OPEN ACCESS Freely available online

PLOS ONE

Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data

Patricio S. La Rosa[1], J. Paul Brooks[2], Elena Deych[1], Edward L. Boone[2], David J. Edwards[2], Qin Wang[2], Erica Sodergren[3], George Weinstock[3], William D. Shannon[1]*

1 Division of General Medical Sciences, Department of Medicine, Washington University School of Medicine, St. Louis, Missouri, United States of America, **2** Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, Virginia, United States of America, **3** The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, United States of America

Abstract

This paper presents new biostatistical methods for the analysis of microbiome data based on a fully parametric approach using all the data. The Dirichlet-multinomial distribution allows the analyst to calculate power and sample sizes for experimental design, perform tests of hypotheses (e.g., compare microbiomes across groups), and to estimate parameters describing microbiome properties. The use of a fully parametric model for these data has the benefit over alternative non-parametric approaches such as bootstrapping and permutation testing, in that this model is able to retain more information contained in the data. This paper details the statistical approaches for several tests of hypothesis and power/sample size calculations, and applies them for illustration to taxonomic abundance distribution and rank abundance distribution data using HMP Jumpstart data on 24 subjects for saliva, subgingival, and supragingival samples. Software for running these analyses is available.

Citation: La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, et al. (2012) Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. PLoS ONE 7(12): e52078. doi:10.1371/journal.pone.0052078

Editor: Ethan P. White, Utah State University, United States of America

Received April 2, 2012; **Accepted** November 13, 2012; **Published** December 20, 2012

Funding: This work was supported by National Institutes of Health (NIH) Grant U54 HG004968 ''Human Microbiome Project Consortium Sequencing of Healthy People'', NIH Grant 1UH2AI083265 ''The Neonatal Microbiome and Necrotizing Enterocolotis'', and St. Louis Children's Hospital and Children Discovery Institute Grant ''The St. Louis Neonatal Gut Microbiome Initiative''. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wshannon@wustl.edu

Introduction

The NIH Human Microbiome Project (HMP) [1] aims at characterizing, using next generation sequencing technology, the genetic diversity of microbial populations living in and on humans, and at investigating their roles in the functioning of the human body, such as their effects in nutrition and susceptibility to disease [2]. In just a few years, much work has been done to optimize the processes for collecting microbiome samples, processing the DNA, running the sequencing technology, and generating taxonomies/phylogenies from these sequences [3]. These developments will facilitate access to microbiome technology for laboratories of all sizes, enabling application in varied fields of biology, from agriculture to human disease research. However, the biostatistical analysis of metagenomic data is still being developed. Several methods to analyze metagenomic data have been proposed based on exploratory cluster analysis, bootstrap or resampling methods, and application of univariate and non-parametric statistics to subsets of the data [4–12]. However, these methods require a significant reduction of information, such as Unifrac [7] which reduces sequence data to pairwise distances, or ignoring correlations and the multivariate structure inherent in microbiome data, such as Metastats [12] which does univariate 'one-taxa-at-a-time' analyses.

Given the multivariate nature of the metagenomic data, having multivariate analysis tools is becoming important in the microbiome research community. Microbiome researchers are interested in testing multivariate hypotheses concerning the effects of treatments or experimental factors on whole assemblages of bacterial taxa, and in estimating sample sizes for such experiments. These types of analyses are useful for studies aiming at assessing the impact of microbiota on human health and on characterizing the microbial diversity in general. Statistical methods to design and analyze such studies will contribute to the translation of microbiome research from technical (bench) development to clinical (bedside) application.

The focus of this work is to develop multivariate methods to test for differences in bacterial taxa composition between groups of metagenomic samples. Multivariate non-parametric methods based on permutation test such as Mantel test [13,14], Analysis of Similarity (ANOSIM) [15], and NP-Manova [16] are widely used among community ecologists for this purpose. However, although these three methods are attractive when a parametric distribution of the data is unknown, we believe they are not always appropriate for analyzing microbiome data. First, although a hypothesis of group difference can be tested, the results of these tests are difficult to interpret since they cannot quantify the size of the difference between the groups in terms of bacterial taxa composition. Second, permutation tests work under the assumption that the dispersion (variability) of samples within groups is the same in all groups [16], a strong assumption which when violated can lead to inflation of type I error. Third, non-parametric methods are usually less powerful than parametric methods, so

when a parametric alternative is available it should be the preferred method to model metagenomic data.

In this paper, we present biostatistical methods for the analysis of microbiome data based on a fully multivariate parametric approach. In particular, the parametric model used in this paper is the Dirichlet-Multinomial distribution which has been shown recently to model metagenomic data well. In [17] the authors apply the Dirichlet-multinomial mixture for the probabilistic modeling of microbial metagenomics data, which was used to successfully cluster communities into groups with a similar composition. However, a multivariate hypothesis testing framework to compare populations using this model was not derived. In this work, we apply a different parameterization of Dirichlet-multinomial model to the one presented in [17], which is suitable to perform hypothesis testing across groups based on difference between location (mean comparison) as well as scales (variance comparison/dispersion). Using this model, we develop methods to perform parameter estimation, multivariate hypothesis testing power and sample size calculation. An open source R statistical software package ('HMP: Hypothesis Testing and Power Calculations for Comparing Metagenomic Samples from HMP') for fitting these models and tests is available [18].

In addition, the methods developed here are not constrained by computational resources and work for any size microbiome dataset (e.g., number of sequence reads and samples). These methods and are also likely applicable to phylogenetic analysis which is currently being investigated.

Materials and Methods

Ethics Statement

Subjects involved in the study provided written informed consent for screening, enrollment and specimen collection. The protocol was reviewed and approved by the Institutional Review Board at Washington University in St. Louis. The data were analyzed without personal identifiers. Research was conducted according to the principles expressed in the Declaration of Helsinki.

Human Microbiome Data

Human microbiome data analyzed in this paper are from the subgingival, supragingival, and saliva oral sites of 24 subjects (male and female), 18–40 years old, from two geographic regions of the US: Houston, TX and St. Louis, MO [19]. The analyses presented here illustrate how the Dirichlet-multinomial biostatistical analysis is used with real data. Approximately 1×10^5 sequences were obtained from the V1–V3 and V3–V5 variable regions of the 16S ribosomal RNA gene, and collapsed into a single sample. The sequencing was performed at one of four genome sequencing centers (J. Craig Venter Institute, Broad Institute, Human Genome Sequencing Center at Baylor, and Genome Sequencing Center at Washington University in St. Louis). Sequence reads were assigned to bacterial taxa using the Ribosomal Database Project (RDP) classifier [20], which provides a confidence score for each taxonomic classification. Only taxa labels with a confidence score $>=80\%$ were retained in this analysis, and taxa labels below this threshold were relabeled as unknown. Although the choice of an 80% threshold on the confidence score is arbitrary, in [21] it was shown that threshold ranging between 50% to 90% provided an average classification performance of between 77% at the genus level up to 97% at the phylum level.

Statistical Model for HMP Data

Dirichlet-multinomial model. Consider a set of microbiome samples measured on P subjects with K distinct taxa at an arbitrary level (e.g., phylum, class, etc.) identified across all samples. Not all taxa need to be found in all samples. Let x_{ik}, $i=1,\ldots,P$; $k=1,\ldots,K$ be the number of reads in subject i for taxon k, and let be the taxa count vector obtained from sample i. Note that x_{ik} is 0 when taxon k is not in sample i. Let $N_{i.}=\sum_{k=1}^{K} x_{ik}$ be the total number of sequence reads in sample i, $N_{.k}=\sum_{i=1}^{P} x_{ik}$ be the total number of sequence reads for taxon k across all samples, and be the total number of sequences over all samples and taxa. Table 1 shows the format of an RDP-mapped microbiome data set.

Count data such as this is routinely analyzed using a multinomial distribution which is appropriate when the true frequency of each category (e.g., each taxon in microbiome data) is the same across all samples. This implies that as the number of sample points increases (i.e., number of reads) within each sample, taxa frequencies in all samples converge to the same value (e.g., all samples converge onto 40% taxa A, 25% taxa B,…) with no variability between samples. When the data exhibit overdispersion this convergence result does not occur (i.e., taxa frequencies in all samples do not converge to the same values), and the multinomial model is incorrect [22]. Hypothesis testing based on the multinomial model in the presence of overdispersion can result in an increased Type I Error (i.e., saying the microbiome samples are different when they are not) [23].

The Dirichlet-multinomial distribution prevents Type I Error inflation by taking into account the overdispersion in count data in the form displayed in Table 1. It can be characterized by the following two set of parameters [24]: $\pi=\{\pi_j,\ j=1,...,K\}$, $0\le\pi_j\le1$, $\sum\pi_j=1$ which is a vector of the expected taxa frequencies, and $\theta\ge0$ which is a number indicating the amount of overdispersion. Using this parameterization, the Dirichlet-multinomial distribution is defined as [24]:

$$\mathrm{P}(\boldsymbol{X}_i=\boldsymbol{x}_i;\boldsymbol{\pi},\theta)= \frac{N_{i.}!}{x_{i1}!\ldots x_{iK}!}\frac{\Pi_{\mathrm{j}=1}^{K}\Pi_{\mathrm{r}=1}^{x_{ij}}\{\pi_{\mathrm{j}}(1-\theta)+(r-1)\theta\}}{\Pi_{r=1}^{N_{i.}}(1-\theta)+(r-1)\theta}. \quad (1)$$

Table 1. Format of a microbiome data set for P subjects and K distinct taxa at an arbitrary level (e.g., Phylum, Class, etc.).

	Taxa				
Sample	**1**	**2**	**…**	***K***	***Total***
1	X_{11}	X_{12}	…	X_{1K}	N_{1*}
2	X_{21}	X_{22}	…	X_{2K}	N_{2*}
⋮	⋮	⋮	⋱	⋮	⋮
P	X_{P1}	X_{P2}	…	X_{PK}	N_{P*}
Total	N_{*1}	N_{*2}	…	N_{*K}	N_{**}

doi:10.1371/journal.pone.0052078.t001

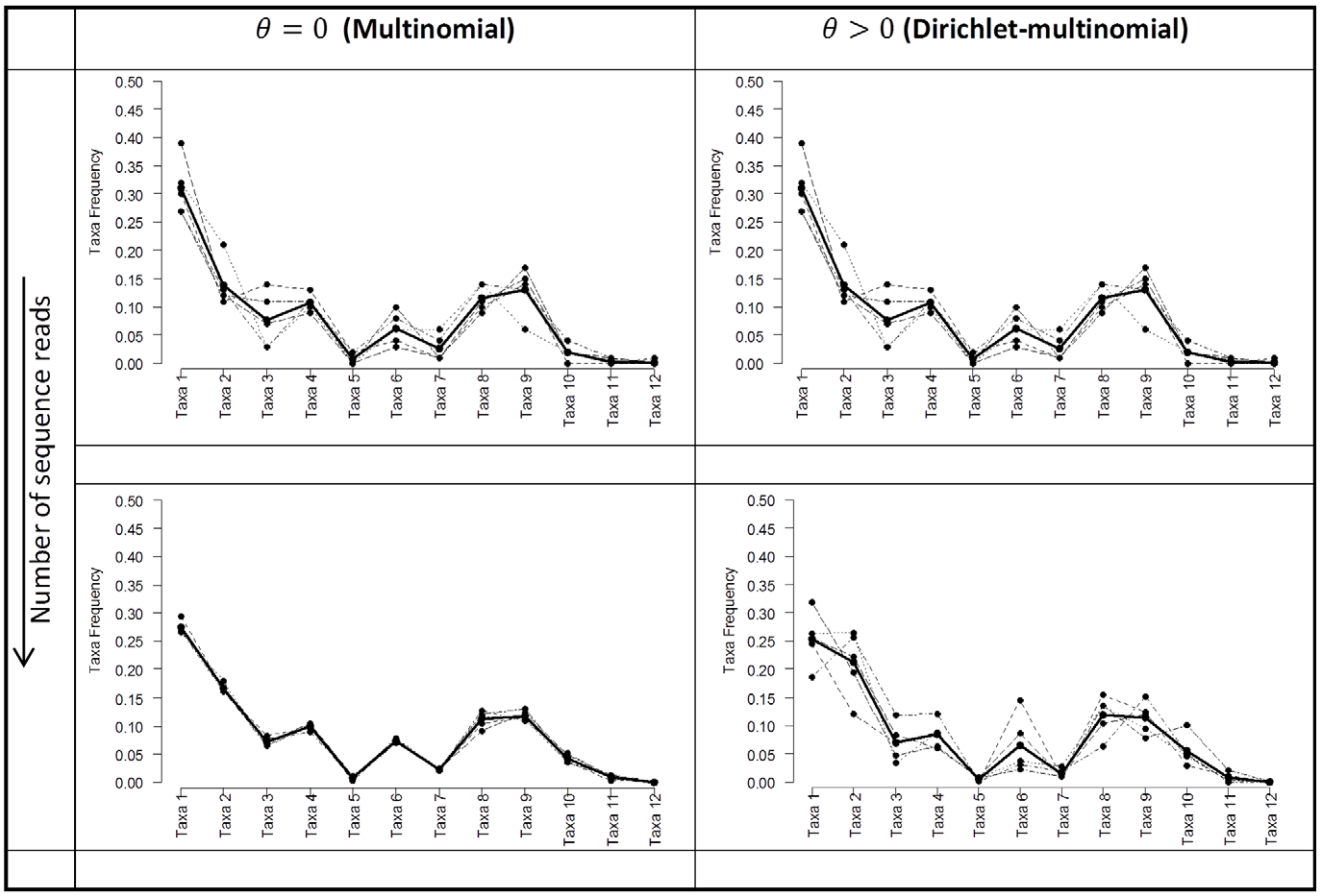

Figure 1. Description of Dirichlet-multinomial parameters. Intuitive description of the meaning of the overdispersion parameter θ. The four plots show the taxa frequencies $\hat{\pi}_{ik}$ for each of the five hypothetical samples (dashed lines) with 12 taxa in each sample, and the corresponding weighted average across the five samples given by the vector of taxa frequencies $\boldsymbol{\pi}$ (solid line). The plots on the left show the taxa frequencies of samples drawn from a Multinomial distribution ($\theta = 0$) and the plots on the right show taxa frequencies of five samples drawn from a Dirichlet Multinomial($\theta > 0$). The top row of plots is for samples with a smaller number of sequence reads, while the bottom row of plots is for samples with a larger number of sequence reads. As the number of reads increases for the multinomial distribution increases each samples taxa frequencies converge onto the mean, while for the Dirichlet-multinomial an increased number of reads is still associated with the same variability between the individual samples.
doi:10.1371/journal.pone.0052078.g001

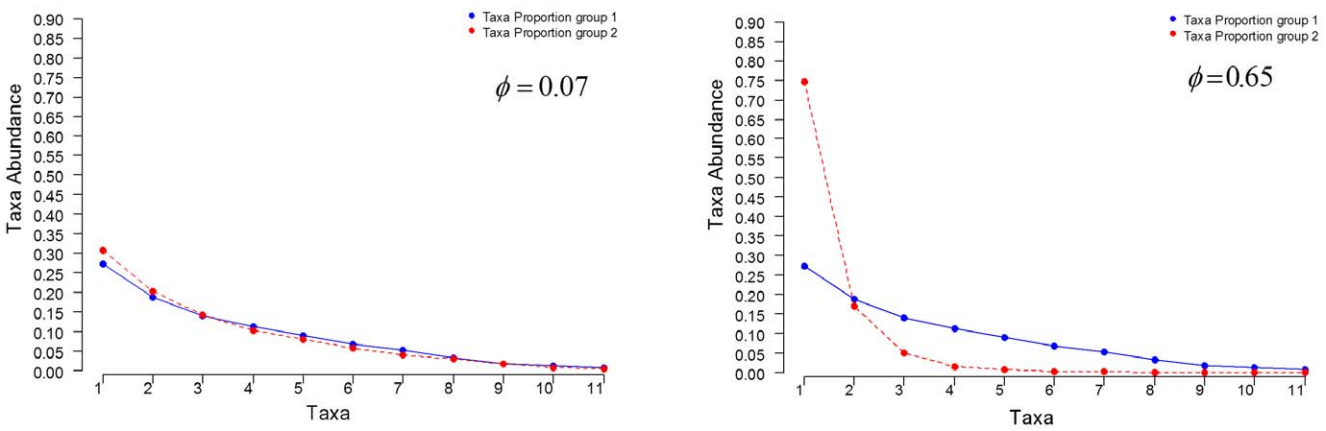

Figure 2. Definition of effect size. Illustration of a small and a large effect size when comparing two groups.
doi:10.1371/journal.pone.0052078.g002

The above parameterization of the Dirichlet-multinomial distribution is suitable to perform hypothesis testing across groups based on difference between locations (comparisons of $\boldsymbol{\pi}$ vectors) as well as scales (comparison of θ values). Other parameterizations of the Dirichlet-multinomial distribution can be found in [23,25]. Note that the Dirichlet-multinomial distribution is a generalization of the multinomial model, which results when $\theta=0$. When $\theta>0$ the data variability is larger than what is expected from the multinomial distribution, and the Dirichlet-multinomial distribution provides a better fit to the data.

On a side note, if the elements of the taxa count vector, $\boldsymbol{x}_i$, obtained from a sample are ranked (i.e., $x_{i1}\geq x_{i2}\geq\cdots\geq x_{iK}$), then the Dirichlet-multinomial can be used to model the rank abundance distributions (RAD) vector across samples. This is useful if the analyst is interested in comparing community structure and complexity across microbiome samples and body sites, but not interested in the names of the community members [26–28]. If the elements of the taxa count vector, $\boldsymbol{x}_i$, obtained from a sample are not ranked (i.e., x_{ik} has the same taxa label across all samples), then we are modeling the abundance of species keeping their labels. This type of analysis is useful to compare community composition across microbiome samples and body sites, and it is usually referred to as analysis of species composition data [29]. Since we are interested in analyzing different taxonomic levels, we will refer to this as analysis of taxa composition data. The interested reader is referred to [26–29] and references therein for more details on the importance and applications of taxa composition data and RAD data analyses to study biodiversity.

Estimating $\boldsymbol{\pi}$ and θ. Referring to the data structure in Table 1 on a set of P samples with counts on K taxa, we compute the frequency of taxon k in sample i as the percentage of reads within that sample that belong to that taxa (i.e., $\hat{\pi}_{ik}=\frac{x_{ik}}{N_{i.}}$). The elements of the parameter $\boldsymbol{\pi}$ are then computed as the weighted average of the taxa frequency from each sample (i.e., $\hat{\pi}_{ik}$) with weights given by proportion of the number of reads in sample i with respect to the total number of sequence reads (i.e., $w_i=\frac{N_{i.}}{N_{..}}$).

To understand the overdispersion parameter θ a graphical example is shown. In Figure 1 we have four plots showing the taxa frequencies $\hat{\boldsymbol{\pi}}_{ik}$ for each of the five hypothetical samples (dashed lines) with 12 taxa in each sample, and the vector of taxa frequencies $\boldsymbol{\pi}$ (solid line). The plots on the left correspond to taxa frequencies of five samples drawn from a multinomial distribution ($\theta=0$) and the plots on the right correspond to taxa frequencies of five samples drawn from a Dirichlet-multinomial ($\theta>0$). The top row of plots is for samples with a smaller number of sequence reads, while the bottom row of plots is for samples with a larger number of sequence reads. As the number of sequence reads increases the multinomial samples get closer and closer to the $\boldsymbol{\pi}$, while the Dirichlet-multinomial samples continue to show variability and no convergence onto $\boldsymbol{\pi}$. This pattern will hold true in the Dirichlet-multinomial distribution no matter how large the number of sequence reads becomes.

Given taxa counts vectors $\boldsymbol{x}_i=[x_{i1},\ldots,x_{iK}]$ for P subjects, denoted in vector form as$\{\boldsymbol{x}_1,\ \ldots,\ \boldsymbol{x}_P\}$ (see Table 1), the set of parameters$\{\pi_k,\ k=1,\ \ldots,\ K\}$ and θ can be estimated using either the method of moments [24,25,30] or maximum likelihood estimation (MLE) [24] computational procedures. The method of moments estimators of $\{\pi_k\}$ are [25]

$$\hat{\pi}_{\kappa}=\sum_{i=1}^{p}\left(\frac{N_i}{N_{..}}\right)\hat{\pi}_{ik}=\frac{\sum_{i=1}^{p}x_{ij}}{N_{..}}=\frac{N_k}{N..},k=1...,k, \qquad (2)$$

and of θ is [24,30]

$$\hat{\theta}=\sum_{j=1}^{K}\frac{S_j-G_j}{\sum_{j=1}^{K}(S_j+(N_c-1)G_j)}, \qquad (3)$$

where $N_c=(P-1)^{-1}\left(N_{..}-(N_{..})^{-1}\sum_{i=1}^{P}N_{i.}{}^2\right)$, and

$S_j=\frac{1}{P-1}\sum_{i=1}^{P}N_{i.}\left(\hat{\pi}_{ij}-\hat{\pi}_j\right)^2$, and

$G_j=\frac{1}{\sum_{i=1}^{P}(N_{i.}-1)}\sum_{i=1}^{P}N_{i.}\hat{\pi}_{ij}\left(1-\hat{\pi}_{ij}\right)$ with $\hat{\pi}_{ij}=\frac{x_{ij}}{N_{i.}}$. Alternatively, the MLEs $\{\hat{\pi}_{\mathrm{j}}\}$ and $\hat{\theta}$ are given by

$$\left(\{\hat{\pi}_j\},\hat{\theta}\right)=\arg\ \max\ L\left(\{\pi_j\},\theta;\boldsymbol{x}_1,...,\boldsymbol{x}_P\right), \qquad (4)$$

where $L\left(\{\pi_j\},\ \theta;\ \boldsymbol{x}_1,\ \ldots,\ \boldsymbol{x}_P\right)=\prod_{i=1}^{P}\mathbb{P}\left(\boldsymbol{X_i}=\boldsymbol{x_i};\ \{\pi_j\},\ \theta\right)$ is the Dirichlet-multinomial likelihood function. The method of moments and MLE estimation procedures perform equally well in terms of statistical properties (e.g., bias, variance) for the number of subjects and reads we routinely encounter in our microbiome studies. These results are available from the authors as a Technical Report.

Multinomial versus Dirichlet-multinomial test. Since the presence of overdispersion increases the Type 1 Error if not controlled for, it is good to test if overdispersion is present in a set of microbiome samples. This can be done by formally testing the null hypothesis $\mathrm{H_o}:\theta=0$ (implying no overdispersion) versus the alternative hypothesis $\mathrm{H_A}:\theta>0$ (implying overdispersion is present). An optimal test-statistic calculated from the raw metagenomic data (see Table 1) for this hypothesis is the following [31]:

$$T=\sum_{k=1}^{K}\sum_{i=1}^{P}\frac{1}{N_{.k}}\left(x_{ik}-\frac{N_{i.}N_{.k}}{N_{..}}\right)^2, \qquad (5)$$

which approaches a Chi-square distribution with $(P-1)\times(K-1)$ degrees of freedom when the number of sequence reads is large and the same in all samples. In the case that the number of reads varies across samples (such as in microbiomes samples) the test statistics converges to a weighted Chi-square with a modified degree of freedom (see [31] for more details). This is a more complicated formulation and is not presented here, but an approximate solution presented in [31] has been included in the R HMP Package. Note that this hypothesis test establishes that the data are better represented by a Dirichlet-multinomial than a multinomial. However, it does not affirm than Dirichlet-multinomial fits the data best. A goodness-of-fit test statistic for doing this is currently being derived.

Hypothesis Testing

Comparing $\boldsymbol{\pi}$ to a previously specified microbiome population. Consider the problem of comparing microbiome samples to a vector of taxa frequencies $\boldsymbol{\pi}_{\mathrm{o}}$ gathered in an earlier study or hypothesized by the investigator. This might be done to test if new samples come from e the same or different population from earlier samples, such as comparing a population to the HMP healthy controls. This test is analogous to a one sample t-test in classical statistics, which, in our case, corresponds to assessing

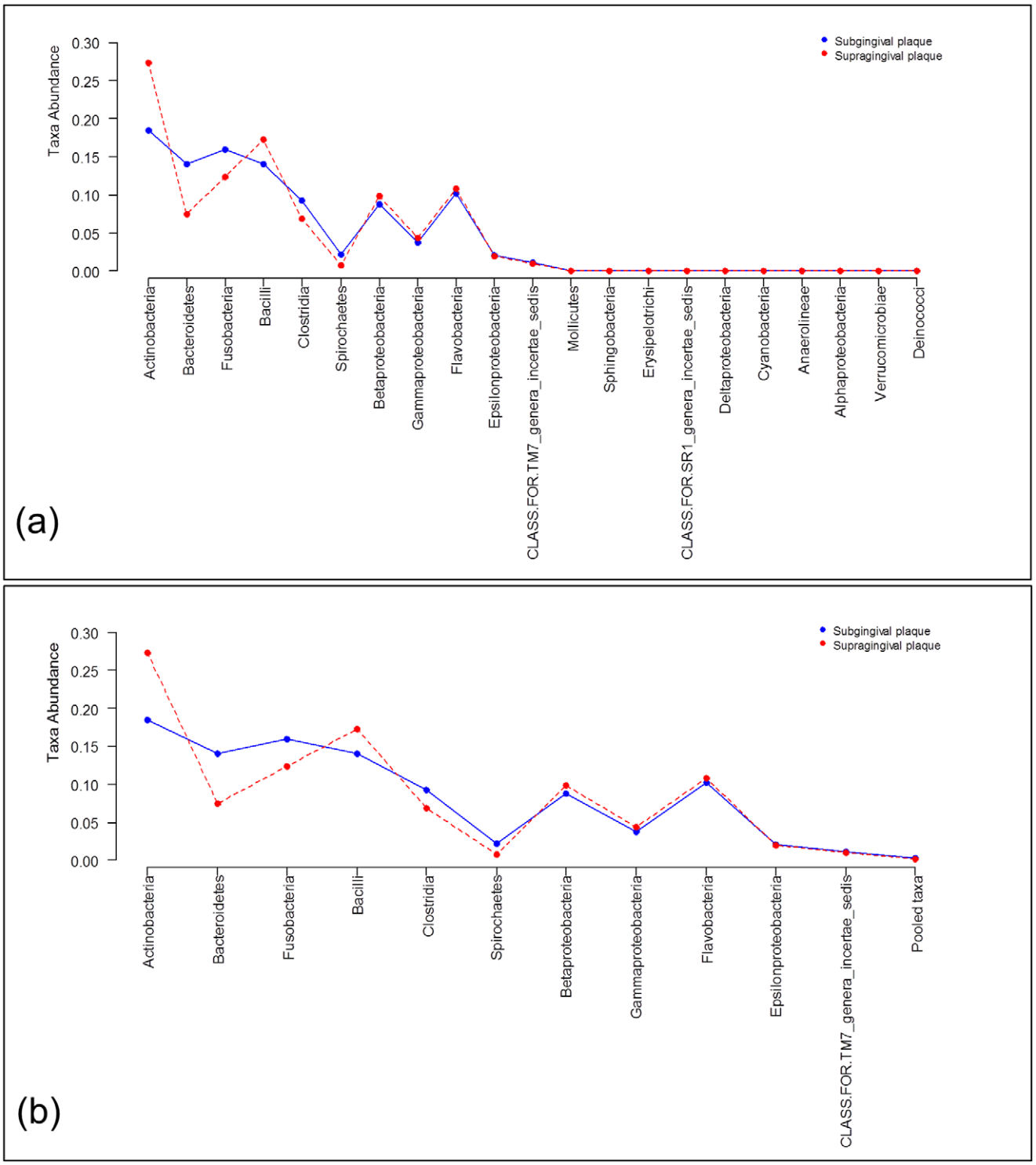

Figure 3. Comparison of two metagenomic groups using a taxa composition data analysis approach. Taxa frequency means at Class level obtained from subgingival plaque samples (blue curve) and from supragingival plaques samples (red curve): a) The mean of all taxa frequencies found in each group, b) The mean of taxa frequencies whose weighted average across both groups is larger than 1%. The remaining taxa are pooled into an additional taxon labeled as 'Pooled taxa'.
doi:10.1371/journal.pone.0052078.g003

whether the vector of taxa frequencies $\boldsymbol{\pi}$ for the new samples, estimated using method of moments or MLE, are equal to the taxa frequencies vector $\boldsymbol{\pi}_0$ from the previously studied population.

The following statistic formally tests the hypothesis $H_o : \boldsymbol{\pi} = \boldsymbol{\pi}_o$ versus the alternative that $H_A : \boldsymbol{\pi} \neq \boldsymbol{\pi}_o$: [32]

Table 2. Power calculation as a function of number of sequence reads and sample size for the comparison of $\boldsymbol{\pi}$ from the subgingiva and supragingiva populations, using as a reference the taxa frequencies obtained from the 24 samples, and 1% and 5% significant levels.

Alpha = 1%	Reads							
Subjects	**500**	**1,000**	**2,500**	**5,000**	**10,000**	**20,000**	**50,000**	**1,000,000**
10	28.67%	29.45%	29.46%	29.83%	29.89%	30.00%	29.80%	29.95%
15	54.25%	55.26%	55.50%	56.16%	56.16%	56.12%	56.57%	56.53%
25	88.48%	89.44%	89.76%	90.03%	90.00%	90.11%	90.06%	90.04%
50	99.95%	99.96%	99.97%	99.98%	99.96%	99.97%	99.97%	99.97%
Alpha = 5%	**Reads**							
Subjects	**500**	**1,000**	**2,500**	**5,000**	**10,000**	**20,000**	**50,000**	**1,000,000**
10	51.96%	52.79%	53.14%	52.91%	53.20%	53.57%	53.16%	53.34%
15	76.01%	77.10%	77.90%	77.88%	77.98%	78.00%	77.92%	78.09%
25	96.50%	96.80%	97.02%	97.02%	97.13%	97.17%	97.09%	97.10%
50	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%

doi:10.1371/journal.pone.0052078.t002

$$X_{1\text{ sample test}} = (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_{\mathrm{o}})^T \left(V\left(\boldsymbol{\pi}_{\mathrm{o}}, \hat{\theta}, N_{\mathrm{g}}\right)\right)^{-} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_{\mathrm{o}}), \qquad (6)$$

which is a generalized Wald test statistic where $\hat{\boldsymbol{\pi}}$ is an unbiased estimator of $\boldsymbol{\pi}$, $(\cdot)^{-}$ is the Moore-Penrose generalized inverse, and $V\left(\boldsymbol{\pi}_{\mathrm{o}}, \hat{\theta}, N_{..}\right) = {N_{..}}^{-2}\, C\left(\hat{\theta}, N_{..}\right)\left(D(\boldsymbol{\pi}_{\mathrm{o}}) - \boldsymbol{\pi}_{\mathrm{o}}{\boldsymbol{\pi}_{\mathrm{o}}}^T\right)$ with $D(\boldsymbol{\pi}_{\mathrm{o}})$ a diagonal matrix with diagonal elements given by $\boldsymbol{\pi}_{\mathrm{o}}$ and $C\left(\hat{\theta}, N_{..}\right) = \hat{\theta}\left(\sum_{i=1}^{P} N_{i.}^2 - N_{..}\right) + N_{..}$, and where $N_{..}$ is the total number of reads in the samples. The asymptotic null distribution of $X_{1\text{ sample test}}$ is a Chi-square with degrees of freedom equal to the rank of the matrix $\left(D(\boldsymbol{\pi}_{\mathrm{o}}) - \boldsymbol{\pi}_{\mathrm{o}}{\boldsymbol{\pi}_{\mathrm{o}}}^T\right)^{-}$, from which the statistical significance (P value) is calculated for the test.

Comparing $\boldsymbol{\pi}$ from two sample sets. Consider the problem of comparing microbiome samples between two groups of subjects (e.g., healthy versus diseased), or two body sites (e.g., oral versus skin). This can be done to test if two sets of microbiome samples are the same or different, such as is in a case-control study. This test is analogous to a two sample t-test in classical statistics, which, in our case, corresponds to evaluate whether the taxa frequencies observed in both groups of metagenomic samples, denoted by $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$, are equal.

The following statistic formally tests the hypothesis $H_o : \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$ versus the alternative that$H_A : \boldsymbol{\pi}_1 \neq \boldsymbol{\pi}_2$ [32,33]

$$X_{2\text{ sample test}} = (\hat{\boldsymbol{\pi}}_1 - \hat{\boldsymbol{\pi}}_2)^T (S)^{-1} (\hat{\boldsymbol{\pi}}_1 - \hat{\boldsymbol{\pi}}_2), \qquad (7)$$

which is a generalized Wald-type test statistics where $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$ are the method of moments estimates, required for Wald-type statistics, of $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$, and S is a diagonal matrix given by

$$S = \left(\sum_{m=1}^{2} \left\{ {N_{..m}}^2\, C\left(\hat{\theta}_m, N_{..m}\right)^{-1} (1 - \varpi_m)^2 \right\} \right)^{-1} D\left(\boldsymbol{\pi}_{\mathrm{p}}\right), \qquad (8)$$

where $N_{..m}$ is the total number of reads in group m, $\hat{\theta}_m$ is the method of moments estimates of the overdispersion parameter of group m, $D\left(\boldsymbol{\pi}_{\mathrm{p}}\right)$ is a diagonal matrix with diagonal elements given by $\boldsymbol{\pi}_{\mathrm{p}} = \sum_{m=1}^{2} \varpi_m \hat{\boldsymbol{\pi}}_m$, a weighted average of estimated group means where $\varpi_m = N_{..m}^2\, C(\theta_m, N_{..m})^{-1} \left(\sum_{r=1}^{J} N_{..r}^2\, C(\theta_r, N_{..r})^{-1} \right)^{-1}$, $C(\theta_m, N_{..m}) = \theta_m \left(\sum_{j=1}^{P_m} N_{j.}^2 - N_{..m} \right) + N_{..m}$, and P_m is the number of subjects in group m. The asymptotic null distribution of $X_{2\text{ sample test}}$ is Chi-square with degrees of freedom equal to $(K-1)$, where K is the number of taxa, from which the statistical significance (P value) is calculated for the test.

Comparing π from more than two groups. Consider the problem of comparing microbiome populations between more than two groups of subjects (e.g., healthy, moderately sick, severely sick), or several body sites (e.g., saliva, subgingival and supragingival). This can be done to test if multiple sets of metagenomic samples are the same or different. This test is analogous to an analysis-of-variance test in classical statistics, which in our case corresponds to inquiry whether the taxa frequencies observed in multiple groups of microbiome samples, denoted by $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \cdots, \boldsymbol{\pi}_J$, are equal.

The following statistic formally tests the hypothesis $H_o : \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 = \cdots = \boldsymbol{\pi}_J$ versus the alternative that$H_A : \boldsymbol{\pi}_m \neq \boldsymbol{\pi}_n$ for at least one pair of groups [32,33]

$$X_{\text{several sample test}} = \sum_{i=1}^{J} (\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_{\mathrm{P}})^T \left(\bar{\boldsymbol{S}}_i\right)^{-1} (\hat{\boldsymbol{\pi}}_i - \boldsymbol{\pi}_{\mathrm{P}}), \qquad (9)$$

which is a generalized Wald-type test statistics given by the weighted difference between each estimated group mean, $\boldsymbol{\pi}_{\mathrm{p}} = \sum_{m=1}^{J} \varpi_m \hat{\boldsymbol{\pi}}_m$, a weighted average of the J estimated group means, with weights $\varpi_m = {N_{..m}}^2\, C(\theta_m, N_{..m})^{-1} \left(\sum_{r=1}^{J} {N_{..r}}^2\, C(\theta_r, N_{..r})^{-1} \right)^{-1}$, and $\bar{\boldsymbol{S}}_i$ a diagonal matrix given by

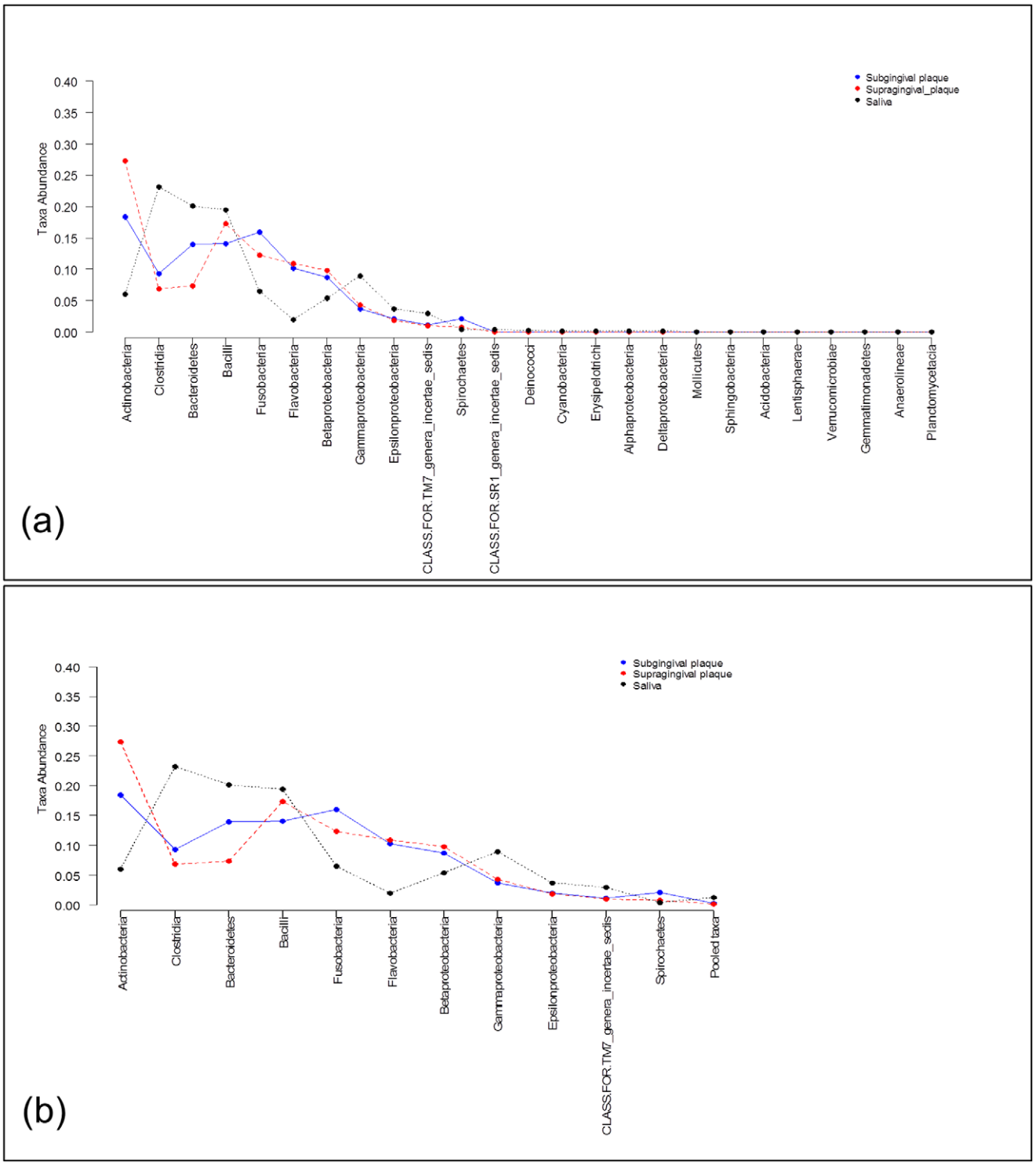

Figure 4. Comparison of three metagenomic groups using a taxa composition data analysis approach. Taxa frequencies at class level obtained from saliva (black line), subgingival plaque (blue line), and from supragingival plaques samples (red line): a) The mean of all taxa frequencies found in each group, b) the mean of taxa frequencies whose weighted average across both groups is larger than 1%. The remaining taxa are pooled into an additional taxon labeled as 'Pooled taxa'.
doi:10.1371/journal.pone.0052078.g004

Table 3. Unadjusted and Bonferroni adjusted p-values for all pairwise comparisons between saliva, supragingiva and subgingiva samples.

	Supraginigiva	Subgingiva
Saliva	P<0.00001 (unadjusted)	P<0.00001 (unadjusted)
	P<0.00003 (Bonferroni)	P<0.00003 (Bonferroni)
Supragingiva		P = 0.0007 (unadjusted)
		P = 0.0021 (Bonferroni)

doi:10.1371/journal.pone.0052078.t003

$$\bar{S}_i = \left(N_{..i}{}^2 C\left(\hat{\theta}_i, N_{..i}\right)^{-1} \right)^{-1} D(\boldsymbol{\pi}_{\mathrm{p}}). \qquad (10)$$

The asymptotic null distribution of $X_{\text{several sample test}}$ is Chi-square with degrees of freedom equal to $(J-1)(K-1)$, where J is the number of groups and K is the number of taxa, from which the statistical significance (P value) is calculated for the test. Note that there does not yet exist a multiple comparisons test analogous to Tukey's Least Significance Difference or Duncan's Range Test [34] routinely used in ANOVA to determine which groups are different when the omnibus rejects the null hypothesis, and is a focus of ongoing work in our lab.

Power and Sample Size

When designing an experiment the goal is to simultaneously reduce the probability of deciding that the groups are different when they are not (Type I Error), and reduce the probability of deciding the groups are not different when in fact they are (Type II Error). From convention we often set the Type I Error = 0.05 (significance or P value) and the Type II Error = 0.2 resulting in power = 0.8, or 80% (power = 1– Type II error). The sample size needed to achieve these error rates depend on the probability model parameters, the hypothesis being tested, and the effect size indicating how different the groups are.

Power can be calculated in the R package for each of the four hypothesis tests discussed above, but for clarity we will only discuss comparison of $\boldsymbol{\pi}$ across two groups. Assume that the model parameters $\boldsymbol{\pi}$ and θ are known for each group, and we are interested in formally testing the hypothesis $H_o: \boldsymbol{\pi}_1 = \boldsymbol{\pi}_2$ versus the alternative that$H_A: \boldsymbol{\pi}_1 \neq \boldsymbol{\pi}_2$. Intuitively, the effect size is defined by how far apart the vector of taxa frequencies $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are from each other. There are several ways to quantify this. For example, a modified Cramer's φ criterion can be used which ranges from 0, denoting the taxa frequencies are the same in both groups, to 1, denoting the taxa frequencies are maximally different (see Appendix S1 for more details). In Figure 2 we show examples of hypothetical data where the effect size is small ($\varphi = 0.07$) and large ($\varphi = 0.65$) across two groups. It would be expected that more samples will be needed to test the 2 group comparison hypotheses for the small effect size than it would be for the large effect size parameters.

Power and sample size calculations are part of the R HMP package for the hypotheses presented in this paper [18]. The technical details of the mathematics for doing this are beyond the scope of this paper. We therefore have included for interested readers the mathematics for power and sample estimation in the Technical Report available from the authors.

Performance Properties of these Tests

Statistical methods need to be tested for their performance to ensure the Type I and II error, P values, power and sample size calculations, and other results from their application are correct. This can be done analytically and proven mathematically, as well as through comprehensive Monte Carlo simulation studies. We chose the latter approach to confirm that these statistics behave as expected and present the results in the Technical Report available from the authors. We elected not to include these results in detail in this paper since it would detract from the primary goal of presenting statistical methods for applied analysis of metagenomic data. However, we briefly discuss those results which showed uniformly that these methods and software are valid.

We simulated Dirichlet-multinomial data for a variety of sample sizes, number of taxa, overdispersion, and effect size, and ran hypothesis tests for one sample, two sample and multiple sample comparisons. These simulations showed the Type I and II Error rates were as expected.

We performed simulated power and sample size calculations and obtained the correct results and show, as expected, the effect size, overdispersion, and sample size influence power. As the effect size increases, overdispersion decreases, or sample size increases, the power goes up. Of particular interest is that in some examples the number of reads also impacts power, with power increasing as the number of reads increases, holding effect size, overdispersion, and sample size constant. This appears to be related to the value of the overdispersion parameter, where for smaller overdispersion the number of reads has the greatest impact on power. Recall that as overdispersion goes to 0, the data converge to a multinomial distribution where the number of reads is known to have significant impact on power.

The Technical Report also presents several other tests of hypothesis that we did not include here since they seem less likely relevant to researchers. This includes comparing the overdispersion parameter across groups, and comparing distributions defined simultaneously by both $\boldsymbol{\pi}$ and θ.

Results of Taxa Composition Data Analysis

In this section, we present results of analyses of metagenomic data from the 24 samples described above for saliva, subgingival and supragingival plaques analyzing the data at the class level. In our experience with metagenomic data analysis two types of analyses are routinely done. When the investigator is interested in community composition (what bacteria are there) the analysis proceeds with taxa labels preserved. In ecology this is usually known as analysis of species composition data [29], and here we will refer to this as taxa-composition data analysis. Alternatively, when the investigator is interested in community structure (what are the high level descriptions of the samples such as richness and diversity) the analysis proceeds without the taxa labels. In ecology this is called as analysis of rank abundance distribution (RAD) data [26–28]. The methods presented in this paper can be applied to both of these situations as illustrated below. In this section the samples are analyzed using a taxa-composition data analysis approach, and in the following section the same analyses are applied using a RAD data analysis approach. It should be noted that for these examples, when the taxa labels are ignored there is a loss of information in the data and the subsequent test of hypotheses show a decrease in power.

One technical issue for the applied data analysis involves the presence of rare taxa. The test statistics proposed are based on the Chi-square distribution and the calculation of the P value is more precise when there are not many rare taxa. This is related to the technical issue of the convergence rate of the test statistic onto its

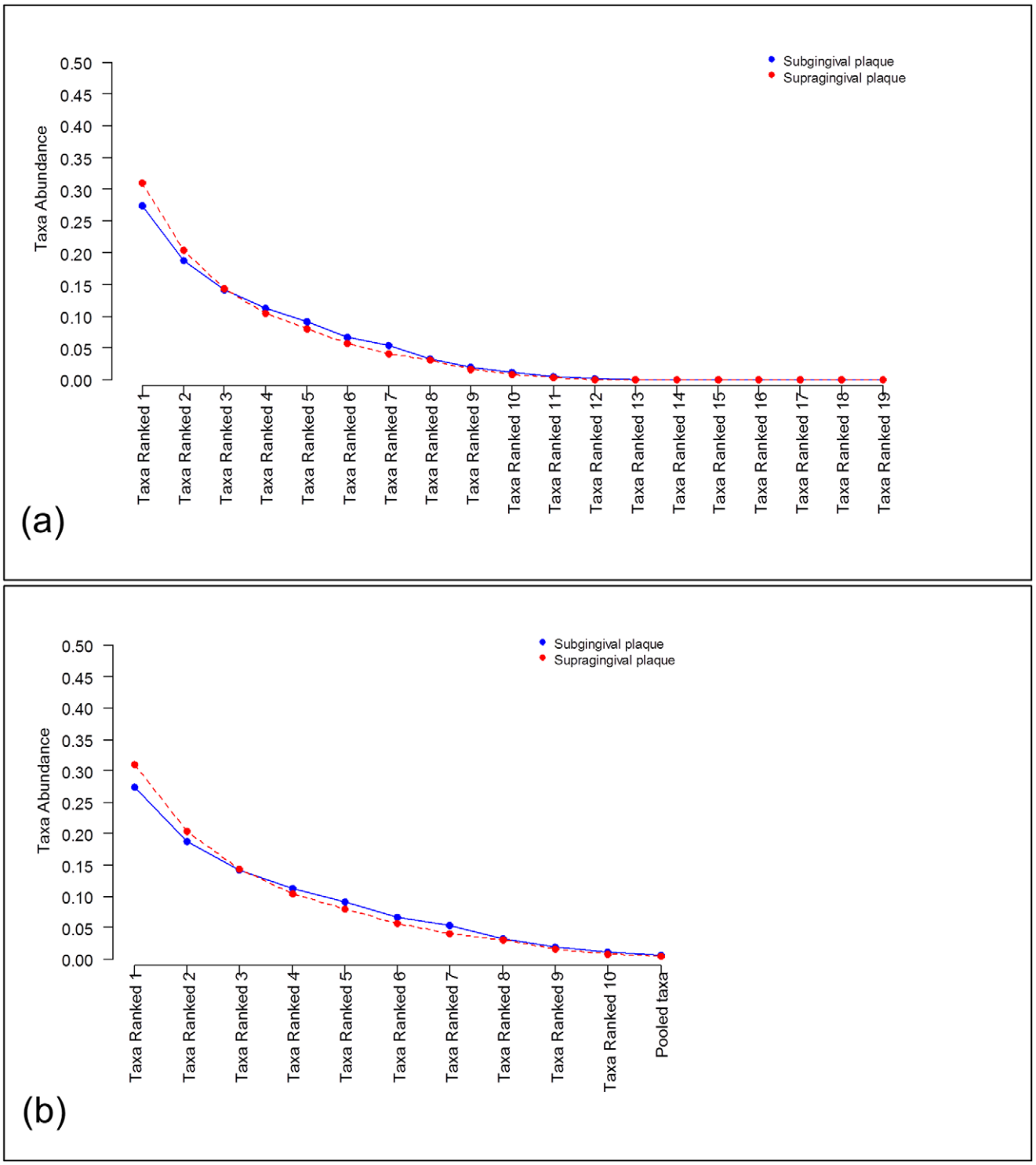

Figure 5. Comparison of two metagenomic groups using rank abundance distribution data. Ranked taxa frequencies mean at class level obtained from subgingival plaque samples (blue curve) and from supragingival plaques samples (red curve): a) The means of all ranked taxa frequencies found in each group; b) The mean of ranked taxa frequencies whose weighted average across both groups is larger than 1%. The remaining taxa are pooled into an additional taxon labeled as 'Pooled taxa'.
doi:10.1371/journal.pone.0052078.g005

Chi-square distribution. To improve the convergence rates of these test statistics all taxa frequencies whose weighted average across all groups is smaller than 1% are combined into a single taxon labeled as 'Pooled taxa'. An illustration of the taxa composition data to be analyzed is shown in Figure 3 a) where we see that taxa from Mollicutes to Deinococci have low prevalence and found that their weighted average across both groups was less than 1%. In Figure 3 b) the same data are shown

Table 4. Power calculation as a function of number of sequence reads and sample size for the comparison of ranked $\boldsymbol{\pi}$ from the subgingiva and supragingiva populations, using as a reference the taxa frequencies obtained from the 24 samples, and 1% and 5% significant levels.

Alpha = 1%	Reads							
Subjects	**500**	**1000**	**2500**	**5000**	**10000**	**20000**	**50000**	**1000000**
10	8.57%	9.56%	10.06%	10.98%	10.51%	10.50%	10.62%	10.17%
15	15.88%	17.42%	18.91%	19.55%	19.85%	19.29%	19.32%	20.10%
25	36.36%	38.81%	41.65%	41.65%	42.91%	42.93%	42.66%	43.54%
50	81.81%	85.60%	87.38%	88.16%	87.50%	87.98%	88.30%	88.59%
Alpha = 5%	**Reads**							
Subjects	**500**	**1000**	**2500**	**5000**	**10000**	**20000**	**50000**	**1000000**
10	23.60%	24.60%	26.30%	22.80%	24.50%	28.20%	25.50%	25.70%
15	32.90%	38.70%	38.60%	40.10%	40.00%	39.10%	37.90%	43.00%
25	61.40%	63.50%	63.90%	65.60%	66.40%	64.90%	66.90%	67.10%
50	93.20%	94.80%	96.50%	95.30%	96.50%	95.40%	96.60%	97.40%

doi:10.1371/journal.pone.0052078.t004

where these rare taxa are pooled, which are the data analyzed in the rest of this section. An alternative approach would be to drop the rare taxa.

Multinomial versus Dirichlet-multinomial Test

Since overdispersion increases the Type 1 Error it is important to test if overdispersion is present in a set of microbiome samples. To do this we use Equation 5 to formally test the null hypothesis $H_o : \theta = 0$ (implying no overdispersion) versus the alternative hypothesis $H_A : \theta > 0$ (implying overdispersion is present). In both subgingival and supragingival plaque samples, the null hypothesis that the data come from a multinomial distribution was rejected in favor of the Dirichlet-multinomial alternative. The overdispersion parameters, using method of moments (see Equation 2), are estimated to be greater than 0 and equal 0.047 for subgingival ($T = 18{,}968$; $df = 11$; $P < 0.00001$), and 0.054 for supragingival ($T = 18{,}953$; $df = 11$; $P < 0.00001$).

Comparing $\boldsymbol{\pi}$ from Two Sample Sets

Consider the problem of comparing microbiome samples between the subgingival and supragingival samples to test if two sets of microbiome samples are different, such as is done in a case-control study. The application of Equation 7 hypothesis test to compare taxa frequencies (see Figure 3 b) $\boldsymbol{\pi}_1$ versus $\boldsymbol{\pi}_2$ corresponding to subgingiva and supragingiva is significant ($X_{2\text{ sample test}} = 25.64$; $df = 11$; $P = 0.007$). From this it is concluded that the null hypothesis that both taxa frequencies are the same is rejected in favor of the alternative that they are different.

Power and Sample Size Calculation

Table 2 shows a power analysis to compare the taxa frequencies of the subgingival plaque versus the supragingival plaque populations from Figure 3b (effect size $\varphi_m = 0.16$) using 1% and 5% significance levels. To calculate power requires the Dirichlet-multinomial parameters, significance level, and specified number of subjects and reads to be defined. In this example the Dirichlet-multinomial parameters are obtained from the subgingival and supragingival 24 sample dataset, the significance levels based on conventional P-values, and a range of subject numbers and reads that could reasonably be obtained in the typical experimental setting.

Table 2 entries are the power achieved for the specified significance level, number of subjects, and number of reads. For example, for significance level = 1%, number of subjects = 15, and number of reads per subject = 10,000, the study has 56% power to detect the effect size observed in the data.

Note that the power is not impacted by increasing the number of reads. In this paper we show the results out to 1,000,000 expected reads per sample, but have conducted experiments running the number of reads out to 10,000,000 and reached the same conclusion. The likely cause of this is that increasing the number of reads does not impact the standard error around $\boldsymbol{\pi}$, while increasing the number of subjects does. However, in experiments based on unlabeled taxa (i.e., rank abundance distributions) the number of reads does impact power.

Comparing $\boldsymbol{\pi}$ from Three Sample Sets

It may be of interest to an investigator to compare three or more groups. Here, for purpose of illustration, we compare the saliva, subgingival and supragingival plaque populations from our 24 subjects. Figure 4 a) shows the taxa frequency to be analyzed where we see that taxa including Deinococci up to Planctomycetacia have very low prevalence. Following the same rationale as for the two sample comparison above, rare taxa were pooled, and the data analyzed is presented in Figure 4 b). It can be seen that the taxa here are the same as used in the comparison of subgingival versus supragingival plaque samples alone. To test if the saliva samples also are better fit to a Dirichlet-multinomial versus multinomial distribution we tested the hypothesis $H_o : \theta = 0$ versus $H_A : \theta > 0$ and conclude that in fact the Dirichlet-multinomial is the better distribution ($P < 0.00001$).

The application of Equation 9 hypothesis test to compare taxa frequencies (see Figure 4) $\boldsymbol{\pi}_1$ versus $\boldsymbol{\pi}_2$ versus $\boldsymbol{\pi}_3$ corresponding to subgingiva, supragingiva, and saliva is significant ($X_{\text{several sample test}} = 258.158$; $df = 22$; $P < 0.00001$). From this it is concluded that the null hypothesis that taxa frequencies across the three groups are the same is rejected in favor of the alternative that they are different.

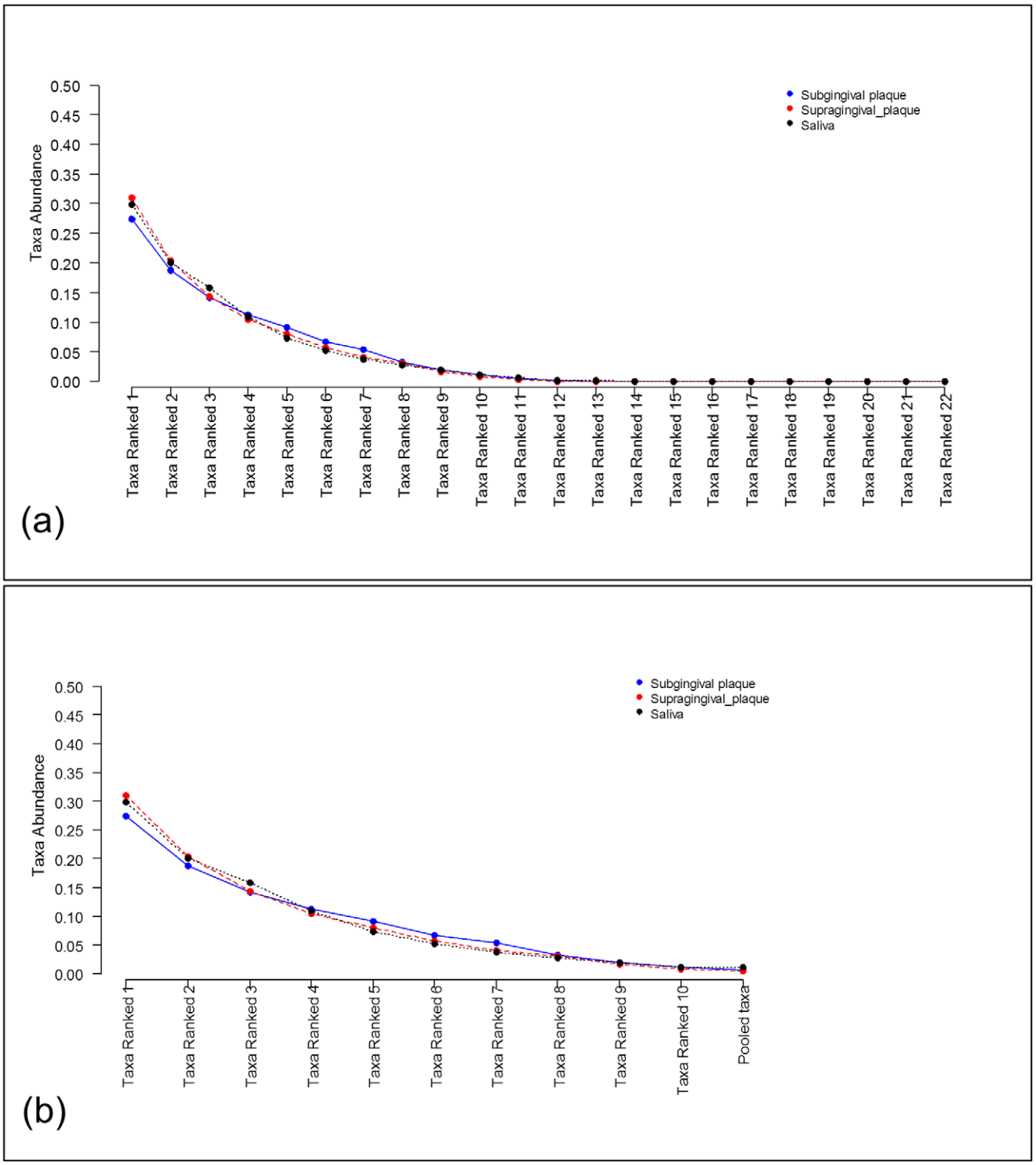

Figure 6. Comparison of three metagenomic groups using rank abundance distribution data. Ranked taxa frequencies mean at class level obtained from subgingival plaque samples (blue curve) and from supragingival plaques samples (red curve): a) The means of all ranked taxa frequencies found in each group; b) The mean of ranked taxa frequencies whose weighted average across both groups is larger than 1%. The remaining taxa are pooled into an additional taxon labeled as 'Pooled taxa'.
doi:10.1371/journal.pone.0052078.g006

The next step in this approach to hypothesis testing is to determine which of the groups are different. In the analysis-of-variance literature this is known as multiple comparisons. A simple approach calculates all pairwise P values and adjusts for the number of tests using a Bonferroni adjustment. In Table 3, we show the p-values (unadjusted and adjusted using Bonferroni) for all pairwise comparisons between saliva, supragingiva and

subgingiva samples. This suggests that all three sample sets are statistically different.

Result of Rank Abundance Distributions Data Analysis

Here we present the same analyses as in the previous example except using rank abundance distributions (RAD) which is of interest when the focus is on community structure (e.g., richness and diversity). Many analysts reduce each sample to a single measure of richness or diversity and then compare these values across groups. However, this results in a significant loss of information which should be avoided when analyzing data. The analyses presented here preserve most of the information (except taxa labels) which should prove to be more valuable for many situations. To illustrate, the RAD data to be analyzed in the following is shown in Figure 5 a) where we see that ranked taxa from 11^{th} to 19^{th} have low prevalence. In Figure 5 b) the same data is shown where these rare ranked taxa are pooled, which are the data analyzed in the rest of this section.

Multinomial versus Dirichlet-multinomial Test

In both subgingival and supragingival plaque samples, the null hypothesis that the data come from a multinomial distribution was rejected in favor of the Dirichlet-multinomial alternative. The overdispersion parameters, using method of moments (Equation 2), are estimated to be greater than 0 and equal 0.008 for subgingival (T normalized = 69945; df = 215; P<0.00001), and 0.02 for supragingival (T normalized = 141301; df = 216; P<0.00001). Note that this hypothesis test establishes that the data are better represented by a Dirichlet-multinomial than a multinomial.

Comparing $\boldsymbol{\pi}$ from Two Sample Sets

The application of the hypothesis test to compare ranked taxa frequencies (see Figure 5 b) $\boldsymbol{\pi}_1$ versus $\boldsymbol{\pi}_2$ corresponding to subgingiva and supragingiva is not significant ($X_{2\text{ sample test}} = 11.08$; df = 10; P = 0.29). From this it is concluded that there is not enough evidence to reject the null hypothesis that ranked taxa frequencies are the same.

Power and Sample Size Calculation

Table 4 shows a power analysis to compare the taxa frequencies of the subgingival plaque versus the supragingival plaque populations from Figure 5 b) (effect size $\varphi_m = 0.07$) using 1% and 5% significant levels, respectively. To calculate power requires the DM parameters, significance level, and specified number of subjects and reads be defined. In this example the Dirichlet-multinomial parameters are obtained from the subgingival and supragingival 24 sample dataset, the significance levels set based on conventional P-values, and a range of subject number and reads that could reasonably be obtained in the typical experimental setting. The table entries are the power achieved for the specified significance level, number of subjects, and number of reads. For example, for significance level = 5%, number of subjects = 15, and number of reads = 10,000, the study has 40% power to detect the effect size observed in the data. Note that compared to the power calculations for the taxa composition data analysis (Table 2) the power is lower for the RAD comparison due to the smaller effect size observed in the data with this analysis.

Comparing $\boldsymbol{\pi}$ from Three Sample Sets

Figure 6 a) shows the ranked taxa frequency to be analyzed where we see that ranked taxa between the 11^{th} to the 22^{nd} most abundant taxa have very low prevalence. Following the same rationale as for the two sample comparison above, ranked rare taxa were pooled, and the data analyzed is presented in Figure 6 b). It can be seen that the taxa here are the same as used in the comparison of subgingival vs supragingival plaque samples alone. To test if the saliva samples also are better fit to a Dirichlet-multinomial versus multinomial distribution we tested the hypothesis $H_o : \theta = 0$ versus $H_A : \theta > 0$ and conclude that in fact the Dirichlet-multinomial is the better distribution (P<0.00001).

The application of Equation 9 hypothesis test to compare taxa frequencies (see Figure 6 b)) $\boldsymbol{\pi}_1$ versus $\boldsymbol{\pi}_2$ versus $\boldsymbol{\pi}_3$ corresponding to subgingiva, supragingiva, and saliva is not significant ($(X_{\text{several sample test}} = 28.048;\ \text{df} = 20;\ P = 0.10)$. From this we concluded that there is not enough evidence to reject the null hypothesis that ranked taxa frequencies across the three groups are the same. Since the test of the three groups does not reject the null hypothesis the multiple comparison tests is not applicable.

Discussion

The major contribution of this work is to begin formulating a biostatistical foundation for the analysis of metagenomic data. The Dirichlet-multinomial model is designed for count data and accounts for over dispersion, which if not adjusted for will result in increased Type I Error. The model gives rise to a broad class of statistical methods, including one sample and multi-sample tests of hypothesis, as well as calculating sample size and power estimates for experimental design. It also provides a set of parameters that can be interpreted analogous to the mean and variance of the bacterial diversity in a population. Computationally this model can accommodate large datasets consisting of multiple samples and essentially unlimited number of reads. For illustration of these methods we presented results of analyses and sample size/power calculations for three body sites for normal healthy individuals collected through the Human Microbiome Project.

Several issues that were referred to in the paper are discussed here. First, the performance of statistical tests depends on their behaving as predicted by statistical theory. For example, a test statistic under the null hypothesis should result in 5% of the tests being significant at the P< = 0.05 level. This and other measures of statistical performance have been confirmed through extensive simulation studies and are in a Technical Report available from the authors.

Second, the Dirichlet-multinomial model can be applied to taxa labeled and unlabeled data corresponding to Taxa composition and Rank Abundance Distribution (RAD) data analyses. In ecology this represents two alternative strategies focused on comparing individual species or diversity (RAD) across communities. The tools proposed here have general use in ecology, but we focused only on metagenomics in this paper. We leave it for others with in- depth experience in ecology to explain how these analyses can best be used in that field [26–29].

Third, in statistics a parametric model is usually preferred over a non-parametric models (e.g., permutation, bootstrapping) when available. In almost all cases parametric models are more efficient and require less data to achieve a given level of power. They also retain more information contained in the data (see the Introduction Section for a detailed discussion). Also, unlike non-parametric methods, our test statistics are appropriate when comparing groups that do not have the same within group variability, a common occurrence in microbiome data.

One of the potential limitations of our method is the incorporation of the rare taxa in the analysis. The performance of the test statistics proposed depends on their convergence to the Chi-square distribution which requires that on having rare taxa

with a minimum frequency across subjects. Though, the proposed approached of 'pooling rare taxa' can be seen as loss of information, it currently stands as a practical approach which avoids giving importance to artificial rare taxa due to the effect of noise in the data. The analysis of rare taxa in metagenomic data is an ongoing topic of discussion and study; it is difficult to identify rare taxa from noise due to sequencing and classification errors, which is not the focus of these methods.

Several methods will be developed extending the Dirichlet-multinomial model for more complex metagenomic research designs and datasets. First, when parameters $\boldsymbol{\pi}$ are shown to be different across groups, it is important to determine which taxa or ranked taxa are causing this difference. To avoid multiple testing problems from doing all univariate comparisons, methods analogous to linear contrasts from analysis-of-variance are being investigated. Second, application of the Dirichlet-multinomial to repeated measures, or mixed models analysis, can be used to monitor changes in the microbiome over time. Third, regression analysis adjusting for covariates can model changes in the microbiome such as how diet, age, or gender affects the stool microbiome. The three topics are current areas of research by the authors.

Supporting Information

Appendix S1 Measure of effect size. Introduction of a modified Cramer's φ criterion such that it does not depend on the sample size when the test statistics takes into account the overdispersion.
(DOCX)

Author Contributions

Conceived and designed the experiments: GW ES. Performed the experiments: GW ES. Analyzed the data: PSL ED WDS. Wrote the paper: PSL WDS. Design Statistical Methods: PSL JPB ELB DJE QW WDS. Design Software: PSL ED WDS.

References

1. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, et al. (2009) The NIH Human Microbiome Project. Genome Research 19: 2317–2323.
2. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449: 804–810.
3. Wooley JC, Godzik A, Friedberg I (2010) A Primer on Metagenomics. PLoS Comput Biol 6: e1000667.
4. Singleton DR, Furlong MA, Rathbun SL, Whitman WB (2001) Quantitative Comparisons of 16S rRNA Gene Sequence Libraries from Environmental Samples. Appl Environ Microbiol 67: 4374–4376.
5. Martin AP (2002) Phylogenetic Approaches for Describing and Comparing the Diversity of Microbial Communities. Appl Environ Microbiol 68: 3673–3682.
6. Schloss PD, Larget BR, Handelsman J (2004) Integration of Microbial Ecology and Statistics: a Test To Compare Gene Libraries. Appl Environ Microbiol 70: 5485–5492.
7. Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71: 8228–8235.
8. Schloss PD, Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol 71: 1501–1506.
9. Schloss PD, Handelsman J (2006) Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. Appl Environ Microbiol 72: 6773–6779.
10. Schloss PD, Handelsman J (2006) Introducing TreeClimber, a test to compare microbial community structures. Appl Environ Microbiol 72: 2379–2384.
11. Hamady M, Lozupone C, Knight R (2009) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J 4: 17–27.
12. White JR, Nagarajan N, Pop M (2009) Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. PLoS Comput Biol 5: e1000352.
13. Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer research 27: 209–220.
14. Mantel N, Valand RS (1970) A technique of nonparametric multivariate analysis. Biometrics: 547–558.
15. Clarke KR (1993) Non-parametric multivariate analyses of changes in community structure. Australian journal of ecology 18: 117–143.
16. Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecology 26: 32–46.
17. Holmes I, Harris K, Quince C (2012) Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One 7: e30126.
18. La Rosa PS, Deych E, Shands B, Shannon WD (2011) HMP: Hypothesis Testing and Power Calculations for Comparing Metagenomic Samples from HMP. R-package.
19. Human Microbiome Project 16S rRNA Clinical Production Pilot (ID: 48335). pp. The NCBI BioProject website. Available: http://www.ncbi.nlm.nih.gov/bioproject?term = 48335. Accessed 18 Sep 2012.
20. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, et al. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Research 33: D294–D296.
21. Vilo C, Dong Q (2012) Evaluation of the RDP Classifier Accuracy Using 16S rRNA Gene Variable Regions. Metagenomics.
22. Cox DR (1983) Some remarks on overdispersion. Biometrika 70: 269–274.
23. Brier SS (1980) Analysis of contingency table under cluster sampling. Biometrika 67: 591–596.
24. Tvedebrink T (2010) Overdispersion in allelic counts and theta-correction in forensic genetics. Theor Popul Biol 78: 200–210.
25. Mosimann JE (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika 49: 65–82.
26. Whittaker R (1965) Dominance and diversity in land plant communities. Science 147: 250.
27. Magurran AE (2004) Measuring biological diversity: Wiley-Blackwell.
28. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, et al. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett 10: 995–1015.
29. Legendre P (1998) Numerical ecology. Developments in environmental modelling.
30. Weir BS, Hill WG (2002) ESTIMATING F-STATISTICS. Annual Review of Genetics 36: 721–750.
31. Kim BS, Margolin BH (1992) Testing Goodness of Fit of a Multinomial Model Against Overdispersed Alternatives. Biometrics 48: 711–719.
32. K. J Koehler, Wilson JR (1986) Chi-square tests for comparing vectors of proportions for several cluster samples. Communications in statistics Theory and Methods 15: 2977–2990.
33. Wilson JR, Koehler KJ (1984) Testing of equality of vectors of proportions for several cluster samples. Proceedings of Joint Statistical Association Meetings Survey Research Methods.
34. Kirk RE (1968) Experimental Design. Belmont: Wadsworth Inc.