A Rarefaction-Without-Resampling Extension of PERMANOVA for Testing Presence-Absence Associations in The Microbiome

Yi-Juan Hu (✉ yijuan.hu@emory.edu)

Emory University <https://orcid.org/0000-0003-2171-9041>

Glen A. Satten

Emory University School of Medicine

Methodology

Keywords: microbiome association test, community-level test, global test, read depth, library size, multiple rarefaction

Posted Date: August 10th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-774776/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

A rarefaction-without-resampling extension of PERMANOVA for testing presence-absence associations in the microbiome

Yi-Juan Hu^{1,*} and Glen A. Satten²

¹Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 30322, USA;

Email: yijuan.hu@emory.edu

²Department of Gynecology and Obstetrics, Emory University School of Medicine, Atlanta,

GA, 30322; Email: gsatten@emory.edu

*Corresponding author

Abstract

Background: PERMANOVA [1] is currently the most commonly used method for testing community-level hypotheses about microbiome associations with covariates of interest. PERMANOVA can test for associations that result from changes in which taxa are present or absent by using the Jaccard or unweighted UniFrac distance. However, such presence-absence analyses face a unique challenge: confounding by library size (total sample read count), which occurs when library size is associated with covariates in the analysis. It is known that *rarefaction* (subsampling to a common library size) controls this bias, but at the potential costs of information loss and the introduction of a stochastic component into the analysis.

Methods: Here we develop a non-stochastic approach to PERMANOVA presence-absence analyses that aggregates information over *all* potential rarefaction replicates without actual resampling, when the Jaccard or unweighted UniFrac distance is used. We compare this new approach to three possible ways of aggregating PERMANOVA over multiple rarefactions obtained from resampling: averaging the distance matrix, averaging the (element-wise) squared distance matrix, and averaging the F -statistic.

Results: Our simulations indicate that our non-stochastic approach is robust to confounding by library size and outperforms each of the stochastic resampling approaches. We also show that, when overdispersion is low, averaging the (element-wise) squared distance outperforms averaging the unsquared distance, currently implemented in the R package **vegan**. We illustrate our methods using an analysis of data on inflammatory bowel disease (IBD) in which samples from case participants have systematically smaller library sizes than samples from control participants.

Conclusions: Our extension of PERMANOVA for presence-absence analyses using a non-stochastic approach that aggregates information over all potential rarefaction replicates without actual resampling is robust to confounding by library size and outperforms stochastic resampling approaches.

Keywords: microbiome association test, community-level test, global test, read depth, library size, multiple rarefaction

Background

PERMANOVA [1] is currently the most commonly used method for testing community-level hypotheses about the microbiome. It is used to perform distance-based analyses of taxonomic count data, in order to find associations between community composition and covariates of interest such as environmental factors or clinical outcomes. By choosing an appropriate distance measure, it can be used to analyze either relative abundance data or presence-absence data, in which counts have been converted to a binary variable that indicates whether a taxon is present or absent from a sample. When studying the ecology of the human microbiome and its relationship to a condition or disease, PERMANOVA analyses that use relative abundance data appear to be more common. However, many associations between presence or absence of taxa and human traits have been reported in the medical literature. For example, increased presence of taxa in the gut was found to be associated with a healthy, diverse diet [2], while increased presence of certain taxa in the vagina was correlated with spontaneous preterm birth [3]. These observations suggest a continuing role for presence-absence analyses in human microbiome studies, especially in situations where it is believed that shifts in rare taxa may drive the disease process.

Presence-absence analyses using PERMANOVA typically adopt either the Jaccard or unweighted UniFrac [4, 5] distance. Presence-absence analyses are particularly sensitive to variability in library size, as the chance of observing rare taxa may depend on sequencing depth. Even though modern 16S rRNA gene sequencing typically yields large library sizes, under-sampling of rare taxa may still occur. Without additional information, it is impossible to distinguish between a taxon that is completely absent from the biological specimen, and a taxon that is actually present but has no reads in a particular sample because of under-sampling. For this reason, if library size is associated with any covariate, that covariate may show a spurious association with presence or absence of rare taxa. In theory, confounding by library size may affect the analysis of relative abundance data as well, but the impact is

typically much smaller.

Ecologists frequently use rarefaction [6], which is to overcome the potential confounding effects of library size. Rarefaction corresponds to subsampling reads to a common *rarefaction depth*, which is often the lowest observed library size (after removing outlying samples with very low library size). As rarefaction produces data with a single library size, there can be no covariates that are associated with library size. On the other hand, rarefaction can result in a substantial loss of reads and hence statistical power [7]. Worse, rarefaction also introduces a stochastic component to the analysis. Multiple rarefaction [8] has been proposed to recover some of the information lost through a single rarefaction, but the results obtained using even multiple rarefaction may depend on the number and specifics of the rarefaction replicates sampled.

Presence-absence analyses have also lost some popularity as they are more sensitive [9] to even small amounts of cross-sample contamination (including “index hopping” or “barcode jumping” [10]), although improvements in laboratory methods (e.g., [11], [12]), as well as use of the error-correction program Decontam [13] in the bioinformatics processing, may ameliorate these problems [14].

Although rarefaction is controversial among statisticians [7], we find investigators continue to use rarefaction to overcome the potential confounding effects of library size. Weiss et al. [15] advocated rarefaction over alternative normalization methods, the latter of which tend to rely on strong, unrealistic assumptions about the distribution of the microbial count data and have failed in one way or another when evaluated in large-scale benchmarking studies [15–17]. Perhaps for these reasons, rarefaction continues to be used with some frequency in recent publications [18–23] when calculating the Jaccard or unweighted UniFrac distance. In fact, rarefaction is recommended by the developers of the unweighted UniFrac distance [4, 15] and is currently the standard normalization method implemented in QIIME [24] and QIIME2 [8] for unweighted Unifrac analyses. Newly developed statistical methods also continue to use rarefaction as an integral part of their approaches [25–29]. These arguments suggest that

improvements in rarefaction-based methods remain an important area.

Here we develop a non-stochastic version of multiple rarefaction that can be used for PERMANOVA presence-absence analyses that use the Jaccard or unweighted UniFrac distance. Our approach is based on an approximation to the result that would be obtained if we could average over *all* potential rarefied datasets. To evaluate the performance of our new approach, we compare it to three possible ways of aggregating the information from multiple rarefied datasets: averaging the distance matrix; averaging the (element-wise) squared distance matrix; and averaging the PERMANOVA F -statistic. Only the first strategy, implemented in the `avgdist` function in the `vegan` R package, is widely available, and uses a default value of 100 rarefaction replicates; the averaged distance matrix is then used in a standard PERMANOVA analysis. Although this strategy has already been adopted in several studies (e.g., [30, 31]), its performance has not been compared to the other strategies. In the results section, we present extensive simulation results in which we numerically compared our non-stochastic PERMANOVA to the three resampling strategies; we also evaluate how much of the information lost by rarefaction can be recovered by (an infinite number of) multiple rarefactions. Finally, we apply the methods we discuss to data from a study on inflammatory bowel disease (IBD). We conclude with a discussion section.

Methods

We briefly review the PERMANOVA approach to establish our notation. Let X denote an $n \times M$ design matrix of covariates, where the rows correspond to the n samples and the columns correspond to M covariates. PERMANOVA can be used to test the effect of a subset of m covariates (m columns of X) using the PERMANOVA F -statistic

$$F = \frac{\text{tr}(H_m \Delta^2 H_m)}{\text{tr}[(I - H_0) \Delta^2 (I - H_0)]},$$

where Δ denotes the distance matrix, Δ^2 the element-wise squared distance matrix, H_m the hat matrix for the m columns we test, H_0 the hat matrix for all columns of X , and $\text{tr}(\cdot)$ is

the trace operation. Note that we drop the multiplier involving the degrees of freedom found in the usual F -statistic as PERMANOVA typically assesses significance via permutation.

For a single rarefaction replicate, F is stochastic because the distance Δ depends on the rarefaction subsample we obtain. The stochastic variability can be reduced by using multiple rarefaction replicates, which raises the question of how to combine information from R rarefaction replicates. One way would be to use the average of the F -statistic over the replicates as a test statistic. The significance of the test statistic is assessed by generating the corresponding average of the F -statistic calculated from each permutation replicate of data. Even though only Δ in the F -statistic varies across rarefaction replicates, averaging the F -statistic is the “safest” approach since the average over replicates is taken after the final step of the analysis of a single rarefaction replicate. We refer to this approach as PERMANOVA- $F(R)$ or $F(R)$. A second, easier, way to combine information would be to separately average the terms in the numerator and denominator of the F -statistic, and then use the ratio of these averages as a test statistic; this corresponds to averaging the *element-wise squared* distance matrices computed for each rarefaction replicate for use in calculating F . We refer to this approach as PERMANOVA-D2(R) or D2(R). A third approach is that implemented in the `avgdist` function in the `vegan` R package, which averages the distance matrices *without squaring* and then uses the average distance to calculate the F -statistic. We refer to this approach as PERMANOVA-D(R) or D(R). PERMANOVA-D2(R) and PERMANOVA-D(R) only calculate the F -statistic once, and are therefore computationally more efficient than PERMANOVA- $F(R)$.

D(R) and D2(R) calculate numerical approximations to $E(\Delta)$ and $E(\Delta^2)$, respectively, where $E(\cdot)$ denotes the expected value over rarefaction replicates. (These expected values are actually conditional on the observed read counts and library sizes; here, we suppress these dependencies for notational convenience.) Non-stochastic versions of these approaches could be provided if these expected values could be calculated analytically. The expectation is taken over the multivariate hypergeometric distribution underlying the generation of a rarefied

taxa count table (see Figure S1). Although the ratio form of the Jaccard and unweighted UniFrac distances makes this difficult, it is possible to use the delta method to develop a series approximation to these expected values. We have found that a two-term truncation of this series is easily calculated and provides a very accurate approximation to the exact expected values.

More specifically, to develop our method, we first note that both Jaccard and unweighted UniFrac distances between samples i and j based on a single (rarefied) dataset can be written as

$$\Delta_{ij} = \frac{K^{-1} \sum_{k=1}^K b_k (\mathcal{I}_{ik} + \mathcal{I}_{jk} - 2\mathcal{I}_{ik}\mathcal{I}_{jk})}{K^{-1} \sum_{k=1}^K b_k (\mathcal{I}_{ik} + \mathcal{I}_{jk} - \mathcal{I}_{ik}\mathcal{I}_{jk})}. \quad (1)$$

For the Jaccard distance, the summation in (1) is over K taxa, the weights $b_k = 1$ for all k , and $\mathcal{I}_{ik} = I(Y_{ik} > 0)$, where Y_{ik} is the entry in a rarefied taxa count table that corresponds to the i th sample and k th taxon (see Figure S1). For the unweighted UniFrac, the summation is over the K nodes of the phylogenetic tree (excluding the root node) used to calculate the UniFrac distance. For each node k , we define $\mathcal{S}_k$ to be the set of leaf nodes (taxa) that are “below” the k th node (i.e., that have node k as an ancestor). If we define $\mathcal{I}_{ik} = I(\sum_{k' \in \mathcal{S}_k} Y_{ik'} > 0)$ and take weights b_k to be the branch length from node k to its parent, then the unweighted UniFrac also has the form given in equation (1).

We define

$$T_{a;ij} = K^{-1} \sum_{k=1}^K b_k (\mathcal{I}_{ik} + \mathcal{I}_{jk} - a\mathcal{I}_{ik}\mathcal{I}_{jk}) \quad (2)$$

for $a = 1, 2$, and note that $\Delta_{ij} = T_{2;ij}/T_{1;ij}$. In the Supplemental Texts, we show using the delta method that we can approximate the expected value of Δ_{ij}^2 over all rarefaction replicates by

$$E(\Delta_{ij}^2) \approx \frac{\mu_{2;ij}^2}{\mu_{1;ij}^2} + \frac{1}{2} \left[\frac{2}{\mu_{1;ij}^2} \text{Var}(T_{2;ij}) - \frac{8\mu_{2;ij}}{\mu_{1;ij}^3} \text{Cov}(T_{1;ij}, T_{2;ij}) + \frac{6\mu_{2;ij}^2 \text{Var}(T_{1;ij})}{\mu_{1;ij}^4} \right], \quad (3)$$

where $\mu_{a;ij}$ is the expected value of $T_{a;ij}$ over all rarefaction replicates. Expressions for the moments of $T_{a;ij}$ with respect to the multivariate hypergeometric distribution that are needed

to evaluate (3) are also provided in the Supplemental Texts.

We propose to use (3), which is a second-order approximation to $E(\Delta_{ij}^2)$, to evaluate the expected values over rarefaction replicates of the numerator and denominator in the PERMANOVA F -statistic. Note that the first-order approximation is simply $\mu_{2;ij}^2/\mu_{1;ij}^2$, which will also be considered in some of our numerical evaluations. To the extent that (3) provides an accurate evaluation of the expected value of Δ_{ij}^2 , this corresponds to averaging the (element-wise) squared distance matrix over *all* rarefaction replicates; we refer to this approach as PERMANOVA-D2-A_o or D2-A_o, where o denotes the order of the approximation used to calculate $E(\Delta_{ij}^2)$. The workflows of D2-A₂, F(R), D2(R), and D(R) are illustrated in Figure 1, which contrasts the non-stochastic nature of D2-A₂ against the other, stochastic resampling approaches.

Results

Simulation Studies

We created simulated data for 50 cases and 50 controls in which a binary confounder and the case-control status were associated with presence or absence of taxa in the microbiome. We based our simulation on two motivating datasets. One is a microbiome dataset from a smoking study that sequenced the 16S rRNA gene of the throat microbiome [32] and yielded 856 taxa. The other is an ecology dataset from a study of dune meadow vegetation [33] from which the hobby farming subset yielded 21 species (which are also referred to as “taxa” in the sequel for simplicity).

To select taxa that were differentially present in cases and controls for simulations based on the throat microbiome data, we considered two association mechanisms. In the first mechanism (referred to as M1), we randomly selected 100 taxa (after excluding abundant taxa with mean relative abundance $> 1\%$), without regard to the phylogeny, to be potentially associated with the case-control status D . In the second mechanism (referred to as M2), we selected a

cluster of taxa based on the phylogenetic tree to be potentially associated. Specifically, we partitioned the taxa into 20 clusters (lineages) using the partitioning-around-medoids algorithm with pairwise distances given by the between-taxa patristic distance matrix. We selected the largest cluster, which consists of 137 taxa (after excluding abundant taxa with mean relative abundance $> 1\%$) to be potentially associated with the case-control status. We independently selected a second set of 100 taxa in M1 and M2 to be associated with a binary confounder C , which has a 70% “success” rate in controls but only 30% in cases; note that the sets of taxa associated with D and C may overlap.

To generate data based on the dune ecology data, we considered only one association mechanism (referred to as M3) in which we randomly selected 5 taxa (after excluding the most abundant taxon) to be potentially associated with the case-control status D (which may represent a binary “treatment” variable in an ecology study). The selected taxa had mean relative abundances 0.073, 0.064, 0.039, 0.029, and 0.010. We independently selected a second set of 5 taxa in M3 to be associated with a confounder C , which had the same binary structure as in M1 and M2. The 5 selected taxa had mean relative abundances 0.094, 0.040, 0.019, 0.010, and 0.0093. In our simulations, one taxon was common to the sets associated with D and C .

To simulate data for the i th sample, we initially set the relative abundances of all taxa, denoted by π_i , to the population mean relative abundances π (all entries of which are non-zero), which were estimated from the throat microbiome or dune ecology dataset. For a sample with $D_i = 1$, we set the entries in π_i to 0, independently for each taxon selected to be associated with D , with probability β ($\in [0, 1]$). We used the value of β as a measure of the *effect size*; note that $\beta = 0$ corresponds to the null hypothesis of no association between the microbiome and case-control status D . For a sample with $C_i = 1$, we further set the entries in π_i to 0, independently for each taxon selected to be associated with C , with fixed probability 0.5. To ensure that all entries in π_i sum up to 1, for each sample we increased the probability mass of the most abundant taxon by the total mass that had been set to 0; since the most abundant taxon has sufficient probability mass to ensure it is always present, this operation does not

change its presence-absence status. Given π_i , we then sampled read count data using the Dirichlet-Multinomial (DM) model. We set the overdispersion parameter to 0.02 in M1 and M2 and 0.0001 in M3, which were estimated from the throat microbiome and dune ecology datasets, respectively. For samples with $D_i = 0$, we used mean library size 10K in M1 and M2 and 100 in M3. To introduce systematic differences in library size, for samples with $D_i = 1$, we varied the mean library size from 10K to 5K in M1 and M2 and from 100 to 50 in M3. Given a mean library size ω , the library size for each sample was drawn from $N(\omega, \sigma = \omega/3)$. Unless otherwise stated, the sampled library size was truncated at 2.5K in M1 and M2 and 20 in M3. The truncation values were used as the rarefaction depth; in this way, no samples were discarded due to rarefaction (corresponding to assuming that samples with problematic library sizes have been removed *a priori* when analyzing real data).

We used the Jaccard distance to analyze data generated using M1 and M3, and the unweighted UniFrac distance for data generated using M2, corresponding to the most appropriate distance for each scenario. Unless otherwise specified, all calculations were carried out using our implementation of PERMANOVA, `permanovaFL` [34], available in the R package LDM. For the new non-stochastic approach, we used the second-order approximation to the expectation of the squared distance unless otherwise stated. In some studies, we included `permanovaFL` applied to the unrarefied distance matrix, referred to as UR. We also evaluated a strategy [that was recommended in the literature \[15\]](#) and that simply adjusts the library size as a covariate in `permanovaFL` applied to unrarefied distance matrix, referred to as UR-L. We calculated the type I error (i.e., size) and power for testing the hypothesis of no case-control differences after adjusting for the confounder; the nominal significance level was set to 0.05. Results for size and power were based on 10,000 and 1,000 replicates of data, respectively.

Evaluating the approximations to the expected Jaccard and unweighted UniFrac distances

We first evaluated the quality of the first- and second-order approximations to the expectation of the (squared and unsquared) Jaccard and unweighted UniFrac distances, by assessing

how many rarefaction replicates were required before we could see evidence that the empirical average over rarefaction replicates was converging to a different limit than our approximation. In Figure 2, we show the log of the maximum percent difference between this empirical average and the first- and second-order approximations, as a function of the log of the number of rarefactions. The linear dependence with slope $-1/2$ is expected given the $\sqrt{R}$ convergence rate for the average over R rarefaction replicates, but only when the number of replicates is small enough that the stochastic error in the average is larger than the error in the o th ($o = 1, 2$) order approximation. When the number of replicates is large enough that the difference between the average over replicates and the true expected value is small compared to the error in the o th order approximation, we expect the empirical log maximum error to switch to a horizontal line as R increases. The (approximate) number of replicates at which this change occurs indicates the number of rarefaction replicates that would be required for the average over replicates to obtain a more accurate approximation to $E(\Delta_{ij}^2)$ than use of the o th order approximation in (3).

For scenarios M1 and M2, we see some evidence that after 10^4 replicates, the first-order approximation is no longer adequate, while we see little evidence that the second-order approximation is inadequate even when as many as 10^6 replicates are used. The situation in scenario M3 is different, commensurate with the much smaller number of taxa in this scenario. The first-order approximation is inadequate for as few as $10^{2.5}$ replicates, while there may be evidence that the empirical average may be outperforming the second-order approximation to the expectation of the squared Jaccard distance when more than 10^6 replicates are used. Further, the overall error in the empirical average is almost an order of magnitude larger than in scenarios M1 and M2. Taken as a whole, the results in Figure 2 suggest that the second-order approximation performs better than using the empirical average over rarefactions unless the number of rarefactions is extremely large. Even the first-order approximation is better than the empirical average unless the number of taxa is very small. While Figure 2 reports results for data that were simulated under the null hypothesis when β was set to 0, we observed

similar patterns for data simulated under the alternative hypothesis, as reported in Figure S2.

We performed ordination by first calculating the second-order approximation to the expected (squared) distance matrix, and then projecting off the effect of the confounder C , for simulated datasets generated using effect size $\beta = 0$. Figure 3 shows that this ordination (referred to as “all rarefactions”) is very similar to ordination based on the average (squared) distance matrix over 100 rarefactions in M1 and M2, and moderately similar in M3; this observation is consistent with the higher error rate for scenario M3 seen in Figure 2. As a comparison, we also provided ordination based on unrarefied and singly rarefied distance matrices. The patterns of similarity between all and 100 rarefactions were also observed when β was non-zero, as shown in Figure S3.

Comparing different versions of PERMANOVA in the presence of systematic differences in library size

We evaluated the size of D2-A₂, D2(100), F(100), UR, and UR-L. The results are summarized in Figure 4 (upper panel). The proposed methods all control size, regardless of how different the library sizes are distributed between cases and controls. As expected, the size of UR becomes more inflated as the difference in library size increases. It should be noted that UR-L controls size in M1 and M2 and approximately controls size in M3 probably because the library size is essentially a simple binary covariate (i.e., differing in mean in cases and controls); UR-L may not work well in more complex settings if the functional form of the dependence on library size were misspecified. To demonstrate that we have induced substantial confounding effects, we show in Figure S4 that the sizes of D2-A₂ are as large as 0.2 when the confounder is not controlled for.

We compared the power of D2-A₂, D2(100), F(100), and UR-L with a fixed differential library size: 10K and 5K for the mean library sizes of controls and cases, respectively, in M1 and M2, and 100 and 50 in M3. The results are displayed in Figure 4 (lower panel). D2-A₂ yields the highest power, with D2(100) and F(100) having only slightly less power but UR-L

substantially less power. Power results for UR are not shown as this method does not control size. We also compared the power of D2-A₂ and D2-A₁ (results not shown). For scenarios M1 and M2, their power was almost identical (less than 0.1% absolute difference); for M3, the power difference was slightly larger, but still less than 0.5%. Figure S5 contrasts the individual p -values of D2-A₂ and D2-A₁ for 1000 replicates of data. The p -values were identical in most cases under M1 and M2 (although differences always favored D2-A₂); for M3 the differences were larger and more random.

We compared the size and power of D2(R) and F(R) over different numbers of rarefactions and summarized the results in Figure 5. We added the results of D2-A₂ as the limit of D2(R) when R goes to infinity. While all methods control size over all R s, aggregating more rarefaction replicates improved power. In M1 and M2 where overdispersion is severe, the bulk of the power increase occurred in the first 5 rarefactions. In M3 where overdispersion is low, power increases were observed even for 100 or more rarefactions. These results indicate the difficulty of choosing R in practical settings. In all scenarios, the power of D2(R) tracked the power of F(R) for all R s, suggesting that the ratio of averages is a good approximation to the average of ratios when calculating the PERMANOVA F -statistic.

In Figure 5, we also considered whether there was a difference in power between methods that replaced the squared distance matrix used in D2(R) and D2-A₂ by the unsquared distance matrix which we refer to as D(R) and D-A₂. We also included results for methods that replaced the `permanovaFL` function by the `adonis2` function, and refer to them as `adonis2-D2( $R$ )`, `adonis2-D2-A2`, `adonis2-D( $R$ )`, and `adonis2-D-A2`. Note that the existing method implemented in `vegan` with the default value of R is `adonis2-D(100)`. We demonstrated that, at least in M3, averaging the unsquared distance matrix performs worse than averaging the squared distance matrix, regardless of the number of rarefactions, with either `permanovaFL` or `adonis2`. We also found that the power of `permanovaFL` is higher than the power of `adonis2` in M3. Note that all methods shown in Figure 5 controlled size for all R s.

The costs of rarefying, when there are no systematic differences in library size

When there is no confounding by library size, the analysis of unrarefied data using UR is valid and can be expected to have optimal power. Then, it is of interest to compare the power of D2-A₂ to UR. We considered a wide range of the mean library sizes (equal for samples having $D = 0$ and $D = 1$) for scenarios M1, M2, and M3. In each setting, we considered two levels of rarefaction depth, 25% and 10% of the mean library size. Note that on average 90% of the reads are lost in each rarefaction when the rarefaction depth is 10%. The results are displayed in Figure 6. We see that, in M1 and M2 where causal taxa are all rare (i.e., having low relative abundances), rarefaction does lead to loss of power when compared to the analysis of the full data, but the power loss diminishes as the mean library size increases. When the mean library size is large enough, the power of D2-A₂ eventually reaches that of UR when the rarefaction depth is 25% of the mean. In M3 where several causal taxa are quite abundant, interestingly, D2-A₂ had higher power than UR. This is not surprising because rarefaction reduces noise at rare taxa that are not associated while having minimal effects on the presence or absence of abundant taxa.

Computation time

The time required to analyze one dataset simulated under M1 using the Jaccard distance was 762 seconds to calculate the average of 100 F -statistics for $F(100)$ in both the observed dataset and the 5,000 permutation datasets. It took 8 seconds to calculate the average of 100 squared Jaccard distance matrices used in D2(100) and then an additional 9 seconds to run `permanovaFL` once (with 5,000 permutations). D2-A_o replaces the average of squared distance in D2(100) by the approximated expectation of squared distance, which took 0.15 or 5 seconds to calculate using the first-order or second-order approximation, respectively. Timings reported here are for a MacBook Pro with 2.9 GHz Intel Core i7 and 16GB memory.

For scenario M2, we used the `GUniFrac` function in the `GUniFrac` R package to calculate the unweighted UniFrac distance for each rarefaction replicate. Analysis of a single dataset using

F(100) required 933 seconds, while D2(100) required 99 seconds (of which 90 seconds were used to calculate the average of 100 squared UniFrac distance matrices and 9 seconds for running `permanovaFL`). D2-A₂ required 673 seconds to calculate the second-order approximation to the expected squared distance, but D2-A₁ required only 0.5 seconds. Note that by permuting covariates rather than taxa count tables, $E(\Delta)$ or $E(\Delta^2)$ only need to be calculated once, even when significance is determined by permutation.

Analysis of the IBD Data

We analyzed data from case participants diagnosed with Inflammatory Bowel Disease (IBD) and control participants that was first reported by [35]. The goal of our analysis was to test for presence-absence association of the rectal microbiome with IBD status, while controlling for two potential confounders: sex and antibiotic use. Following a common practice, we filtered out taxa that were present in fewer than five samples. We further removed samples with library sizes less than 5,000, which resulted in a loss of 3.4% samples and allowed for a rarefaction depth of 5,221 reads. Our final dataset comprised 284 samples and 2,565 taxa.

Motivated by other researchers [15] who reported substantial confounding of PERMANOVA results due to library size in the same study, we confirmed that the library size distributions were indeed systematically different (Figure 7). Thus, we rarefied the read count data of all 284 samples to the minimum depth 5,221. Similarly to Figure 3, we constructed ordination plots using the squared Jaccard and unweighted UniFrac distances (after removing the effects of sex and antibiotic use [34]) in Figure S6, without rarefaction and with 1, 100, and all rarefactions. These plots demonstrate a clear shift in cases compared with controls both before and after rarefaction. In particular, the ordination based on our calculated expected distance matrix (i.e., all rarefactions) is almost identical to the ordination based on the average distance matrix of 100 rarefactions and even similar to the ordination based on one rarefaction; all the rarefaction-based ordinations show some differences from ordinations based on the unrarefied data.

We applied $F(R)$ and $D2(R)$ with R increasing from 1 to 100 and $D2-A_2$, using both the Jaccard and unweighted UniFrac distances. Using the full dataset, the p -values of all tests were equal to the minimum possible value (in this case $1/5001$, where we chose 5,000 to be the maximum number of permutations), regardless of the number of rarefactions. These p -values indicated a very strong presence-absence association of the IBD status with the rectal microbiome at the community level, which is not surprising given the clear shift of case ordinations relative to controls even with only one rarefaction.

Because the association in the IBD data was too strong to illuminate differences in analysis methods, we analyzed subsets of the IBD data, each consisting of 50 randomly selected samples. For each of 1000 such subsets, we obtained p -values using $F(R)$, $D2(R)$, and $D2-A_2$, and reported the $-\log$ of the average of the p -values in Figure 8. The $-\log(p\text{-value})$ obtained using $F(R)$ and $D2(R)$ are in close agreement for all values of R . As R increases, the $-\log(p\text{-value})$ using $F(R)$ and $D2(R)$ increase and eventually converge to the $-\log(p\text{-value})$ obtained using $D2-A_2$. These patterns are consistent with our observations using simulated data, reported in Figure 5. While the results of $F(R)$ and $D2(R)$ depend on the value of R , $D2-A_2$ produces a single value which is, additionally, the largest $-\log(p\text{-value})$ (i.e., the smallest p -value) among the three methods. Although $F(100)$ and $D2(100)$ provide similar p -values to $D2-A_2$ in this data example, this coincidence is not guaranteed to occur at $R = 100$ in general. Without having results from $D2-A_2$, we would not know if the p -values by $F(R)$ and $D2(R)$ have converged by $R = 100$; many more rarefactions would be needed to evaluate this question. In summary, we conclude that $D2-A_2$ provides the best presence-absence analyses of these data.

Discussion

We have presented a non-stochastic method for a rarefaction-based analysis of presence-absence data using PERMANOVA. This approach uses the delta method to approximate the analytic average of the squared distance matrix over all rarefaction replicates. We showed that our non-stochastic method has higher power than each of the three stochastic approaches to

aggregating information over multiple rarefaction replicates. Further, our analytical average of the squared distance matrix can be used as if it were an ordinary distance matrix, for example, for performing ordination, constructing an omnibus test that combines results from multiple distances [25], or plugging into any of the users’ favorite distance-based tests (e.g., MiRKAT [36]). We have implemented all the methods described here, including the function for calculating the analytical average of the squared or unsquared distance matrix, in LDM, our R package for microbiome data analysis, available on GitHub at <https://github.com/yijuanhu/LDM>.

We have previously developed the linear decomposition model (LDM) that unifies the community-level and taxon-level tests into one framework [34]. We also extended the LDM for testing presence-absence associations based on multiple rarefactions [37]; some of those results mirror the findings we obtained here. Specifically, we compared a stochastic approach that averages the LDM F -statistic from R rarefaction replicates to a non-stochastic approach that separately calculates the expected values of the residual sum-of-squares (RSS) terms in the numerator and denominator of the LDM F -statistic with respect to the rarefaction operation. Parallel to the results we found here, the non-stochastic approach had superior performance. Unlike the case of PERMANOVA, the expected values over rarefaction replicates for the LDM could be calculated exactly.

When the distributions of library size are similar between cases and controls, the relative power of the unrarefied (UR) and rarefied analysis (D2-A₂) depends on whether rare or abundant taxa are involved in the association. Although unrarefied analysis typically gives the best power when only rare taxa are associated, the power lost by rarefaction is minimal, as long as the overall sequencing depth is moderately high and the rarefaction depth is not too low. When abundant taxa are associated, rarefaction can even improve power. Given the robustness of rarefied analysis to any systematic variation in library size, it seems prudent to use D2-A₂ even when there is no obvious difference in the distributions of library size.

The efficiency of rarefied analysis depends on the rarefaction depth. In general, the guideline for choosing a rarefaction depth is similar to other applications that use rarefaction: the

depth should be low enough to include most samples, but can exclude samples with anomalously low library sizes. We also saw that the power of D2-A₂ could be slightly higher than that of D2-A₁, with the largest differences occurring when the number of taxa is small. Thus we recommend use of D2-A₂ whenever it is computationally feasible; fortunately the situations in which D2-A₂ is time-consuming to calculate are exactly the situations for which the performance of D2-A₁ is nearly identical to D2-A₂. When the number of taxa is small, the second-order correction D2-A₂ should always be used. Finally, since PERMANOVA typically permutes covariates rather than residuals, we also note that the second-order approximation only needs to be calculated once for each analysis.

In addition to the sensitivity to differences in library sizes, presence-absence metrics (e.g., Jaccard and unweighted UniFrac distances) are also more sensitive to classification or sequencing error [9] and a phenomenon called “index hopping” (or “barcode jumping”) [10], compared to abundance-weighted metrics (e.g., Bray-Curtis and weighted UniFrac distances). Index hopping occurs when non-ligated adapters from one sample (typically low-biomass or negative control samples where more adapters were in the sample than actual DNA) randomly ligate to free DNA from another (high-biomass) sample on the same sequencing run, causing sequences assigned to incorrect samples. Two general approaches can be used to ameliorate these problems. First, laboratory methods can be improved to reduce or remove index hopping; dual barcoding seems to be an effective strategy [12] although other strategies have been proposed (e.g., using known spike-in taxa [11]). Second, the unique pattern of occurrence of contamination sequences, which have relative abundance that increases as DNA concentration decreases and have higher relative abundance in negative controls than in experimental samples [38], can be detected using the software package Decontam [13], which has been shown to be effective in removing contamination in controlled experiments [14]. The growing awareness of the importance of this problem [39] should speed the development and adoption of laboratory and analysis methods to prevent, or at least remove, these experimental artifacts.

Conclusions

We present a non-stochastic approach to PERMANOVA presence-absence analyses that aggregates information over *all* potential rarefaction replicates without actual resampling, when the Jaccard or unweighted UniFrac distance is used. Our approach is robust to confounding by library size and outperforms the three possible stochastic resampling approaches we have considered. As PERMANOVA is currently the most commonly used method for testing community-level hypotheses about the microbiome, our extension of PERMANOVA for robust and powerful analysis of presence-absence associations represents substantial advances of PERMANOVA to be better used to advance scientific knowledge.

Funding

This research was supported by the National Institutes of Health awards R01GM116065 (Hu).

Availability of data and materials

The R package LDM is available on GitHub at <https://github.com/yijuanhu/LDM> in formats appropriate for Macintosh, Linux, or Windows systems.

Authors' contributions

YJH and GAS conceived the study, developed the method, and wrote the manuscript. YJH performed the simulation studies and analyzed the data. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publications

No consent for publication was required for this study; all microbiome datasets used here are publicly available.

Ethics approval and consent to participate

This study only involved secondary analyses of existing, de-identified datasets; as such it does not require separate IRB consent.

Acknowledgements

Not applicable.

References

1. McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*. 2001;82(1):290–297.
2. Claesson MJ, Jeffery IB, Conde S, Power SE, O’connor EM, Cusack S, et al. Gut microbiota composition correlates with diet and health in the elderly. *Nature*. 2012;488(7410):178–184.
3. Freitas AC, Bocking A, Hill JE, Money DM. Increased richness and diversity of the vaginal microbiota and spontaneous preterm birth. *Microbiome*. 2018;6(1):1–15.
4. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*. 2005;71(12):8228–8235.
5. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*. 2007;73(5):1576–1585.
6. Hughes JB, Hellmann JJ. The application of rarefaction techniques to molecular inventories of microbial diversity. *Methods in Enzymology*. 2005;397:292–308.

7. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*. 2014;10(4):e1003531.
8. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*. 2019;37(8):852–857.
9. Minich JJ, Sanders JG, Amir A, Humphrey G, Gilbert JA, Knight R. Quantifying and understanding well-to-well contamination in microbiome research. *MSystems*. 2019;4(4).
10. Hornung BV, Zwittink RD, Kuijper EJ. Issues and current standards of controls in microbiome research. *FEMS microbiology ecology*. 2019;95(5):fiz045.
11. Zinter MS, Mayday MY, Ryckman KK, Jelliffe-Pawlowski LL, DeRisi JL. Towards precision quantification of contamination in metagenomic sequencing experiments. *Microbiome*. 2019;7:62.
12. Pjevac P, Hausmann B, Schwarz J, Kohl G, Herbold CW, Loy A, et al. An economical and flexible dual barcoding, two-step PCR approach for highly multiplexed amplicon sequencing. *Frontiers in Microbiology*. 2021;12:669776.
13. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. 2018;6(1):226.
14. Karstens L, Asquith M, Davin S, Fair D, Gregory WT, Wolfe AJ, et al. Controlling for contaminants in low-biomass 16s rRNA gene sequencing experiments. *mSystems*. 2019;4(4):e00290–19.
15. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. 2017;5(1):27.
16. Thorsen J, Brejnrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, et al. Large-scale benchmarking reveals false discoveries and count transformation sensitiv-

- ity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome*. 2016;4(1):62. PMID: PMC5123278.
17. Hawinkel S, Mattiello F, Bijmens L, Thas O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in Bioinformatics*. 2017;20(1):210–221.
 18. Stewart CJ, Ajami NJ, O’Brien JL, Hutchinson DS, Smith DP, Wong MC, et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature*. 2018;562(7728):583–588.
 19. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature*. 2018;555(7695):210–215.
 20. Durack J, Huang YJ, Nariya S, Christian LS, Ansel KM, Beigelman A, et al. Bacterial biogeography of adult airways in atopic asthma. *Microbiome*. 2018;6(1):1–16.
 21. Armstrong AJ, Shaffer M, Nusbacher NM, Griesmer C, Fiorillo S, Schneider JM, et al. An exploration of Prevotella-rich microbiomes in HIV and men who have sex with men. *Microbiome*. 2018;6(1):1–16.
 22. Darcy JL, Washburne AD, Robeson MS, Prest T, Schmidt SK, Lozupone CA. A phylogenetic model for the recruitment of species into microbial communities and application to studies of the human microbiome. *The ISME journal*. 2020;14(6):1359–1368.
 23. Fouquier J, Huizar NM, Donnelly J, Glickman C, Kang DW, Maldonado J, et al. The Gut Microbiome in Autism: Study-Site Effects and Longitudinal Analysis of Behavior Change. *Msystems*. 2021;6(2).
 24. Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, et al. Advancing our understanding of the human microbiome using QIIME. In: *Methods in Enzymology*. vol. 531. Elsevier; 2013. p. 371–444.
 25. Tang ZZ, Chen G, Alekseyenko AV. PERMANOVA-S: association test for microbial

- community composition that accommodates confounders and multiple distances. *Bioinformatics*. 2016;32(17):2618–2625.
26. Zhang J, Wei Z, Chen J. A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics*. 2018;34(11):1875–1883.
 27. Brill B, Amir A, Heller R. Testing for differential abundance in compositional counts data, with application to microbiome studies. *arXiv*. 2019;1904.08937.
 28. Plantinga AM, Chen J, Jenq RR, Wu MC. pldist: Ecological dissimilarities for paired and longitudinal microbiome association analysis. *Bioinformatics*. 2019;35(19):3567–3575.
 29. Koh H, Tuddenham S, Sears CL, Zhao N. Meta-analysis methods for multiple related markers: Applications to microbiome studies with the results on multiple α -diversity indices. *Statistics in Medicine*. 2021;.
 30. Glassman SI, Martiny JB. Ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *bioRxiv*. 2018;p. 283283.
 31. Glassman SI, Weihe C, Li J, Albright MB, Looby CI, Martiny AC, et al. Decomposition responses to climate depend on microbial community composition. *Proceedings of the National Academy of Sciences*. 2018;115(47):11994–11999.
 32. 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):e15216.
 33. Jongman E. Data analysis in community and landscape ecology. Cambridge University Press; 1995.
 34. Hu YJ, Satten GA. Testing hypotheses about the microbiome using the linear decomposition model (LDM). *Bioinformatics*. 2020;36(14):4106–4115.
 35. Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, et al. The treatment-naïve microbiome in new-onset Crohn’s disease. *Cell Host and Microbe*.

2014;15(3):382–392.

36. 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. *The American Journal of Human Genetics*. 2015;96(5):797–807.
37. Hu YJ, Lane A, Satten GA. A rarefaction-based extension of the LDM for testing presence-absence associations in the microbiome. *Bioinformatics*. 2021;p. <https://doi.org/10.1093/bioinformatics/btab012>.
38. Jervis-Bardy J, Leong LEX, Marri S, Smith RJ, Choo JM, Smith-Vaughan HC, et al. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome*. 2015;3:19.
39. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, S WL. Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends in Microbiology*. 2019;27(2):105–117.

Figure titles and legends

Figure 1. Workflows of the proposed methods, $F(R)$, $D2(R)$, and $D2-A_2$, as well as the existing method $D(R)$.

Figure 2. Convergence of the average distance matrix over R rarefactions to the first-order (gray) and second-order (black) approximations to the expected distance matrix for three simulated data scenarios. The y -coordinate for each point is the maximum over all matrix elements of the absolute value of the percent difference between the average distance matrix and the approximate expected distance matrix, relative to the approximate expected distance matrix. The upper and lower panels correspond to squared and unsquared distance matrices, respectively. The reference line in red is the least-squares fit to the points corresponding to the second-order approximation, with the slope constrained to be -0.5 corresponding to a $\sqrt{R}$ convergence rate. The effect size β was set to 0. In M1 and M2, the mean library size was set to 10K for controls and 5K for cases; in M3, it was set to 100 for controls and 50 for cases.

Figure 3. Ordination based on the (squared and centered) distance matrices (after projecting off the effect of the confounder) for the data simulated under the null when the effect size β was set to 0. In M1 and M2, the mean library size was set to 10K for controls and 5K for cases; in M3, it was set to 100 for controls and 50 for cases. The ellipses are 90% confidence limits.

Figure 4. Size (i.e., type I error) and power of the tests, when there are differential library sizes in cases and controls. For assessing power, the mean library sizes are 5K in cases under M1 and M2, and 50 in cases under M3. The gray dotted line represents the nominal significance level of 0.05.

Figure 5. Power of the tests with rarefaction over increasing number of rarefactions. The effect size β was set to 0.4, 0.35, and 0.2 under M1, M2, and M3, respectively.

Figure 6. Comparing unrarefied analysis with rarefied analysis, when there are no systematic differences in library size between cases and controls. The effect size β was set to 0.4,

0.35, and 0.2 under M1, M2, and M3, respectively.

Figure 7. Differential distributions of library sizes in cases and controls of the IBD data. The p -value was generated by the Wilcoxon rank-sum test.

Figure 8. The $-\log$ of average p -values for testing presence-absence association of the rectal microbiome with the IBD status for 1000 subsets of the original data, each consisting of 50 randomly selected samples.

Figure 1. Workflows of the proposed methods, $F(R)$, $D2(R)$, and $D2-A_2$, as well as the existing method $D(R)$.

Figure 2. Convergence of the average distance matrix over R rarefactions to the first-order (gray) and second-order (black) approximations to the expected distance matrix for three simulated data scenarios. The y -coordinate for each point is the maximum over all matrix elements of the absolute value of the percent difference between the average distance matrix and the approximate expected distance matrix, relative to the approximate expected distance matrix. The upper and lower panels correspond to squared and unsquared distance matrices, respectively. The reference line in red is the least-squares fit to the points corresponding to the second-order approximation, with the slope constrained to be -0.5 corresponding to a $\sqrt{R}$ convergence rate. The effect size β was set to 0. In M1 and M2, the mean library size was set to 10K for controls and 5K for cases; in M3, it was set to 100 for controls and 50 for cases.

Figure 3. Ordination based on the (squared and centered) distance matrices (after projecting off the effect of the confounder) for the data simulated under the null when the effect size β was set to 0. In M1 and M2, the mean library size was set to 10K for controls and 5K for cases; in M3, it was set to 100 for controls and 50 for cases. The ellipses are 90% confidence limits.

Figure 4. Size (i.e., type I error) and power of the tests, when there are differential library sizes in cases and controls. For assessing power, the mean library sizes are 5K in cases under M1 and M2, and 50 in cases under M3. The gray dotted line represents the nominal significance level of 0.05.

Figure 5. Power of the tests with rarefaction over increasing number of rarefactions. The effect size β was set to 0.4, 0.35, and 0.2 under M1, M2, and M3, respectively.

Figure 6. Comparing unrarefied analysis with rarefied analysis, when there are no systematic differences in library size between cases and controls. The effect size β was set to 0.4, 0.35, and 0.2 under M1, M2, and M3, respectively.

Figure 7. Differential distributions of library sizes in cases and controls of the IBD data. The p -value was generated by the Wilcoxon rank-sum test.

Figure 8. The $-\log$ of average p -values for testing presence-absence association of the rectal microbiome with the IBD status for 1000 subsets of the original data, each consisting of 50 randomly selected samples.

Included in a rarefaction?	Taxon k	Taxon k'	All other taxa	
Yes	Y_{ik}	$Y_{ik'}$	$N_0 - Y_{ik} - Y_{ik'}$	N_0
No	$C_{ik} - Y_{ik}$	$C_{ik'} - Y_{ik'}$	$N_i - N_0 - C_{ik} + Y_{ik} - C_{ik'} + Y_{ik'}$	$N_i - N_0$
	C_{ik}	$C_{ik'}$	$N_i - C_{ik} - C_{ik'}$	N_i

Figure S1. Schematic 2×3 table illustrating the multivariate hypergeometric distribution underlying the generation of counts during rarefaction of sample i . N_i is the original library size of sample i , N_0 is the rarefaction depth, C_{ik} is the original read count of taxon k in sample i , and Y_{ik} is the read count of taxon k in sample i after rarefaction.

Figure S2. Counterpart to Figure 2, for data that were simulated under the alternative when the effect size β was set to 0.6 in M1 and M2 and 0.3 in M3.

Figure S3. Counterpart to Figure 3, for data that were simulated under the alternative when the effect size β was set to 0.6 in M1 and M2 and 0.3 in M3.

Figure S4. Size of the tests, when the confounder is not controlled for. The gray dotted line represents the nominal significance level of 0.05.

Figure S5. Contrasting p -values generated by D2- A_2 and D2- A_1 for 1000 replicates of data. The mean library sizes are 10K and 5K in controls and cases, respectively, under M1 and M2, and 100 and 50 in controls and controls under M3. The effect size β was set to 0.4 under M1 and M2 and 0.2 under M3.

Figure S6. Ordination based on the (squared and centered) Jaccard and unweighted UniFrac distance matrices (after projecting off sex and ³⁷antibiotic use variables) for the IBD data. The ellipses are 95% confidence limits for case and control clusters.

Supplemental Texts: Calculating the Expected Jaccard and Unweighted UniFrac Distances

S1. Statistics of Rarefaction

To calculate the expectation of the squared Jaccard distance, we will need the following quantities. From the multivariate hypergeometric (MH) distribution illustrated in Figure S1, we have

$$\Pr(Y_{ik} = 0) = \frac{\binom{C_{ik}}{0} \binom{N_i - C_{ik}}{N_0}}{\binom{N_i}{N_0}} := \bar{p}_{ik} := 1 - p_{ik},$$

and

$$\Pr(Y_{ik} = 0 \ \& \ Y_{ik'} = 0) = \frac{\binom{C_{ik}}{0} \binom{C_{ik'}}{0} \binom{N_i - C_{ik} - C_{ik'}}{N_0}}{\binom{N_i}{N_0}} := \bar{P}_{i;kk'},$$

where $:=$ is used to define a new quantity. In terms of these quantities, it is easy to show that

$$\Pr(Y_{ik} = 1 \ \& \ Y_{ik'} = 1) := P_{i;kk'} = p_{ik} + p_{ik'} + \bar{P}_{i;kk'} - 1.$$

Since $\mathcal{I}_{ik} = I(Y_{ik} > 0)$, we have $E(\mathcal{I}_{ik}) = p_{ik}$ and $E(\mathcal{I}_{ik}\mathcal{I}_{ik'}) = P_{i;kk'}$.

For the unweighted UniFrac distance, we define $\mathcal{I}_{ik} = I(\sum_{k' \in \mathcal{S}_k} Y_{ik'} > 0)$. We have from the collapsibility of the MH distribution that

$$\Pr(\mathcal{I}_{ik} = 0) = \Pr\left(\sum_{k' \in \mathcal{S}_k} Y_{ik'} = 0\right) = \frac{\binom{N_i - \sum_{k' \in \mathcal{S}_k} C_{ik'}}{N_0}}{\binom{N_i}{N_0}} := \bar{p}_{ik} := 1 - p_{ik}.$$

Similar to the Jaccard case, let

$$\Pr(\mathcal{I}_{ik} = 0 \ \& \ \mathcal{I}_{ik'} = 0) = \Pr\left(\sum_{k' \in \mathcal{S}_k \cup \mathcal{S}_{k'}} Y_{ik'} = 0\right) := \bar{P}_{i;kk'}.$$

Then, as with the Jaccard distance, we have

$$\Pr(\mathcal{I}_{ik} = 1 \ \& \ \mathcal{I}_{ik'} = 1) := P_{i;kk'} = p_{ik} + p_{ik'} + \bar{P}_{i;kk'} - 1,$$

and so $E(\mathcal{I}_{ik}) = p_{ik}$ and $E(\mathcal{I}_{ik}\mathcal{I}_{ik'}) = P_{i;kk'}$. Note that the precise definitions of $\mathcal{I}_{ik}$, p_{ik} and $P_{i;kk'}$ are different for the Jaccard and unweighted UniFrac distances.

S2. Delta-Method Approximation to the Expected Jaccard and Unweighted UniFrac Distances

We use the Delta method to derive a second-order approximation to $E(\Delta^2)$, as well as to $E(\Delta)$ which is also included in our numerical evaluation. For simplicity we drop the indices i, j . If we expand $\Delta = T_2/T_1$ for T_1 around μ_1 and T_2 around μ_2 to second order, we obtain

$$E(\Delta) \approx \frac{\mu_2}{\mu_1} + \frac{1}{2} \left[-\frac{2}{\mu_1^2} \text{Cov}(T_1, T_2) + \frac{2\mu_2 \text{Var}(T_1)}{\mu_1^3} \right] = \frac{\mu_2}{\mu_1} + \frac{\mu_2 m_{11} - \mu_1 m_{12}}{\mu_1^3}$$

where

$$m_{ab} = E(T_a T_b).$$

Similarly, if we expand $\Delta^2 = T_2^2/T_1^2$ for T_1 around μ_1 and T_2 around μ_2 to second order, we obtain

$$E(\Delta^2) \approx \frac{\mu_2^2}{\mu_1^2} + \frac{\mu_1^2 m_{22} - 4\mu_1 \mu_2 m_{12} + 3\mu_2^2 m_{11}}{\mu_1^4}.$$

To justify use of the delta method, recall from equation (2) that T_1 and T_2 can be considered as means over the stochastic process that generates data at a random taxon (for the Jaccard distance) or a random branch (for the unweighted UniFrac distance). Thus, we can expect that the second moments are of order $\mathcal{O}(K^{-1})$ and that each successive moment will be smaller by a factor of $K^{-\frac{1}{2}}$.

S3. Calculation of Moments

We show the details for the UniFrac distance; the results for the Jaccard distance are easily obtained from those of the UniFrac by setting all weights $b_k = 1$ and choosing the sum over k to index taxa rather than nodes in the phylogenetic tree. The first moments are easy, and we find

$$\mu_a = E \left\{ \sum_{k=1}^K b_k \left(\mathcal{I}_{ik} + \mathcal{I}_{jk} - a \mathcal{I}_{ik} \mathcal{I}_{jk} \right) \right\} = \sum_k b_k \left(p_{ik} + p_{jk} - a p_{ik} p_{jk} \right) = p_{i\cdot}^{(1)} + p_{j\cdot}^{(1)} - a \sum_k b_k p_{ik} p_{jk},$$

where we define $p_{i\cdot}^{(1)} = \sum_{k=1}^K b_k p_{ik}$. The second moment requires some work and we first expand

$$\begin{aligned} m_{a_1 a_2} &= E \left\{ \sum_{k=1}^K \sum_{k'=1}^K b_k b_{k'} \left(\mathcal{I}_{ik} + \mathcal{I}_{jk} - a_1 \mathcal{I}_{ik} \mathcal{I}_{jk} \right) \left(\mathcal{I}_{ik'} + \mathcal{I}_{jk'} - a_2 \mathcal{I}_{ik'} \mathcal{I}_{jk'} \right) \right\} \\ &= E \left\{ \sum_{k=1}^K b_k^2 \left(\mathcal{I}_{ik} + \mathcal{I}_{jk} - a_1 \mathcal{I}_{ik} \mathcal{I}_{jk} \right) \left(\mathcal{I}_{ik} + \mathcal{I}_{jk} - a_2 \mathcal{I}_{ik} \mathcal{I}_{jk} \right) \right\} \end{aligned} \quad (\text{S1})$$

$$+ E \left\{ \sum_{k,k'=1, k \neq k'}^K b_k b_{k'} \left(\mathcal{I}_{ik} + \mathcal{I}_{jk} - a_1 \mathcal{I}_{ik} \mathcal{I}_{jk} \right) \left(\mathcal{I}_{ik'} + \mathcal{I}_{jk'} - a_2 \mathcal{I}_{ik'} \mathcal{I}_{jk'} \right) \right\}. \quad (\text{S2})$$

The first term (S1) can be factored out to give, by noting that $\mathcal{I}_{ik}^2 = \mathcal{I}_{ik}$,

$$\begin{aligned} &E \left\{ \sum_{k=1}^K b_k^2 \left[\mathcal{I}_{ik} + \mathcal{I}_{jk} + 2\mathcal{I}_{ik} \mathcal{I}_{jk} - 2(a_1 + a_2) \mathcal{I}_{ik} \mathcal{I}_{jk} + a_1 a_2 \mathcal{I}_{ik} \mathcal{I}_{jk} \right] \right\} \\ &= \sum_{k=1}^K b_k^2 \left\{ p_{ik} + p_{jk} + \left[2 - 2(a_1 + a_2) + a_1 a_2 \right] p_{ik} p_{jk} \right\} \\ &= p_{i\cdot}^{(2)} + p_{j\cdot}^{(2)} + \left[2 - 2(a_1 + a_2) + a_1 a_2 \right] \sum_{k=1}^K b_k^2 p_{ik} p_{jk}, \end{aligned}$$

where we define $p_{i\cdot}^{(2)} = \sum_{k=1}^K b_k^2 p_{ik}$. The second term (S2) can be factored out to give

$$\begin{aligned} &\sum_{k,k'=1, k \neq k'}^K b_k b_{k'} \left\{ E \left(\mathcal{I}_{ik} \mathcal{I}_{ik'} + \mathcal{I}_{ik} \mathcal{I}_{jk'} + \mathcal{I}_{ik'} \mathcal{I}_{jk} + \mathcal{I}_{jk} \mathcal{I}_{jk'} \right) \right. \\ &\quad \left. - E \left(a_2 \mathcal{I}_{ik} \mathcal{I}_{ik'} \mathcal{I}_{jk'} + a_2 \mathcal{I}_{jk} \mathcal{I}_{ik'} \mathcal{I}_{jk'} + a_1 \mathcal{I}_{ik'} \mathcal{I}_{ik} \mathcal{I}_{jk} + a_1 \mathcal{I}_{jk'} \mathcal{I}_{ik} \mathcal{I}_{jk} \right) + a_1 a_2 E \left(\mathcal{I}_{ik} \mathcal{I}_{ik'} \mathcal{I}_{jk} \mathcal{I}_{jk'} \right) \right\} \\ &= \sum_{k,k'=1, k \neq k'}^K b_k b_{k'} \left\{ P_{i;kk'} + 2p_{ik} p_{jk'} + P_{j;kk'} - \left(a_2 p_{jk'} + a_1 p_{jk} \right) P_{i;kk'} - \left(a_2 p_{ik'} + a_1 p_{ik} \right) P_{j;kk'} + a_1 a_2 P_{i;kk'} P_{j;kk'} \right\}. \end{aligned}$$

The term that involves $p_{ik} p_{jk'}$ can be simplified by noting that

$$2 \sum_{k,k'=1, k \neq k'}^K b_k b_{k'} p_{ik} p_{jk'} = 2 \sum_{k=1}^K \sum_{k'=1}^K b_k b_{k'} p_{ik} p_{jk'} - 2 \sum_{k=1}^K b_k^2 p_{ik} p_{jk} = 2p_{i\cdot}^{(1)} p_{j\cdot}^{(1)} - 2 \sum_{k=1}^K b_k^2 p_{ik} p_{jk}.$$

Combining the results of the two terms (S1) and (S2), we have the second moment to be

$$\begin{aligned} m_{a_1 a_2} &= p_{i\cdot}^{(2)} + p_{j\cdot}^{(2)} + 2p_{i\cdot}^{(1)} p_{j\cdot}^{(1)} + \left[a_1 a_2 - 2(a_1 + a_2) \right] \sum_{k=1}^K b_k^2 p_{ik} p_{jk} \\ &\quad + \sum_{k,k'=1, k \neq k'}^K b_k b_{k'} \left\{ \left(1 - a_1 p_{jk} - a_2 p_{jk'} \right) P_{i;kk'} + \left(1 - a_1 p_{ik} - a_2 p_{ik'} \right) P_{j;kk'} + a_1 a_2 P_{i;kk'} P_{j;kk'} \right\}. \end{aligned}$$

To speed up computation of the three terms in $m_{a_1 a_2}$ that involve $\sum_{k,k'=1,k \neq k'}^K$, we modify the sum over (k, k') to only consider $k < k'$ values; this will allow us to cut the number of terms in the sum by a factor of 2. This requires a bit of care, since the case $a_1 \neq a_2$ has some asymmetry, so we cannot just impose the restriction and multiply by a factor of 2 to compensate. To start this, note that $\sum_{k,k'=1,k \neq k'}^K = \sum_{k=1}^{K-1} \sum_{k'=k+1}^K + \sum_{k=2}^K \sum_{k'=1}^{k-1}$. For the first term involving $\sum_{k,k'=1,k \neq k'}^K$, we define the first double summation to be

$$\phi_{ij}(a_1, a_2) = \sum_{k=1}^{K-1} \sum_{k'=k+1}^K b_k b_{k'} \left(1 - a_1 p_{jk} - a_2 p_{jk'}\right) P_{i;kk'}.$$

Then the second double summation can be written as

$$\sum_{k=2}^K \sum_{k'=1}^{k-1} b_k b_{k'} \left(1 - a_1 p_{jk} - a_2 p_{jk'}\right) P_{i;kk'} = \sum_{k'=1}^{K-1} \sum_{k=k'+1}^K b_k b_{k'} \left(1 - a_1 p_{jk} - a_2 p_{jk'}\right) P_{i;kk'} = \phi_{ij}(a_2, a_1),$$

where the first equation follows from exchanging the order of the sums. Thus, the first term involving $\sum_{k,k'=1,k \neq k'}^K$ is $\phi_{ij}(a_1, a_2) + \phi_{ij}(a_2, a_1)$. The second term involving $\sum_{k,k'=1,k \neq k'}^K$ only differs by exchanging i and j , and gives $\phi_{ji}(a_1, a_2) + \phi_{ji}(a_2, a_1)$. The situation with the third term is easier, as the summation does not involve a_1 and a_2 :

$$a_1 a_2 \sum_{k,k'=1,k \neq k'}^K b_k b_{k'} P_{i;kk'} P_{j;kk'} = 2a_1 a_2 \sum_{k=1}^{K-1} \sum_{k'=k+1}^K b_k b_{k'} P_{i;kk'} P_{j;kk'}.$$

The calculation for the Jaccard distance, obtained by setting $b_k = 1$ for any k , is nearly identical. The only difference is that with the weights removed, we have $p_i^{(1)} = p_i^{(2)}$.