Genome Medicine

METHOD

Open Access

CrossMark

An adaptive association test for microbiome data

Chong Wu[1], Jun Chen[2], Junghi Kim[1] and Wei Pan[1*]

Abstract

There is increasing interest in investigating how the compositions of microbial communities are associated with human health and disease. Although existing methods have identified many associations, a proper choice of a phylogenetic distance is critical for the power of these methods. To assess an overall association between the composition of a microbial community and an outcome of interest, we present a novel multivariate testing method called aMiSPU, that is joint and highly adaptive over all observed taxa and thus high powered across various scenarios, alleviating the issue with the choice of a phylogenetic distance. Our simulations and real-data analyses demonstrated that the aMiSPU test was often more powerful than several competing methods while correctly controlling type I error rates. The R package `MiSPU` is available at https://github.com/ChongWu-Biostat/MiSPU and CRAN.

Background

A variety of microbial communities (i.e., microbiotas) and their genomes (i.e., microbiome) exist throughout the human body [1] and play an important role in one's overall health, such as food digestion, nutrition, development and regulation of the immune system, and prevention of the invasion and growth of pathogens [2]. On the other hand, disruptions of the human microbial communities are associated with a wide range of human diseases, such as liver cancer [3], obesity [4], colorectal cancer [5], inflammatory bowel disease (IBD) [6], type 2 diabetes [7], and antibiotic-associated diarrhea [8]. Understanding the association between human microbiotas and diseases might help in diagnosing disease and developing personalized medicine [9] that restores a disturbed microbial ecosystem to a healthy state, for instance, using a personalized synthetic community and complementary set of nutrients [2].

Recent advances in sequencing technologies have made it feasible to profile microbiotas in a large number of samples via targeted sequencing of the 16S rRNA gene [10], and extend the study of the human genome to the human microbiome, which consists of the collection of the microbial genomes at various sites of the human body

and is seen as an extended human genome [11]. Many human microbiome studies aim to detect a possible association of the human microbiome with a phenotype, such as a disease status, called an outcome (of interest) here, after adjusting for potential confounders. These association studies not only can improve our understanding of the non-genetic components of complex traits and diseases, but also might open up an entirely new way for drug development. Although univariate tests (on a single taxon one by one) are widely used in the analysis of differential abundance, multivariate tests (on multiple taxa jointly and simultaneously) have become increasingly popular due to their higher statistical power in aggregating multiple weak associations and reducing the burden of multiple testing. Furthermore, many univariate tests critically depend on some strong parametric assumptions on the distributions or mean-variance functional forms for microbiome data, leading to inflated type I errors when the assumptions are violated [12]. In contrast, no such assumption is imposed in our proposed multivariate test, which, coupled with a proposed permutation procedure for p value calculation, is essentially semi-parametric and applicable to even small sample size problems. In this paper, we mainly focus on multivariate tests.

One popular method for testing the association between an overall microbiome composition and an outcome of interest is to use a distance- or dissimilarity-based test,

*Correspondence: weip@biostat.umn.edu
[1] Division of Biostatistics, University of Minnesota, 420 Delaware St. SE, 55455 Minneapolis, USA
Full list of author information is available at the end of the article

Wu *et al. Genome Medicine* (2016) 8:56

Page 2 of 12

such as PERMANOVA [13]. Via the standard pipelines such as QIIME and mothur [14, 15], the 16S sequence tags are usually clustered into operational taxonomic units (OTUs), which can be considered surrogates for biological taxa within a specified amount of sequence divergence allowed for each OTU. At 97 % similarity, these OTUs represent common species. A specific distance measure is chosen to measure the dissimilarity between each pair of samples, taking into account the phylogeny among taxa. Then the pairwise distance is compared to the distribution of the outcome of interest for evaluating the association between the overall microbiome composition and the outcome. Recently, a new method called the microbiome regression-based kernel association test (MiRKAT) was proposed [16]. Incorporating phylogenetic relationships among taxa, MiRKAT transforms a phylogenetic distance metric into a kernel to measure similarities among samples. Then a semi-parametric kernel machine regression framework is applied to evaluate the association. MiRKAT allows for an easy covariate adjustment and extensions to other types of outcome. By the correspondence between the distance-based association testing and kernel machine regression [16, 17], MiRKAT is closely related to distance-based methods, such as PERMANOVA. In addition, MiRKAT provides an omnibus test that combines several relevant kernels making it more robust across different scenarios. However, the choice of kernels has to be decided by the end user, and more importantly, no automatic taxon selection or weighting is implemented in the framework.

Up till now, numerous distance measures have been developed to depict community differences between two samples. Among many possible distance metrics, the UniFrac-type distance metrics are most popular. They account for phylogenetic relationships among microbial taxa [18–20]. There are several different versions of UniFrac. The unweighted UniFrac distance [18], which is defined as the fraction of the branch length of the tree that leads to descendants from either sample, but not both, is a qualitative diversity measure and is very efficient in detecting abundance changes in rare taxa given that more prevalent species are likely to be present in all individuals. In contrast, the weighted UniFrac distance [19], which weights the branches of a phylogenetic tree based on the abundance differences, is more sensitive to changes in abundant taxa. The generalized UniFrac distance [20] was introduced to unify the weighted and unweighted versions by striking a balance in weighting between relative differences and absolute differences. Many other distances ignoring phylogenetic information are also available. The Bray–Curtis distance [21], for example, quantifies the taxonomic dissimilarity between two samples on the basis of the OTU counts only.

Noise accumulation is a vital problem for high-dimensional data. For example, due to noise accumulation in estimating the population centroids in a high-dimensional feature space, classification using all features can be as bad as a random guess [22]. A severe limitation of kernel- or distance-based methods is that they do not conduct variable selection or variable weighting, which can alleviate noise accumulation and is crucial for high-dimensional microbiome data. In particular, with the dimension much larger than the sample size, some and even most microbial taxa may not be associated with the outcome. Without variable selection or weighting, using all the taxa for distance or kernel calculations simply contributes noise, leading to power loss as to be shown. Therefore, differential weighting of the microbial taxa according to their importance can potentially improve the power of a microbiome association test. We, thus, propose a data-driven approach to achieve adaptive weighting of the taxa based on the data. The proposed method is based on a generalized taxon proportion combining microbial abundance information with phylogenetic tree information and an adaptive test called the aSPU test, which is based on a family of a sum of powered score (SPU) tests [23], incorporating variable weighting. Each SPU test is indexed as SPU(γ) by an integer $\gamma > 0$ that controls the extent of weighting on the variables. We call the corresponding tests MiSPU(γ) (microbiome-based sum of powered score) and aMiSPU (adaptive MiSPU) when applied to microbiome data. We will demonstrate through numerical simulations and analysis of real data that aMiSPU can be easily applied and is much more powerful than existing tests in most scenarios with well-controlled type I error rates. Although aMiSPU was inspired by the aSPU test, the two differ in whether and how to accommodate unique features of microbial data. In particular, we propose the generalized taxon proportion to combine microbial abundance information and phylogenetic tree information simultaneously. As shown in numerical simulations, directly applying the aSPU test with OTU abundances generally failed to achieve high power. Finally, an R package `MiSPU` that implements MiSPU with a C++ version of UniFrac distance calculation has been developed.

Methods

Data and notation

Suppose n samples have been collected, each with a microbial community profile. For sample i, let Y_i denote an outcome of interest, which can be binary (e.g., disease status) or continuous. Let $X_i = (X_{i1}, \ldots, X_{ip})$ be the p covariates, such as age, gender, and other clinical and environmental variables that we want to adjust for. Let $Z_i = (Z_{i1}, \ldots, Z_{im})$ be the abundances of m taxa derived from the observed q OTUs for the ith sample. Note that

Wu *et al. Genome Medicine* (2016) 8:56

Page 3 of 12

an OTU represents a common species while a taxon is a group of one or more species. Here, we assume that each of $Z_{i1}, \ldots, Z_{iq}$ is the count of the OTU in sample i and Z_{ik}, $q + 1 \leq k \leq m$, is the sum of the counts of the OTUs belonging to taxon k in sample i. The evolutionary relationships among these OTUs and taxa are given by a rooted phylogenetic tree, which contains all q OTUs (as leaf nodes) and $m - q$ taxa (as internal nodes). Suppose b_k is the distance from the root of the phylogenetic tree to taxon k, and $p_{ik} = Z_{ik} / \sum_{j=1}^{q} Z_{ij}$ is the proportion of taxon k in sample i. The goal is to test for a possible association between the overall microbial community composition and the outcome of interest after adjusting for the covariates.

A new class of tests: MiSPU

The MiSPU and aMiSPU tests are introduced in this section. Figure 1 illustrates the overall structure of the tests, detailing the input (a rooted phylogenetic tree, a sample of OTU counts, an outcome of interest, and possibly some covariates) and the three key steps: calculating a generalized taxon proportion for each taxon, calculating the test statistics, and applying a residual permutation scheme to obtain the p values.

One major characteristic of microbial composition data is that taxa are related as described by a phylogenetic tree. Phylogenetic distance measures that account for phylogenetic relationships among taxa can be much more powerful than those ignoring evolutionary information [20]. Among these, UniFrac distances are most popular. Consider two samples i and j. The unweighted

UniFrac distance, which considers only species presence or absence, is a qualitative measure and is defined as [18]:

$$d_{ij}^U = \frac{\sum_{k=1}^{m}\{b_k|I(p_{ik} > 0) - I(p_{jk} > 0)|\}}{\sum_{k=1}^{m} b_k},$$

where $I(\cdot)$ is the indicator function. In contrast, weighted UniFrac, which uses OTU abundance information, is a quantitative measure [19]:

$$d_{ij}^W = \frac{\sum_{k=1}^{m} b_k|p_{ik} - p_{jk}|}{\sum_{k=1}^{m} b_k|p_{ik} + p_{jk}|}.$$

Our basic observation is that phylogenetic distance metrics, which account for the relationship among taxa via a phylogenetic tree, measure the distance among samples using all the variables (i.e., taxa) without variable selection or variable weighting. Since the dimension of microbial data is usually high, much larger than the number of samples, many taxa may provide only weak or no signals. Using a phylogenetic distance without variable weighting or variable selection may or may not be powerful. Instead, corresponding to the unweighted and weighted UniFrac distances, for each sample i and taxon k, we define the corresponding generalized taxon proportions as

$$Q_{ik}^u = b_k I(p_{ik} > 0), \qquad Q_{ik}^w = b_k p_{ik},$$

respectively. Note that the raw weighted UniFrac distance [19] between two samples is exactly the same as the L_1 distance of the weighted generalized taxon proportion between the two samples.

Inspired by a multivariate test for association analysis of rare variants [23], we construct a class of versatile
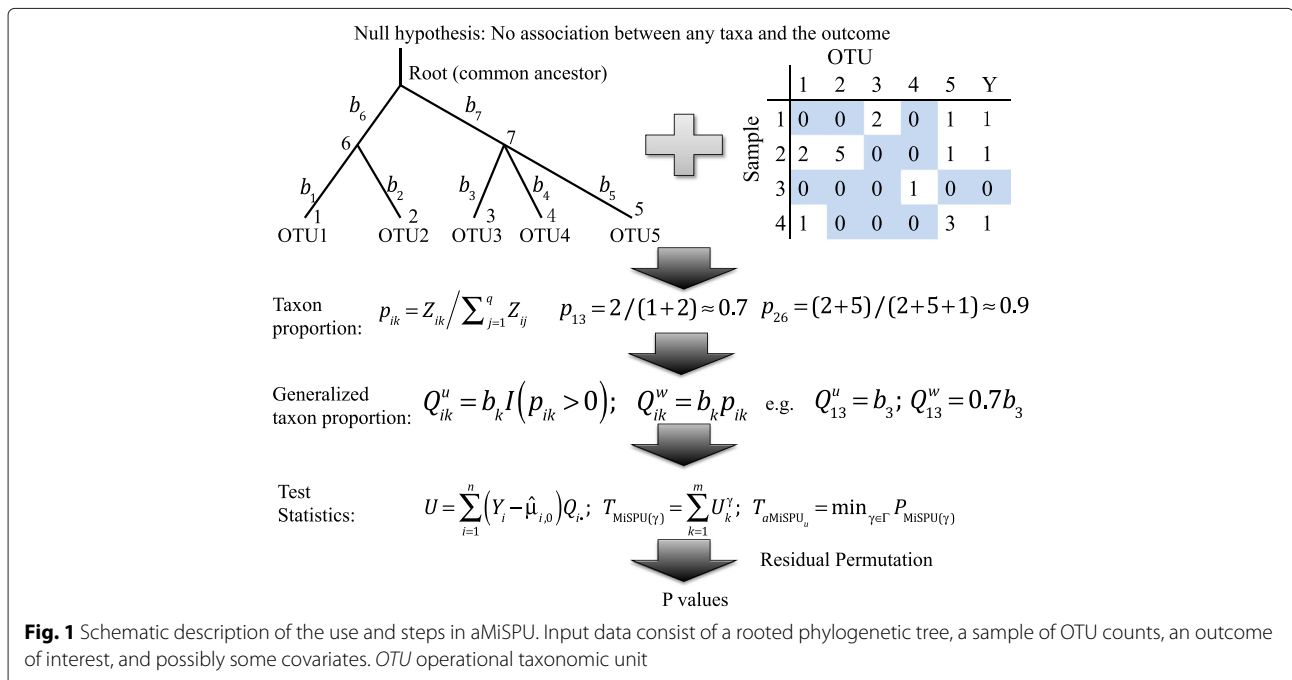

Fig. 1 Schematic description of the use and steps in aMiSPU. Input data consist of a rooted phylogenetic tree, a sample of OTU counts, an outcome of interest, and possibly some covariates. *OTU* operational taxonomic unit

Wu *et al. Genome Medicine* (2016) 8:56

Page 4 of 12

score-based tests such that for a given scenario, at least one of the tests is powerful. Then we combine these tests to maintain high power across a wide range of scenarios. Specifically, for a binary outcome, we use a logistic regression model:

$$\text{Logit}[\Pr(Y_i = 1)] = \beta_0 + \beta' X_i + \sum_{k=1}^{m} Q_{ik} \varphi_k,$$

where Q_{ik} is either Q_{ik}^u or Q_{ik}^w.

For a continuous outcome, we use a linear model:

$$Y_i = \beta_0 + \beta' X_i + \sum_{k=1}^{m} Q_{ik} \varphi_k + \epsilon_i,$$

where ϵ_i is an error term with mean 0 and variance σ^2.

We are interested in testing the null hypothesis $H_0: \varphi = (\varphi_1, \ldots, \varphi_m)' = 0$. That is, there is no association between any taxa and the outcome of interest under H_0. The score vector $U = (U_1, \ldots, U_m)'$ for φ is [17, 23–25]:

$$U = \sum_{i=1}^{n} (Y_i - \hat{\mu}_{i,0}) Q_{i \cdot},$$

where $Q_{i \cdot} = (Q_{i1}, Q_{i2}, \ldots, Q_{im})$ and $\hat{\mu}_{i,0}$ is the predicted mean of the outcome of interest (Y_i) under H_0. Note that a general weighted score-based test can be written as

$$T_G = w'U = \sum_{k=1}^{m} w_k U_k,$$

where $w = (w_1, \ldots, w_m)'$ is a vector of weights for the m generalized taxon proportions. Most existing association tests use the score vector U to construct a test statistic, because of the closed form of the score vector U and because most of the information in the data is contained in U. Therefore, we use U to construct the weights for the score vector U. Under H_0, we have $U \sim N(0, \text{Cov}(U|H_0))$ asymptotically, suggesting that a larger $|U_k|$ offers stronger evidence to reject $H_{0,k}: \varphi_k = 0$. Specifically, we choose $w = (U_1^{\gamma-1}, \ldots, U_m^{\gamma-1})'$ to weight the score vector for the generalized taxon proportions, leading to a MiSPU test:

$$T_{\text{MiSPU}(\gamma)} = w'U = \sum_{k=1}^{m} U_k^{\gamma}.$$

Since $\gamma = 1$ essentially treats all the variables as equally important while association directions of the generalized taxon proportions may vary, $\gamma = 1$ often yields low power and thus is excluded here. Importantly, as γ increases, the MiSPU(γ) test puts more weight on the larger components of U while gradually ignoring the remaining components. As γ goes to infinity, we have

$$T_{\text{MiSPU}(\infty)} \propto ||U||_\infty = \max_{k=1}^{m} |U_k|.$$

We simply define $T_{\text{MiSPU}(\infty)} = \max_{k=1}^{m} |U_k|$. Note that the two versions of Q_{ik}, i.e., Q_{ik}^w and Q_{ik}^u, yield weighted MiSPU$_\text{w}$ and unweighted MiSPU$_\text{u}$, respectively.

We use a permutation scheme [23] to calculate the p value as the following:

1. Fit the null linear or logistic regression model by regressing Y on the covariates X under H_0 to obtain $\hat{\mu}_{i,0} = E(Y_i|H_0)$ and residuals $r_i = Y_i - \hat{\mu}_{i,0}$.
2. Permute the residuals $r = \{r_i | i = 1, \ldots, n\}$ to obtain a permuted set $r^{(b)}$.
3. Regress Q on the covariates X to obtain the residuals $\hat{Q}$.
4. Calculate the new score vector based on the permuted residuals as $U^{(b)} = \sum_{i=1}^{n} \hat{Q}_i . r_i^{(b)}$ and the corresponding null statistic $T_{\text{MiSPU}}^{(b)} = T_{\text{MiSPU}}(U^{(b)})$.
5. Calculate the p value as $\left[\sum_{b=1}^{B} I\left(|T_{\text{MiSPU}}^{(b)}| \geq |T_{\text{MiSPU}}| \right) + 1 \right] / (B + 1)$ after B permutations.

It would be desirable to data-adaptively choose the value of γ and the version of the generalized taxon proportion since the optimal choice of them depends on the unknown true association patterns. Like the adaptive SPU (aSPU) test [23], we propose an adaptive MiSPU (aMiSPU) test, which combines the p values of multiple MiSPU tests with various values of γ and two versions of Q_{ik}. Suppose that we have some candidate values of γ in Γ, e.g., $\Gamma = \{2, 3, \ldots, 8, \infty\}$, as used in our later simulations and real-data analysis. Then, our combining procedure is to take the minimum p value:

$$T_{\text{aMiSPU}_\text{u}} = \min_{\gamma \in \Gamma} P_{\text{MiSPU}_\text{u}(\gamma)},$$

$$T_{\text{aMiSPU}_\text{w}} = \min_{\gamma \in \Gamma} P_{\text{MiSPU}_\text{w}(\gamma)},$$

$$T_{\text{aMiSPU}} = \min \left\{ P_{\text{aMiSPU}_\text{u}}, P_{\text{aMiSPU}_\text{w}} \right\}.$$

Note that we take the minimum p value of aMiSPU$_\text{u}$ and aMiSPU$_\text{w}$ to form the final aMiSPU test. $T_{\text{aMiSPU}_\text{u}}$, $T_{\text{aMiSPU}_\text{w}}$, and T_{aMiSPU} are no longer a genuine p value, but we can use the permutation to estimate its p value, using the same set of null statistics used to calculate the p values for the MiSPU tests [23].

We comment on the choice of Γ and the version of the generalized taxon proportion. Depending on how many taxa are truly associated with the outcome of interest, one may use a smaller or larger γ. For example, if more of the taxa are not associated, a larger γ would be desirable. In our numerical simulations and real-data analysis, we have found that $\Gamma = \{2, 3, \ldots, 8, \infty\}$ often suffices. MiSPU(8) often gives almost the same results as those of MiSPU(∞), suggesting there is no need to use other larger γ's. In practice, we suggest using the aMiSPU test, which combines the strengths (and possibly weaknesses) of various

Wu *et al. Genome Medicine* (2016) 8:56

Page 5 of 12

MiSPU tests. The aMiSPU test can be regarded as a rigorous means for multiple testing adjustment with the use of several MiSPU tests, while the results of MiSPU tests may shed light on the underlying association patterns. For example, if a MiSPU with the unweighted generalized taxon proportion gives the most significant p value, it may indicate the outcome of interest is more likely to be associated with the abundance changes in rare taxa. If some odd γ's yield more significant results than even γ's, then most or all of the large associations are in the same direction.

Although we focus on rRNA sequencing data, the proposed method can be applied to metagenomic whole-genome shotgun sequencing data as well. Via MEGAN [26], DNA reads (or contigs) can be summarized as OTUs and their counts. Using a standard algorithm, species-specific sequences are assigned to OTUs or taxa near the leaves of a phylogenetic tree, whereas widely conserved sequences are assigned to taxa closer to the root [26]. Once we have OTU abundance data and a phylogenetic tree, aMiSPU can be applied as before.

Taxon selection
A limitation of most multivariate tests is their inability to select variables: even if the null hypothesis is rejected, they may not give any information on which taxa are (or are not) likely to be associated with the outcome of interest. We note that the aMiSPU test can be used to rank the importance of the taxa. First, if $P_{\text{aMiSPU}_u} < P_{\text{aMiSPU}_w}$, we use the unweighted generalized taxon proportion in the subsequent analysis; otherwise, we use the weighted one. For ease of exposition, suppose we choose the weighted one. Second, we estimate the optimal value of $\hat{\gamma} = \text{argmin}_{\gamma \in \Gamma} P_{\text{MiSPU}_w(\gamma)}$ chosen by the aMiSPU$_w$ test. If $\hat{\gamma} = \infty$, we can easily find the most significant taxon. Third, suppose $\hat{\gamma} < \infty$, then we assess the relative contribution of each taxon r to the aMiSPU$_w$ test as $C_r = |U_r|^{\hat{\gamma}} / \sum_{j=1}^{m} |U_r|^{\hat{\gamma}}$. Fourth, we rank the taxa based on their C_r values, and we can select a few top k_1 taxa, such as $k_1 = 1$, or such that the sum of their relative contributions $\sum_{r=1}^{k_1} C_r \geq \alpha_1$ with $\alpha_1 = 0.7$, say. The choice of k_1 or α_1 determines the trade-off between increasing true positives and increasing false positives.

The `MiSPU` package and implementation
We implemented the MiSPU and aMiSPU tests in an R statistical software package called `MiSPU`, in which a C++ version of UniFrac distances faster than the `GUniFrac` R package is also provided. The package is available on GitHub (https://github.com/ChongWu-Biostat/MiSPU) and CRAN. We applied MiRKAT from the `MIRKAT` R package developed by Ni Zhao and Michael Wu at website http://research.fhcrc.org/wu/en/software.html. The SPU and aSPU tests are available in the R package `aSPU` on CRAN.

Simulation settings
We used a phylogenetic tree of OTUs from a real throat microbiome data set [27], which consists of 856 OTUs after discarding singleton OTUs. The simulation settings were similar to that used in [16]. Specifically, we generated the OTU counts for each individual via the following steps:

1. Based on a real throat microbiome data set [27], the estimated OTU proportions $(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{856})$ as well as the estimated overdispersion parameter $\hat{\theta}$ were obtained via maximum likelihood.
2. For sample i, the observed OTU proportions were randomly generated from a Dirichlet distribution: $(p_{1i}, p_{2i}, \ldots, p_{856i}) \sim \text{Dirichlet}(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{856}, \hat{\theta})$.
3. The total count of OTUs for sample i, say n_i, was randomly drawn from a negative binomial distribution with mean 1000 and size 25. This step mimicked varying total reads per sample.
4. For sample i, the observed OTU counts were randomly generated from a multinomial distribution: $(Z_{i1}, Z_{i2}, \ldots, Z_{i856}) \sim \text{Multinomial}(n_i; p_{1i}, p_{2i}, \ldots, p_{856i})$.

The procedure for generating simulated data is available as a function in R package `MiSPU`. We considered several simulation scenarios that differed in how some OTUs were related to the outcome of interest.

Under simulation scenario 1, we partitioned the 856 OTUs into 20 clusters (lineages) by partitioning around medoids based on the cophenetic distance matrix. The abundance of these 20 OTU clusters varied tremendously, such that each OTU cluster corresponded to some possible bacterial taxa. We assumed that the outcome of interest depended on the abundance cluster that constituted 6.7 % of the total OTU reads. Then we simulated dichotomous outcomes as follows:

$$\text{Logit}\left(E(Y_i|X_i, Z_i)\right) = 0.5\,\text{scale}(X_{1i} + X_{2i}) + \beta\,\text{scale}\left(\sum_{j \in A} Z_{ij}\right),$$

where β was the effect size and scale(Z_{i1}) standardized the sample mean of Z_i's to 0 and the standard deviation to 1. For continuous outcomes, we simulated under the model

$$Y_i = 0.5\,\text{scale}(X_{1i} + X_{2i}) + \beta\,\text{scale}\left(\sum_{j \in A} Z_{ij}\right) + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$. X_{1i} and X_{2i} were the covariates to be adjusted for, and A was the index set of the selected OTU cluster. X_{1i} was generated from a Bernoulli distribution Bin(1, 0.5), while X_{2i} was from a standard normal distribution $N(0, 1)$. To consider the effect of potential confounders, we studied the case where X_{2i} and Z_i were correlated, specifically, $X_{2i} = \text{scale}\left(\sum_{j \in A} Z_{ij}\right) + N(0, 1)$.

Wu *et al. Genome Medicine* (2016) 8:56

Page 6 of 12

We varied the effect size β to mimic different magnitudes of association.

Under simulation scenario 2, we partitioned all the OTUs into 40 clusters and assumed the outcome was associated with the abundance cluster with only three OTUs. Under simulation scenarios 3, 4, and 5, we assumed that the outcome of interest was associated with the abundance cluster with 24.8 %, 16.6 %, and 1.5 % of the total OTU reads, respectively. Under simulation scenario 6, we assumed the outcome was associated with 50 randomly selected OTUs.

For all the simulation scenarios, we considered using $MiSPU_u$ and $MiSPU_w$ with $\gamma = 2, 3, \ldots, 8$. We combined the MiSPU tests to get $aMiSPU_u$, $aMiSPU_w$, and aMiSPU. We compared aMiSPU with MiRKAT with the weighted and unweighted UniFrac kernels (K_w and K_u, respectively), the Bray–Curtis kernel (K_{BC}), and a generalized UniFrac kernel with $\alpha = 0.5$ (K_5). Additionally, we also applied the optimal MiRKAT, which combines the above four kernels.

Throughout the simulations, the sample size and test significance level were fixed at 100 and $\alpha = 0.05$, respectively. The results were based on 1000 independent replicates for $\beta \neq 0$ and 10,000 independent simulations for $\beta = 0$.

Results

In this section, we present the simulation results from MiSPU and MiRKAT, as well as the results for three real-data sets. We used several published and de-identified human datasets without any confidential information, thus, accordingly there was no human subject involvement and no need for an IRB approval.

Numerical simulation results for type I error and power

To save space, we focus on a few simulation set-ups with a binary outcome. The extensive simulation results with different association patterns and outcomes were similar to those presented below and, thus, are relegated to the supporting information in Additional file 1.

First, the type I error rates of MiSPU and aMiSPU across different simulation set-ups were satisfactorily controlled when the confounders were suitably adjusted (Table 1).

When the covariates were independent of the microbiome composition, MiSPU and aMiSPU controlled type I error rate well no matter whether we adjusted for X or not. In comparison, when X and Z were correlated, adjusting for X was necessary: failing to adjust for X led to an inflated type I error. To save space, we only show some results. The type I errors at the nominal $\alpha = 0.01$ level were also investigated with the same conclusion (Additional file 1: Table S1).

Figure 2 shows statistical power with a binary outcome in simulation scenario 1, in which a phylogenetic cluster with 6.7 % OTUs was associated with the outcome. For all the tests considered, the power increased when the effect size increased. Due to the upweighting of the microbial taxa more likely to be informative, a $MiSPU_w$ test was much more powerful than a MiRKAT test, regardless of whether X and Z were correlated or not. Because only a few taxa were related to the outcome of interest, a $MiSPU(\gamma)$ test with a larger γ performed slightly better than that with a smaller γ. Nevertheless, $MiSPU_w(2)$ still performed much better than any MiRKAT. Compared to $MiSPU_w(\infty)$, $aMiSPU_w$ combining different weights with various γ values lost some power but still maintained power considerably higher than that of many other tests. As expected, by ignoring the phylogenetic information of the microbiome data, the SPU and the aSPU tests [23] failed to achieve high power (not shown). Since there were some abundant OTUs in the informative cluster A, unweighted UniFrac suffered from a loss of power and led to the failure of $aMiSPU_u$ to improve power. However, aMiSPU combining $aMiSPU_u$ and $aMiSPU_w$ lost only little power compared to $aMiSPU_w$. Note that when X and Z were independent, adjusting for the covariates X or not had a minimum effect on the power (Additional file 1: Figure S1). The simulation results for continuous outcomes were similar (Additional file 1: Figures S2 and S3).
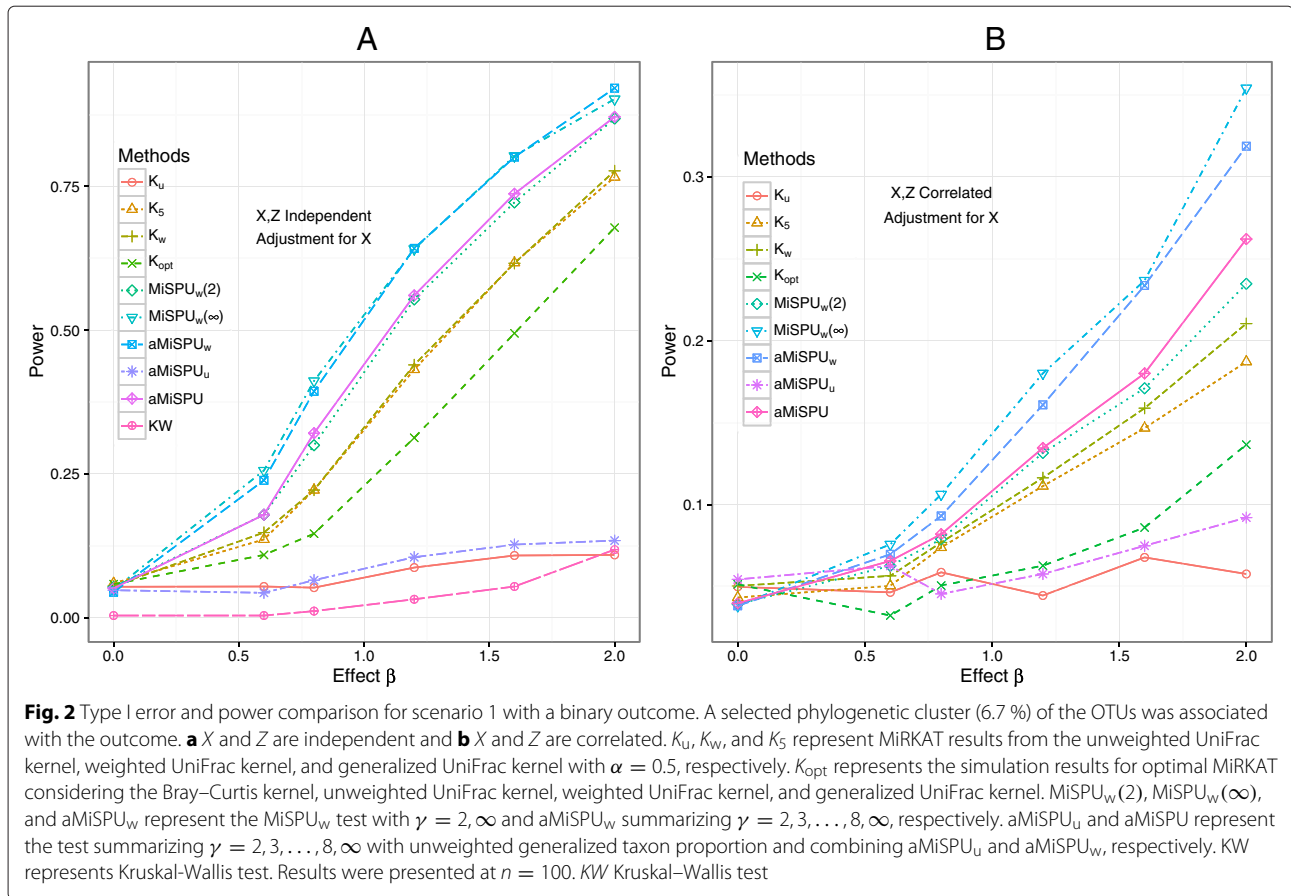
Figure 3 shows the statistical power with a binary outcome in simulation scenario 2, where a small phylogenetic cluster that contains only three OTUs was associated with the outcome. We again show the empirical power curves when X and Z were independent (Fig. 3a) and when X and Z were correlated (Fig. 3b). The results are similar to those

Table 1 Empirical type I error rates for MiSPU and aMiSPU for scenario 1 with a binary outcome

	$MiSPU_w(2)$	$MiSPU_w(\infty)$	$aMiSPU_w$	$aMiSPU_u$	aMiSPU
$X \perp\!\!\!\perp Z$, adjust X	0.052	0.050	0.052	0.049	0.048
$X \perp\!\!\!\perp Z$, no adjust X	0.051	0.051	0.052	0.049	0.049
$X \sim Z$, adjust X	0.043	0.038	0.043	0.049	0.040
$X \sim Z$, no adjust X	0.091[a]	0.119[a]	0.112[a]	0.053	0.088[a]

The type I error rate was evaluated for situations in which the covariates were independent of the OTUs ($X \perp\!\!\!\perp Z$) or correlated with the OTUs ($X \sim Z$) based on 10,000 simulated data sets at $\alpha = 0.05$
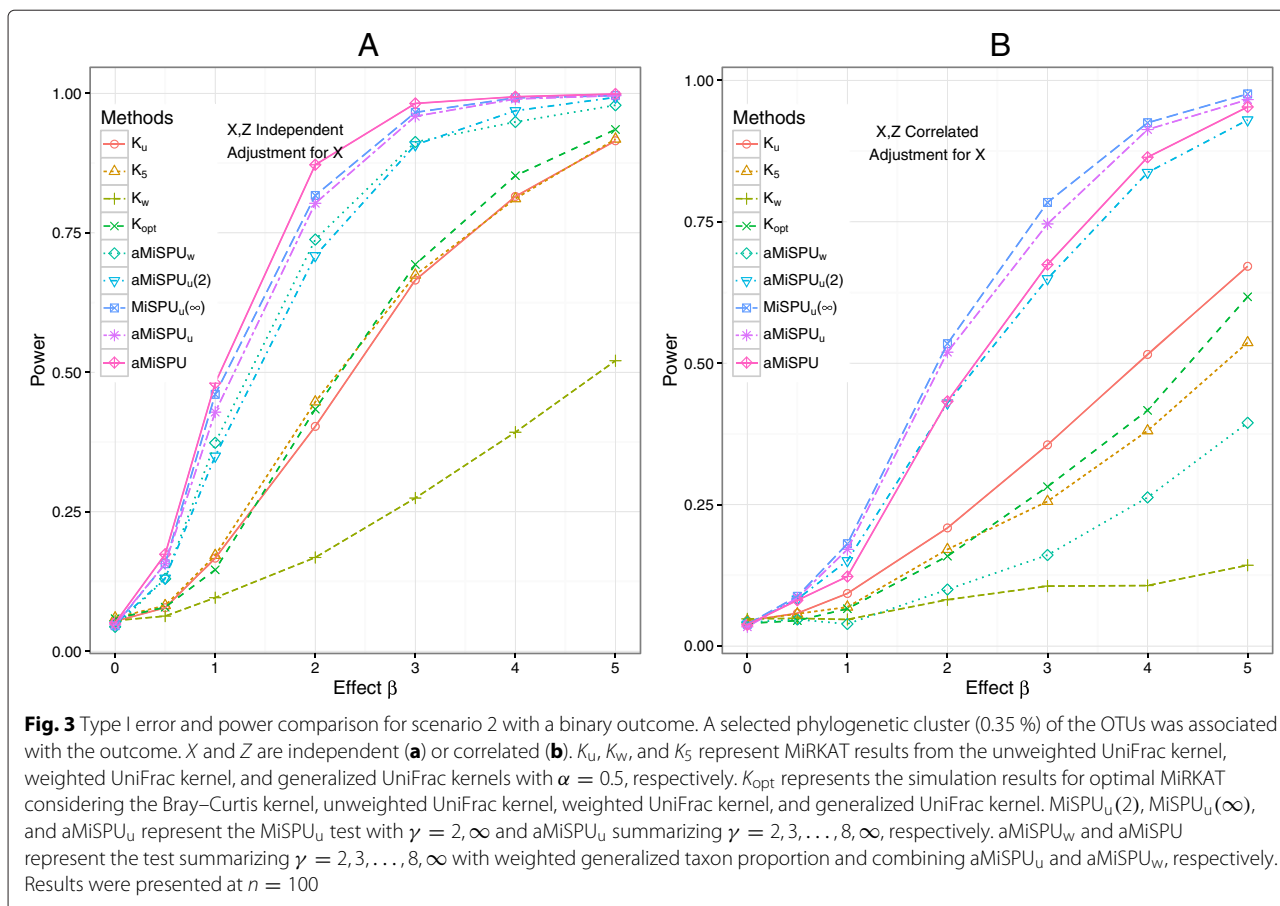[a]Inflated type I error rates

Wu *et al. Genome Medicine* (2016) 8:56

Page 7 of 12

Fig. 2 Type I error and power comparison for scenario 1 with a binary outcome. A selected phylogenetic cluster (6.7 %) of the OTUs was associated with the outcome. **a** X and Z are independent and **b** X and Z are correlated. K_u, K_w, and K_5 represent MiRKAT results from the unweighted UniFrac kernel, weighted UniFrac kernel, and generalized UniFrac kernel with $\alpha = 0.5$, respectively. K_{opt} represents the simulation results for optimal MiRKAT considering the Bray–Curtis kernel, unweighted UniFrac kernel, weighted UniFrac kernel, and generalized UniFrac kernel. $MiSPU_w(2)$, $MiSPU_w(\infty)$, and $aMiSPU_w$ represent the $MiSPU_w$ test with $\gamma = 2, \infty$ and $aMiSPU_w$ summarizing $\gamma = 2, 3, \ldots, 8, \infty$, respectively. $aMiSPU_u$ and $aMiSPU$ represent the test summarizing $\gamma = 2, 3, \ldots, 8, \infty$ with unweighted generalized taxon proportion and combining $aMiSPU_u$ and $aMiSPU_w$, respectively. KW represents Kruskal-Wallis test. Results were presented at $n = 100$. *KW* Kruskal–Wallis test

of simulation scenario 1, except that $aMiSPU_u$ performed better than $aMiSPU_w$. aMiSPU, which combines $aMiSPU_u$ and $aMiSPU_w$, lost only little power compared to the best choice MiSPU, but remained much more powerful than any of MiRKAT. As expected, the weighted UniFrac kernel was the least powerful.

Other simulations showed consistently that aMiSPU generally outperformed MiRKAT and aSPU when a phylogenetic cluster was associated with the outcome (Additional file 1: Figures S4, S5, and S6). However, when some randomly selected OTUs were associated with the outcome (scenario 6), the aSPU test was the winner (Additional file 1: Figure S7); however, we comment that this scenario may not be realistic.

In practice, the true state of nature can vary from case to case. The simulation results show that the power of MiRKAT essentially depends on the chosen kernel; a poor choice of the kernel leads to a tremendous loss of power. In contrast, MiSPU uses the generalized taxon proportion Q_{ik} and puts higher weight on taxa more likely to be informative, achieving much higher power than MiRKAT in most situations. The performance of MiSPU is also dependent on the choice of γ and the version of the generalized taxon proportion: a better choice leads to higher power.

However, aMiSPU alleviates this problem by combining MiSPUs with different γ's and the two versions of the generalized taxon proportion, and it is the overall winner over a wide range of different scenarios.

Univariate testing on each OTU or taxon one by one incurs a heavy burden for a correction for multiple testing. Often the easy-to-use but conservative Bonferroni method is applied, leading to reduced power. Compared to multivariate testing methods, such as MiSPU and MiRKAT, the power of the nonparametric Kruskal–Wallis test [28, 29] was very low (Fig. 2a). Even worse, many parametric univariate tests, due to their strong parametric assumptions on the distributions or parametric specifications on the mean-variance forms for the OTU counts, may have inflated false positive rates, as pointed out by others [12, 30]. For example, in our simulations under scenario 1, the empirical type I error rates for DESeq2 [31] and metagenomeSeq-fitZig [32] were inflated. Accordingly, we did not further investigate their power properties. Relevantly and importantly, univariate tests encounter the so-called curse of the compositionality problem: since the increased (or decreased) relative abundance of some OTUs necessarily leads to other (null or unmodified) OTUs having opposite changes in their

Wu *et al. Genome Medicine* (2016) 8:56

Page 8 of 12

Fig. 3 Type I error and power comparison for scenario 2 with a binary outcome. A selected phylogenetic cluster (0.35 %) of the OTUs was associated with the outcome. X and Z are independent (**a**) or correlated (**b**). K_u, K_w, and K_5 represent MiRKAT results from the unweighted UniFrac kernel, weighted UniFrac kernel, and generalized UniFrac kernels with $\alpha = 0.5$, respectively. K_{opt} represents the simulation results for optimal MiRKAT considering the Bray–Curtis kernel, unweighted UniFrac kernel, weighted UniFrac kernel, and generalized UniFrac kernel. MiSPU$_u$(2), MiSPU$_u$(∞), and aMiSPU$_u$ represent the MiSPU$_u$ test with $\gamma = 2, \infty$ and aMiSPU$_u$ summarizing $\gamma = 2, 3, \ldots, 8, \infty$, respectively. aMiSPU$_w$ and aMiSPU represent the test summarizing $\gamma = 2, 3, \ldots, 8, \infty$ with weighted generalized taxon proportion and combining aMiSPU$_u$ and aMiSPU$_w$, respectively. Results were presented at $n = 100$

relative abundance, there are false positives for some null OTUs. In contrast, multivariate joint testing methods, such as PERMANOVA, MiRKAT, and aMiSPU, do not suffer from this curse of the compositionality problem.

Numerical simulation results for taxon selection

Beyond an overall assessment of association, several methods [28, 29, 31–33] have been developed for identifying specific OTUs driving a detected association. For example, since the compositions of potentially pathogenic bacteria across healthy and disease populations might be different, identifying such bacteria is of interest. One by-product of the aMiSPU test is a ranking of the importance of the taxa. We evaluated taxon selection using simulated data under scenario 1 with an effect size equal to 2, and compared the results to those of the other metagenomic tools, metagenomeSeq-fitZig [32], a Kruskal–Wallis test as used in LEFSe (linear discriminant analysis effect size) [28] and STAMP [29], and DESeq2 [31], a representative for RNA-seq analysis.

The simulation results under scenario 1 are summarized in Table 2. The informative OTU set contained 57 OTUs. On average, the taxon set selected by aMiSPU contained 58.5 OTUs, 27.2 of which were truly informative.

In contrast, fitZig [32] selected 157 OTUs and only 12.3 OTUs were truly informative. Perhaps due to the failure to consider the fact that most OTUs in a microbiome association study are rare, DESeq2 and the KW test performed poorly with a too small mean number of true positives. Under scenario 1, we chose a relatively abundant OTU cluster that contained 57 OTUs to be related to the outcome. As expected, incorporating phylogenetic tree information helped us select truly informative abundant

Table 2 Sample means (standard deviations in parentheses) of the total number of selected OTUs (Total), and of the numbers of true positives and false positives

	Method	Total	TP	FP
$X \perp\!\!\!\perp Z$	fitZig	157.0 (49.4)	12.3 (4.7)	144.6 (45.7)
	DESeq2	20.4 (4.8)	3.4 (1.5)	17.0 (4.4)
	KW	25.8 (6.1)	3.5 (1.6)	22.3 (5.7)
	aMiSPU with $\alpha_1 = .7$	234.1 (241.5)	42.6 (18.5)	191.6 (234.8)
	aMiSPU with $k_1 = 1$	58.5 (76.9)	27.2 (20.6)	31.3 (78.1)

Based on 1000 simulation replications under scenario 1, by fitZig [32], DESeq2 [31], KW test, aMiSPU with $\alpha_1 = .7$, or aMiSPU with $k_1 = 1$. For fitZig, DESeq2 and KW test, cutoff 5×10^{-5}, 0.05, 0.05 were chosen, respectively
FP number of false positives, *KW* Kruskal–Wallis test, *OTU* operational taxonomic unit, *TP* number of true positives

Wu *et al. Genome Medicine* (2016) 8:56

Page 9 of 12

OTUs, thus aMiSPU performed better. In contrast, with only a moderate effect size for each informative OTU, a univariate association test was much less powerful in identifying informative OTUs.

Analysis of a gut microbiome data set for gender and diet effects

Diet strongly affects human health, partly by modulating gut microbiome composition. Wu et al. [34] investigated the association of dietary and environmental variables with the gut microbiota, where the diet information was converted into a vector of micro-nutrient intakes. In this cross-sectional study, 98 healthy volunteers were enrolled and habitual long-term diet information was collected using a food frequency questionnaire. The questionnaires were converted to intake amounts of 214 micro-nutrients, which was further normalized via a residual method to standardize for caloric intake. Stool samples were collected, from which DNA samples were analyzed and denoised prior to taxonomic assignment. The denoised sequences were then analyzed by the QIIME pipeline [15] with the default parameter settings, yielding 3071 OTUs after discarding the singleton OTUs.

Increasing evidence suggests that there is a sex difference in the human gut microbiome, which in turn modulates many pathological and physiological processes [35, 36]. However, no significant sex effect was detected using PERMANOVA based on this data set [34]. We thus re-analyzed the data set for the gender effect by applying MiRKAT and MiSPU with 100,000 permutations. Using MiRKAT, we found the p values from weighted UniFrac, unweighted UniFrac, and the Bray–Curtis kernel to be 0.035, 0.039, and 0.087, respectively. The optimal MiRKAT generated a p value of 0.080, failing to reject the null hypothesis even at the $\alpha = 0.05$ significance level. In comparison, $MiSPU_w(2)$, $MiSPU_w(3)$, $MiSPU_w(8)$, and $MiSPU_w(\infty)$ provided p values of 0.011, 0.0018, 0.0022, and 0.0022, respectively. $MiSPU_w(3)$ provided the most significant p value, suggesting that there is a sparse association pattern between gut microbiome composition and gender status, and the large associations between gender and and one or few microbial taxa were in the same direction. aMiSPU, combining the weighted and unweighted generalized taxon proportions and $\gamma = \{2, 3, \ldots, 8, \infty\}$, yielded a p value of 0.0058, rejecting the null hypothesis at the $\alpha = 0.01$ significance level, suggesting an association between gender status and microbiome composition. Note that perhaps due to the relatively high signal sparsity, previous studies [34, 37] using distance-based methods [13] failed to find any association. Unlike MiRKAT and distance-based analyses, the aMiSPU test can be used for taxon selection. Since $MiSPU_w(3)$ provided the most significant p value, we used the weighted generalized taxon proportion and

$\hat{\gamma} = 3$. We found that a taxon in *Bacteroides* explained more than 90 % of the relative contributions. The top four taxa all came from *Bacteroides*, suggesting that gender was likely associated with *Bacteroides*, but independent of other enterotypes (Fig. 4).

One goal of the study is to identify nutrients that are associated with the gut microbiome composition. We re-analyzed the data from the gut samples by using MiRKAT [16] and aMiSPU. Specifically, we applied the optimal MiKRAT test to analyze the association between each nutrient and microbial community composition by combining the weighted and unweighted UniFrac distances, generalized UniFrace distance with $\alpha = 0.5$, and the Bray–Curtis distance (after being transformed to the corresponding similarity matrices). We further applied $aMiSPU_u$ and $aMiSPU_w$ with $\gamma = 2, 3, \ldots, 8, \infty$. Then we combined $aMiSPU_u$ and $aMiSPU_w$ for aMiSPU. Figure 5 shows that there was no uniformly most powerful test. Depending on the unknown truth, including specific association directions and effect sizes, a given test may or may not be the most powerful. Perhaps due to the sparse association between some of the nutrients and microbial community composition, $aMiSPU_u$ detected some signals undiscovered by others.

Analysis of a gut microbiome data set for association with inflammatory bowel disease

The disruption of the gut microbiota is thought to have an important effect on the etiology of IBDs such as Crohn's disease (CD) and ulcerative colitis (UC). Willing et al. [6] explored the composition of the IBD gut microbiome and
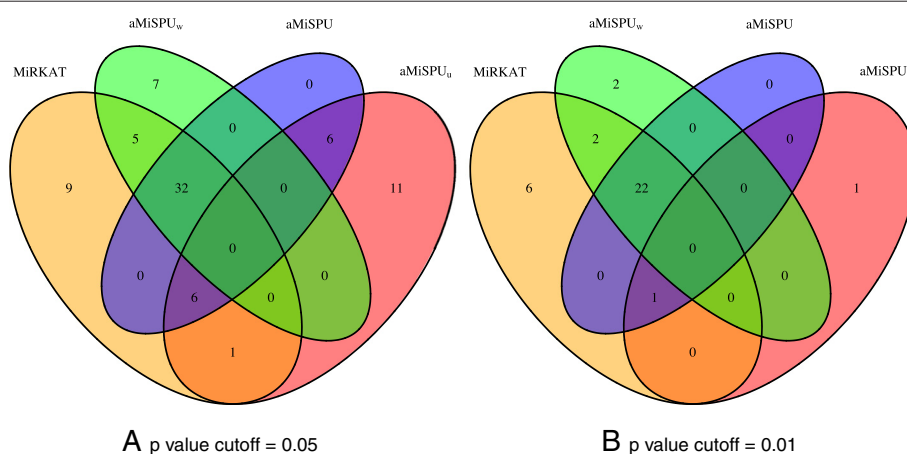

Fig. 4 Phylogenetic tree of *Bacteroides* enterotypes for a gut microbiome data set. *Black edges* stand for non-associated signals, while *red edges* stand for the associated signals. The width of the edges stands for the magnitude of the association

Wu *et al. Genome Medicine* (2016) 8:56

Page 10 of 12
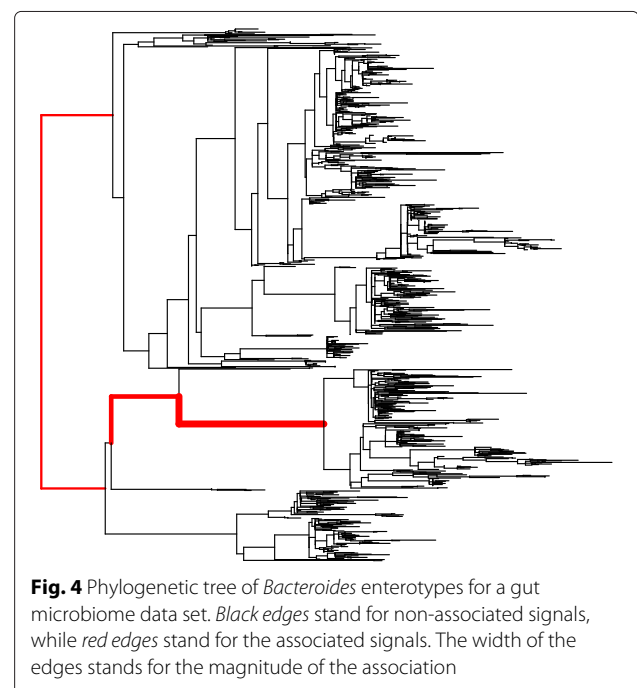
Fig. 5 Venn diagram of detected associations for the gut microbiome data set. In the testing, 214 nutrients are included. Results are shown for a *p* value cutoff of 0.05 (**a**) and 0.01 (**b**). MiRKAT represents the results for optimal MiRKAT considering the Bray–Curtis kernel, unweighted UniFrac kernel, weighted UniFrac kernel, and generalized UniFrac kernel. aMiSPU$_w$ represents a test combining MiSPU$_w$ with $\gamma = 2, \infty$. aMiSPU$_u$ and aMiSPU represent the test summarizing $\gamma = 2, 3, \ldots, 8, \infty$ and combining aMiSPU$_u$ and aMiSPU$_w$, respectively

identified some IBD-associated bacterial signatures. In this cohort study, 40 twin pairs who were concordant or discordant for CD or UC were collected and the compositions of microbial communities in feces samples were determined via 454 pyrotag sequencing. Sequences were checked for quality and those that were less than 200 base pairs in length, contained incorrect primer sequences, or contained more than one ambiguous base were discarded [6].

We tested the association between disease status and the overall microbiome composition via MiRKAT and MiSPU using 10,000 permutations. MiRKAT yielded *p* values from weighted UniFrac, unweighted UniFrac, and Bray–Curtis kernels of 0.223, 0.059, and 0.475, respectively. The optimal MiRKAT generated a *p* value of 0.144, failing to reject the null hypothesis even at the $\alpha = 0.10$ significance level. In comparison, MiSPU$_u$(2), MiSPU$_u$(3), and MiSPU$_u$(∞) provided *p* values of 0.036, 0.053, and 0.084, respectively. The aMiSPU test, combining the weighted and unweighted generalized taxon proportions and $\gamma \in \{2, 3, \ldots, 8, \infty\}$, yielded a *p* value of 0.097, slightly smaller than 0.10, rejecting the null hypothesis at the 0.10 significance level. None of these tests could reject the null hypothesis at the $\alpha = 0.05$ significance level, perhaps due to the small sample size. Note that, perhaps because disease status was more likely to be associated with abundance changes in rare taxa, MiSPU$_u$ provided a more significant *p* value than MiSPU$_w$.

Analysis of a throat microbiome data set for smoking effects

Cigarette smokers have an increased risk of infection involving the respiratory tract. Recently, a microbiome-profiling study was conducted to investigate the effect of smoking on the oropharyngeal and nasopharyngeal bacterial communities [27]. In brief, they analyzed bacterial colonization in the upper airway in 29 healthy cigarette smokers compared with 33 non-smokers. For each DNA sample, 102 of the bacterial rRNA genes were PCR-amplified using individually barcoded primer sets. Then pyrosequences were denoised prior to taxonomic assignment [38]. Using the QIIME pipeline [15], sequences were clustered at 97 % similarity level into OTUs. They excluded the samples with fewer than 500 reads and OTUs with only one read, leading to 60 samples remaining and 856 OTUs. Gender ($p < 0.05$) and antibiotic use within the last 3 months were collected.

In a previous analysis [16], MiKRAT was applied to test the association between smoking and microbial community composition while adjusting for the effect of gender and antibiotic status. Using MiRKAT, we found the *p* values from weighted UniFrac, unweighted UniFrac, and Bray–Curtis kernels to be 0.0048, 0.014, and 0.002, respectively. The optimal MiRKAT generated a *p* value of 0.0031 [16]. In comparison, MiSPU$_w$(2), MiSPU$_w$(7), MiSPU$_w$(8), and MiSPU$_w$(∞) yielded *p* values of 0.0147, 0.0011, 0.0013, and 0.0012, respectively. MiSPU(8) and MiSPU(∞) provided almost the same *p* values, further confirming that there was no need to use other larger γ's. MiSPU$_w$(7) provided the most significant *p* value, suggesting that there was a sparse association pattern and the large associations between smoking status and one or few microbial taxa were in the same direction. aMiSPU$_w$, combining all the MiSPU$_w$ tests with $\gamma = 2, 3, \ldots, 8, \infty$, yielded a *p* value of 0.0029. aMiSPU$_u$, combining all the MiSPU$_u$ tests with $\gamma = 2, 3, \ldots, 8, \infty$, yielded a *p* value of 0.0431, less significant than that from aMiSPU$_w$ and suggesting that some abundant taxa may be correlated with

Wu *et al. Genome Medicine* (2016) 8:56

Page 11 of 12

smoking status. The aMiSPU test, combining aMiSPU$_w$ and aMiSPU$_u$, yielded a p value of 0.0050, confirming the results of the previous analysis, though it was slightly larger than that of the optimal MiRKAT.

Discussion

We have proposed and studied a class of MiSPU tests and an adaptive version (aMiSPU) for an overall association between a microbial community and an outcome of interest. The aMiSPU test is based on the score vector for a new variable called generalized taxon proportion, which combines taxon abundance information with phylogenetic tree information, rendering it both computationally efficient and general to cover a wide range of applications with binary or quantitative outcomes and possible covariates. Our major contribution is that, by recognizing the limitation of the existing methods without variable selection or variable weighting, we propose the use of the two versions of the generalized taxon proportion to account simultaneously for the effects of relative abundances of microbial taxa and that of branch lengths in a phylogenetic tree, and apply many possible weights indexed by a single parameter $\gamma \geq 2$ to weight the taxa differentially. This approach can maintain high power in a wide range of scenarios.

Besides assessing the overall association with a microbial community, one may be interested in finding possible taxa driving a detected association. Unlike MiRKAT [16] and other distance-based methods [13, 20, 39], which are unable for taxon selection, the proposed aMiSPU test can be used to rank the importance of taxa and thus, provide some insights on which taxa are likely to be associated with the outcome of interest.

A few modifications or extensions are possible. First, in our current implementation of MiSPU, we propose the use of a generalized taxon proportion and weight it based on its corresponding score component; we may explicitly consider some interactions among the taxa. Second, we take the minimum p value to combine the results of multiple MiSPU tests. Instead, we may apply other methods that may perform better in some scenarios [40]. Finally, though we focused on a binary and continuous outcome of interest, it might be of interest and possible to extend MiSPU to cases with a multivariate, longitudinal or survival outcome in a general framework of regression.

Conclusions

We have evaluated the MiSPU and aMiSPU tests extensively using both simulated and real data, revealing their excellent performance across many situations. As noted, aMiSPU maintains high power across a wide range of scenarios, though the identity of the most powerful MiSPU test is expected to change with the varying scenarios. In comparison with other multivariate joint tests, we found that aMiSPU was often much more powerful, and thus we recommend its use in practice. An R package `MiSPU` implementing the aMiSPU test and a C++ version of the UniFrac distance calculation are available on GitHub (https://github.com/ChongWu-Biostat/MiSPU) and CRAN.

Additional file

Additional file 1: Seven supporting figures and one supporting table. A description of each is given within the file. (PDF 3778 kb)

Author details
[1]Division of Biostatistics, University of Minnesota, 420 Delaware St. SE, 55455 Minneapolis, USA. [2]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, 200 First St. SW, 55905 Rochester, USA.

References

1. Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 2012;486(7402):215–21.
2. Relman DA. The human microbiome and the future practice of medicine. JAMA. 2015;314(11):1127–8.
3. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nat Biotechnol. 2007;25(6):675–80.
4. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457(7228):480–4.
5. Ahn J, Sinha R, Pei Z, Dominianni C, Wu J, Shi J, et al. Human gut microbiome and risk of colorectal cancer. J Natl Cancer Inst. 2013;105(24):1907–11.

Wu *et al. Genome Medicine* (2016) 8:56

Page 12 of 12

6. Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, Zheng Z, et al. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology. 2010;139(6):1844–54.
7. Karlsson FH, Tremaroli V, Nookaew I, Bergström G, Behre CJ, Fagerberg B, et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature. 2013;498(7452):99–103.
8. Willing BP, Russell SL, Finlay BB. Shifting the balance: antibiotic effects on host–microbiota mutualism. Nat Rev Microbiol. 2011;9(4):233–43.
9. Sonnenburg JL, Fischbach MA. Community health care: therapeutic opportunities in the human microbiome. Sci Transl Med. 2011;3(78): 12–17.
10. Lasken RS. Genomic sequencing of uncultured microorganisms from single cells. Nat Rev Microbiol. 2012;10(9):631–40.
11. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012;13(4):260–70.
12. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 2015;26:27663.
13. McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology. 2001;82(1):290–7.
14. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.
15. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.
16. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, et al. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. Am J Hum Genet. 2015;96(5): 797–807.
17. Pan W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. Genet Epidemiol. 2011;35(4):211–16.
18. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol. 2005;71(12): 8228–35.
19. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol. 2007;73(5): 1576–85.
20. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics. 2012;28(16):2106–13.
21. Beals EW. Bray–Curtis ordination: an effective strategy for analysis of multivariate ecological data. Adv Ecol Res. 1984;14(1):1–55.
22. Fan J, Fan Y. High dimensional classification using features annealed independence rules. Ann Stat. 2008;36(6):2605–37.
23. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. Genetics. 2014;197(4):1081–95.
24. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009;33(6):497–507.
25. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case–control genome-wide association studies. Am J Hum Genet. 2010;86(6):929–42.
26. Huson DH, Auch AF, Qi J, Schuster SC. Megan analysis of metagenomic data. Genome Res. 2007;17(3):377–86.
27. Charlson ES, Chen J, Custers-Allen R, Bittinger K, Li H, Sinha R, et al. Disordered microbial communities in the upper respiratory tract of cigarette smokers. PLoS ONE. 2010;5(12):15216.
28. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):60.
29. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. Stamp: statistical analysis of taxonomic and functional profiles. Bioinformatics. 2014;30(21):3123–4.
30. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10(4):1003531.
31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):1–21.
32. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013;10(12):1200–2.
33. Peng X, Li G, Liu Z. Zero-inflated beta regression for differential abundance analysis with metagenomics data. J Comput Biol. 2016;23(2): 102–10.
34. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011;334(6052):105–18.
35. Markle JG, Frank DN, Mortin-Toth S, Robertson CE, Feazel LM, Rolle-Kampczyk U, et al. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. Science. 2013;339(6123):1084–8.
36. Bolnick DI, Snowberg LK, Hirsch PE, Lauber CL, Parks B, Lusis AJ, et al. Individual diet has sex-dependent effects on vertebrate gut microbiota. Nat Commun. 2014;5:4500.
37. Moeller AH, Degnan PH, Pusey AE, Wilson ML, Hahn BH, Ochman H. Chimpanzees and humans harbour compositionally similar gut enterotypes. Nat Commun. 2012;3:1179.
38. Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, et al. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods. 2009;6(9):639–41.
39. Kelly BJ, Gross R, Bittinger K, Sherrill-Mix S, Lewis JD, Collman RG, et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics. 2015;31:2461–8.
40. Pan W, Han F, Shen X. Test selection with application to detecting disease association with multiple snps. Hum Hered. 2010;69(2):120–30.