Impact of experimental bias on compositional analysis of microbiome data

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

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

[2]Department of Gynecology and Obstetrics, Emory University School of Medicine, Atlanta, GA, 30322, USA.

[*]Corresponding author email: yijuan.hu@emory.edu

Keywords: compositional effect, false-discovery rate (FDR), LOCOM, interaction bias, main effect bias, taxon ratios, test differential abundance

Abstract

Microbiome data are subject to experimental bias that is caused by DNA extraction, PCR amplification among other sources, but this important feature is often ignored when developing statistical methods for analyzing microbiome data. McLaren, Willis and Callahan (2019) proposed a model for how such bias affects the observed taxonomic profiles, which assumes main effects of bias without taxon-taxon interactions. Our newly developed method, LOCOM (logistic regression for compositional analysis) for testing differential abundance of taxa, is the first method that accounted for experimental bias and is robust to the main effect biases. However, there is also evidence for taxon-taxon interactions. In this report, we formulated a model for interaction biases and use simulations based on this model to evaluate the impact of interaction biases on the performance of LOCOM as well as other available compositional analysis methods. Our simulation results indicated that LOCOM remained robust to a reasonable range of interaction biases. The other methods tended to have inflated FDR even when there were only main effect biases. LOCOM maintained the highest sensitivity even when the other methods cannot control the FDR. We thus conclude that LOCOM outperforms the other methods for compositional analysis of microbiome data considered here.

Introduction

Experimental bias introduced during DNA extraction and PCR amplification, among other steps in the experimental pipeline, is a pervasive feature of microbiome data. Absent experimental procedures that produce unbiased data, it is therefore necessary to account for experimental bias when analyzing microbiome data. Fortunately, McLaren, Willis and Callahan (MWC) [1] have proposed a model to explain how experimental bias affects microbiome data in which the observed relative abundance of each taxon is a product of the taxon's true relative abundance and a taxon-specific bias factor, normalized over all taxa observed in the sample. Each taxon-specific bias factor represents the accumulation of multiplicative biases over all the steps in the experimental pipeline, so that multiple sources of bias are described by a single factor. With the MWC model in mind, we developed LOCOM [2], a logistic-regression-based compositional analysis method for detecting differentially abundant taxa, that only estimates parameters that are free from bias, i.e. are not affected by bias factors. To the best of our knowledge, LOCOM was the first method that accounted for experimental bias and was shown both analytically and numerically to be fully robust to any bias that follows the MWC model; none of the other existing methods for testing taxon differential abundance have considered experimental bias.

However, the MWC model assumes no between-taxon "interaction" bias, i.e., the presence or abundance of one taxon does not affect the bias factors of any other taxa in the sample. LOCOM may not be bias-robust if the "main effect" biases of the MWC model were supplemented by interaction biases. In evaluating the MWC model, Zhao and Satten [3] in fact found a small amount of evidence for interaction biases, although with smaller magnitude than the main effect biases in the MWC model; this finding is the motivation for this study. Here, we formulate a model for interaction biases and use simulations based on this model to evaluate the impact of interaction biases on the performance of LOCOM, as well as a number of existing compositional analysis approaches (ANCOM [4], ANCOM-BC [5], fastANCOM [6],

ALDEx2 [7], WRENCH [8], DACOMP (v1.26) [9], LinDA [10], and the Wilcoxon rank-sum test of log-ratio transformed data after adding pseudocounts of either 0.5 or 1). Thus, this report provides the first assessment of the impact of interaction biases on analysis of microbiome data.

Methods

We generalize the MWC model using the framework of Zhao and Satten [3] by adding an interaction bias $\theta_{jj'}$ to the log-linear model that relates the expected value of the observed relative abundances and true relative abundances:

$$\log(p_{ij}) = \log(\pi_{ij}) + \gamma_j + \sum_{j'} \theta_{jj'}\pi_{ij'} + \alpha_i \ , \tag{1}$$

where π_{ij} is the true relative abundance of taxon j in sample i, p_{ij} is the expected value of the observed relative abundance, γ_j is the main effect bias for taxon j in the MWC model, and α_i is the sample-specific normalization factor that ensures the compositional constraint on p_{ij}. We followed [3] to introduce the effects of covariates X_i on taxon j by replacing $\log(\pi_{ij})$ in (1) by $\log(\pi_j^0) + X_i^{\mathrm{T}}\beta_j$, where β_j contains the effect sizes and π_j^0 is the true relative abundance of taxon j when $X_i = 0$. The interaction $\theta_{jj'}$ determines the extent to which taxon j' affects the bias factor for taxon j. The assumption underlying the MWC model is $\theta_{jj'} = 0$ for all j and j'. To generate interactive biases in our simulation studies, we additionally express $\theta_{jj'}$ relative to γ_j as

$$\theta_{jj'} = -\phi\epsilon_{jj'}\gamma_j \ , \tag{2}$$

where ϕ is a constant for all taxa pairs that we call the magnitude of interaction biases, and $\epsilon_{jj'}$ is a non-negative error term having mean one. The value 0.612 for ϕ and the negative sign in (2) reflect the findings of Zhao and Satten [3], who considered a model like (1) but with $\pi_{ij'}$ replaced by $I[\pi_{ij'} > 0]$ in the interaction term; more details are found in Supporting Information.

Our simulations for evaluating the effects of interaction biases on compositional analysis methods were based on models (1) and (2) and data on 856 taxa of the upper-respiratory-tract microbiome by Charlson et al. [11]. We considered both binary and continuous traits of interest without any confounding covariates; we also considered a binary confounder when the trait was binary. We used the two sets of causal taxa (i.e., taxa that are associated with the trait) that were used in [2], namely, a random sample of 20 taxa (referred to as M1) and the five most abundant taxa (M2); we assumed a common effect size β of the trait on all causal taxa. We designed three scenarios for the distribution of interaction biases: sampling the error term $\epsilon_{jj'}$ non-differentially based on $N(0.5, 0.1^2)$ for all taxa (referred to as S-nondiff), differentially based on $N(0.5, 0.1^2)$ and $Beta(0.5, 0.5)$ (a bimodal distribution with larger variance than 0.1^2) for null and causal taxa, respectively (S-diff-causal), and differentially for a random selection of half of the taxa and the remaining taxa (S-diff-half). Unlike S-nondiff, S-diff-causal has modest (but trait-related) variation in $\epsilon_{jj'}$ while S-diff-half has large variation in $\epsilon_{jj'}$ that is not trait-related. Additional details of the simulation settings are found in Supporting Information, but generally follow the simulations in [2].

Results

The empirical FDR and sensitivity of all aforementioned compositional analysis methods for detecting the causal taxa, for simulations with a binary trait and no confounder under scenarios M1 and M2, are displayed in Figures 1 and 2, respectively (results for simulations with confounders and continuous traits showed similar patterns and are therefore omitted here). Note that $\phi = 0$ corresponds to no interaction bias at any taxa (i.e., main effect biases only) and $\beta = 0$ corresponds to no differential abundance at any taxa (i.e., the global null). In all cases, the FDR of LOCOM remained at or close to the nominal level as long as the magnitude of interaction bias $\phi < 1$, which is substantially larger than the empirical value 0.612 observed in the Brooks data [3]; under the global null in particular, LOCOM always controlled the FDR regardless of the value of ϕ. It was only when both ϕ and β became

unrealistically large that we observed moderate inflation in the FDR of LOCOM. The FDR inflation of LOCOM was similar in S-nondiff and S-diff-causal, because the interaction biases were similarly distributed at the majority of taxa which were null taxa; the inflation was (slightly) the largest in S-diff-half, because the interaction biases had the largest variability among the three scenarios at a given ϕ. The interaction biases caused some loss of sensitivity for LOCOM, but the drop was relatively small and LOCOM maintained the highest sensitivity among all methods in all of our simulation scenarios. The results of the other methods showed a similar trend to LOCOM with FDR increasing and sensitivity decreasing as ϕ increased, and increasing FDR inflation as β increased.

The other methods had very different performance even when there were only main effect biases ($\phi = 0$). ANCOM-BC and fastANCOM performed the best among the other methods, controlling FDR for the full range of β values we have considered when ϕ was small and having similar FDR inflation as LOCOM when both β and ϕ became large; however, this performance came at the cost of a substantially lower sensitivity than LOCOM. ANCOM had moderately inflated FDR when β was small but performed better when β was increased; nevertheless, its sensitivity was among the lowest. ALDEx2 had inflated FDR and poor sensitivity when β was large, in M1 but not in M2. WRENCH, Wilcox-alr-half, and Wilcox-alr-one had highly inflated FDR except in the global null case and LinDA had highly inflated FDR in all cases including the global null. These results were based on data after applying the filter of rare taxa as recommended by LOCOM; in general, ANCOM, ANCOM-BC, fastANCOM and LinDA had worse FDR control when a less stringent filter was applied. Note that, unlike the LOCOM paper that used the older version (v1.1) of DACOMP, we used the latest version (v1.26) here, which yielded highly inflated FDR and low sensitivity in M2 (because some reference taxa were incorrectly selected to be causal taxa).

Discussion

In this study, we found that LOCOM was robust to not only main effect biases but also

a reasonable range of interaction biases. The other methods tended to have inflated FDR even when there were only main effect biases; many of them did not control the FDR even when there was no experimental bias at all (results shown in [2]). LOCOM maintained the highest sensitivity among all methods even when the other methods did not control the FDR. Therefore, we conclude that LOCOM outperforms most (if not all) existing methods for compositional analysis of microbiome data.

The robust performance of all methods to the interaction biases is likely owing to three reasons. First, each term $\theta_{jj'}\pi_{ij'}$ in (1) that governs the effect of taxon j' on the bias factor of taxon j is small because the relative abundance $\pi_{ij'}$ is generally small. Second, the contributions from different taxa to the total interaction bias $\sum_{j'} \theta_{jj'}\pi_{ij'}$ have different directions and tend to cancel out in the sum. Third, any non-zero mean interaction bias across taxa (or individuals) would be automatically included in the normalization factor α_i (or the taxon bias γ_j); this "automatic centering" may also decrease the apparent effect of the interaction bias.

We used the Brooks mock community data to motivate our simulation studies, which may have two limitations. First, the Brooks samples contain at most seven taxa, all having equal true relative abundance; other mock community datasets, especially datasets that contain a large number of taxa and that mimic real microbiome data, should be used to study the interaction bias. Second, bacteria in a real microbial community may have different interactions from bacteria in a mock community. Therefore, study of experimental biases and their impact on downstream analysis continues to be an important topic in the foreseeable future.

Acknowledgments

This research was supported by the National Institutes of Health awards R01GM116065 (Hu, Satten) and R01GM141074 (Hu, Satten).

References

1. McLaren MR, Willis AD, Callahan BJ. Consistent and correctable bias in metagenomic sequencing experiments. Elife. 2019;8:e46923.
2. Hu Y, Satten GA, Hu YJ. LOCOM: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control. Proceedings of the National Academy of Sciences. 2022;119(30):e2122788119.
3. Zhao N, Satten GA. A log-linear model for inference on bias in microbiome studies. In: Statistical Analysis of Microbiome Data. Springer; 2021. p. 221–246.
4. 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.
5. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. Nature communications. 2020;11(3514):https://doi.org/10.1038/s41467–020–17041–7.
6. Zhou C, Wang H, Zhao H, Wang T. fastANCOM: a fast method for analysis of compositions of microbiomes. Bioinformatics. 2022;38(7):2039–2041.
7. 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.
8. Kumar MS, Slud EV, Okrah K, Hicks SC, Hannenhalli S, Bravo HC. Analysis and correction of compositional bias in sparse sequencing count data. BMC genomics. 2018;19(1):799.
9. Brill B, Amir A, Heller R. Testing for differential abundance in compositional counts data, with application to microbiome studies. The Annals of Applied Statistics. 2022;16(4):2648–2671.

10. Zhou H, He K, Chen J, Zhang X. LinDA: linear models for differential abundance analysis of microbiome compositional data. Genome biology. 2022;23(1):1–23.

11. 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.

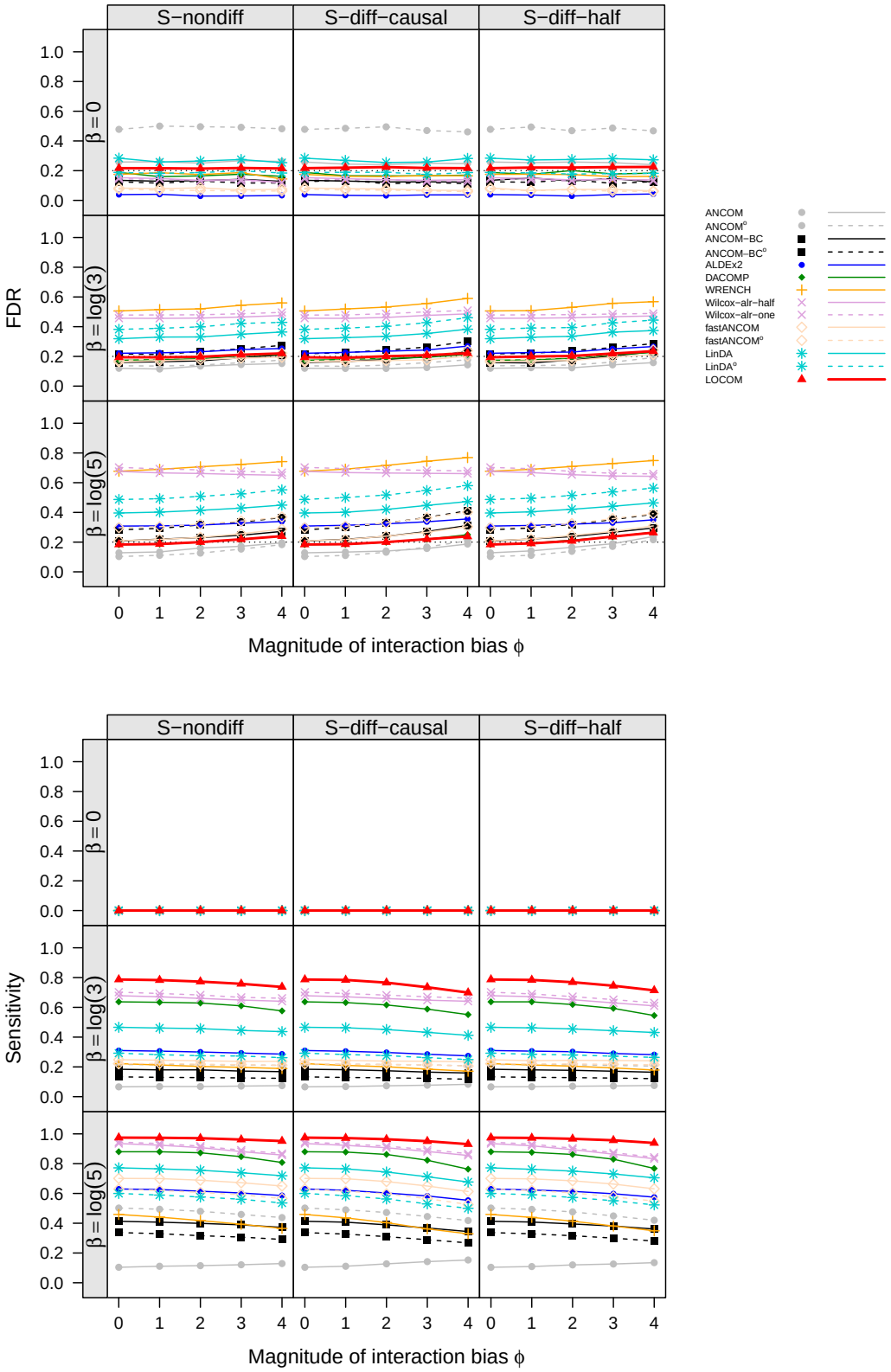

Figure 1: FDR and sensitivity at the nominal FDR 20% (black dotted line) of LOCOM and other compositional methods for data simulated under M1 (randomly selecting 20 taxa to be causal with effect size β) and had a binary trait. With superscript ^{o}, the filter for rare taxa that are present in less than 10% of samples as adopted by the original programs was used; without ^{o}, the more stringent filter with 20% cutoff as recommended by LOCOM was used. Details of Wilcox-alr-half and Wilcox-alr-one are found in [2]. All results were based on 1000 replicates of data.

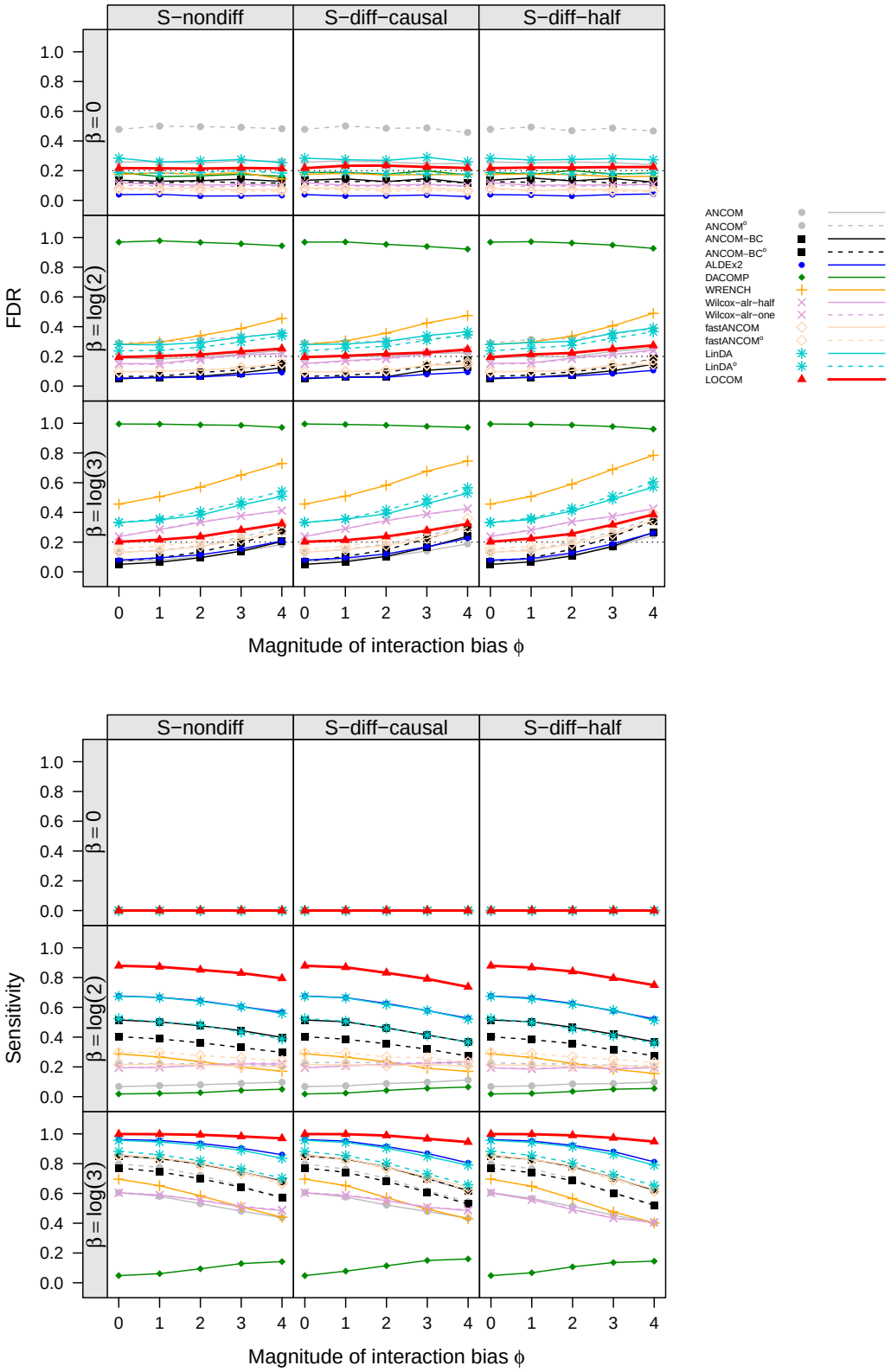

Figure 2: Same as the caption to Figure 1 except that the data were simulated under M2 (choosing the top five most abundant taxa to be causal with effect size β).