Statistical methods for detecting differentially abundant taxa in metagenomic samples

James Robert White, whitej@umd.edu
Mihai Pop, mpop@umiacs.umd.edu

Abstract

Motivation: Numerous studies are currently underway to characterize the microbial communities inhabiting our world. These studies will dramatically expand our understanding of the microbial biosphere and, more importantly, will reveal the secrets of the complex symbiotic relationship between us and our commensal bacterial communities. An important prerequisite for such discoveries are computational tools able to rapidly and accurately compare large datasets generated from complex bacterial communities.

Results: We describe a statistical method for detecting differentially abundant organisms between two populations using count data (e.g. 16S rRNA surveys). In high-complexity environments, our method employs the false discovery rate to improve specificity and properly handles low abundance taxa. We demonstrate the use of our tool by comparing publicly available human and mouse gut microbiome datasets to identify differences between these bacterial populations at different levels of resolution. We additionally re-analyze the data generated in a recent study on obesity and identify a previously uncharacterized difference between the gut flora of obese and lean human subjects. To illustrate the flexibility of our methods, we further assess differentially abundant metabolic subsystems from 85 newly generated microbial and viral metagenomes.

Availability: A web server implementation of our methods is available at http://www.cbcb.umd.edu/~whitej/metastats/detection.shtml. Source code is freely available at (sourceforge site).

Introduction

The increasing availability of high-throughput, inexpensive sequencing technologies has led to the birth of a new scientific field, metagenomics, encompassing large-scale analyses of the microbial communities that inhabit our bodies and our planet. Large-scale sequencing of bacterial populations allows us a first glimpse at the many microbes that cannot be analyzed through traditional means (only 1-5% of all bacteria can be isolated and independently cultured with current methods). Studies of environmental samples have initially focused on targeted sequencing of individual genes, in particular the 16S subunit of ribosomal RNA [1-7]. This gene is commonly used to characterize the diversity of an environment in studies that involve the random sampling of an environment's genomic content.

Several software tools have been developed in recent years for comparing different environments on the basis sequence data. DOTUR [8], Libshuff [9], S-libshuff [10], SONs [11], MEGAN [12], UniFrac [13], and TreeClimber [14] all focus on different aspects of such an analysis. DOTUR clusters sequences into operational taxonomic units (OTUs) and provides estimates of the diversity of a microbial population thereby providing a coarse measure for comparing different communities. SONs extends DOTUR with a statistical test for estimating

the similarity between two environments, specifically, the fraction of OTUs shared between two communities. Libshuff and *∫*-libshuff provide a hypothesis test for deciding whether two communities are different, and TreeClimber and UniFrac frame this question in a phylogenetic context. Note that these methods aim to assess **whether**, rather than **how** two communities differ. The latter question is particularly important as we begin to analyze the contribution of the microbiome to human health. Metagenomic analysis in clinical trials will require information at individual taxonomic levels to aid the direction of future experiments and treatments. As an example, we would like to identify bacteria whose presence or absence contributes to human disease and develop antibiotic or probiotic treatments. The software MEGAN of Huson *et al.* is one of the first to addresses the nature of taxonomic differences between two environments, albeit at a qualitative level.

A statistical bootstrap approach designed by Rodriguez-Brito *et al.* compares the abundances of *subsystems* (e.g. biochemical pathways, clusters of functionally related genes) in two environments using a difference of medians calculation [15]. Though this method does not depend on the distribution of the subsystems, it ignores variation between multiple subjects from a single environment, and lacks power, often requiring a prohibitive number of samples to achieve statistical significance. Accounting for inter-subject variation is essential for clinical trials when dozens or hundreds of subjects may be taken from each treatment.

In this paper, we describe a rigorous statistical method for detecting differentially abundant taxa in two microbial populations and assess the significance of observed differences. Such rigor is particularly necessary as metagenomic studies are increasingly applied in a clinical setting (e.g. Human Microbiome Project [16]), as well as to cope with the increasing size and complexity of the datasets being analyzed. In high-complexity environments, our method estimates the false discovery rate (FDR) and separately evaluates low abundance taxa. While current microarray analysis packages implement the FDR, they are designed for continuous data rather than discrete counts, and therefore will fail to properly account for the significance of sparse observations in metagenomic data.

We demonstrate the use of our tool by comparing publicly available human and mouse gut microbiome datasets to identify differences between these microflora at different levels of resolution. Furthermore, we re-analyze the data generated in a recent study on obesity and identify a previously uncharacterized difference between the gut flora of obese and lean human subjects. Finally, we apply our methods to metabolic data and determine differentially abundant subsystems between 85 microbial and viral metagenomes. The methods described in this paper have been implemented as a web server (www.cbcb.umd.edu/~whitej/metastats/detection.shtml) and are available as source code at (sourceforge site).

Methods

Our method relies on the following assumptions: (i) we are given data corresponding to two populations (e.g. sick and healthy human gut communities, or individuals exposed to different treatments), each consisting of multiple individuals (or samples); (ii) for each sample we are provided with a list of taxonomic units (taxa) present in the sample (whether individual organisms or phylogenetic groupings) as well as an estimate of the relative abundance of these

taxa in the sample. Our goal is to identify individual taxa in such datasets that "explain" the difference between the two populations, i.e. taxa whose abundance in the two populations is different. Furthermore, we develop a statistical measure of our confidence in the observed differences. In this paper we focus on data generated through 16S rRNA surveys, however the methods can be applied to any other experimental technique that provides abundance data.

The taxa abundance matrix

The input to our method consists of taxonomic counts from multiple subjects in two populations. A taxa abundance matrix (TAM) can be created using the frequency of each taxon observed within each individual. The i^{th} row of this matrix represents a specific taxon, while the j^{th} column represents a single individual. Thus, the cell in the i^{th} row and j^{th} column is the total number of observations of taxon *i* in subject *j* (fig. 1). Every distinct observation is represented only once in the matrix, i.e. overlapping taxa are not allowed. If there are *g* subjects in the first population, they are represented by the first *g* columns of the matrix; the remaining columns represent subjects from the second population.

	S1	**S2**	.....	**S(*N*-1)**	**S*N***
T1	*f(1,1)*	*f(1,2)*	.....	*f(1,N-1)*	*f(1,N)*
T2	*f(2,1)*	*f(2,2)*			
.	.				.
.	.				.
.	.				.
T(*M*-1)	*f(M-1,1)*				
T*M*	*f(M,1)*		.....		*f(M,N)*

Figure 1 Format of the taxa abundance matrix. Each row represents a specific taxon, while each column represents a subject (replication). The frequency of the i^{th} taxon in the j^{th} subject (*f(i,j)*) is recorded in the corresponding cell of the matrix. If there are *g* subjects in the first population, they are represented by the first *g* columns of the matrix, while the remaining columns represent subjects from the second population.

T statistic computation

To allow the comparison of abundance numbers across multiple individuals, we convert the raw abundance measure to a fraction representing the relative contribution of each taxon to each of the individuals (columns in TAM). This results in a relative proportion matrix (RPM) of the same dimensions as the TAM, but the cell in the i^{th} row and the j^{th} column (which we shall denote a_{ij}) is the proportion of taxon *i* observed in individual *j*. Note when there are differences in the total number of observations from each subject, it is necessary to use a normalization procedure prior to calculation of sample mean and variance.

For each taxon *i*, we compare its abundance across the two populations using the standard two-sample *t* statistic. Specifically, we calculate the mean proportion $\bar{x}_{i1}$, and variance s^2_{i1} of treatment 1:

$$\bar{x}_{i1} = \frac{1}{n_1} \sum_{j \in \text{treatment 1}} a_{ij}$$

$$s^2_{i1} = \frac{1}{n_1 - 1} \sum_{j \in \text{treatment 1}} \left(a_{ij} - \bar{x}_{i1} \right)^2$$

Similarly, we calculate $\bar{x}_{i2}$ and s^2_{i2} for treatment 2. Note n_1 and n_2 are the number of subjects in treatment 1 and treatment 2, respectively. Finally, the two-sample *t* statistic for each taxon *i* is calculated:

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

A challenge in analyzing count data is that the *t* statistic is not accurate in the case of low frequency organisms. We performed multiple simulation studies to uncover the limits of this technique. The first simulation involved 10 subjects from the same population in which the true mean proportion of an organism (X) is known, which we'll denote p_X. The population is normally distributed with a standard deviation p_X*0.1. A single experiment *k* begins by choosing 10 subjects randomly from the population and simulating 50 taxonomic observations for each subject. Each sample is classified as organism X or not organism X in a taxa abundance matrix based on the true proportion of the organism in each subject. The resulting table is converted to proportions from which we calculate the one-sample *t* statistic for relative abundances of organism X:

$$t_k = \frac{\bar{x}_k - p_X}{\sqrt{\frac{s^2_k}{50}}}.$$

As the population is normally distributed, t_k follows an approximate *t*-distribution with 9 degrees of freedom for adequately large values of p_X. We ran simulations of 250,000 experiments using p_X = {0.5, 0.2, 0.1, 0.05, 0.01} and found the *t*-distribution to be reasonably valid for all values above p_X= 0.05. This particular proportion implies we expect to see 2.5 observations per subject on average, and so, 25 observations over all subjects. Our simulations indicate that accurate results can be obtained for taxa containing 25 or more observations within each population, therefore, as a heuristic, taxa rows corresponding to fewer than 25 observations in both populations are removed from the RPM described above, and analyzed separately as described below. Additional simulations varying sampling rates between subjects and increasing sampling depth also supported this heuristic.

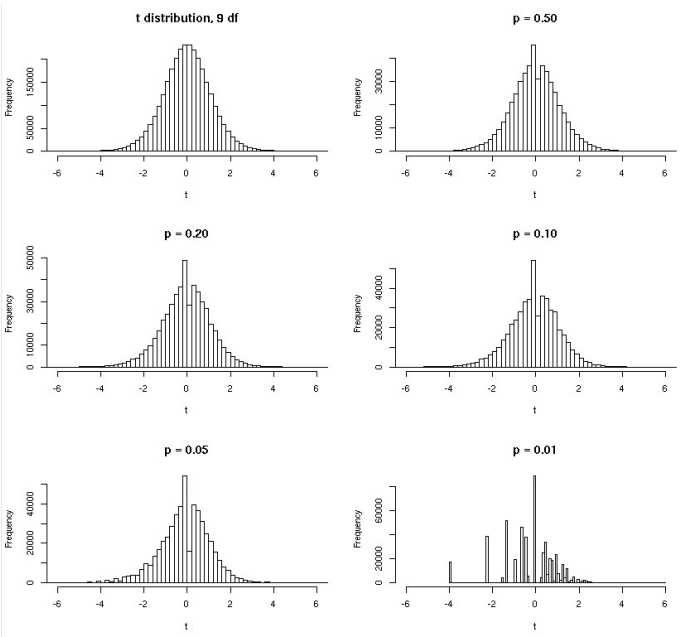

Figure 2 Distribution of one-sample *t* statistics for simulated values of p_X. As the value of p_X becomes small, the *t*-distribution no longer accurately approximates the true distribution. One can see that the distribution of *t* statistics becomes sparse and asymmetric due to the limited number of observations in each experiment.

Permuted p-values

To determine a threshold for detecting differentially abundant taxa, and to assign accurate confidence values to the observed differences between populations, we estimate the null distribution of t_i nonparametrically using a permutation method as described in Storey and Tibshirani (2003). Specifically, we randomly permute the treatment labels of the columns of the RPM and recalculate the *t* statistics. Note that the permutation maintains that there are n_1 replications for treatment 1 and n_2 replications for treatment 2. Repeating this procedure for *B* trials, we obtain *B* sets of *t* statistics: t_1^{0b}, …, t_M^{0b}, b = 1, …, *B*, where *M* is the number of taxa in the RPM.

Finally, the p-values for each taxon *i*, (*i* = *1*, …, *M*) are calculated as the fraction of permuted tests with a higher *t* statistic than the original:

$$p_i = \frac{1}{BM} \sum_{b=1}^{B} \#\left\{ j : \left| t_j^{0b} \right| \geq |t_i|, j = 1,\ldots,B \right\}.$$

All experiments described below set *B* = 1000, and so the precision of the p-values will be at worst on the order of 1/*B*. One should cautiously set the number of permutations so that the precision of the p-values is well below the significance threshold used to call taxa differentially abundant. In our studies, 1000 permutations are appropriate because our significance thresholds are greater than 0.01.

Low frequency taxa

We do not include low frequency taxa (< 25 observations in either treatment) in the analysis presented above because the null distribution of the *t* statistic for these rare organisms varies widely depending on their relative abundance. However, the probability of observing a rare taxon is approximately equal for all individuals in a treatment, therefore, we can test rare taxa for differential abundance using Fisher's exact test. Fisher's exact test models sampling infrequent categories according to a hypergeometric distribution (sampling without replacement), rather than a binomial distribution. The frequencies of the TAM for each low frequency taxon are pooled to create a 2x2 contingency table (fig. 3), which acts as input for a two-tailed test. Using the notation from figure 3, the null hypergeometric probability of observing a 2x2 contingency table is:

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

By calculating this probability for a given table, and all tables more extreme than that observed, one can calculate the exact probability of obtaining the original table by chance assuming that the null hypothesis (i.e. no difference abundance) is true [17]. There have been decades of debate over Fisher's exact test as possibly being too conservative [18, 19]. However, for practical purposes, a conservative significance test is preferable over encountering additional type I error.

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

Figure 3 Format of a 2x2 contingency table used in testing for differential abundance between rare taxa. f_{11} is the number of observations of taxon *i* in all individuals from treatment 1. f_{21} is the number of observations that are not taxon *i* in all individuals from treatment 1. f_{12} and f_{22} are similarly defined for treatment 2.

The false discovery rate

For complex environments (many taxa), the direct application of the *t* statistic as described above is inappropriate as we are faced with a multiple hypothesis testing scenario. An intuitive correction involves decreasing the p-value cutoff proportional to the number of tests performed (a Bonferroni correction), thereby reducing the number of false positives. This approach, however, results in a significant decrease in statistical power, making detection of differential abundance difficult.

An alternative approach aims to control the false discovery rate (FDR), which is defined as the proportion of false positives within the set of predictions [20], in contrast to the false positive rate defined as the proportion of false positives within the entire set of tests. In this context, the significance of a test is measured by a q-value, an individual measure of the FDR for each test [21].

We implemented the following algorithm, adapted from Storey and Tibshirani (2003), for the automated computation of q-values:

Given an ordered list of p-values, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(M)}$, and a range of values $\lambda = 0, 0.01, 0.02, \ldots, 0.90$, we compute

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

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

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

Thus, the hypothesis test with p-value $p_{(i)}$ has a corresponding q-value of $\hat{q}(p_{(i)})$. Note that this method yields conservative estimates of the true q-values, i.e. $\hat{q}(p_{(i)}) \geq q(p_{(i)})$.

Data used in this paper

Human gut 16S rRNA sequences were prepared as described in Eckburg *et al.* and Ley *et al.* (2006) and are available in GenBank, accession numbers: DQ793220-DQ802819, DQ803048, DQ803139-DQ810181, DQ823640-DQ825343, AY974810-AY986384. Mouse gut 16S rRNA sequences were prepared as described in Ley *et al.* (2005) and obtained from GenBank accession numbers: DQ014552-DQ015671, AY989911-AY993908. We acquired metabolic functional profiles of 85 metagenomes from the online supplementary materials of Dinsdale *et al.* (2008) (http://www.theseed.org/DinsdaleSupplementalMaterial/).

16S rRNA taxonomic assignment

There are several widely used methods for taxonomic assignment of 16S rRNA. Different approaches include sequence comparison, sequence composition, and phylogenetic analysis. In our experiments we assigned all 16S sequences to taxa using a naïve Bayesian classifier currently employed by the Ribosomal Database Project II (RDP) [22]. This software rapidly classifies sequences from kingdom to genus according to Bergey's *Taxonomic Outline of the Prokaryotes* [23]. Trained on ~23,000 pre-classified 16S sequences, the RDP classifier provides a statistical confidence for each classification, and is available for use online (http://rdp.cme.msu.edu/classifier/classifier.jsp).

Results

Taxa associated with human obesity

Recently Ley *et al.* (2006) published a study of gut microbes associated with obesity in humans and concluded that obesity has a microbial element, specifically that Firmicutes and Bacteroidetes are differentially abundant between lean and obese humans. Obese subjects had a significantly higher relative abundance of Firmicutes and a lower relative abundance of Bacteriodetes than the lean subjects. Furthermore, obese subjects were placed on a calorie-restricted diet for one year, after which the subjects' gut microbiota more closely resembled that of the lean individuals.

We obtained the 20,609 16S rRNA genes sequenced in Ley *et al.* and assigned them to taxa at different levels of resolution (note that 2,261 of the 16S sequences came from a previous study [2]). We initially sought to re-establish the primary result from this paper using our methodology. Figure 4 illustrates the shift in Firmicutes and Bacteroidetes abundances before and after the obese subjects' diets, and our method agreed with the results of the original study: Firmicutes are significantly more abundant in obese subjects ($P = 0.003$) and Bacteroidetes are significantly more abundant in the lean population ($P < 0.001$). Furthermore, our method also detected Actinobacteria to be differentially abundant, a result not reported by Ley *et al.* Approximately 5% of the gut population was composed of Actinobacteria in obese subjects and was significantly less frequent in lean subjects ($P = 0.004$) (fig. 5). This result indicates our method is more sensitive than the approach used in the original study.

To explore whether we could refine the broad conclusions of the initial study, we re-analyzed the data at the class level. We identified four classes of organisms that were differentially abundant: Clostridia ($P = 0.006$), Bacteroidetes ($P < 0.001$), Actinobacteria ($P = 0.003$), and Delta-proteobacteria ($P = 0.003$) (fig. 6). The first three were the dominant members of the corresponding phyla (Firmicutes, Bacteroides, Actinobacteria, respectively) and followed the same distribution as observed at a coarser level. At the phylum level, Proteobacteria (the phylum parent to Delta-proteobacteria) were not found to be differentially abundant between lean and obese individuals, however, at the class level, delta-proteobacteria were significantly enriched in lean individuals. Proteobacteria were not detected as differentially abundant due to a severe bloom of Gamma-proteobacteria in a single obese subject (15% of the individual's 16S sequences), leading to a high sample variance in obese subjects and a small overall *t* statistic.

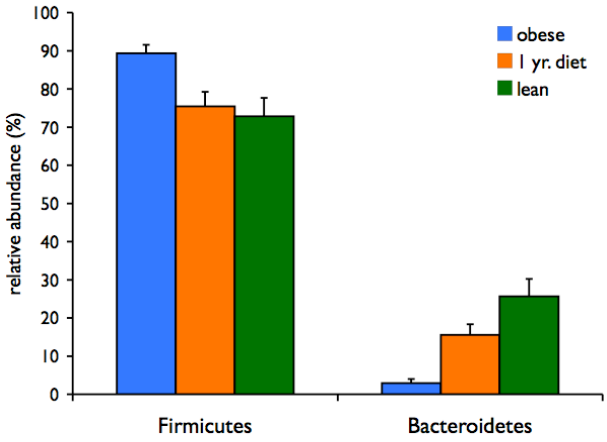

Figure 4 Mean relative abundances (%) of Firmicutes and Bacteroidetes (± s.e.) in lean and obese subjects, as well as obese subjects after a 52-week calorie-restricted diet. After the obese subjects finished the diet, their microbial communities began to resemble that of the lean individuals.

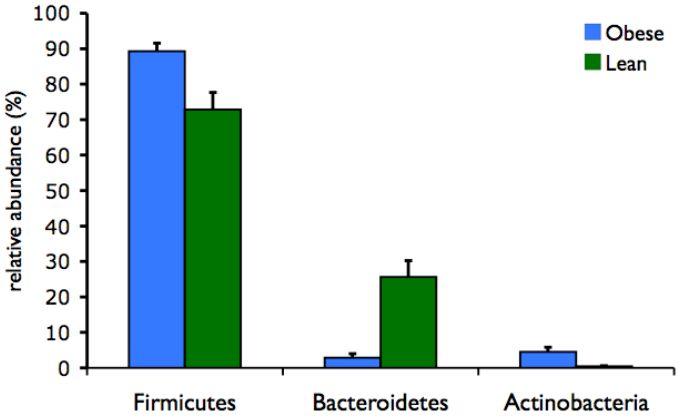

Figure 5 Differentially abundant phyla detected using our method (mean percentage ± s.e., p-value ≤ 0.05). No p-value correction for multiple hypothesis tests was employed. We successfully re-established the major result of Ley *et al.*, and discovered that Actinobacteria are also differentially abundant. Both Firmicutes and Actinobacteria have significantly higher relative abundances in obese people than lean people. In the lean population, Bacteroidetes make up a higher proportion of the gut microbiota than in the obese population.

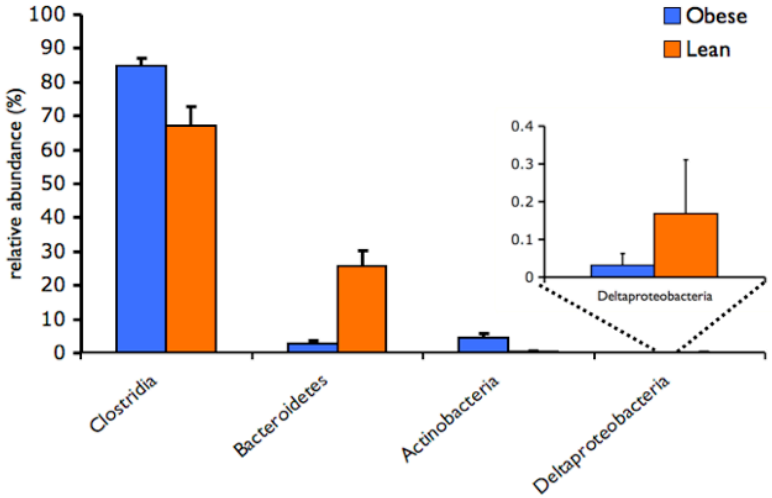

Figure 6 Differentially abundant classes detected (mean percentage ± s.e., p-value ≤ 0.05). No p-value correction for multiple hypothesis tests was employed. Mean proportion of Deltaproteobacteria in obese and lean subjects was 0.03% and 0.44%, respectively. Clostridia are responsible for the differential abundance of Firmicutes.

Human vs. mouse gut microbial communities

Mouse models are important research tools in many biomedical areas, including the study of the commensal microbial populations [24]. To evaluate the differences between human and mouse gut microbial communities, we applied our methods to 6,250 16S rRNA sequences from the 7 human and 12 mouse control subjects of two obesity studies [4, 5]. We discovered seven differentially abundant classes (see table 1). Two of the three most abundant classes, Clostridia and Bacilli, were differentially abundant: humans maintained higher levels of Clostridia (P = 0.018) while mice had more abundant Bacilli (P = 0.003).

We also applied our methods at the genus level (101 genera were identified in these two datasets), employing the false discovery rate method for assessing the significance of our results using a q-value cutoff of 0.05. We identified 21 differentially abundant genera, several of which were well-represented taxa (see table 2). Surprisingly, we found virtually no *Lactobacillus* in the human samples, a genus previously characterized in the human gut [25, 26]. However, our result is consistent with a prior study which also found a low abundance of Lactobacilli in the distal human gut [27]. These results indicate substantial interspecies variation between human and mouse gut microbial communities, and these differences should be taken into account when using the mouse models in microbiome studies.

Class name	Human	Mouse	p-value
Clostridia	66.9 ± 5.8	49.1 ± 3.2	0.019
Bacilli	4.27 ± 0.97	12.1 ± 1.9	0.003
Actinobacteria (class)	0.447 ± 0.18	0.979 ± 0.17	0.041
Verrucomicrobiae	0.162 ± 0.14	0	0.006
Alphaproteobacteria	0.115 ± 0.12	0	0.026
Epsilonproteobacteria	0	0.261 ± 0.17	0.002
TM7 genera incertae sedis	0	0.220 ± 0.10	0.032

Table 1 Differentially abundant classes of organisms from human and mouse gut microbiota (p-values ≤ 0.05). Human and mouse columns display mean relative abundance (%) ± standard error. Cells containing '0' indicate that no observations of the taxa were found. Clostridia and Bacilli, two of the three most abundant classes observed were differentially abundant.

Genera	Human	Mouse	q-value
Bacteroides	14.6 ± 4.44	0.51 ± 0.40	0.025
Faecalibacterium	12.4 ± 2.27	0	< 0.001
Ruminococcus	10.7 ± 2.09	0.78 ± 0.30	< 0.001
Roseburia	8.58 ± 2.05	2.00 ± 0.44	0.025
Dorea	1.97 ± 0.59	6.35 ± 0.89	0.002

Tannerella	0.99 ± 0.68	31.9 ± 3.18	< 0.001
Sporobacter	0.73 ± 0.42	0.03 ± 0.03	< 0.001
Syntrophococcus	0.67 ± 0.34	2.77 ± 0.62	0.038
Mahella	0.50 ± 0.22	0	< 0.001
Succiniclasticum	0.35 ± 0.35	0	< 0.001
Phascolarctobacterium	0.22 ± 0.17	0	0.025
Akkermansia	0.16 ± 0.14	0	0.040
Bryantella	0.15 ± 0.10	9.75 ± 1.30	< 0.001
Eggerthella	0.10 ± 0.07	0.84 ± 0.16	0.003
Parasporobacterium	0.09 ± 0.09	1.92 ± 0.62	0.040
Hespellia	0.07 ± 0.06	0.46 ± 0.09	0.047
Lactobacillus	0	4.76 ± 1.50	0.025
Sporobacterium	0	0.51 ± 0.21	< 0.001
Acetitomaculum	0	0.47 ± 0.23	< 0.001
Oribacterium	0	0.37 ± 0.15	0.011
Helicobacter	0	0.26 ± 0.17	0.020

Table 2 Differentially abundant genera of organisms from human and mouse gut microbiota (q-values ≤ 0.05). Human and mouse columns display mean relative abundance (%) ± standard error. Cells containing '0' indicate that no observations of the taxa were found. Since we thresholded using q-values, we expect that only one of these 21 genera is a false positive.

Differentially abundant metabolic subsystems in microbial and viral metagenomes

While assessing the population microbes in a community is useful, it does not provide a detailed description of the metabolic potential of the microbial community. The discovery of rapidly changing elements in the genome such as CRISPRs [28-30] has shown that although two organisms have identical 16S genes, their functions may be variable. Thus, recent studies have proposed examining the *pan-genome* of an environment rather than organisms individually [31, 32]. Recently, Dinsdale *et al.* profiled 87 different metagenomic samples (~15 million sequences) using the SEED platform (http://www.theseed.org) [33]. We obtained functional profiles from 45 microbial and 40 viral metagenomes analyzed in this study to identify differentially abundant subsystems. Of the 26 subsystems detected in the profiles, 11 were found to be significantly different (p-values ≤ 0.02) (fig. 7). Thus, we expect less than one false positive overall. Subsystems for nucleotides and DNA metabolism were significantly more abundant in viral metagenomes, while nitrogen metabolism, membrane transport, and carbohydrates were all enriched in microbial communities. In contrast to the original study, virulence subsystems were less abundant than previously reported, and they were not differentially abundant between the microbial and viral metagenomes.

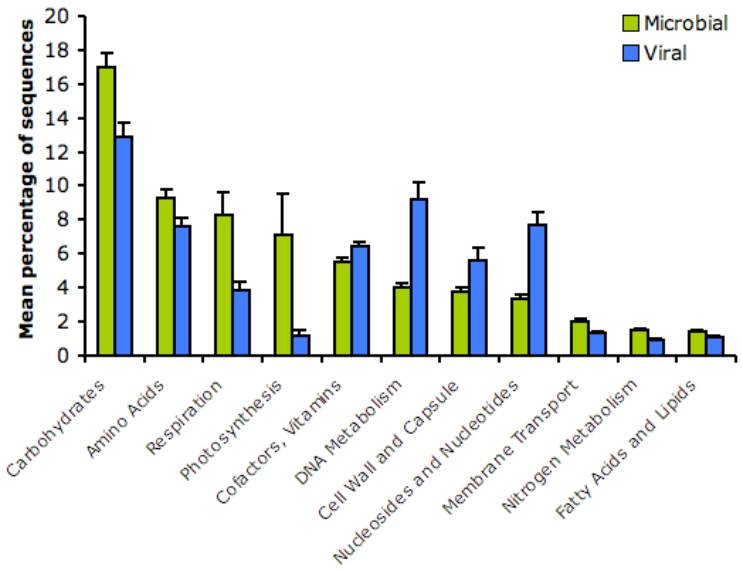

Figure 7 Differentially abundant metabolic subsystems between microbial and viral metagenomes (mean percentage ± s.e., p-values ≤ 0.02). We find that viral metagenomes are significantly enriched for nucleotides and nucleosides ($P < 1e\text{-}6$) and DNA metabolism ($P < 1e\text{-}4$). Processes for respiration, photosynthesis, and carbohydrates are all overrepresented in microbial metagenomes.

Discussion and Conclusions

We have presented a statistical method for handling frequency data to detect differentially abundant categories between two populations. While this method has been described in the context of 16S data, it can be applied to the analysis of frequency data generated through other means, including random shotgun sequencing of environmental samples (binning tools could provide the abundance information in this case), or microarray-based comparisons between environments (e.g. using the PhyloChip [6]). Detection of differentially abundant subsystems with multiple subjects from each environment will also benefit from this approach. Our method can also be generalized to experiments with more than two populations by substituting the *t* test with a one-way ANOVA test. Furthermore, if only a single sample from each treatment is available, a chi-squared test could be easily substituted for a *t* test, which is known to be appropriate for cases in which a category is observed ≥ 10 times for each treatment [17].

In the coming years metagenomic studies will increasingly be applied in a clinical setting, requiring new algorithms and software tools to be developed that can exploit data from hundreds to thousands of patients. The methods described above represent an initial step in this direction by providing a robust and rigorous statistical method for identifying organisms whose differential abundance correlates with disease. These methods are available via web server through www.cbcb.umd.edu/~whitej/metastats/detection.shtml.

Acknowledgments

The authors were funded in part by a grant from the Bill and Melinda Gates Foundation.

References

1. Bik EM, Eckburg PB, Gill SR, Nelson KE, Purdom EA, Francois F, Perez-Perez G, Blaser MJ, Relman DA: **Molecular analysis of the bacterial microbiota in the human stomach**. *Proc Natl Acad Sci U S A* 2006, **103**(3):732-737.
2. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the human intestinal microbial flora**. *Science* 2005, **308**(5728):1635-1638.
3. Gao Z, Tseng CH, Pei Z, Blaser MJ: **Molecular analysis of human forearm superficial skin bacterial biota**. *Proc Natl Acad Sci U S A* 2007, **104**(8):2927-2932.
4. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI: **Obesity alters gut microbial ecology**. *Proc Natl Acad Sci U S A* 2005, **102**(31):11070-11075.
5. Ley RE, Turnbaugh PJ, Klein S, Gordon JI: **Microbial ecology: human gut microbes associated with obesity**. *Nature* 2006, **444**(7122):1022-1023.
6. Palmer C, Bik EM, Digiulio DB, Relman DA, Brown PO: **Development of the Human Infant Intestinal Microbiota**. *PLoS Biol* 2007, **5**(7):e177.
7. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ: **Microbial diversity in the deep sea and the underexplored "rare biosphere"**. *Proc Natl Acad Sci U S A* 2006, **103**(32):12115-12120.
8. Schloss PD, Handelsman J: **Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness**. *Applied and environmental microbiology* 2005, **71**(3):1501-1506.
9. Singleton DR, Furlong MA, Rathbun SL, Whitman WB: **Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples**. *Applied and environmental microbiology* 2001, **67**(9):4374-4376.
10. Schloss PD, Larget BR, Handelsman J: **Integration of microbial ecology and statistics: a test to compare gene libraries**. *Applied and environmental microbiology* 2004, **70**(9):5485-5492.
11. Schloss PD, Handelsman J: **Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures**. *Applied and environmental microbiology* 2006, **72**(10):6773-6779.
12. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome Res* 2007, **17**(3):377-386.
13. Lozupone C, Knight R: **UniFrac: a new phylogenetic method for comparing microbial communities**. *Applied and environmental microbiology* 2005, **71**(12):8228-8235.
14. Schloss PD, Handelsman J: **Introducing TreeClimber, a test to compare microbial community structures**. *Applied and environmental microbiology* 2006, **72**(4):2379-2384.
15. Rodriguez-Brito B, Rohwer F, Edwards RA: **An application of statistics to comparative metagenomics**. *BMC Bioinformatics* 2006, **7**:162.
16. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI: **The human microbiome project**. *Nature* 2007, **449**(7164):804-810.
17. Zar JH: **Biostatistical analysis**, 4th edn. Upper Saddle River, N.J.: Prentice Hall; 1999.

18. Agresti A: **A Survey of Exact Inference for Contingency Tables**. *Statistical Science* 1992, **7**(1):131-153.
19. Yates F: **Tests of Significance for 2 X 2 Contingency-Tables**. *J Roy Stat Soc a Sta* 1984, **147**:426-463.
20. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing**. *J Roy Stat Soc B Met* 1995, **57**(1):289-300.
21. Storey JD, Tibshirani R: **Statistical significance for genomewide studies**. *Proc Natl Acad Sci U S A* 2003, **100**(16):9440-9445.
22. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy**. *Applied and environmental microbiology* 2007, **73**(16):5261-5267.
23. **Bergey's Taxonomic Outline of the Prokaryotes**, 2nd edn. New York, NY: Springer-Verlag; 2004.
24. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI: **An obesity-associated gut microbiome with increased capacity for energy harvest**. *Nature* 2006, **444**(7122):1027-1031.
25. Conway PL, Gorbach SL, Goldin BR: **Survival of lactic acid bacteria in the human stomach and adhesion to intestinal cells**. *Journal of dairy science* 1987, **70**(1):1-12.
26. Kullen MJ, Sanozky-Dawes RB, Crowell DC, Klaenhammer TR: **Use of the DNA sequence of variable regions of the 16S rRNA gene for rapid and accurate identification of bacteria in the Lactobacillus acidophilus complex**. *Journal of applied microbiology* 2000, **89**(3):511-516.
27. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome**. *Science* 2006, **312**(5778):1355-1359.
28. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes**. *Science* 2007, **315**(5819):1709-1712.
29. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes**. *PLoS computational biology* 2005, **1**(6):e60.
30. Kunin V, Sorek R, Hugenholtz P: **Evolutionary conservation of sequence and secondary structures in CRISPR repeats**. *Genome Biol* 2007, **8**(4):R61.
31. Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC, Jr., Rohwer F: **Using pyrosequencing to shed light on deep mine microbial ecology**. *BMC genomics* 2006, **7**:57.
32. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: **The microbial pan-genome**. *Current opinion in genetics & development* 2005, **15**(6):589-594.
33. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L *et al*: **Functional metagenomic profiling of nine biomes**. *Nature* 2008, **452**(7187):629-632.