inflated log-normal strategy against a number of methods using simulations developed in McMurdie et al. [100]. Counts for each simulated microbiome dataset were generated from a multinomial derived from a dataset available in the Phyloseq package [109, 110]. Further details and simulation code are available in [100]. Figure **44**, Figure **45**, and Figure **46** highlight the performance of the zero-inflated log-normal model as compared to DESeq, DESeq2, edgeR, and metagenomeSeq. We observed that metagenomeSeq2 (MSEQ2+) with a shrunken positive component and even solely a zero-inflated log-normal model (MSEQ2) have a lower false positive rate compared to a zero-inflated Gaussian mixture model (MSEQ).

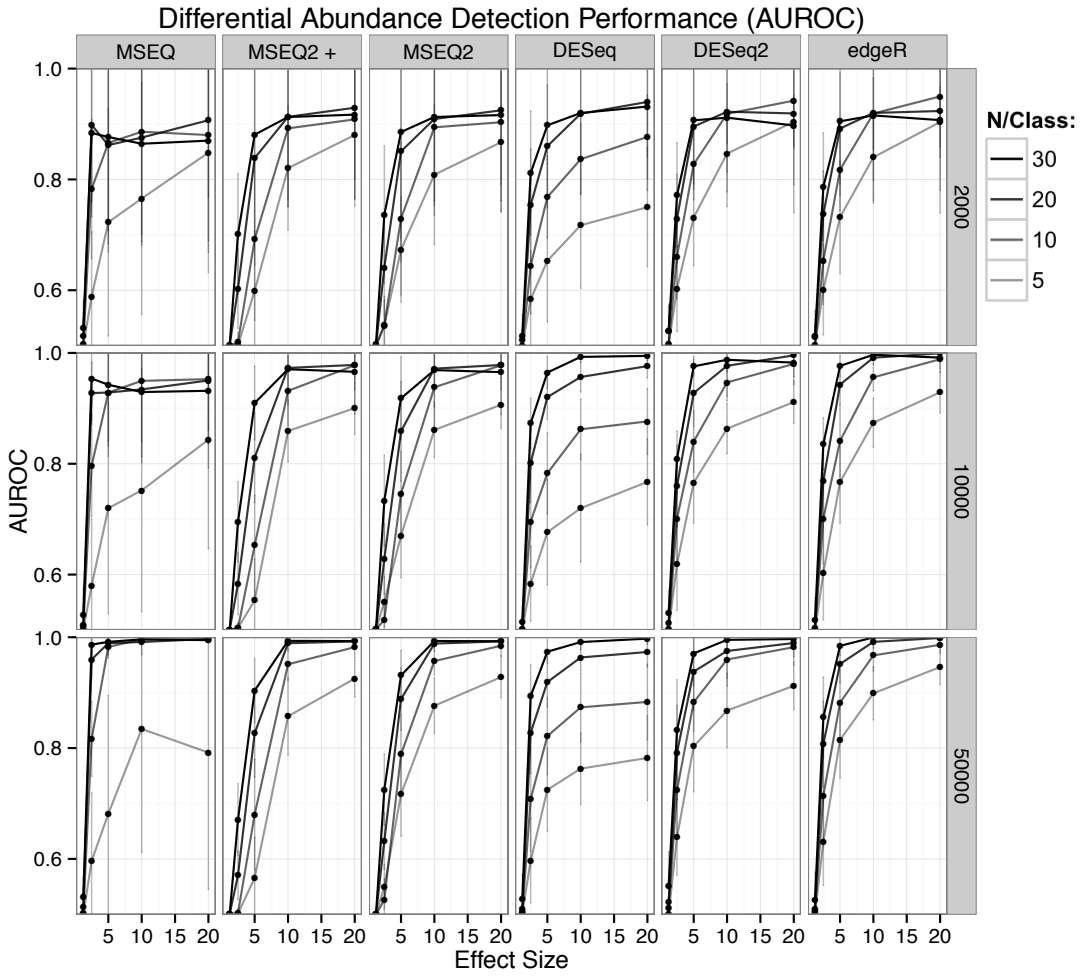

Figure 44: AUROC for independent simulation.

AUROC values (y-axis) for a number of simulations of varying effect sizes (x-axis), average depth of coverage (row), and method (column). We compared metagenomeSeq (MSEQ), metagenomeSeq2 with the positive component shrunken (MSEQ2+), metagenomeSeq2 without shrinkage (MSEQ2), DESeq, DESeq2, and edgeR. All methods appear to perform fairly well with MSEQ outperforming all methods in simulations with high-depths of coverage and larger than 5 samples per class.

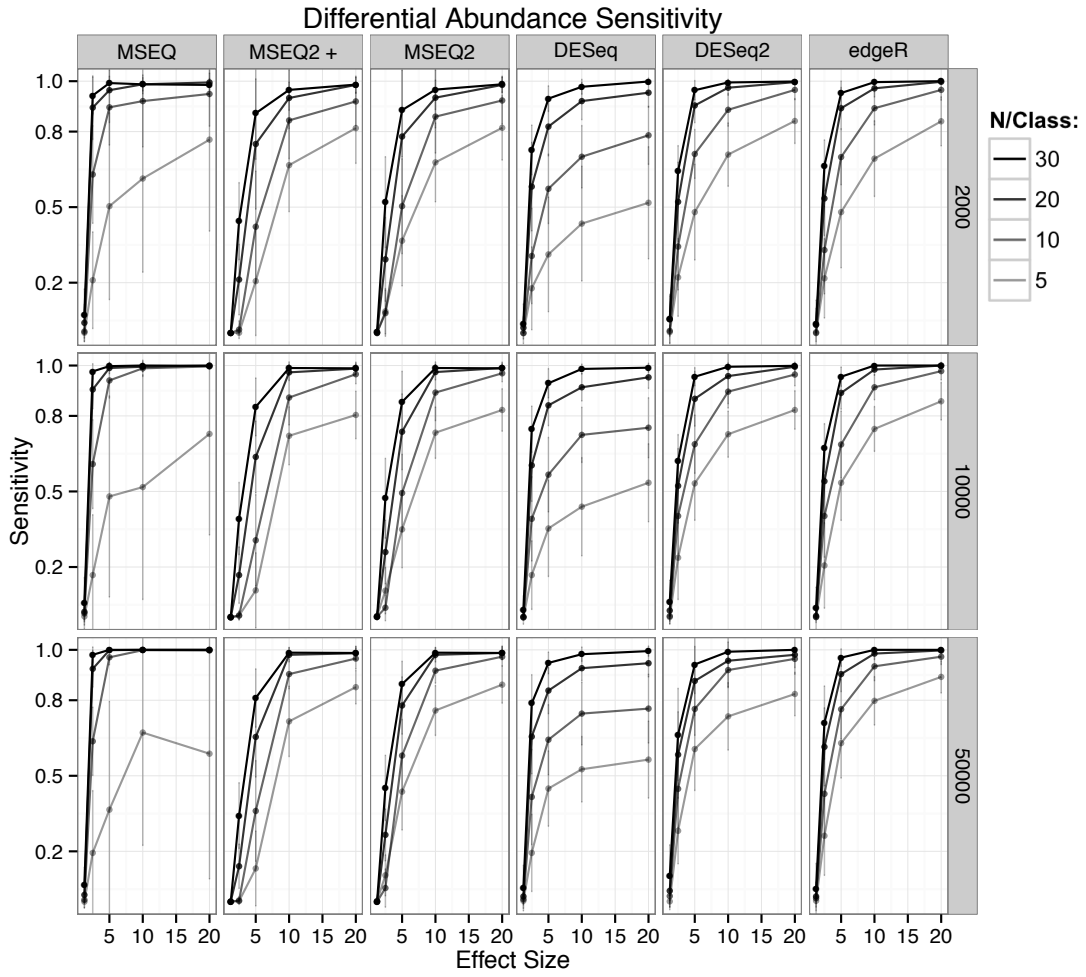

Figure 45: Sensitivity analysis for independent simulation.

Sensitivity (y-axis) for a number of simulations of varying effect sizes (x-axis), average depth of coverage (row), and method (column). We compared metagenomeSeq (MSEQ), metagenomeSeq2 with the positive component shrunken (MSEQ2+), metagenomeSeq2 without shrinkage (MSEQ2), DESeq, DESeq2, and edgeR. All methods appear to perform fairly well.

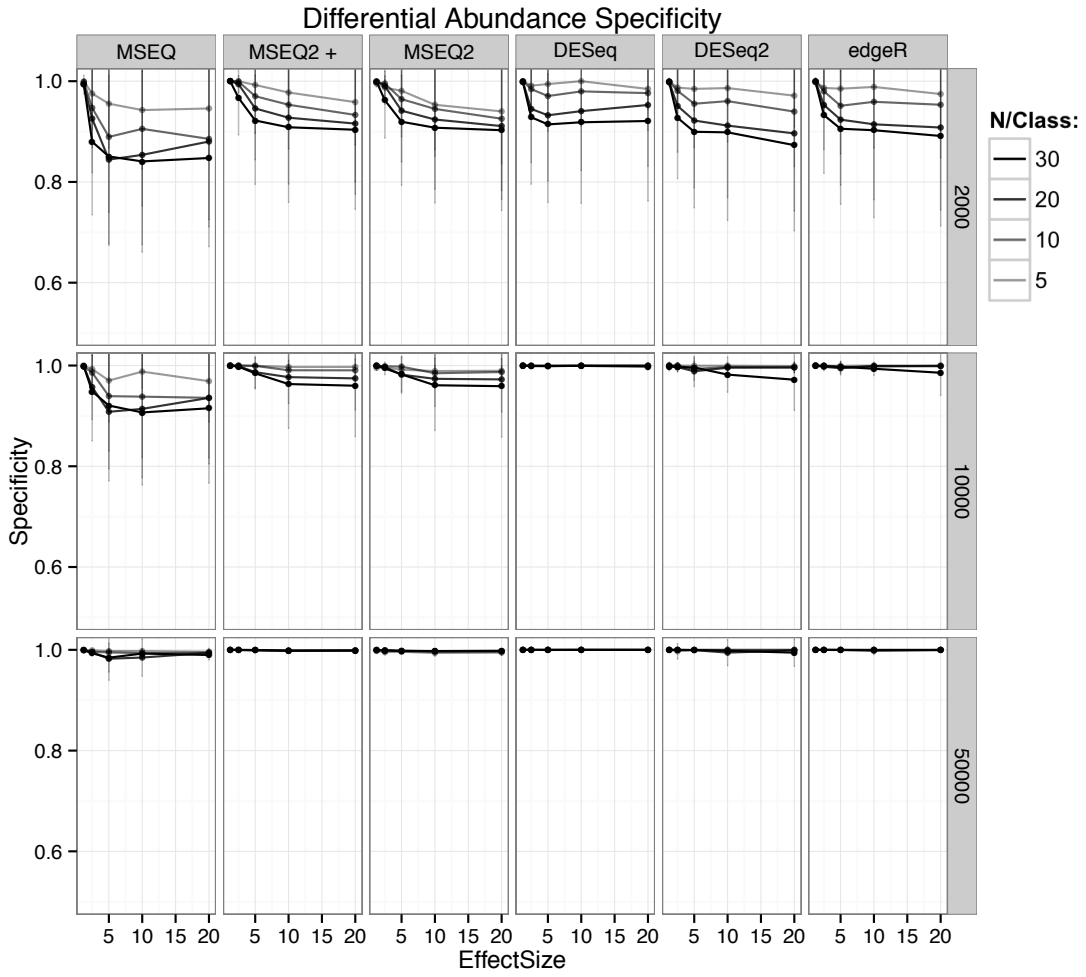

Figure 46: Specificity for independent simulation.

Specificity values (y-axis) for a number of simulations of varying effect sizes (x-axis), average depth of coverage (row), and method (column). We compared metagenomeSeq (MSEQ), metagenomeSeq2 with the positive component shrunken (MSEQ2+), metagenomeSeq2 without shrinkage (MSEQ2), DESeq, DESeq2, and edgeR. All methods, except for metagenomeSeq (the zero-inflated Gaussian) appear to perform fairly well for lower depths of coverage.

4.5.2 Classification AUROC of Type II diabetes is greatly increased using genes picked by metagenomeSeq2

To build a predictive model for T2D based on gut microbiome gene features, we followed a similar procedure to Qin et al. for feature selection and 50-gene biomarker estimation (see Methods). We performed this analysis on CSS-normalized,

log-transformed counts for genes detected as significantly abundant using the method presented earlier. We observed significantly improved classification performance, AUROC=0.95 CI=(0.92-0.98) compared to a reported AUROC=0.81 CI=(0.76-0.85) in Qin et al.

Following the same procedure as Qin et al. we generated a biomarker index for Type II diabetes replacing genes chosen by metagenomeSeq2. We analyzed the set of genes used in generating the Type II index by blasting against the NCBI nucleotide database and compare to the list generated by Qin et al. Of the 50 markers used to discriminate Type II diabetic samples from controls, most of the control enriched genes are included in the Lachnospiraceae family, with a majority coming from *Roseburia intestinalis*, a butyrate producing bacteria recently described by Schloissnig et al. [111] as prevalent and significantly retained in healthy European and American intestinal stool samples.

4.6 Discussion

This chapter examined sparsity, normalization, and differential abundance testing of whole metagenomic shotgun sequencing data. Metagenomic data has unique properties compared to microarrays and other high-throughput data requiring careful attention in reporting differential abundance due to clinical or phenotypic variability. We first characterize sparsity in metagenomic data. We explore various counting strategies and determine that even the most lenient strategy in counting reads results in a sparse matrix. Additionally, we highlight the positive relationship sequencing depth and gene presence share. While there has been discussion, but no validation, of a core microbiome present in at least 50% of individuals observed it is

quite possible the core is present in all similar phenotypic individuals, but not observed due to low depths of coverage [36, 112].

While recent literature has characterized the issues regarding TSS normalization in marker-gene surveys, TSS is still the norm for metagenomic studies. We performed an analysis comparing samples' count distributions across depths of coverage to test the assumptions behind TSS. We observed that the count distributions do not follow the scaling normalization assumptions. While TSS normalization is particularly problematic due to extreme deviations at the tails, we highlight DESeq and CSS normalization as alternative normalization approaches.

Finally, while the majority of studies do not currently model under-sampling and sparsity in metagenomic data we highlight the utility of a zero-inflated log-normal model that allows for an accurate analysis of sparse metagenomic and marker-gene surveys. We have shown how certain other methods are biased in the analysis of a healthy/diseased gut microbiome study, but that we are able to achieve superior disease classification with metagenomeSeq2. Comparing metagenomeSeq2 to metagenomeSeq we improved on both computational efficiency and a high false positive rate while retaining it's sensitivity. Additionally, by fitting a feature specific model one can now estimate the probability a sample will realize any gene based on group membership and sampling rate.

4.7 Materials & Methods

4.7.1 Data acquisition

The Qin et al. study consisted of consisted of two stages for which the first

stage was used in the following analyses. Stage I samples consisted of 26 female non-diabetic, 25 female diabetic, 48 male non-diabetic and 46 male diabetic stool whole-genome shotgun paired-end reads with insert sizes of ~350 base pairs. Samples were sequenced on HiSeq2000. Further information regarding read generation and phenotypic information for the Chinese Type II WMS study is described in [5]. The Karlsson study consisted of 145 female European samples with either T2D (n = 53), impaired glucose tolerance (IGT; n = 49) or normal glucose tolerance (NGT; n = 43). Analyses were performed on the same count matrix Qin et al. study used. Since the authors do not provide Wilcoxon W-statistics for the full set of genes, we reanalyzed their data, following procedures as described in their manuscript. Unfortunately, we were not able to reproduce their results exactly (Figure **47**). As the Karlsson's study count matrix was not available, reads were mapped to the reference developed by Qin et al. Sequences by Qin et al. and Karlsson et al. were downloaded from the short read archive, accession numbers SRA045646 and ERP002469. Human Microbiome Project sample sequences were downloaded from the convenient HMP DACC at http://www.hmpdacc.org/.

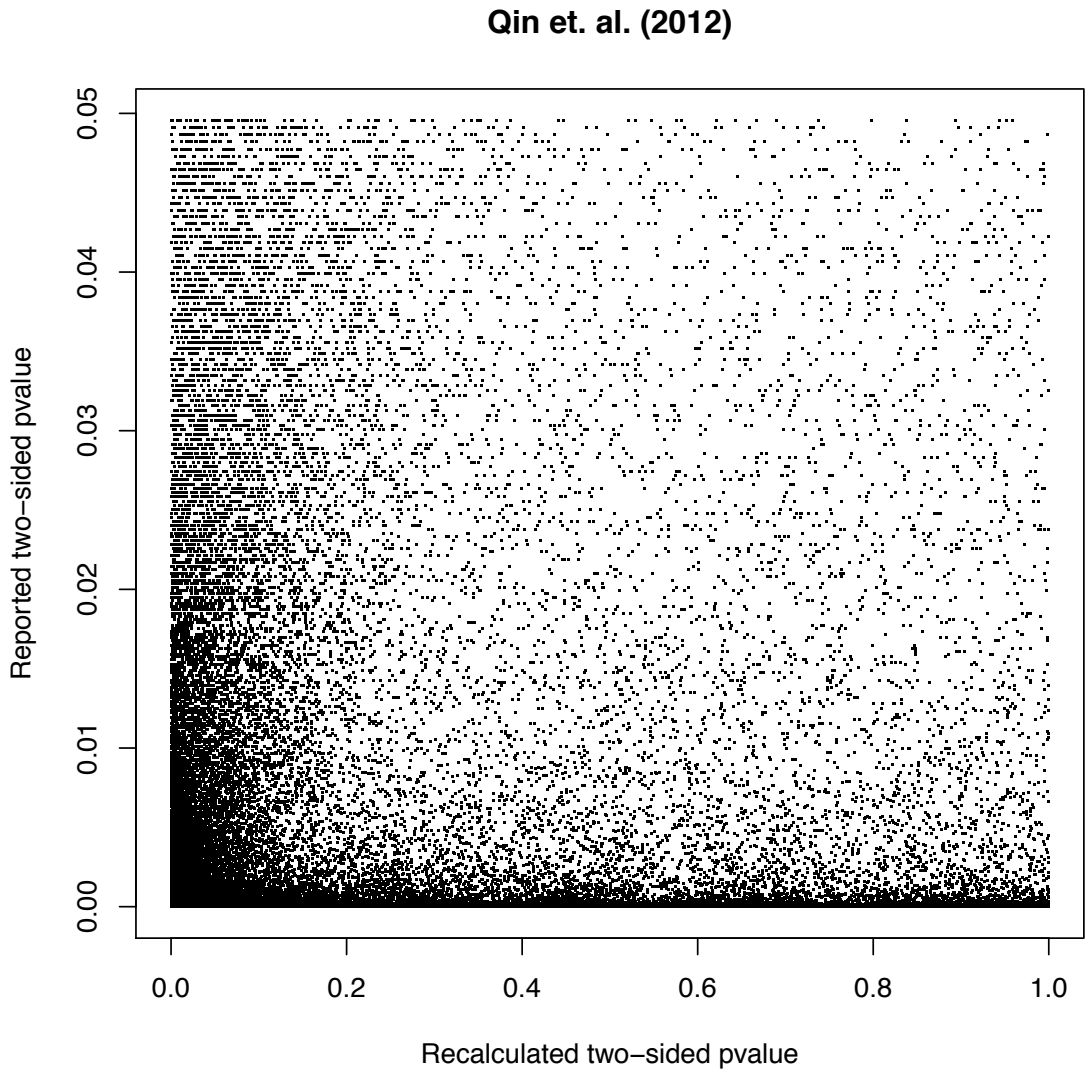

Figure 47: Comparison of Stage I two-sided Wilcoxon rank-sum statistic with continuity correction.

Scatter plots comparing the reported p-values (y-axis) from the two-sided Wilcoxon rank-sum statistic estimated p-values using the base R command wilcox.test (x-axis).

The two T2D datasets had high rates of alignment with an average of 81.2% (2.7% sd) in the Chinese T2D study and an average of 70.8% (5.1% sd) reads aligned in the European T2D study implying high coverage of the genes represented in the catalogue. In contrast the oral portion of the human microbiome project had an

average of 15.3% (9.7% sd) reads aligned, indicating the gene catalogue reference is incomplete; the oral portion of the human microbiome had a very few percentage of reads mapped and potential biases due to a reference from the gut microbiome.

4.7.2 Abundance count matrix generation

We generated the count matrices using five different methods. The first, termed unique, only counts read that uniquely aligned to a particular gene. All plots used this matrix, unless specified. The second, proportional, is a method for handling a read that aligns to multiple genes. In this method, we assign the read to a gene with a probability based on the proportion of *uniquely* aligned reads within the gene relative to the other genes. The third method, all, assigned a count to every mapped gene. The fourth method, random, assigned a count to a randomly chosen gene if there were multiple alignments. The final method, fractional, evenly splits a single count amongst the genes to which it aligns.

Ignoring mate information, using Bowtie 2 with preferences, –very-fast -a –no-unal –no-hd –omit-sec-seq –no-sq we mapped reads from the European and Chinese T2D study as well as the HMP oral microbiome subset to the updated MetaHit Gene catalogue downloaded from http://www.gigadb.org.

4.7.3 Statistical Methods

Filtered genes to have an average of at least 1 count per million in at least 35 samples. Tables describing sparsity were based on filtering raw gene counts present in at least 10 samples. P-values for each test were corrected to control the false discovery rate using the Benjamini-Hochberg procedure [98].

4.7.4 Software

The results presented in this article were obtained using R version 3.1.0 and the software packages metagenomeSeq 1.7.12, DESeq 1.16.0 and edgeR 3.6.2. These packages are available through the Bioconductor project. Additionally, sideChannelAttack 1.0-6 and pROC 1.7.3 available through CRAN were used in the classification analysis.

4.7.5 Classification

We identified a set of 50 genes from 1000 differentially abundant genes based on metagenomeSeq2 log fold-changes with the maximum relevance minimum redundancy (mRMR) feature selection framework developed in sideChannelAttack. We calculated a T2D-index using the CSS-normalized log-transformed gene abundances as features for a support vector machine with a radial kernel using Leave-One-Out cross validation.

5 The gut microbiome strongly predicts phenotype X

5.1 Background

5.1.1 Context

Many microbiome studies have begun to link the microbiome with various diseases and produce biomarkers to classify a phenotype or disease status. With Wikum Dinalankara as well as Lee Mendelowitz, Sean Kross, Héctor Corrada Bravo and Mihai Pop we asked the question as to whether or not these biomarkers are reflective of any biology or merely due to the high-dimensional nature of the data. This work is in the initial stages, but provides an important cautionary tale to those in the field that the small subset of features used are potentially merely strong predictors and not relevant.

5.1.2 Abstract

Recent studies have characterized and compared the microbial communities in a number of healthy and diseased individuals with an emphasis on phenotypic classification. In an attempt to classify healthy and diseased individuals researchers often define phenotypic markers based on the sample's microbial community. There is a potentially large diagnostic market in accurately classifying disease with an assay of a small set of genes or species. Many of the studies make use of common machine learning techniques, e.g. cross-validation, in an attempt to select biologically relevant markers. We investigate a number of machine learning techniques and feature set selection procedures on two whole metagenomic shotgun sequencing datasets and a marker-gene survey. We reveal that it is possible to achieve potentially biologically

irrelevant markers by not including feature set selection in the cross-validation procedure due to an over-fitting.

5.2 Methods in calculating AUROC

To classify samples by group labels we used four different classification methods. Specifically, we classified samples using support vector machines (SVM) with a radial basis kernel, Linear Discriminant Analysis (LDA), Naïve Bayes (NB), and random forests (RF). Each method was chosen to be representative of popular machine learning classifiers from non-linear (SVM), linear (LDA), probabilistic (NB) and rule-based (RF). We characterized performance by calculating area under the receiver operator curve (AUROC). For SVMs a decision value can be obtained directly from the decision function. Here the decision value for any given vector is the signed distance from the maximum margin hyper-plane to the given vector. With LDA, either the posterior probabilities for a given class or a score based on the discriminant variables can be used as decision values; the latter was selected for our experiments. Random forest and naïve Bayes decision values come from the sample probabilities of an accurate classification. These decision values, along with a binary labeling of the tested data, allow us to calculate an AUROC value.

We built and compared predictive models for each of the gut microbiomes analyzed by performing two-stage feature selection and testing four commonly used machine learning techniques (Figure **48**) provides a visual overview of our procedure. Variants of these models are commonly used in a number of disease studies [5, 9, 101].

Specifically, we performed a two-stage feature set selection. The first stage

consisted of defining a feature set of 50 genes by associating genes with phenotype by either performing differential presence (Fisher's presence-absence odds ratios) or abundance (t-test or Wilcoxon rank-sum test). The second stage consisted of defining the index to limit the number of genes, either by statistic/pvalue or a mutual information optimization methodology. The mutual information optimization methodology used here and in Qin et al. was minimum Redundancy Maximum Relevance (mRMR). Subsequently we performed either leave-one-out cross validation or 5-fold cross validation on the classification techniques listed above. This resulted in six methods for defining our feature set, two types of cross validation, and four classification techniques totaling. In total 48 methods to compare on a number of null phenotypes for each of our microbiome datasets.

We utilized kernel densities in visualizing the distribution of AUROC values of 1000 randomly generated phenotypes for each dataset and classifier. For leave-one-out cross validation, a decision value for each sample of the dataset was calculated by training the classifier with all other samples, excluding the sample tested. For five-fold cross validation, five AUROC values were obtained corresponding to each fold, and the mean of these AUROC values was selected.

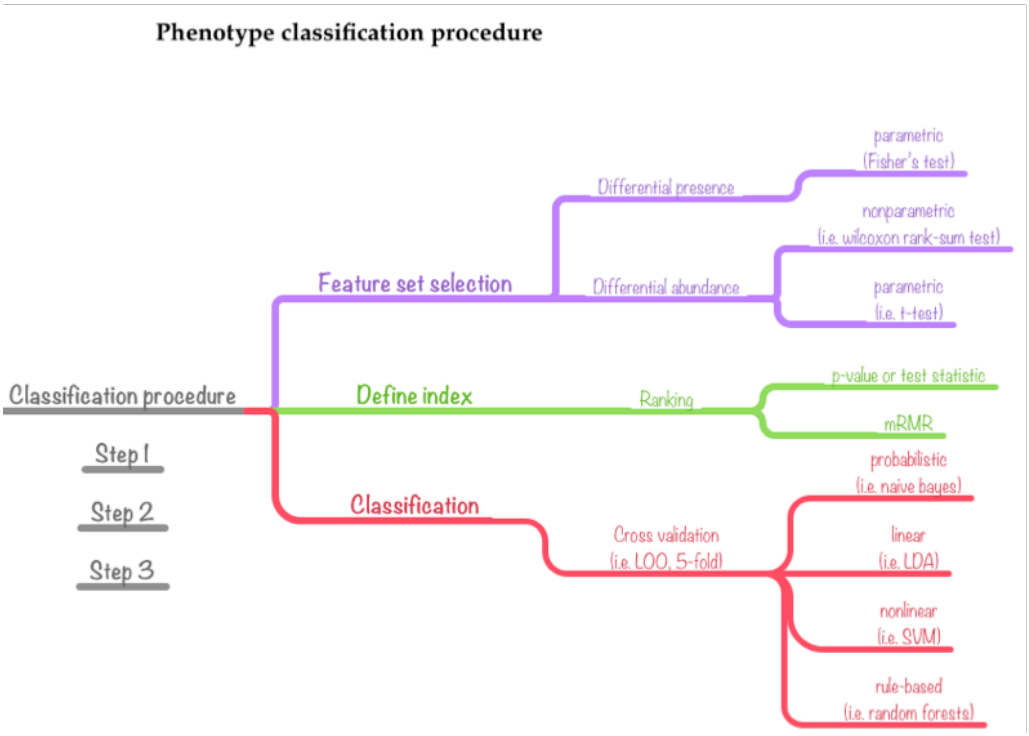

Figure 48: Phenotype classification procedure.

We performed classification using two-stage feature set selection. The first stage consisted of defining a feature set by either performing differential presence or abundance testing to determine features associated with a phenotype. The second stage consisted of defining an index, a set of 50 genes for classification purposes. Subsequently we performed classification with either leave-one-out cross validation or 5-fold cross validation on a number of popular machine learning techniques.

5.3 Results

5.3.1 Case study using a large infant gut-cohort

Making use of a similar procedure to the Qin et al. study described earlier we tested our ability to predict whether a sample belongs to "Phenotype X", a randomly assigned phenotype on our infant cohort of 992 samples from four countries described in Chapter 3. Each sample was randomly assigned label 1 or 0, indicating whether the sample was classified as having phenotype X. The statistic of choice was presence/absence odds ratio. Using the 50 OTU abundances with greatest odds ratios and a cross-validation to mitigate over-fitting, specifically leave-one-out (LOO), we

were able to achieve high AUROC values using a radial kernel SVM. In one particular run, we were able to accuracy predict phenotype X with an accuracy measure of 0.88 AUROC and 'infer' biological meaning for phenotype X despite the null labeling. In contrast the true phenotype measured lower. We analyze three other studies' AUROC distribution for other 'phenotype X's.

5.3.2 Effect of feature set selection

To understand the effect of feature set selection methods on various classifiers we analyzed the distribution of AUROC using six different methods for feature set selection. The methods for feature selection consist of two stages as consistent with recent methods used in a number studies producing biomarker indices [5, 6, 101]. The first stage consisted of calculating associations between features and phenotype with either a parametric, non-parametric, or presence-absence based test and described in detail in Methods. We analyzed two WMS studies and one marker-gene study and report AUROC values for the three combined despite slight differences between the WMS datasets and the marker-gene survey [6, 16, 50]. We report on leave-one-out cross-validated data in this section.

In analyzing the distribution of AUROC values for features selected by *t*-test and Wilcoxon's rank sum, both perform very similarly with the average AUROC value for features selected by *t*-test [0.90 (0.13 sd)] and Wilcoxon [0.88 (0.13)]. In comparison, feature selection by odds ratio often produced much lower AUROC values [0.78 (0.15)].

Comparing AUROC values for features chosen in the second stage by either the top 50 rankings of features or mRMR yielded the greatest difference in values.

For each association test, filtering features by mRMR produced higher AUROC values *t*-test [0.93 vs. 0.87], Wilcoxon [0.85 vs. 0.91], and OR [0.87 vs. 0.69]. This is a result of the mRMR heuristic procedure producing features that have the least mutual information with each other.

5.3.3 Cross-validation post feature-set selection does not mitigate the effects of over fitting for null phenotype classification.

To characterize the robustness of classification on null phenotypes we employed two types of cross-validation, leave one out (LOO) and 5-fold. The results obtained (Figure **49**, Figure **50**) show that all the chosen classifiers had high AUROC in predicting a sample's randomly chosen phenotype. The only method robust to over-fitting was the use of first stage odds ratios in the marker-gene survey. This is a poor comparison though due to no prior filtering of the dataset and the selection of features absent in at least 98% of the dataset, however, after filtering this effect is mitigated and higher AUROC values are obtained (not shown).

All methods performed extremely well in classifying null phenotypes. In particular, we observed that we consistently high achieve AUROC values with a median of [SVM- 0.94, LDA- 0.88, NB- 0.88, RF- 0.91] using LOO cross-validation across datasets with a median 0.90 AUROC. Additionally, when using a more stringent type of cross validation, 5-fold we also perform extremely well. We consistently achieve median AUROC values of [SVM- 0.93, LDA- 0.84, NB- 0.89, RF- 0.91] with a median of 0.90. In comparison to the null phenotypes, the true label classification AUROC values performed equally well in both cases with an average AUROC of 0.89 and 0.93 (CV, LOO).

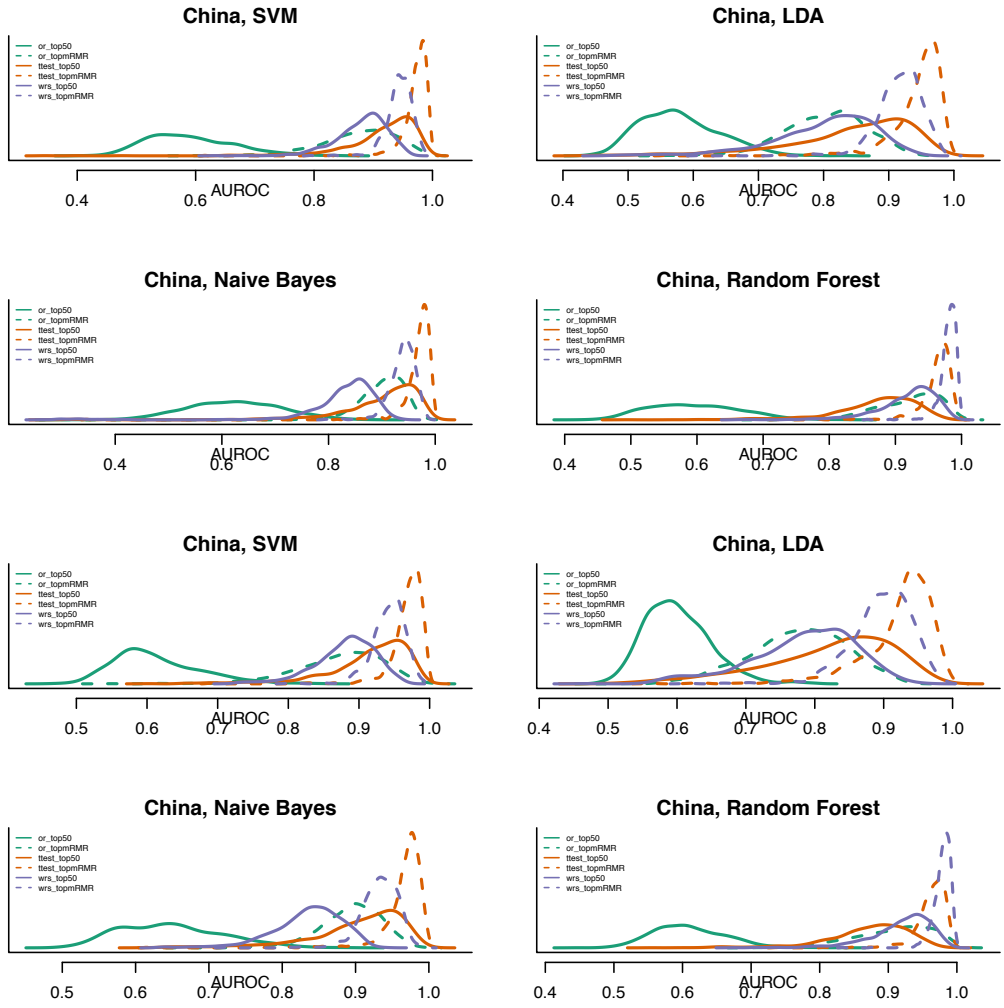

Figure 49: Cross validation post feature-set selection does not mitigate the effects of over-fitting post feature selection for WMS studies.

Kernel density plots of the AUROC values for each of the six feature set selection methods for the Chinese gut microbiome dataset produced in [5]. Each panel denotes the classifier used with the top four employing LOO cross-validation and the bottom four 5-fold cross-validation. Even with the more stringent cross-validation most null phenotypes are well predicted with the minimal feature set, except for features chosen by presence-absence.

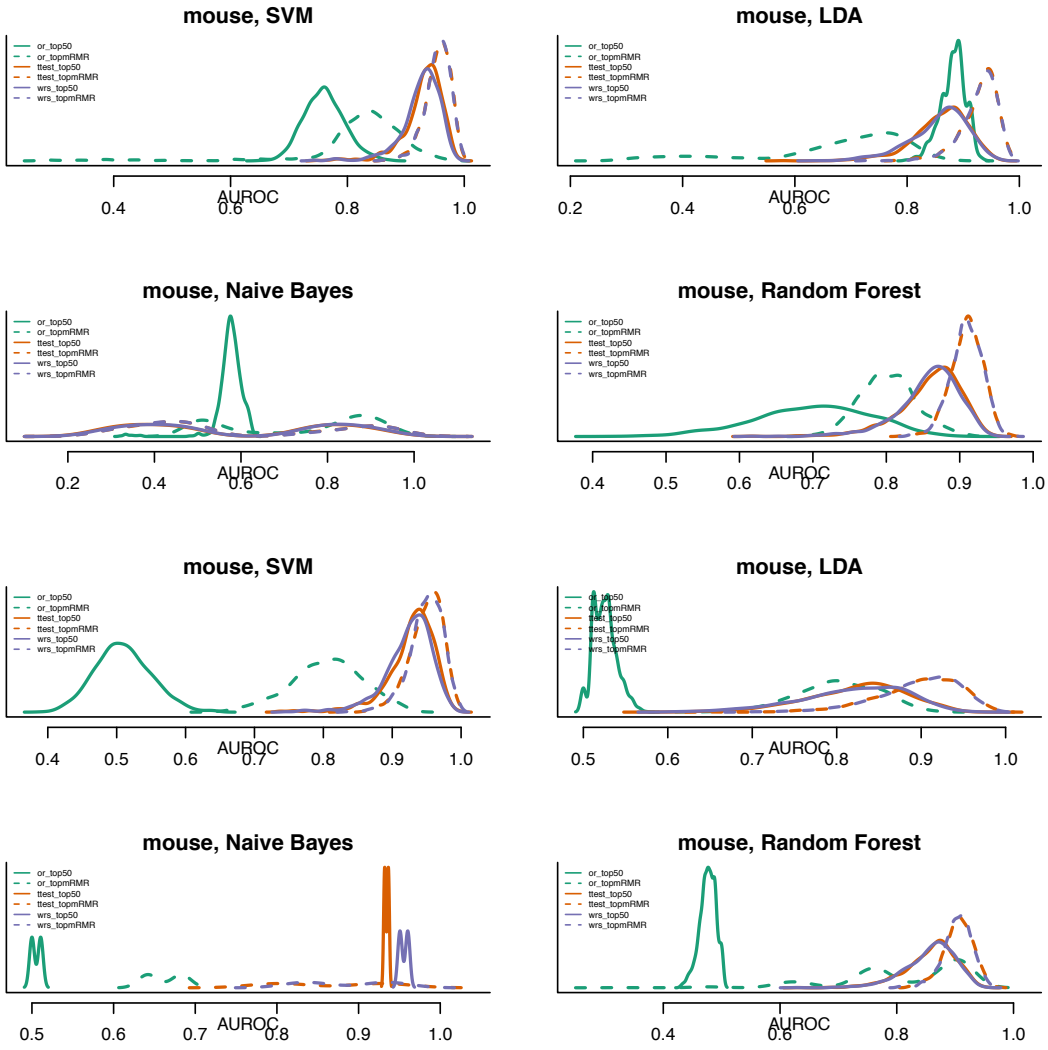

Figure 50: Cross validation post feature-set selection does not mitigate the effects of over-fitting post feature selection for marker-gene studies.

Kernel density plots of the AUROC values for each of the six feature set selection methods for the mouse gut marker-gene dataset produced in [50]. Each panel denotes the classifier used with the top four employing LOO cross-validation and the bottom four 5-fold cross-validation. Even with the more stringent cross-validation most null phenotypes are well predicted with the minimal feature set, except for features chosen by presence-absence.

5.3.4 Averaging feature set statistics for selection does not mitigate over-fitting

In an attempt to mitigate the effects of over-fitting prior to classification we partitioned the data into five equally sized bins and performed our three methods of

Stage I filtering. We averaged the statistics for each method to produce one set and then proceed with Stage II filtering. Analyzing the mouse gut marker-gene survey data we observed that our null AUROC distributions perform much more as expected with a median AUROC of 0.58 with the true phenotype all achieving an AUROC of 1.0 (Figure 51). An AUROC of 1.0 is expected in this simple study where the groups are easily distinguished.

However, for the WMS datasets, binning the data does not mitigate the effects of over-fitting due to feature selection (Figure 52). The difference between Stage II procedures of selecting the 50 features with greatest statistic or using the mRMR procedure is negligible. However, the median AUROC is 0.92. In this situation we would expect that including feature set selection in the cross-validation approach should mitigate the effect and drive densities mean towards a more expected value (closer to 0.5).

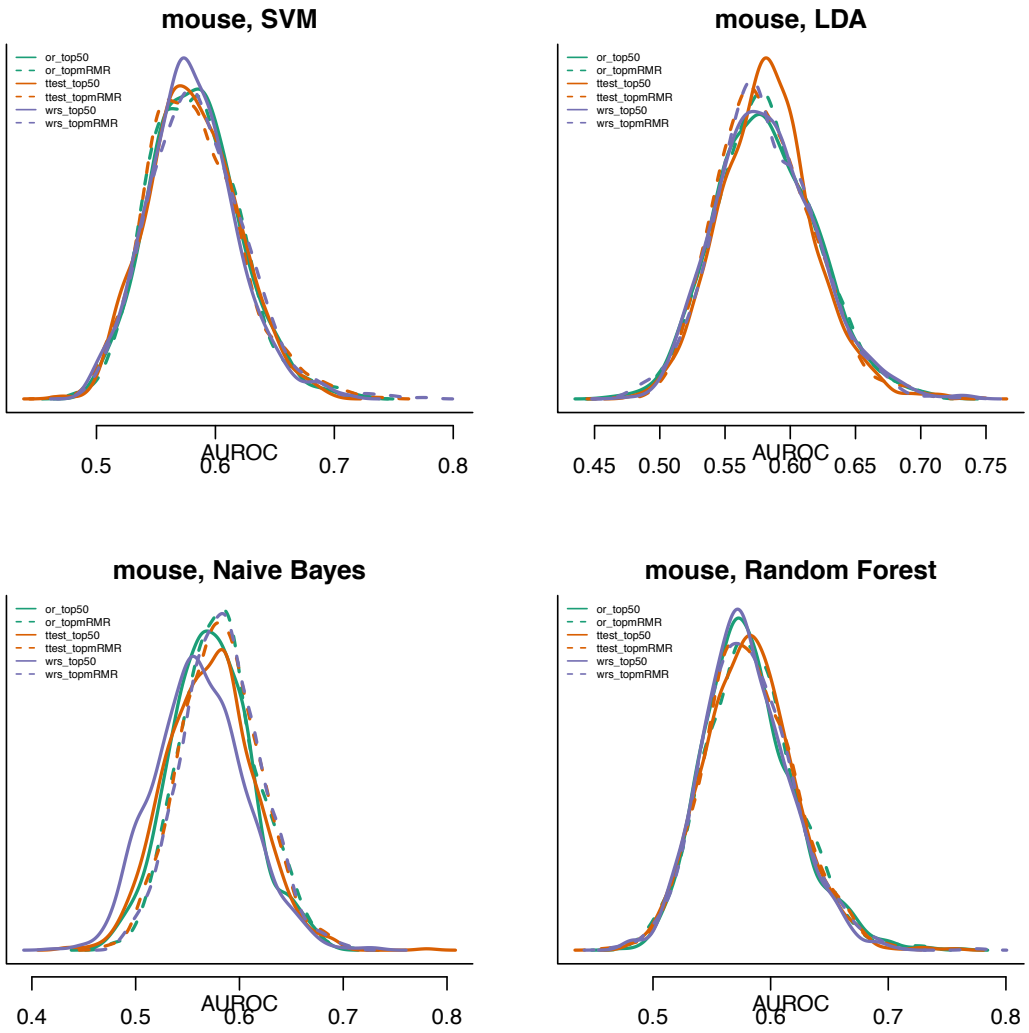

Figure 51: Performing feature-set selection on averaged statistics mitigates the effect of over-fitting.

Kernel density plots of the AUROC values for each of the six feature set selection methods for the mouse gut marker-gene dataset produced in [50] following an averaging of statistics calculated on binned samples. Each panel denotes the classifier used with 5-fold cross-validation.

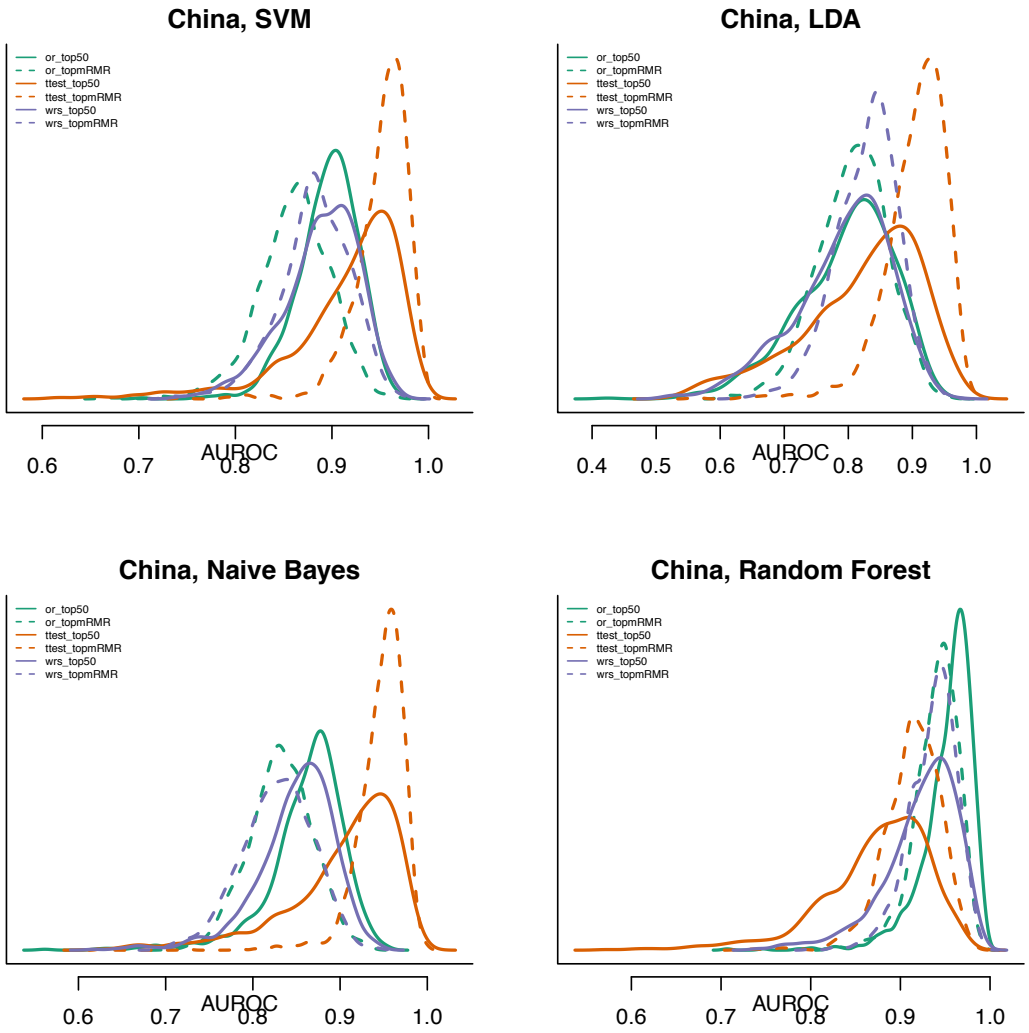

Figure 52: Performing feature-set selection on averaged statistics does not mitigate the effect of over-fitting on WMS studies.

Kernel density plots of the AUROC values for each of the six feature set selection methods for the Chinese gut microbiome dataset produced in [5] following an averaging of statistics calculated on binned samples. Each panel denotes the classifier used with 5-fold cross-validation.

5.3.5 Including feature set selection within cross-validation mitigates over-fitting

In an attempt to mitigate the effects of over-fitting prior to cross-validation we incorporated feature set selection within the 5-fold cross-validation framework.

Incorporating feature set selection within the cross-validation leads to an average of

0.58 AUROC for the *t*-test feature set selection procedure (top50 and mRMR) in the European whole metagenomic shotgun dataset. Preliminary analyses suggest similar results for the other whole metagenomic shotgun dataset and other feature set selection methods.

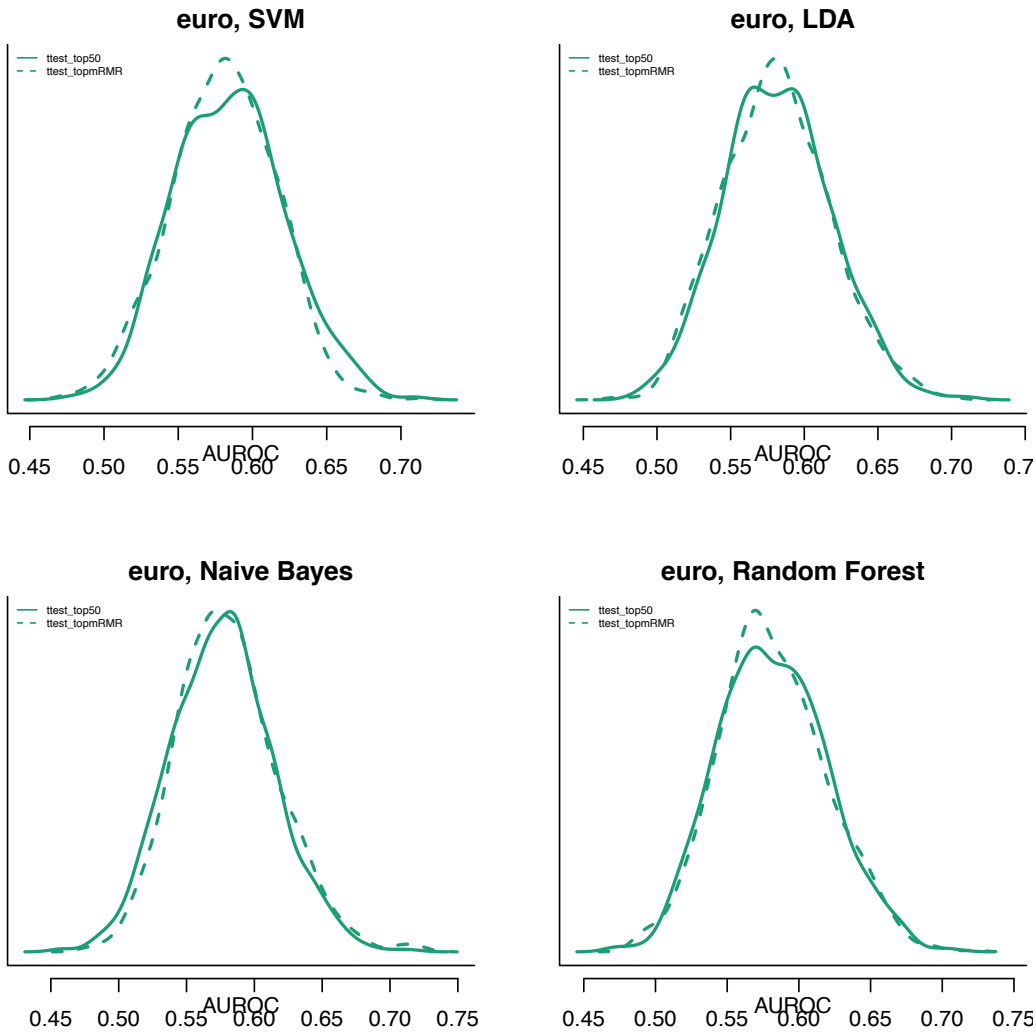

Figure 53: Performing feature-set selection within a cross-validation framework mitigates the effect of over-fitting on WMS studies.

Kernel density plots of the AUROC values for two feature set selection methods for the European gut microbiome dataset. Each panel denotes the classifier used with 5-fold cross-validation.

6. Discussion and conclusion

6.1 Contributions

This work constitutes a number of improvements to the field of scientific computation and metagenomics in the application of sparsity aware methods for normalization and differential abundance analysis. Chapter 1 provides a brief overview of the methods used to analyze microbial communities prior to this dissertation. Chapter 2 introduced the need for sparsity-aware normalization and testing methods and a novel normalization method that significantly improved sample clustering and a zero-inflated Gaussian mixture model that improved fold-change estimates. Chapter 3 highlighted the application of the method one of the largest gut cohorts at the time on an important disease and our ability to find novel potential pathogenic and probiotic associations for children with diarrhea in the developing world. Chapter 4 highlights an improved sparse-aware testing methodology and an overview of the sparsity in whole-metagenomic shotgun sequencing data. Chapter 5 continues with the target of many microbiome experiments, production of biomarkers that can be used to detect a disease or phenotype. The chapter warns of the many pitfalls using these biomarkers, namely many are potentially useful for classification, but are not much more. I reiterate that each chapter was a culmination of work presented mostly in the following papers published or in preparation:

Chapter 2:

- Paulson, JN., Stine, OC., Bravo, HC., Pop, M. (2013). Differential abundance analysis for marker-gene surveys. Nature Methods, Volume 10, pg. 1200-1202.
- Paulson, JN., Bravo, HC., Pop, M. (2014). Reply to: "a fair comparison". Nature Methods, Volume 11, pg. 359-360.

Chapter 3:

- Pop, M.*, Walker, AW.*, Paulson, JN.*, Lindsay, B.*, Antonio, M.*, Hossain MA.*, Oundo J.*, Tamboura B.*, Mai V.*, Astrovskaya I., Bravo, HC., Rance R., Stares M., Levine MM., Panchalingam S., Kotloff K., Ikumapayi UN., Ebruke C., Adeyemi M., Ahmed D., Ahmed F., Alam MT., Amin R., Siddiqui S., Ochieng JB., Ouma E., Juma J., Mailu E., Omore R., Morris JG., Breiman RF., Saha D., Parkhill J., Nataro JP., Stine, OC. (2014). Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. Genome Biology, Volume 15, pg. R76.

Chapter 4:

- Paulson, JN., Hill, CM., Almeida, M., Bravo, HC., Pop, M. Differential abundance analysis for whole metagenomic shotgun sequencing. In preparation.

Chapter 5:

- Paulson, JN.*, Wikum, D.*, Mendelowitz, L., Kross, S., Bravo, HC, Pop, M. (2015). The gut microbiome strongly predicts phenotype X. In preparation.

6.2 Future work

6.2.1 Microbial co-abundance networks for whole metagenomic shotgun sequencing data and marker-gene surveys accounting for sparsity and under-sampling.

Even in very deep sequencing surveys, the nature of the "counting experiment" that is a metagenomic analysis can skew representative population estimates for community members. To address this issue, I propose methods to aid in defining operon structure, novel organism discovery and assembly with improved clustering methods. The central idea is to weight zero values of a gene or taxa according to the posterior probability of not being observed due to under-sampling. These methods will not only have broad general applicability in the analysis of large, relatively sparse data sets, but they will provide better insight into the biological properties of complex microbial communities and their potential roles in environmental niches and human health.

I have shown that biases due to sparsity induced by under-sampling can be

successfully accounted for in marker-gene surveys (Chapter "Differential abundance analysis for marker-gene surveys") and for whole metagenomic shotgun sequencing data (Chapter "4 Differential abundance analysis of WMS sequencing data") with the use of a zero-inflated Gaussian (ZIG) mixture to model species [33]. It is important to investigate the appropriateness of a ZIG model in WMS data. In estimating the ZIG model a posterior probability is calculated for every zero value as belonging to an 'under-sampling' technical artifact distribution or being truly absent.

The Co-Abundant Gene clustering algorithm from Nielsen, et al. is as follows:

1. Pick a gene at random as a 'seed.'
2. Calculate the Pearson correlation coefficient (PCC) of the seed with all other proportionally normalized genes.
3. Merge into one 'canopy' genes with a PCC > 0.9.
4. Define abundance profile as median of gene abundances in canopy.
5. Iterate steps 1-4 until stabilization.

I propose a posterior probability approach to expound upon Nielsen, et al.'s clustering algorithm for the *de novo* segregation of metagenomic data [113]. The clustering algorithm can aid in assembly, organismal discovery, operon structure determination, and understanding phenotype-specific pathways [7, 101]. The gene cluster count matrices from objective 1 are the substrate for the clustering.

In particular, the Co-Abundant Gene (CAG) clustering algorithm [113] is a random seed selection algorithm that picks a gene at random and calculates pair-wise Pearson correlation coefficients (PCC) clustering those within a threshold. The algorithm is iterative and continues until all genes have been assigned to a cluster followed by a filtering process of singletons and presence in a minimum number of samples. Using the first objective's posterior probability we can calculate weighted PCCs to explore robustly generated clusters.

Glossary

OTU – Operational Taxonomic Unit

DNA – Deoxyribonucleic acid

RNA – Ribonucleic acid

rRNA – Ribosomal RNA component

rDNA – Ribosomal DNA component

RDP – Ribosomal Database Project

EM – Expectation Maximization

CAG – Co-Abundant Gene

PCC – Pearson correlation coefficient

Bibliography

1. Handelsman J, Tiedje J, National Research Council (US) Committee on Metagenomics: Challenges and Functional, Functional NRC (Us) C on MC and: *THE NEW SCIENCE OF METAGENOMICS: Revealing the Secrets of Our Microbial Planet*. 2007.

2. Pop M, Walker AW, Paulson J, Lindsay B, Antonio M, Hossain M, Oundo J, Tamboura B, Mai V, Astrovskaya I, Bravo H, Rance R, Stares M, Levine MM, Panchalingam S, Kotloff K, Ikumapayi UN, Ebruke C, Adeyemi M, Ahmed D, Ahmed F, Alam M, Amin R, Siddiqui S, Ochieng JB, Ouma E, Juma J, Mailu E, Omore R, Morris J, et al.: **Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition.** *Genome Biol* 2014, **15**:R76.

3. Liu B, Faller LL, Klitgord N, Mazumdar V, Ghodsi M, Sommer DD, Gibbons TR, Treangen TJ, Chang YC, Li S, Stine OC, Hasturk H, Kasif S, Segrè D, Pop M, Amar S: **Deep sequencing of the oral microbiome reveals signatures of periodontal disease**. *PLoS One* 2012, **7**.

4. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, others, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ, others, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ: **Vaginal microbiome of reproductive-age women.** *Proc Natl Acad Sci U S A* 2011, **108 Suppl**:4680–4687.

5. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, et al.: **A metagenome-wide association study of gut microbiota in type 2 diabetes**. *Nature* 2012:55–60.

6. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, Nielsen J, Bäckhed F: **Gut metagenome in European women with normal, impaired and diabetic glucose control.** *Nature* 2013, **498**:99–103.

7. Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le Chatelier E, Almeida M, Quinquis B, Levenez F, Galleron N, Gougis S, Rizkalla S, Batto J-M, Renault P, Doré J, Zucker J-D, Clément K, Ehrlich SD: **Dietary intervention impact on gut microbial gene richness.** *Nature* 2013, **500**:585–8.

8. Greenblum S, Turnbaugh PJ, Borenstein E: **Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease.** *Proc Natl Acad Sci U S A* 2012, **109**:594–9.

9. Qin N, Yang F, Li A, Prifti E, Chen Y, Shao L, Guo J, Le Chatelier E, Yao J, Wu L, Zhou J, Ni S, Liu L, Pons N, Batto JM, Kennedy SP, Leonard P, Yuan C, Ding W, Chen Y, Hu X, Zheng B, Qian G, Xu W, Ehrlich SD, Zheng S, Li L: **Alterations of the human gut microbiome in liver cirrhosis**. *Nature* 2014.

10. Abrahamsson TR, Jakobsson HE, Andersson AF, Björkstén B, Engstrand L, Jenmalm MC: **Low diversity of the gut microbiota in infants with atopic eczema**. *J Allergy Clin Immunol* 2012, **129**.

11. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods* 2010:335–336.

12. Lozupone C, Knight R: **UniFrac: a New Phylogenetic Method for Comparing Microbial Communities**. *Appl Environ Microbiol* 2005, **71**:8228–8235.

13. Ghodsi M, Liu B, Pop M: **DNACLUST: accurate and efficient clustering of phylogenetic marker genes**. *BMC Bioinformatics* 2011, **12**:271.

14. Maidak BL, Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, Woese CR: **The RDP (Ribosomal Database Project).** *Nucleic Acids Res* 1997, **25**:109–111.

15. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes J a, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C: **Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.** *Nat Biotechnol* 2013, **31**:814–21.

16. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li SSSSSS, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, et al.: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464**:59–65.

17. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC Bioinformatics* 2010, **11**:94.

18. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot NS, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F: **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**. *Brief Bioinform* 2013, **14**:671–683.

19. Pearson K: **Mathematical Contributions to the Theory of Evolution . --On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs**. *Proc R Soc London* 1897, **60**:489–498.

20. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**:249–264.

21. Bolstad BM, Irizarry R a, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19**:185–93.

22. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **32 Suppl**:496–501.

23. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**:R106.

24. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11**:R25.

25. Rodriguez-Brito B, Rohwer F, Edwards RA: **An application of statistics to comparative metagenomics.** *BMC Bioinformatics* 2006, **7**:162.

26. White JR, Nagarajan N, Pop M: **Statistical methods for detecting differentially abundant features in clinical metagenomic samples.** *PLoS Comput Biol* 2009, **5**:e1000352.

27. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C: **Metagenomic biomarker discovery and explanation.** *Genome Biol* 2011, **12**:R60.

28. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100**:9440–5.

29. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26**:139–140.

30. Langmead B, Hansen KD, Leek JT: **Cloud-scale RNA-sequencing differential expression analysis with Myrna**. *Genome Biol* 2010, **11**:R83.

31. Smyth G: **limma: Linear Models for Microarray Data**. In *Bioinforma Comput Biol Solut Using R Bioconductor*; 2005:397–420.

32. Law CWC, Chen Y, Shi W, Smyth GGK: **Voom! Precision weights unlock linear model analysis tools for RNA-seq read counts**. *Genome Biol* 2013, **15**:1–17.

33. Paulson JN, Stine OC, Bravo HC, Pop M: **Differential abundance analysis for microbial marker-gene surveys.** *Nat Methods* 2013, **10**:1200–2.

34. Paulson JN, Bravo HC, Pop M: **Reply to: "a fair comparison".** *Nat Methods* 2014, **11**:359–60.

35. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward D V, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C: **Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment**. *Genome Biol* 2012:R79.

36. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI: **A core gut microbiome in obese and lean twins.** *Nature* 2009, **457**:480–484.

37. Frazee AC, Langmead B, Leek JT: **ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets**. *BMC Bioinformatics* 2011:449.

38. Consortium HMP, Human T, Project M: **A framework for human microbiome research**. *Nature* 2012, **486**:215–221.

39. Charlson ES, Bittinger K, Haas AR, Fitzgerald AS, Frank I, Yadav A, Bushman FD, Collman RG: **Topographical continuity of bacterial populations in the healthy human respiratory tract**. *Am J Respir Crit Care Med* 2011, **184**:957–963.

40. Hall DB: **Zero-inflated Poisson and binomial regression with random effects: a case study.** *Biometrics* 2000, **56**:1030–1039.

41. Soneson C, Delorenzi M: **A comparison of methods for differential expression analysis of RNA-seq data**. *BMC Bioinformatics* 2013, **14**:91.

42. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 2009, **26**:493–500.

43. Wang X, Wu Z, Zhang X: **Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq.** *J Bioinform Comput Biol* 2010, **8 Suppl 1**:177–192.

44. Salzman J, Jiang H, Wong WH: **Statistical Modeling of RNA-Seq Data**. *Stat Sci* 2011, **26**:62–83.

45. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq.** *Nat Biotechnol* 2013, **31**:46–53.

46. White JR, Navlakha S, Nagarajan N, Ghodsi M-RR, Kingsford C, Pop M: **Alignment and clustering of phylogenetic markers--implications for microbial diversity studies.** *BMC Bioinformatics* 2010, **11**:152.

47. Langendijk PS, Hanssen JT, Van der Hoeven JS: **Sulfate-reducing bacteria in association with human periodontitis.** *J Clin Periodontol* 2000, **27**:943–950.

48. Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA, Sahasrabudhe A, Dewhirst FE: **Bacterial diversity in human subgingival plaque**. *J Bacteriol* 2001, **183**:3770–3783.

49. Colombo AP V, Boches SK, Cotton SL, Goodson JM, Kent R, Haffajee AD, Socransky SS, Hasturk H, Van Dyke TE, Dewhirst F, Paster BJ: **Comparisons of subgingival microbial profiles of refractory periodontitis, severe periodontitis, and periodontal health using the human oral microbe identification microarray.** *J Periodontol* 2009, **80**:1421–1432.

50. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI: **The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice.** *Sci Transl Med* 2009, **1**:6ra14.

51. Hughes JB, Hellmann JJ: **The application of rarefaction techniques to molecular inventories of microbial diversity**. *Methods Enzymol* 2005, **397**:292–308.

52. Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R: **Human-associated microbial signatures: Examining their predictive value**. *Cell Host Microbe* 2011:292–296.

53. Faust K, Raes J: **Microbial interactions: from networks to models**. *Nat Rev Microbiol* 2012:538–550.

54. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI: **Human gut microbiome viewed across age and geography.** *Nature* 2012, **486**:222–7.

55. Friedman J, Alm EJ: **Inferring correlation networks from genomic survey data**. *PLoS Comput Biol* 2012, **8**:e1002687.

56. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C: **Microbial Co-occurrence Relationships in the Human Microbiome**. *PLoS Comput Biol* 2012, **8**:e1002606.

57. Lozano R, Naghavi M, Foreman K, Lim S, Aboyans V, Abraham J, et al.: **Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010**. *Lancet* 2012, **380**:2095–2128.

58. Black RE, Cousens S, Johnson HL, Lawn JE, Rudan I, Bassani DG, Jha P, Campbell H, Walker CF, Cibulskis R, Eisele T, Liu L, Mathers C, of WHO CHERG, UNICEF: **Global, regional, and national causes of child mortality in 2008: a systematic analysis**. *Lancet* 2010, **375**:1969–1987.

59. Walker CL, Aryee MJ, Boschi-Pinto C, Black RE: **Estimating diarrhea mortality among young children in low and middle income countries**. *PLoS One* 2012, **7**:e29151.

60. Levine MM, Kotloff KL, Nataro JP, Muhsen K: **The Global Enteric Multicenter Study (GEMS): impetus, rationale, and genesis**. *Clin Infect Dis* 2012, **55**:S215–S224.

61. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, Wu Y, Sow SO, Sur D, Breiman RF, Faruque AS, Zaidi AK, Saha D, Alonso PL, Tamboura B, Sanogo D, Onwuchekwa U, Manna B, Ramamurthy T, Kanungo S, Ochieng JB, Omore R, Oundo JO, Hossain A, Das SK, Ahmed S, Qureshi S, Quadri F, Adegbola RA, Antonio M: **Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study**. *Lancet* 2013, **382**:209–222.

62. Mackie RI, Sghir A, Gaskins HR: **Developmental microbial ecology of the neonatal gastrointestinal tract**. *Am J Clin Nutr* 1999, **69**:1035s–1045s.

63. Cebra JJ: **Influences of microbiota on intestinal immune system development**. *Am J Clin Nutr* 1999, **69**:1046S–1051S.

64. Sjogren YM, Tomicic S, Lundberg A, Bottcher MF, Bjorksten B, Sverremark-Ekstrom E, Jenmalm MC: **Influence of early gut microbiota on the maturation of childhood mucosal and systemic immune responses**. *Clin Exp Allergy* 2009, **39**:1842–1851.

65. Rajilic-Stojanovic M, Smidt H, de Vos WM: **Diversity of the human gastrointestinal tract microbiota revisited**. *Env Microbiol* 2007, **9**:2125–2136.

66. Walker AW, Lawley TD: **Therapeutic modulation of intestinal dysbiosis**. *Pharmacol Res* 2013, **69**:75–86.

67. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C: **Metagenomic microbial community profiling using unique clade-specific marker genes**. *Nat Methods* 2012:811–814.

68. Parks DH, Beiko RG: **Identifying biologically relevant differences between metagenomic communities**. *Bioinformatics* 2010, **26**:715–721.

69. Girvan MS, Campbell CD, Killham K, Prosser JI, Glover LA: **Bacterial diversity promotes community stability and functional resilience after perturbation**. *Env Microbiol* 2005, **7**:301–313.

70. McCann KS: **The diversity-stability debate**. *Nature* 2000, **405**:228–233.

71. Shields TM, Chen KD, Gould JM: **Pediatric Case Report of Chronic Colitis Associated With an Unusual Serotype of Streptococcus pneumoniae**. *Infect Dis Clin Pract* 2012, **20**:357–358.

72. Jin D, Chen C, Li L, Lu S, Li Z, Zhou Z, Jing H, Xu Y, Du P, Wang H, Xiong Y, Zheng H, Bai X, Sun H, Wang L, Ye C, Gottschalk M, Xu J: **Dynamics of fecal microbial communities in children with diarrhea of unknown etiology and genomic analysis of associated Streptococcus lutetiensis**. *BMC Microbiol* 2013, **13**:141.

73. Taniuchi M, Sobuz SU, Begum S, Platts-Mills JA, Liu J, Yang Z, Wang XQ, Petri WA, Haque R, Houpt ER: **Etiology of diarrhea in bangladeshi infants in the first year of life analyzed using molecular methods**. *J Infect Dis* 2013, **208**:1794–1802.

74. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poultain J, Qin J, Sicheritz-Ponten T, Tims S: **Enterotypes of the human gut microbiome**. *Nature* 2011, **473**:174–180.

75. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P: **Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa**. *Proc Natl Acad Sci U S A* 2010, **107**:14691–14696.

76. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD: **Linking long-term dietary patterns with gut microbial enterotypes**. *Science (80-)* 2011, **334**:105–108.

77. Taweechotipatr M, Iyer C, Spinler JK, Versalovic J, Tumwasorn S: **Lactobacillus saerimneri and Lactobacillus ruminis: novel human-derived probiotic strains with immunomodulatory activities**. *FEMS Microbiol Lett* 2009, **293**:65–72.

78. Jangi S, Lamont JT: **Asymptomatic colonization by Clostridium difficile in infants: implications for disease in later life**. *J Pediatr Gastroenterol Nutr* 2010, **51**:2–7.

79. Sears CL, Myers LL, Lazenby A, Van Tassell RL: **Enterotoxigenic Bacteroides fragilis**. *Clin Infect Dis* 1995, **20**:S142–S148.

80. Mazmanian SK, Round JL, Kasper DL: **A microbial symbiosis factor prevents intestinal inflammatory disease**. *Nature* 2008, **453**:620–625.

81. Lin A, Bik EM, Costello EK, Dethlefsen L, Haque R, Relman DA, Singh U: **Distinct distal gut microbiome diversity and composition in healthy children from Bangladesh and the United States**. *PLoS One* 2013, **8**:e53838.

82. Sim K, Cox MJ, Wopereis H, Martin R, Knol J, Li MS, Cookson WO, Moffatt MF, Kroll JS: **Improved detection of bifidobacteria with optimised 16S rRNA-gene based pyrosequencing**. *PLoS One* 2012, **7**:e32543.

83. Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ: **Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes**. *Appl Env Microbiol* 2008, **74**:2461–2470.

84. Lindsay B, Pop M, Antonio M, Walker AW, Mai V, Ahmed D, Oundo J, Tamboura B, Panchalingam S, Levine MM, Kotloff K, Li S, Magder LS, Paulson JN, Liu B, Ikumapayi U, Ebruke C, Dione M, Adeyemi M, Rance R, Stares MD, Ukhanova M, Barnes B, Lewis I, Ahmed F, Alam MT, Amin R, Siddiqui S, Ochieng JB, Ouma E, et al.: **Survey of culture, GoldenGate assay, Universal Biosensor assay, and 16S rRNA gene sequencing as alternative methods of bacterial pathogen detection.** *J Clin Microbiol* 2013, **51**:3263–3269.

85. Lay C, Rigottier-Gois L, Holmstrom K, Rajilic M, Vaughan EE, de Vos WM, Collins MD, Thiel R, Namsolleck P, Blaut M, Dore J: **Colonic microbiota signatures across five northern European countries**. *Appl Env Microbiol* 2005, **71**:4153–4155.

86. Lawley TD, Clare S, Walker AW, Stares MD, Connor TR, Raisen C, Goulding D, Rad R, Schreiber F, Brandt C, Deakin LJ, Pickard DJ, Duncan SH, Flint HJ, Clark TG, Parkhill J, Dougan G: **Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing Clostridium difficile disease in mice**. *PLoS Pathog* 2012, **8**:e1002995.

87. Ubeda C, Bucci V, Caballero S, Djukovic A, Toussaint NC, Equinda M, Lipuma L, Ling L, Gobourne A, No D, Taur Y, Jeng RR, van den Brink MR, Xavier JB, Pamer EG: **Intestinal microbiota containing Barnesiella species cures vancomycin-resistant Enterococcus faecium colonization**. *Infect Immun* 2013, **81**:965–973.

88. Senior K: **Faecal transplantation for recurrent C difficile diarrhoea**. *Lancet Infect Dis* 2013, **13**:200–201.

89. Kotloff KL, Blackwelder WC, Nasrin D, Nataro JP, Farag TH, van Eijk A, Adegbola RA, Alonso PL, Breiman RF, Faruque AS, Saha D, Sow SO, Sur D, Zaidi AK, Biswas K, Panchalingam S, Clemens JD, Cohen D, Glass RI, Mintz ED, Sommerfelt H, Levine MM: **The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control study**. *Clin Infect Dis* 2012, **55**:S232–S245.

90. Panchalingam S, Antonio M, Hossain A, Mandomando I, Ochieng B, Oundo J, Ramamurthy T, Tamboura B, Zaidi AK, Petri W, Houpt E, Murray P, Prado V, Vidal R, Steele D, Strockbine N, Sansonetti P, Glass RI, Robins-Browne RM, Tauschek M, Svennerholm AM, Berkeley LY, Kotloff K, Levine MM, Nataro JP: **Diagnostic microbiologic methods in the GEMS-1 case/control study**. *Clin Infect Dis* 2012, **55**:S294–S302.

91. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG: **The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome**. *Gigascience* 2012, **1**:7.

92. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: **Accuracy and quality of massively parallel DNA pyrosequencing**. *Genome Biol* 2007, **8**:R143.

93. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF: **Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.** *Appl Environ Microbiol* 2009, **75**:7537–7541.

94. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM: **The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis**. *Nucleic Acids Res* 2005, **33**:D294–D296.

95. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, others: **Vaginal microbiome of reproductive-age women**. *Proc Natl Acad Sci* 2010:201002611.

96. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R: **UCHIME improves sensitivity and speed of chimera detection**. *Bioinformatics* 2011, **27**:2194–2200.

97. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB**. *Appl Env Microbiol* 2006, **72**:5069–5072.

98. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *J R Stat Soc Ser B* 1995, **57**:289 – 300.

99. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ: **Integration of biological networks and gene expression data using Cytoscape**. *Nat Protoc* 2007, **2**:2366–2382.

100. McMurdie PJ, Holmes S: **Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible**. *PLoS Comput Biol* 2014, **10**:e1003531.

101. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M, Arumugam M, Batto J-M, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N, Jørgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F, Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S, Clément K, Doré J, Kleerebezem M, et al.: **Richness of human gut microbiome correlates with metabolic markers.** *Nature* 2013, **500**:541–6.

102. Mills ED: **Adjusting for covariates in zero-inflated gamma and zero-inflated log-normal models for semicontinuous data**. 2013.

103. Costea PI, Zeller G, Sunagawa S, Bork P: **A fair comparison**. *Nat Methods* 2014, **11**:359.

104. Duan N, Manning WG, Morris CN, Newhouse JP: **Comparison of for Alternative Care Models for the Demand Medical**. *J Bus Econ Stat* 1983, **1**:115–126.

105. Venables WN, Ripley BD: **Modern Applied Statistics with S Fourth edition by**. *World* 2002, **53**(March):86.

106. Golub GH, Heath M, Wahba G: **Generalized cross-validation as a method for choosing a good ridge parameter**. *Technometrics* 1979, **21**:215–223.

107. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning**. *Elements* 2009, **1**:337–387.

108. Hallstrom AP: **A modified Wilcoxon test for non-negative distributions with a clump of zeros**. *Stat Med* 2010, **29**:391–400.

109. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R: **Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample.** *Proc Natl Acad Sci U S A* 2011, **108 Suppl** :4516–4522.

110. McMurdie PJ, Holmes S: **Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data**. *PLoS One* 2013, **8**.

111. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K, Sunyaev SR, Weinstock GM, Bork P: **Genomic variation landscape of the human gut microbiome.** *Nature* 2013, **493**:45–50.

112. Tap J, Mondot S, Levenez F, Pelletier E, Caron C, Furet J-PP, Ugarte E, Muñoz-Tamayo R, Paslier DLE, Nalin R, Dore J, Leclerc M, Mu??oz-Tamayo R, Paslier DLE, Nalin R, Dore J, Leclerc M: **Towards the human intestinal microbiota phylogenetic core.** *Environ Microbiol* 2009, **11**:2574–2584.

113. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto J-M, Quintanilha Dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, et al.: **Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes.** *Nat Biotechnol* 2014, **32**:822–828.