Testing Presence-Absence Associations in the Microbiome using the LDM and PERMANOVA

Short Title: Testing Presence-Absence Associations

Yi-Juan Hu[1,*], Andrea Lane[1], and Glen A. Satten[2]

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

[2]Centers for Disease Control and Prevention, Atlanta, GA, 30341, USA

*Email: yijuan.hu@emory.edu

Keyword: microbiome association test; read depth; library size

Abstract

Background: Most methods for testing microbiome associations with covariates of interest (e.g., clinical outcomes or environmental factors) assumed that the associations are driven by changes in relative abundance of species. However, it is also well known that many associations are driven by changes in presence-absence of species. The analysis of presence-absence associations faces a unique challenge, which is confounding by library size (i.e., sequencing depth). Currently, there is a lack of robust and efficient methods for testing presence-absence associations.

Methods: PERMANOVA is a commonly used distance-based method for testing hypotheses at the community level. We have also developed the linear decomposition model (LDM) that unifies the community-level and OTU-level tests into one framework. Here we present a strategy that can be used with both PERMANOVA and the LDM for testing presence-absence associations. We propose to rarefy the OTU table such that all samples have the same library size, and then apply PERMANOVA to a presence-absence-type of distance based on the rarefied table or the LDM directly to the rarefied table; we repeat the process for a number of randomly rarefied tables and sum up the test statistics over the replicates. The flexible nature of PERMANOVA and the LDM allows discrete or continuous traits or interactions to be tested and extra confounding covariates to be adjusted.

Results: Our simulations indicate that the proposed strategy is robust to any systematic differences in library size. Using simulation, we also explored the optimal number of rarefactions that maintains a good balance between statistical power and computation time. Additionally, we provided practical guidelines on how to select the rarefaction depth. Finally, we illustrated our strategy through an analysis of the data on inflammatory bowel disease (IBD) that have differential library sizes between cases and controls.

Introduction

Most existing methods for testing association of the microbiome with covariates of interest (e.g., clinical outcomes or environmental factors) assumed that the association is driven by changes in relative abundance of species; these methods include LDM [1], metagenomeSeq [2], ANCOM [3, 4], ALDEX2 [5], and PERMANOVA [6, 7] and MiRKAT [8] with Bray-Curtis and weighted UniFrac distances. However, it is also well known that many associations are driven by changes in presence-absence of species. For example, in the human gut, increased species richness are known to be associated with more stable ecosystems [9, 10], which tend to be resistant to environmental pressures such as diet, antibiotic use and pathogen invasion [11, 12]. In contrast, a healthy vaginal microbiome is often characterized by low diversity that is dominated by *lactobacilli* [13].

Several distance measures were designed for quantifying presence-absence dissimilarities. For example, the unweighted UniFrac uses the difference of the species presence-absence status between two samples to determine the inclusion of a particular tree branch [14, 15]. The Jaccard distance is defined as the number of unique species present in either sample, divided by the number of species present in any of the two samples. So, a common approach to testing whether the community-level (i.e., global) presence-absence pattern is associated with covariates of interest is to apply a distance-based method such as PERMANOVA with a presence-absence-type of distance.

The LDM [1] is based on linear models that treat (transformed) relative abundances of species (or OTUs and other features) as the response. It is possible to extend the LDM to testing presence-absence associations by simply replacing the relative abundance data with the presence-absence data. The test will thus inherit the nice features of the original LDM, namely providing unified testing of community-level and OTU-level associations, accommodating both continuous and discrete variables (e.g., clinical outcomes, environmental factors) as well as interaction terms to be tested either singly or in combination, allowing for adjustment of

confounding covariates, and using permutation-based p-values that can control for correlation (e.g., matched-set data [ref]).

The analysis of presence-absence data, including the approaches above, faces a unique challenge, which is confounding by library size (i.e., sequencing depth). Modern 16S rRNA gene sequencing, although having a high yield of library size, still results in *under-sampling* of many species, especially rare ones. It means that a species may appear to be *absent* due to insufficient coverage of sequencing reads. Thus, the observed zeros in the OTU table are a mixture of "structural zeros" (due to physical absence) and "sampling zeros" (due to under-sampling). While structural zeros are the subject of interest in the analysis of presence-absence association, it is impossible to separate them from sampling zeros, which are strongly dependent on library size. In addition, library size is subject to experimental conditions, often varies over several ranges of magnitude, and thereby highly correlates with experimental batches. When the batch assignment correlates with the covariate of interest (as sample randomization over batches can be difficult in medical contexts with complicated process of recruitment), library size will be established as a confounder in the association between the presence-absence of species and the covariate. Failure to account for the variation of library size can easily lead to false positive results. In theory, it also affects the analysis of abundance data, but the impact would be much smaller.

To overcome the potential confounding effects of library size, normalization is needed. Of note, the normalization methods developed for RNA-seq data, such as DESeq2 [16] and EdgeR [17], would not work because they all derive sample-specific constants to rescale count data of each sample, which have no impact on the presence-absence statuses. A commonly used normalization approach in Ecology is *rarefaction* [18, 19]. A rarefaction procedure selects a rarefaction depth (often the lowest library size after excluding outliers) and subsample the reads of each sample without replacement such that all samples have the same library size; samples that have fewer reads than the rarefaction depth are discarded. Rarefaction can be effective in removing the confounding effects of library size, because sampling zeros are evened

up across the covariate of interest (e.g., case-control groups). For example, rarefaction more clearly clusters samples according to biological origin than other normalization techniques do for ordination metrics based on presence or absence [20]. On the other hand, rarefaction typically results in a big loss of sequence reads and hence statistical power. To address this issue, multiple rarefaction [19] has been proposed to salvage the information. However, it remains unclear how to aggregate the information from multiple rarefied datasets, and what is the optimal number of rarefactions that balances computational cost and statistical power.

In this article, we develop a new strategy for using PERMANOVA and the LDM to analyze presence-absence data that accounts for the variation of library size. In the methods section, we will adopt rarefaction as our choice of normalization and identify the correct way to combine multiple rarefactions. In the results section, we present the simulation studies and the application to a real study on inflammatory bowel diseases (IBD). We conclude with a discussion section.

Methods

Both PERMANOVA and the LDM are linear models for which the covariates (metadata) are summarized in a design matrix X (where the rows correspond to n samples and the columns correspond to the covariates). We may partition X by columns into K groups (which we call "submodels") such that $X = (X_1, X_2, \ldots, X_K)$, where each X_k denotes a variable or set of variables we wish to test jointly. For example, X_k may consist of indicator variables for levels of a single categorical variable, or a group of potential confounders that we wish to adjust for simultaneously. Both PERMANOVA and the LDM make the columns of X_k *orthonormal* to the columns of $X_{k'}$ for $k' < k$ using projection (i.e., the Gram-Schmidt process).

Let Y be the OTU table of read counts and $Y^{(r)}$ the rarefied OTU table after the rth rarefaction. For PERMANOVA, we construct a presence-absence-type of distance matrix from $Y^{(r)}$, based on which we calculate the F statistic $F^{(r)}_{k,\text{PERMANOVA}}$ according to formula (12) of Hu and Satten [1]. For the LDM, we convert $Y^{(r)}$ into a presence-absence matrix

$Y_{\text{pa}}^{(r)} = \mathbb{I}(Y^{(r)} > 0)$, where $\mathbb{I}(.)$ is the indicator function that performs element-wise operation. Then, we calculate the OTU-level F statistic $F_{kj}^{(r)}$ and the community-level F statistic $F_{k,\text{global}}^{(r)}$ according to formulas (7) and (8) of Hu and Satten [1] by replacing Y with $Y_{\text{pa}}^{(r)}$; note that we do not scale the rows of $Y_{\text{pa}}^{(r)}$ by row totals in the context of presence-absence analysis. Since rarefaction removes any effect of library size, we expect each test statistic is free of confounding by library size.

With R rarefaction replicates, we aggregate their information by summing up $F_{k,\text{PERMANOVA}}^{(r)}$, $F_{kj}^{(r)}$, or $F_{k,\text{global}}^{(r)}$ over $r = 1, 2, \ldots, R$ to form the final observed statistics. To assess significance of these statistics using Freedman-Lane-based permutation, we operate on the same set of rarefied OTU tables as for the observed statistics. We permute the covariates as in Hu and Satten [1], with a specific permutation of integers $1, 2, \ldots, n$ denoted by π, and calculate the statistics $F_{k,\text{PERMANOVA}}^{(r,\pi)}$, $F_{kj}^{(r,\pi)}$, or $F_{k,\text{global}}^{(r,\pi)}$ according to formulas (13), (10), or (11) of Hu and Satten [1]. Then, we sum up the statistics over $r = 1, \ldots, R$ as for the observed statistics. Since a test statistic based on one replicate is free of confounding by library size, the one based on R replicates is also free of confounding. Overall, the computation time is expected to increase linearly with the number of replicates R.

Finally, we consider an alternative strategy for PERMANOVA that we first sum up distance matrices across rarefied OTU tables and then apply PERMANOVA to the aggregated distance matrix. We show that the aggregated distance matrix is not theoretically satisfactory. Consider a simple case in which samples in group 1 and group 2 have fixed library sizes 10K and 1K, respectively. Then the rarefaction depth is always 1K and *all* reads in group 2 are included in any rarefied OTU table. Consider an extreme sampling process that generated 10 rarefied tables, where the 1K reads of each group-1 sample have no overlap across the 10 tables. Suppose that all reads are independent of each other, which can occur under the multinomial or Poisson sampling distribution with no overdispersion. Also suppose that there is no difference in microbiome profile between any samples and that the variance of the distance between any samples based on any rarefied OTU table is d. As a result, the variance of

the aggregated distance based on the 10 rarefied OTU tables is $10d$ between group-1 samples and $100d$ between group-2 samples. This aggregated distance matrix is unsatisfactory because the distances have systematic differences despite the homogeneous samples. In reality, library sizes are not fixed, reads have overlaps across subsamples, and read count data have overdispersion, but similar differences exist albeit with a weakened strength. Because it is difficult to analytically evaluate the impact of such an aggregated distance matrix on association testing, we will assess the impact via simulation.

Results

Simulation studies

We assumed 50 case samples and 50 control samples in each dataset, and generated their read count data from the Dirichlet-Multinomial (DM) model. We used the relative abundances of 856 OTUs that were estimated from the upper-respiratory-tract (URT) microbiome data [21], and replaced part of the relative abundances by zeros to introduce physical absence; more details are deferred to the next paragraph. We set the overdispersion parameter to 0.02, which was also estimated from the URT data. For library size, we considered a low-throughput setting with mean 1.5K as observed in the URT data, as well as a high-throughput setting with mean 10K. We used these values for controls and varied them for cases so as to introduce systematic differences. Given a mean library size μ, the library size for each sample was drawn from $N(\mu, \mu/3)$. In most cases, the sampled library size was truncated at 2500 in the high-throughput setting and 500 in the low-throughput setting. The truncation values were used as the rarefaction depth unless otherwise stated; in this way, no samples were discarded due to rarefaction. In the last simulation study, we modified this setup and allowed samples to be discarded.

Now we describe how we derived relative abundances of 856 OTUs for cases and controls. Denote the case-control status by Y. We also simulated a confounder C to be a binary variable that has 70% "success" rate in controls but only 30% in cases. We considered two

complementary scenarios for the association mechanism. The first scenario (S1) assumed that a large number of moderately abundant and rare OTUs were *differentially present* between cases and controls, and the second scenario (S2) assumed that a few abundant OTUs were differentially present. While the presence-absence analysis is usually focused on rare OTUs, it is also plausible for an abundant OTU to become absent under a certain condition (reference?). Specifically, in S1, we (uniformly and independently) sampled two sets of 100 OTUs (after excluding those with relative abundance $> 1\%$) to be associated with Y and C; note that the two sets of OTUs may overlap. In S2, after excluding the most abundant OTU, we assumed the next ten most abundant OTUs were associated with Y and the next forty most abundant OTUs were associated with C. Denote the OTU relative abundances estimated from the URT data by the vector π, the entries of which are all positive, and denote the OTU relative abundances in sample i by π_i. For each sample, we set $\pi_i = \pi$ as a baseline. For a case, we set the entries for each OTU selected to be associated with Y to 0 with a probability β ($\in [0, 1]$). Here β is referred to as *effect size* and $\beta = 0$ corresponding to the null hypothesis of no association. For a sample with $C = 1$, we further set the entries in π_i for each OTU selected to be associated with C to 0 with a fixed probability 0.5. To ensure that all entries in π_i sum up to 1, we increased the entry of the most abundant OTU by the total mass that were set to 0; since the most abundant OTU is always present, this operation does not change its presence-absence status.

We applied PERMANOVA and the LDM with the proposed strategy of aggregating multiple rarefactions over the F statistics, referred to as PERMANOVA-F and LDM-F. In some studies, we included PERMANOVA and the LDM without rarefaction, referred to as PERMANOVA-UR and LDM-UR. We also evaluated the strategy that simply adjusted the library size as a covariate in PERMANOVA and the LDM, referred to as PERMANOVA-L and LDM-L. For PERMANOVA, we additionally compared to the strategy that aggregates over the distance matrix, referred to as PERMANOVA-D. Here we use PERMANOVA to refer to the new implementation whose permutation is based on the Freedman-Lane scheme [1]. In some studies,

we also considered the implementation by the `adonis2` function in the R package `vegan` and the strategy that aggregates over the distance matrix, and refer to the method as `adonis2`-D. We evaluated the type I error (i.e., size) and power for testing the global hypothesis of no case-control differences after adjusting for the confounder C; the nominal significance level was set to 0.05. We assessed empirical sensitivity and empirical FDR for detecting individual OTUs that were differentially present between cases and controls after adjusting for C; the nominal FDR was set to 10%. Results for size were based on 10000 replicates; all other results were based on 1000 replicates. The results of PERMANOVA were based on the Jaccard distance unless otherwise specified.

Without rarefaction

We first investigated the consequences of not performing rarefaction in the presence of systematic differences in library size. We displayed the results of size in Figure 1, which shows that the size of PERMANOVA-UR and LDM-UR became increasingly inflated as the difference enlarged. The strategy of adjusting the library size as a covariate seemed to lead to accurate or only slightly inflated size. In the following simulation studies, we focused on the two settings: (1) mean library sizes 10K and 5K for cases and controls, respectively, and (2) mean library sizes 1.5K and 1K.

With rarefaction

Regardless of the number of rarefaction, PERMANOVA-F and LDM-F yielded correct size (Figure 2, upper panel). PERMANOVA-D and `adonis2`-D yielded correct or slightly conservative size. Note that all sizes would have been inflated (> 0.1) if the confounder was not adjusted for (Figure 2, lower panel), indicating that we have induced a substantial confounding effect and all methods were effective in adjusting for that.

For evaluating power, we focused on 5 rarefactions and will justify this number in the next section. Compared to PERMANOVA-F and LDM-F, PERMANOVA-L and LDM-L have

substantially reduced power, PERMANOVA-D has almost identical power, and `adonis2`-D has identical or slightly reduced power (Figure 3, upper panel). We found that, as the sequencing throughput was increased, the power increased under S1, which matched with our expectation, but the power slightly decreased under S2. The latter is because the association signals existed in the abundant OTUs whose presence-absence statuses were minimally affected by the library size, and a larger library size added noise at rare OTUs. With the Jaccard distance, PERMANOVA-F achieved as much power as LDM-F, but the power of PERMANOVA-F fell off quickly when the unweighted UniFrac distance was used. We note that the unweighted UniFrac distance is inappropriate as the association was induced without reference to the phylogenetic tree. For detecting differentially present OTUs, LDM-F yielded higher sensitivity than LDM-L (Figure 3, middle panel) and both controlled FDR (Figure 3, lower panel).

Although we pointed out that the aggregated distance matrix is theoretically unsatisfactory, we have not noted any differences between PERMANOVA-F and PERMANOVA-D in size and power from our simulation studies so far. Thus, we explored over more extreme settings. We focused on scenario S1, Jaccard distance, and 5 rarefactions, fixed all library sizes at 10K for cases and 1K for controls, and set the rarefaction depth at 1K. We varied the overdispersion parameter θ from 0.01, 0.001, to 0.0001. The ordination plot (Figure S1, upper panel) under the null hypothesis $\beta = 0$ (after adjusting for the confounder [1]) showed that, unlike the control samples, the case samples gradually concentrated at the center as the overdispersion reduced, which corroborated our derivation that the cases have smaller variance in the aggregated distance. As a consequence, the size of PERMANOVA-D became conservative (0.031 at overdispersion 0.0001) and the power reduced compared to PERMANOVA-F (by 6% at overdispersion 0.0001) (Figure 4, lower panel).

Optimal number of rarefactions

We explored the optimal number of rarefactions that can maintain a good balance between statistical power and computation time. We see from Figure 4 (upper panel) that, in all

scenarios and for both methods, the power curves grew steeply as the first 3 rarefactions were accumulated, then slowed down significantly when adding the 4th and 5th rarefaction, and nearly stabilized after 5. Note that even though the power was around 60% or 70% (hence not saturated) in several scenarios, it still stabilized after 5 rarefactions. We have further lowered the rarefaction depth from 2.5K to 0.5K in the high-throughput setting and from 0.5K to 0.25K in the low-throughput setting, so that each rarefied table contains a smaller proportion of reads. We found from Figure 4 (lower panel) that the power curves increased more quickly for the first 3 rarefactions, but still stabilized at 5. All results suggested 5 to be a practically optimal number of rarefactions.

Power of rarefying against no rarefying

When there are no systematic differences in library size, the analysis of unrarefied data is valid and gives the full power. In this case, it is of interest to assess power of the LDM (or PERMANOVA) with rarefaction relative to the full power. We first fixed the number of rarefactions at 5 and varied the mean library size, which was set equal in cases and controls, and we set the truncation value and the rarefaction depth to 25% of the mean library size (Figure 5, upper panel). Under S1, we see that, rarefaction led to loss of power against no rarefaction, which was more significant (10%) when the mean depth was low (1K) but shrunk to less than 5% when the mean depth became large (10K). We have further set the truncation value and the rarefaction depth to merely 10% of the mean library size, in order to create a more difficult situation for rarefaction. Indeed, the power loss was enlarged to 15% at 1K; nevertheless, it declined to ∼10% at 10K and ∼5% at 20K. Therefore, with modern sequencing techniques that typically yield high throughput data, we expect the loss of power to be very moderate. Under S2, interestingly, rarefaction even helped to improve power, due to the de-noising nature at rare OTUs which were not associated with the outcome. Next, we fixed the mean library size at 10K, truncation and rarefaction depth at 1K, and gradually increased the number of rarefactions (Figure 5, lower panel). We found that asymptotically we could

eventually recover the full power in S1, although if the rarefaction is severe, it may take a huge number of rarefactions.

Rarefaction depth and power

Finally, we considered the most realistic situation in which the simulated library sizes were merely truncated at 100, the rarefaction depth was chosen to be much higher than 100, and samples with library sizes lower than the rarefaction depth were discarded. Under S1, we have seen a positive relationship between rarefaction depth and power in Figure 4 (by comparing the upper and lower panels), when no samples were discarded. Now we re-evaluate this relationship when raising the rarefaction depth is at the cost of losing samples. Figure 6 shows that the power was first improved as we lifted the rarefaction depth from a very low level and removed samples with the lowest library sizes (hence minimum information). As we continued to raise the rarefaction depth and start losing a significant number of samples, the power started to drop. Under S2, the power always worsened as we raised the rarefaction depth, not to mention the loss of samples. These results suggest we should set rarefaction depth with caution and only remove samples whose library sizes are regarded as outliers.

Analysis of the IBD data

The IBD data we analyzed is a subset of the data generated from a study of a large cohort (called the RISK cohort) of pediatric patients with new-onset Crohn's Disease (CD) [22] as well as non-IBD controls. For each individual, samples from multiple gastrointestinal locations were collected at the time of diagnosis of CD before treatment initiation, and profiled by 16S rRNA gene sequencing on the Illumina MiSeq platform. We focused on a subset of the original data that pertain to samples from mucosal tissue biopsies at the rectum site. Further, we filtered out samples with library size less than 10,000, which cutoff value resulted in a small loss of 10% discarded samples but yielded a very high rarefaction depth (10,081) to be used later. In addition, we adopted a common filter by removing OTUs that were present in less than

five samples. Finally, we obtained data for 267 samples and 2,565 OTUs. In our analysis, we defined IBD cases as those with Crohn's Disease, Indeterminate Colitis, or Ulcerative Colitis, and obtained 169 cases and 98 controls. Our goal is to test the presence-absence association of the rectal microbiome with IBD status, while controlling for two potential confounders, sex and antibiotic use; there is an imbalance in the proportion of male subjects by case status (62% in cases, 44% in controls), indicates potential for confounding. It is of interest to test the association at the community level as well as detecting individual OTUs which contribute significantly to the community-level association.

Motivated by other researchers [20] who reported substantial confounding of PERMANOVA results due to library size based on data from the same study, we checked the library size distribution in cases and controls for our selected data. We indeed found that the library size distributions are systematically different (Figure 7, left). Thus, we rarefied the read count data of all 267 samples to the minimum depth, 10081. We constructed ordination plots in Figure S2 (left column) using the Jaccard distance after removing the effects of sex and antibiotic use, without rarefaction and with one rarefaction. These plots demonstrated a clear shift in cases compared with controls even after rarefaction. Without rarefaction, the two groups seem to separate further apart (which can be seen more clearly in the left two columns of Figure S3), corroborating the confounding effect of library size.

We applied the LDM-F and PERMANOVA-F (using the Jaccard and Unweighted UniFrac distances) with various numbers of rarefaction. The global p-values of all tests as shown in Figure 8 (top left) have achieved the minimum possible value (i.e., 1/5001 where we chose 5000 to be the maximum number of permutations for the global test) regardless of the number of rarefaction. These p-values indicate very strong association of 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 (Figure S2, bottom left). The number of detected OTUs by LDM-F (at FDR 10%) increased dramatically as the number of rarefaction increased from 1 to 5 and largely stabilized after 5 (Figure 8, middle left), further

confirming 5 to be an optimal number. As a comparison, we also applied LDM-UR (without rarefaction), which detected 476 OTUs. The Venn diagram in Figure 8 (bottom left) showed that the sets of OTUs detected by LDM-F with 5, 20, and 100 rarefactions overlap considerably, while the set of OTUs by LDM-UR include a large number (91) that do not overlap with any other set, suggesting excessive false positive findings due to confounding by library size.

As the global p-values based on the full sample showed an uninteresting, monotonic pattern over different numbers of rarefaction, we also analyzed a random subsample of 60 subjects (38 cases and 22 controls) and displayed the results in the right panel of Figures 7–8 and S2–S3. We observed that the global p-values again stabilized after 5 rarefactions (and they look discrete because of the total 5000 permutations). Other results follow similar patterns as those from analyzing the full sample. Notably, the set of OTUs detected by LDM-F with 5 rarefactions is a big subset of those with 20 rarefactions, which is further a big subset of those with 100 rarefactions (Figure 8, bottom right); thus the list of OTUs detected with merely 5 rarefactions is trustable.

Acknowledgments

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

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

1. Hu YJ, Satten GA. Testing hypotheses about the microbiome using the linear decomposition model (LDM). Bioinformatics. 2020;bbtaa260:https://doi.org/10.1093/bioinformatics/btaa260.

2. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nature methods. 2013;10(12):1200–1202.

3. 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. Microbial ecology in health and disease. 2015;26(1):27663.

4. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. Frontiers in microbiology. 2017;8:2114.

5. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014;2(1):15.

6. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral ecology. 2001;26(1):32–46.

7. McArdle BH, Anderson MJ. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology. 2001;82(1):290–297.

8. 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.

9. Legendre P, Legendre LF. Numerical ecology. vol. 24. Elsevier; 2012.

10. Relman DA. The human microbiome: ecosystem resilience and health. Nutrition reviews. 2012;70(suppl_1):S2–S9.

11. Cox MJ, Cookson WO, Moffatt MF. Sequencing the human microbiome in health and disease. Human molecular genetics. 2013;22(R1):R88–R94.

12. Shade A, Peter H, Allison SD, Baho D, Berga M, Bürgmann H, et al. Fundamentals of microbial community resistance and resilience. Frontiers in microbiology. 2012;3:417.

PMCID: PMC3525951.

13. Witkin S, Linhares I. Why do lactobacilli dominate the human vaginal microbiota? BJOG: An International Journal of Obstetrics & Gynaecology. 2017;124(4):606–611.

14. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology. 2005;71(12):8228–8235.

15. 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.

16. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550.

17. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–140. PMCID: PMC2796818.

18. Hughes JB, Hellmann JJ. The application of rarefaction techniques to molecular inventories of microbial diversity. Methods in enzymology. 2005;397:292–308.

19. 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.

20. 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.

21. 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.

22. Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, et al. The

Treatment-Naive Microbiome in New-Onset Crohn's Disease. Cell Host and Microbe. 2014;15(3):382–392.

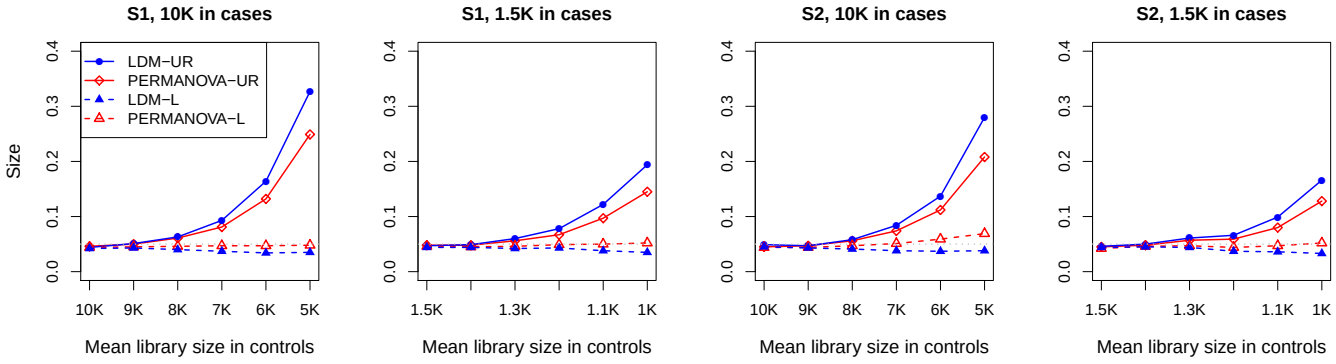

Figure 1. Size of the global tests without rarefaction at varying degrees of differential library sizes. The gray dotted line represents the nominal significance level of 0.05.

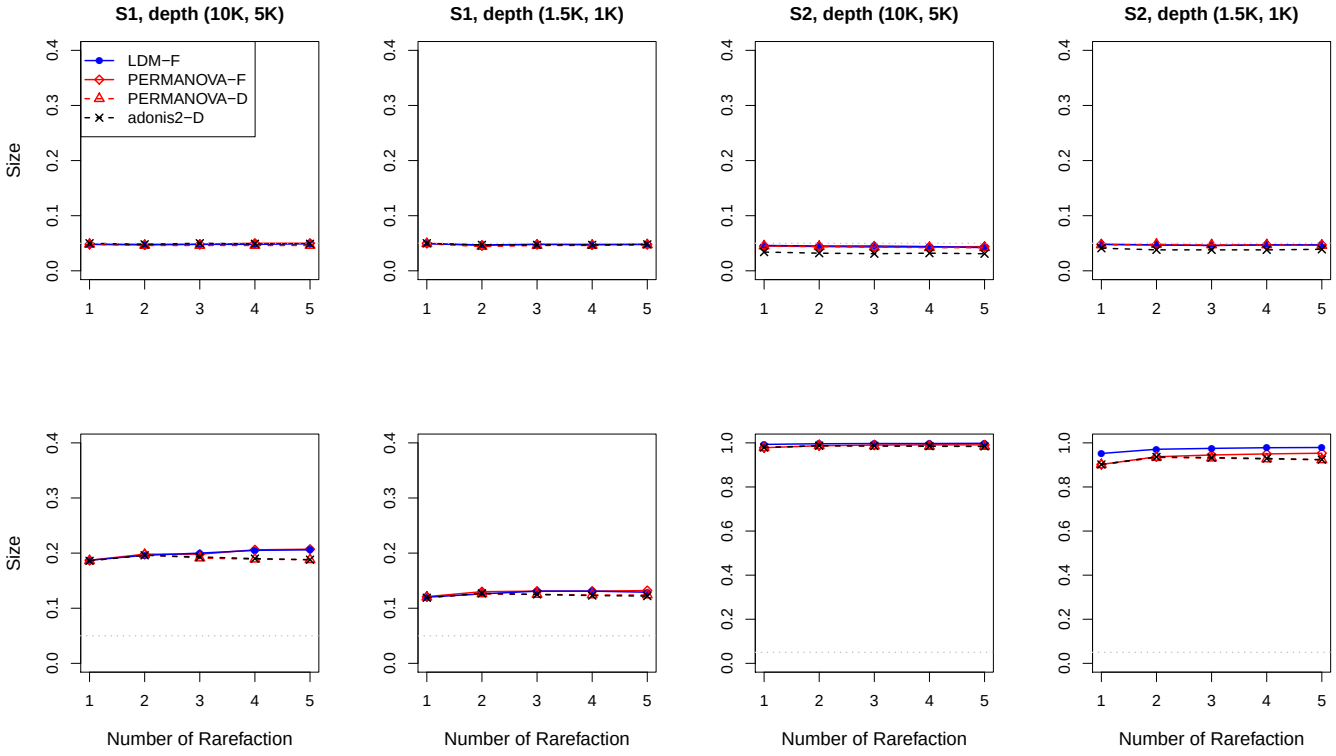

Figure 2. Size of the global tests with different numbers of rarefactions. The top panel pertains to results after adjustment of the confounder. The lower panel pertains to results without adjustment of the confounder.

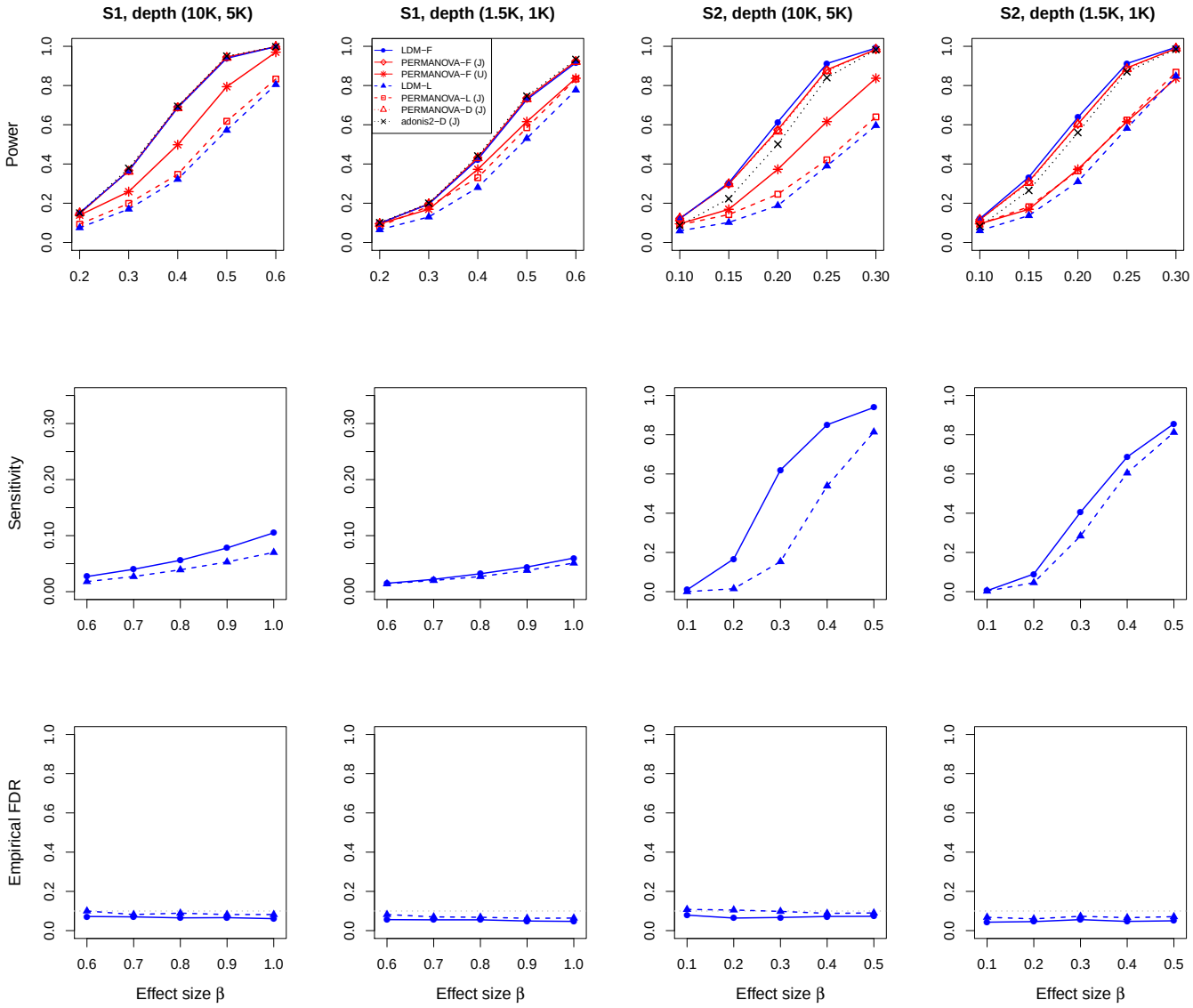

Figure 3. Power, sensitivity, and empirical FDR of tests based on 5 rarefactions. J:Jaccard. U:unweighted UniFrac.

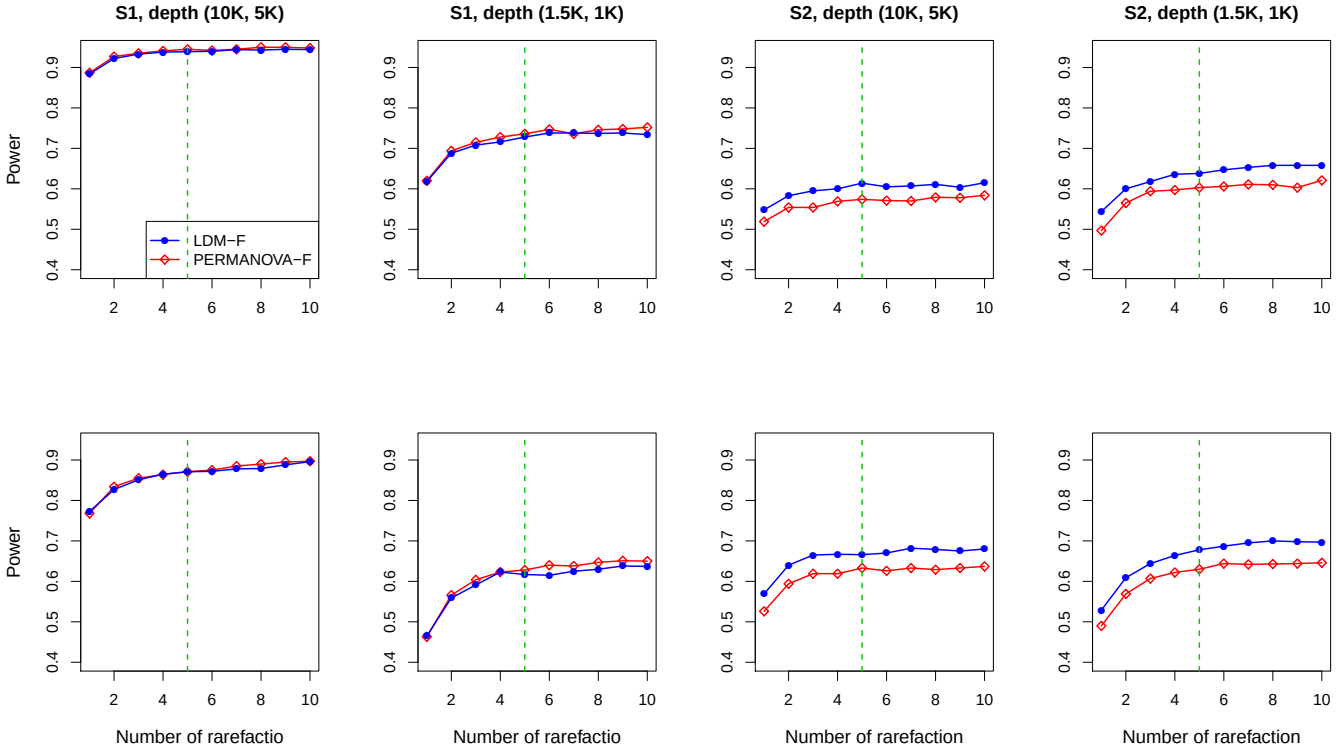

Figure 4. Power of the global tests with increasing number of rarefactions. The results in the upper and lower panels pertain to high and low rarefaction depth, respectively. The effect size β is 0.5 under S1 and 0.2 under S2. The green vertical dashed line represents 5 rarefactions.

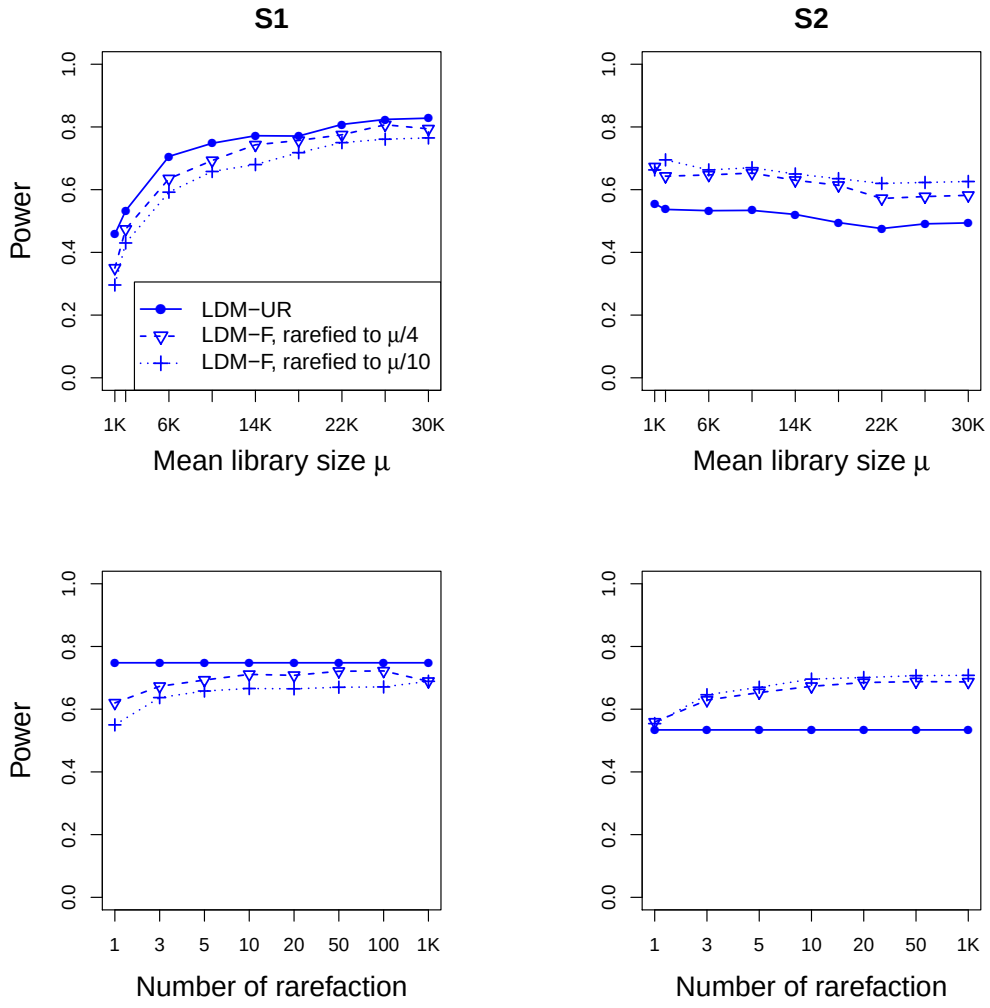

Figure 5. Power of the global tests with and without rarefaction in the absence of differential library sizes. The effect size β is 0.4 under S1 and 0.2 under S2. Results in the upper panel were based on 5 rarefactions. Results in the lower panel were based on mean library size 10K.

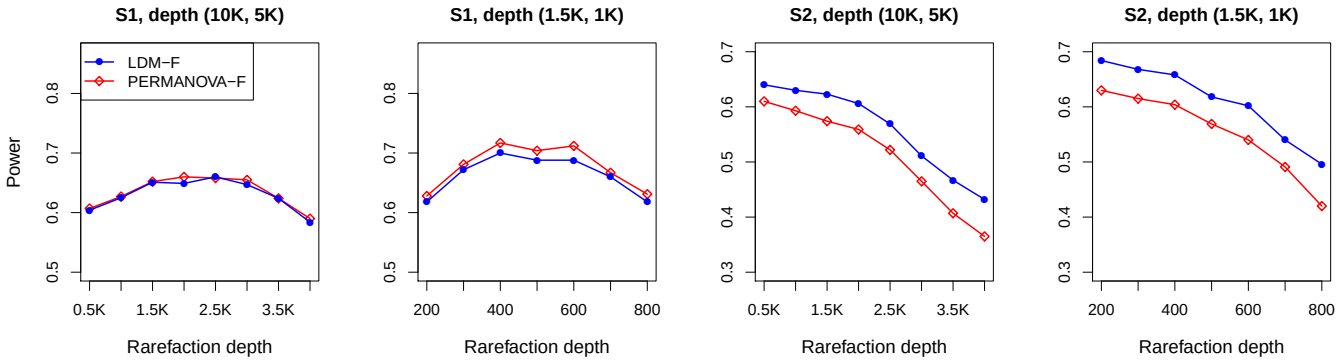

Figure 6. Power of the global tests at varying rarefaction depths. The effect sizes β were 0.4, 0.5, 0.2, 0.2 for the four plots (from left to right). All results were based on 5 rarefactions.

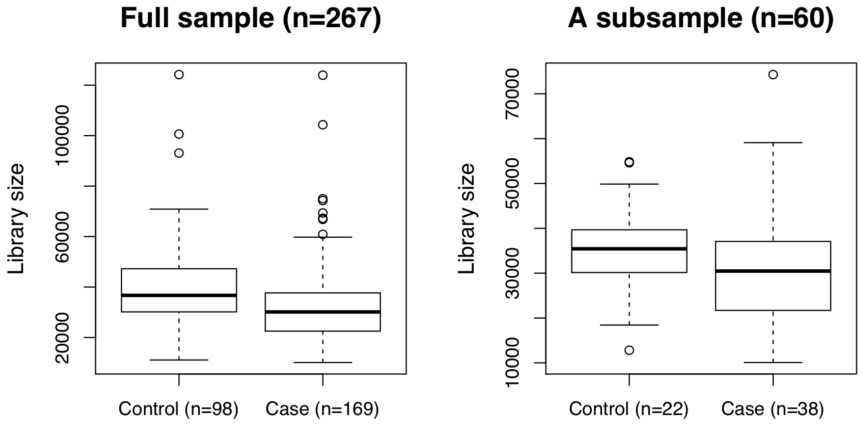

Figure 7. Differential library size of cases and controls in the IBD data.

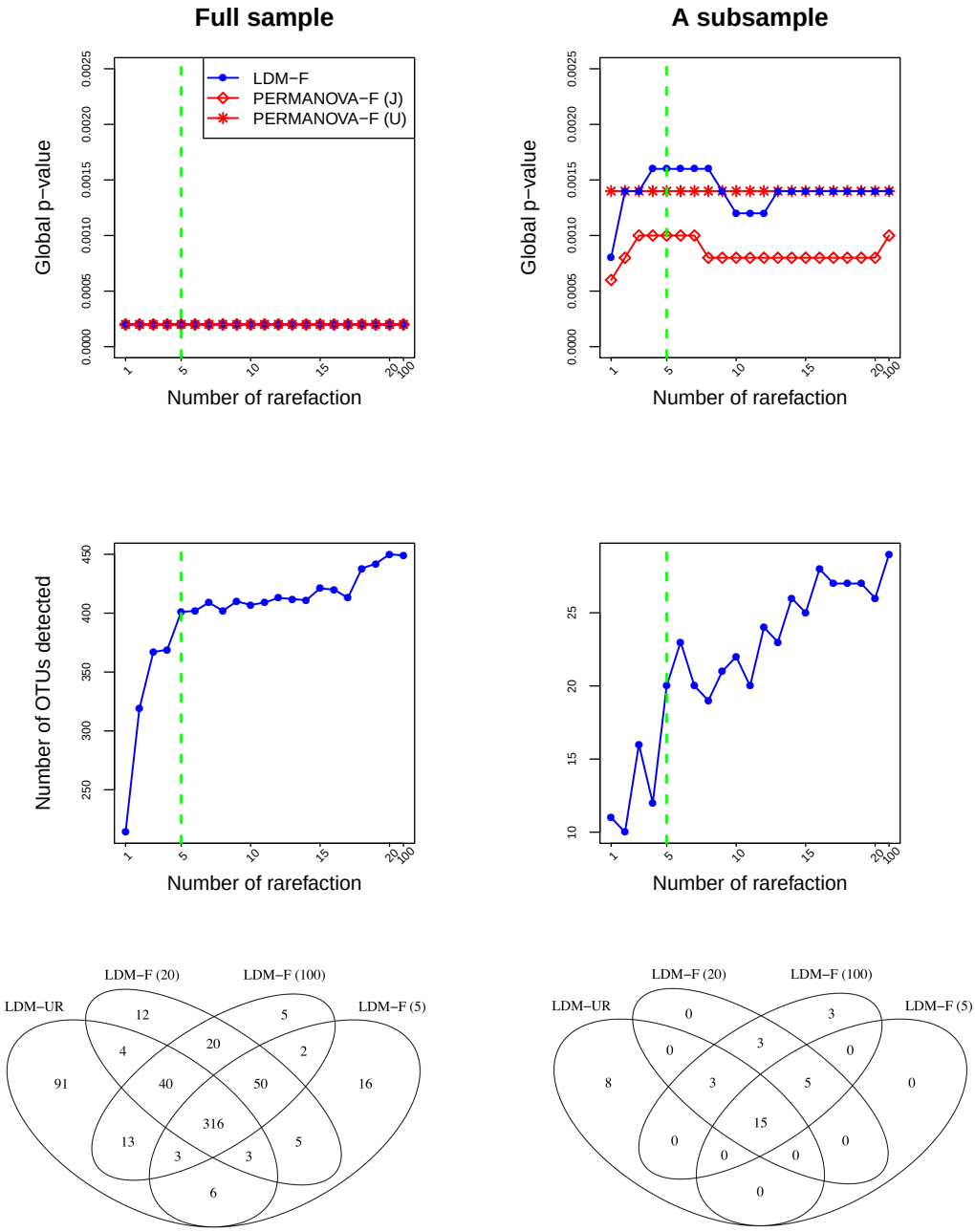

Figure 8. Results in analysis of the IBD data. (Upper panel) P-values for the global test of association between presence-absence of species and case-control status. (Middle panel) Number of OTUs detected as associated at nominal FDR 10%. (Lower panel) Venn diagram for the numbers of detected OTUs. The number is the parenthesis after LDM-F is the number of rarefaction.