Taxonomic bias can create spurious results in microbiome differential-abundance analyses unless it is properly accounted for

Michael R. McLaren* Karen G. Lloyd† Benjamin J. Callahan‡

2021-08-04

Contents

*North Carolina State University; send correspondence to m.mclaren42@gmail.com
†University of Tennessee
‡North Carolina State University

Preface

This in-progress manuscript is not intended for general scientific use. It is incomplete, has not been carefully reviewed, and may contain mistakes or other inaccuracies. Please post comments or questions on the GitHub Issues page or email Mike.

This manuscript addresses the effect that the taxonomic bias inherent in microbiome measurement has on microbial differential-abundance analysis. We describe the basic problem posed by taxonomic bias for measuring changes in the abundance of particular taxa across conditions and describe new strategies for mitigating the errors it induces. Analyses of both relative and absolute abundances are considered. In its current form, the manuscript sits somewhere between a standard scientific article and a monograph; it consists of an article followed by a series of appendices which together give a comprehensive treatment of the implications of the McLaren, Willis, and Callahan (2019) model of taxonomic bias for differential-abundance analysis and experimental design. See the Zenodo record for how to cite the latest version.

1 Introduction

Most microbiome research aims to go beyond qualitative descriptions to make quantitative conclusions about associations between specific microbes or community states and key host or environmental properties. But there is serious disagreement about whether the microbiome measurements made by marker-gene and metagenomic sequencing (MGS) are or can ever be truly quantitative. A primary reason is that taxa vary in how they respond to each step in an MGS protocol, from sample collection to bioinformatic classification. As a result taxa can differ dramatically (e.g. 10-1000X) in how efficiently they are measured—that is, converted from cells into taxonomically classified sequencing reads—making the relative abundances we observe by MGS inaccurate representations of the actual sample compositions. Although often associated with variation in primer binding and amplification rates and marker-gene copy-number, large variation in DNA extraction efficiency and in the ability to correctly classify reads among taxa make this taxonomic bias a feature of shotgun metagenomic measurements as much as marker-gene measurements. Taxonomic bias is protocol specific (McLaren, Willis, and Callahan (2019)) and can even vary among batches within an experiment (Yeh et al. (2018)); it thus not only makes MGS measurements inaccurate, but also incomparable between studies. Taxonomic bias thereby poses a major challenge for any seeking to draw quantitative conclusions from MGS studies.

The field's current answer to this challenge is methods standardization, prioritizing consistency

over accuracy. By treating each sample within a study in exactly the same manner, we can ensure that observed differences between samples are not simply due to methodological differences. Yet standardization does not guarantee that these observed differences reflect the actual differences in taxonomic composition among samples. Consider what might be the simplest and most common inference made from MGS: How does the proportion of a taxon vary between samples? An observation that *Escherichia* coli makes up 10% of the sequencing reads in a fecal sample might, thanks to bias, have arisen from a sample in which the proportion of *E. coli* is actually 50% or 1%. Yet many hope that if we observe the proportion of *E. coli* to increase from 10% to 20% between two samples, its proportion must have actually doubled—by fixing the taxonomic bias of our measurements through standardization, we have (so the argument goes) fixed the measurement error in both samples so that it will cancel when estimate the fold change. Unfortunately, the results of McLaren, Willis, and Callahan (2019) show that reality is not so neat. Because the taxa compete for sequencing effort within a sample, the effect that taxonomic bias has on the observed proportions depends on the taxonomic composition of the sample. As a result, MGS measurements can lead to spurious inferences about changes across samples even when taxonomic bias is constant for each taxon across samples.

This problem posed by taxonomic bias intersects with a second, growing concern within the microbiome field: that the relative abundances measured by MGS can present a misleading view of absolute-abundance dynamics. Without making further assumptions (which are often difficult or impossible to verify) one cannot tell whether a doubling in the proportion of a taxon reflects an increase in its abundance or a decrease in the abundances of other taxa. This concern has motivated the development of various experimental methods to enable researchers to convert the relative abundances measured by MGS into absolute abundances (methods to be reviewed in the Appendix in a future version). An increasingly popular approach in the study of both environmental and host-associated microbiomes involves making an independent measurement of total microbial abundance in each sample (Props et al. (2017),Kevorkian et al. (2018),Lloyd et al. (2020),Tettamanti Boshier et al. (2020),Contijoch et al. (2019),Vandeputte et al. (2017),Vieira-Silva et al. (2019)). This estimate is then multiplied by the MGS proportions to convert them to estimates of absolute abundance. This simple calibration procedure directly transfers the error in the MGS proportions due to taxonomic bias to the estimated absolute abundances, leading to estimates that are not only individually inaccurate but can imply inaccurate variation in abundance across samples.

Here we summarize recent advances in our understanding of how taxonomic bias affects our ability to accurately determine how taxa vary in relative and absolute abundance across samples from different conditions—the inference problem known as *differential abundance (DA) analysis*. We first explain how taxonomic bias can lead to spurious DA results and consider the biological conditions in which such spurious results are likely. But we then describe several experimental and computational methods that can be used to avoid or correct these errors. Collectively, these methods may provide practical solutions to mitigating the effects of taxonomic bias on DA analysis for a majority of microbiome studies.

2 Bias can cause spurious results in proportion-based DA analysis

```
#> Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please
#> use `guide = "none"` instead.

#> Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please
#> use `guide = "none"` instead.
```

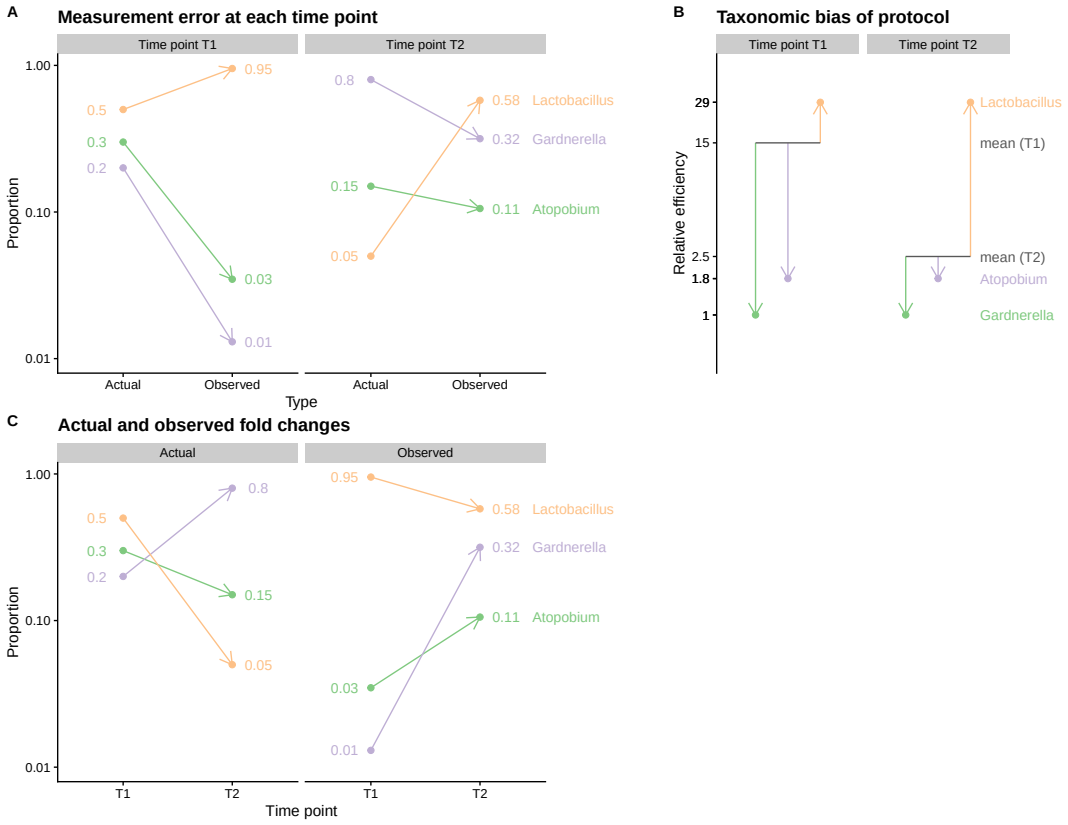

Figure 1: **Taxonomic bias can distort differential abundance results even when it is consistent for each taxon across samples.** Panel A shows the actual and observed proportions for hypothetical community samples from two time points, which differ in their relative abundance of three taxa. Panel B shows taxonomic bias in terms of the relative efficiencies of the three taxa against the mean efficiency of each sample; the difference between the taxon's efficiency and the sample's mean (vertical arrows) determines the fold error seen in Panel A. Panel C rearranges the plot from Panel A to show the actual and observed fold changes between time points. The efficiencies of individual taxa were estimated by McLaren, Willis, and Callahan (2019) from mock community data from Brooks et al. (2015). The abundances are hypothetical but inspired by observations from the human vaginal microbiome; see main text.

We illustrate the problems posed by taxonomic bias for differential abundance using an example where the bias of an MGS protocol was explicitly measured using mock communities. To analyze the effects of taxonomic bias on measurements of the human vaginal microbiome, Brooks et al. (2015) performed 16S rRNA gene sequencing of mock communities of seven key vaginal species using a similar protocol to that of the Vaginal Human Microbiome Project (VaHMP). McLaren, Willis, and Callahan (2019) used this data to estimate the relative measurement efficiency of the seven species. They found that the most efficiently measured species, *Lactobacillus iners*, had an efficiency that was 29X that of the least efficiently measured species, *Gardnerella vaginalis*, primarily due to a greater DNA extraction efficiency and 16S copy number. A third species, *Atopobium vaginae*, had an efficiency 1.8X that of *G. vaginalis*. A common aim of human vaginal microbiome studies, including the VaHMP, is to understand the causes and consequences of shifts in communities from being dominated by a particular *Lactobacillus* species to dominance by a non-*Lactobacillus* species such as *G. vaginalis*, which have been associated with deleterious health outcomes including an increased risk of preterm births in pregnant women. It is therefore important to know whether the efficiency variation found by McLaren, Willis, and Callahan (2019) affects the observed differential abundances between *Lactobacillus*-dominated and *Gardnerella*-dominated communities.

Figure 1 illustrates the effects of taxonomic bias on the measured proportions for hypothetical vaginal samples from a single woman at two points in time, during which the community shifts from being dominated by *Lactobacillus* to being dominated by *Gardnerella*. In each case, we observe a greater proportion of *Lactobacillus* than is truly present, and less of the other two species (Figure 1A). This positive multiplicative error arises because *Lactobacillus* is more efficiently measured than the average species in each sample. McLaren, Willis, and Callahan (2019) showed that the observed and actual proportions of a taxon i in a sample s are related by the equation

$$\text{observed}_i(s) = \text{actual}_i(s) \cdot \frac{\text{efficiency}_i}{\text{mean efficiency}(s)} \tag{1}$$

where $\text{mean efficiency}(s) = \text{actual}_j(s) * \text{efficiency}_j$ in the denominator is the average efficiency of all taxa in that sample. Intuitively, Equation (1) says that the taxon will be overrepresented by the extent to which its measurement efficiency is greater than the average efficiency of cells in the sample. Figure 1B shows the efficiencies of the three taxa and the mean efficiency in each sample. The (multiplicative) difference between the taxon efficiencies and the mean efficiency determines the (multiplicative) error in each sample; for example, the mean efficiency in the second sample is 2.5, leading to a 29/2.5=11.6X error in the proportion of *Lactobacillus*. The error in the second sample is larger—specifically, by 6X—than the first sample because the high proportion of *Gardnerella* has reduced the sample mean efficiency (and thus increased the efficiency of *Lactobacillus* relative to the mean) by 6X.

The enrichment depends on the taxon's efficiency relative to the sample average, rather than the efficiency itself, because the requirement that the proportions of all taxa sum to one locks the taxa in a zero-sum competition. Notably, competition occurs (though in a less absolute form) even at the level of the raw sequencing reads: DNA extractions and PCR reactions have limited output capacity; high-concentration samples are typically subsampled to fixed, lower concentrations; and sequencing libraries are given limited sequencing effort; so that the experimental process itself typically imposes competition to be sequenced. And it is competition that prevents the effects of taxonomic bias from canceling when we analyze how the proportions of individual taxa change across samples.

Figure 1C shows the consequences of measurement error on an analysis of the fold change in proportions between the two samples. In reality, *Gardnerella* changes by 4X, *Lactobacillus* by 0.1X, and *Atopobium* by 0.5X. Yet what is observed is that *Gardnerella* changes by 32X, *Lactobacillus* by 0.6X, and *Atopobium* by 3X. In other words, the change in each case is

multiplied by 6X, causing a much larger increase in *Gardnerella* and a smaller decrease *Lactobacillus* than actual, and the appearance that *Atopobium* increased when it in fact decreased. Why do such errors occur? The multiplicative error in Equation (1) has two parts: a taxon-specific term in the numerator that is the same in each sample, and a taxon-independent term in the denominator that depends on the composition of the sample through the mean efficiency. When we compute the fold change in a taxon i between samples s and t,

$$\frac{\text{observed}_i(t)}{\text{observed}_i(s)} = \frac{\text{actual}_i(t)}{\text{actual}_i(s)} * \frac{\text{efficiency}_i}{\text{efficiency}_i} * \frac{\text{mean efficiency}(s)}{\text{mean efficiency}(t)} \tag{2}$$

the constant efficiency in the numerator of Equation (1) cancels, but the varying mean efficiency does not, leaving an error that is equal to the inverse change in mean efficiency. As a result, we are here left inferring that the fold-changes in proportions have all increased less (or decreased more) than they truly have. In our example, the mean efficiency decreased by 6X (from 15 to 2.5) due to the shift in dominant taxon from the high-efficiency *Lactobacillus* to the low-efficiency *Gardnerella*, causing the observed fold changes to increase by 6X above the actual. This increase distorted the magnitudes of the changes in *Lactobacillus* and *Gardnerella* (decreasing the first and increasing the second) and changed the sign of *Atopobium* from negative to positive. Thus partial cancelling in the effect of a constant multiplicative bias may not be sufficient to protect against spurious results.

Spurious results can occur for the same basic reason in proportion-based absolute-abundance inference. Suppose that the true total abundance in each sample can be determined experimentally. Because the absolute abundance estimate simply multiplies the erroneous observed proportions by this total, the estimated change in absolute abundance is wrong by the same factor as the fold change in proportions. In our Figure 1 example, if the change in total abundance between samples were 2X, then we would observed both *Lactobacillus* and *Atopobium* to have increased in absolute abundance (by 1.2X and 6X), when in fact *Lactobacillus* decreased (by 5X) and *Atopobium* remained the same.

2.1 Effect on regression of multiple samples

We showed how fold changes of individual samples may be unreliable. But microbiome analyses often do not interpret such individual differences, and instead look for average patterns across many samples, such as how the abundance of a taxon varies with some covariate of interest. The covariate may be discrete, such as whether the sample is from a healthy or sick person, or it may be continuous, such as a measure of space, time, or temperature. Often these analyses can be framed as a regression problem. For example, we might hypothesize that the log absolute abundance of taxon i changes with a variable x according to the simple linear regression,

$$E[\log \text{abundance}_i \mid x] = a_0 + a_1 x \tag{3}$$

where x is either continuous (e.g., sediment depth) or binary (e.g., $x = 1$ for treated patients and $x = 0$ for controls). Much of the effect of bias will be absorbed by the model intercept, which is typically not of interest to the researcher. However, variation in the mean efficiency creates additional variation (reducing statistical power to detect variation in the slope) and can also create a statistical bias in our slope estimate equal to the inverse slope of $E[\log \text{mean efficiency} \mid x]$ as a function of x. Therefore if the mean efficiency varies across samples, but is not associated with the covariate x, its effect may simply be to create larger standard errors in our slope estimate. But the larger worry may be the systematically distorted slope (a_1) estimates that arise in situations where the mean efficiency changes with x. Notably, the error in an absolute sense is the same for all taxa—the slope is always reduced by the same amount—but the implications differ for various taxa depending on their slopes, causing magnitude or sign errors depending how the coefficient derived from the (unknown) true abundances compares to that of the mean efficiency (Figure 2).

```
#> Warning: Package `magick` is required to draw images. Image not drawn.
```

3 Implications for real-world inference

These observations show that the effects of taxonomic bias do in fact cause error in proportion-based differential abundance analyses. Yet if the mean efficiency varies much less across samples than the abundances of individual taxa, this error will be negligible. Alternatively, if variation in the mean efficiency is effectively random (not associated with regression covariates), the effect of bias may be to increase the residual variance and thus reduce our statistical power to assay differential abundance—unfortunate to be sure, but perhaps not as much as confidently inferring changes with the wrong sign or magnitude. But there are salient biological scenarios where substantial associations between the mean efficiency and covariate can in fact lead to spurious differential abundance inference.

3.1 Scenarios in which bias may cause spurious results

Figure 1 illustrates how a change in dominance by from a high- to low- efficiency taxon (or vice versa) can easily create spurious DA results. Indeed, this scenario may be particularly common in vaginal microbiome studies. Human microbiomes often dominated by *Lactobacillus*, but can become dominated by non-*Lactobacillus* species such as *Gardnerella* vaginalis and *Atopobium vaginae*, and these low-*Lactobacillus* states are associated with the disease bacterial vaginosis and, in pregnant women, an increased risk of preterm birth. Using mock community measurements from Brooks et al. (2015), McLaren, Willis, and Callahan (2019) showed that the efficiency of two common *Lactobacillus* species were 20-30X greater than *G. vaginalis* and *A. vaginae*. Therefore we should expect samples dominated by *Lactobacillus* to have a much larger mean efficiency than those those dominated by *G. vaginalis* and *A. vaginae*, which would distort the observed associations of all taxa with any covariate that is also associated with *Lactobacillus* dominance. The extraction and 16S sequencing protocol used by Brooks et al. (2015) is similar to that used in the VaHMP MOMs-PI study also led by these researchers to investigate the associations of these taxa with preterm birth (Fettweis et al. (2019)), demonstrating the potential for serious real-world consequences for these statistical/technical concerns.

Similar dynamics occur in a plant-fungal interactions experiment performed by Leopold and Busby (2020) (Figure 2). The authors inoculated commensally-colonized trees with a fungal pathogen. Using DNA mock communities, they showed that (excluding bias from DNA extraction) the pathogen was 10X more efficiently measured than the median commensal and 40X more efficiently measured than the lowest-efficiency commensal. In most hosts, the pathogen increased rapidly between timepoints, driving an increase in the sample mean efficiency that in turn leads to larger decreases being observed in the proportions of commensal taxa than what is predicted using calibrated (bias-corrected) values (REF APPENDIX). For example, the calibrated proportions indicate that the commensal *Penicillium* slightly increased in hosts from the Eastern US and slightly decreased in hosts from the Western US. Yet the observed (uncalibrated) proportions show moderate and large decreases, respectively (Figure 2 above).

These examples involved scenarios where individual taxa can constitute a large proportion of the community. But in samples from highly-diverse ecosystems such as human gut, soil, and ocean sediment, one species rarely dominates. In these cases it might seem intuitive that the mean efficiency should be relatively stable across samples: As diversity increases, the mean efficiency effectively averages over efficiencies from a greater number of taxa and so (by the central limit theorem) will converge to a constant value—if taxon efficiencies are statistically independent of taxon abundances (REF APPENDIX). In practice, associations between efficiencies and abundances may be common and can allow large shifts in mean efficiency to occur even in highly diverse samples.

Figure 2: **Taxonomic bias distorts multi-sample differential abundance inference when the mean efficiency of samples is associated with the covariate of interest.** Details to synthesize into a caption: Regression of log2(Proportion) of the commensal fungus Penicillium versus timepoint; timepoints 1 and 2 are pre- and post-challenge with the pathogen Melampsora. Data is split by the region the host plants are derived from (Eastern and Western US). Calibrated proportions = Observed proportions in the real, experimental samples after adjustment for the bias measured in mock communities. Mean efficiency of each community is inferred by treating the calibrated proportions as the truth, and multiplying by efficiencies estimated from the mocks. The pathogen Melampsora has a high measurement efficiency; thus once it infects the plants, the mean efficiency of the sample increases (purple points). Efficiency is here taken as relative to the focal taxon Penicillium. West plants tend to be more resistant to the pathogen, which likely explains why the mean efficiency doesn't increase as much in the West plants. Penicillium is observed to decrease in log proportion, in both the East and West plants. But the calibrated measurements show that it actually slightly increases in the East plants, and has a lesser decrease in the West plants than what was observed before bias correction. The difference between the Calibrated and Observed data points and regression lines equals the regression line of the mean efficiency: Orange = Green + Purple; Green = Orange - Purple. The absolute error in regression coefficients is the same for all taxa. I picked Penicillium for illustration since it has the smallest observed decrease, which makes the error due to bias have a particularly significant impact.

The human gut provides a clear example of how such associations might arise. Gut microbiomes typically have high diversity at the species level but are dominated by just a small number of phyla, and two in particular: The Bacteroidetes and the Firmicutes. The ratio of Bacteroidetes to Firmicutes can shift substantially between individuals and has been linked to a number of host traits and health conditions. In addition, many DNA extraction protocols more efficiently lyse Gram-negative Bacteroidetes species than Gram-positive Firmicutes species (though there can also be significant variation in extraction efficiency among species within these phyla (McLaren, Willis, and Callahan (2019)). For such protocols, we should expect Bacteroidetes-dominated samples to have substantially larger mean efficiencies than Firmicutes-dominated samples, regardless of species-level diversity. This association of efficiency with phylum abundance will distort DA inferences at any taxonomic level if the covariate is also associated with these phyla.

This example illustrates a general mechanism by which association between the abundances and efficiencies of taxa are created by the common influence of evolutionary history. Just as shared evolutionary history creates positive associations in ecological traits that drives positive correlations in how related taxa vary in abundance across samples, so does it create positive associations in bias-related traits that can lead to similar efficiencies among closely related taxa. Such phylogenically-associated, bias-affecting traits include cell-wall toughness, ribosomal-operon copy number, binding affinity for a given set of primers, and representativeness in taxonomic databases. Ecology and efficiency are both affected by the same evolutionary history, leading to positive associations between the two: A change in diet that increases the relative abundances of many Bacteroidetes species and in so doing also increases the relative abundance of many easy-to-lyse species.

Associations between measurement efficiency and relative abundance can also arise because a single trait affects both. For example, microbes at the ocean floor are slowly buried, sinking into a low nutrient, low oxygen environment. Lloyd et al. (2020) estimate the log fold change in the estimated absolute abundance of various taxa with sediment depth (as a proxy for time) to determine which taxa are able to persist and even grow in this difficult environment. It is plausible that microbes with tougher cell walls would tend to persist longer (alive or dead) in the sediment, while at the same time being more difficult to extract DNA from than microbes with weaker cell walls. As the relative abundance of tougher species increases with depth, the mean extraction efficiency decreases. This decrease would increase the inferred log fold changes and could lead to inferred growth of taxa that are actually just persisting or even slowly dying off. Another example of a trait that might simultaneously affect efficiency and relative abundance is ribosomal copy number, which increases the measurement efficiency in ribosomal amplicon experiments and is also linked to differences in ecology and population dynamics among species.

3.1.1 Error in absolute abundance measurements

So far we have ignored error associated with our absolute abundance measurements. In fact, these measurements can also have a tendency to measure contributions from some taxa more efficiently than others. **TODO: State how this could create problems. Might also note the ambiguity in what the truth/target is (e.g. live cells, or all cells; cells, or 16S rRNA gene copies; observations in Jian, Salonen, and Korpela 2021).** Interestingly, taxonomic bias may actually make qPCR a better method for absolute quantification than cell counting for 16S rRNA gene sequencing experiments. 16S rRNA gene qPCR measures the concentration of 16S rRNA gene copies in the extracted DNA and is therefore necessarily affected by three large sources of bias in 16S rRNA gene experiments: extraction, amplification, and copy-number variation. Yet these biases are shared by the sequencing measurement. Though they make qPCR measurements a bad proxy for total cell density, when used for absolute differential abundance inference the shared bias in the qPCR and sequencing measurements can cancel, leaving relative accurate fold changes.

(Jian, Salonen, and Korpela 2021 suggests some of these ideas and APPENDIX gives a mathematical justification.) If common primers are used by qPCR and 16S rRNA gene sequencing, the bias due to primer mismatches and other sources of variation in amplification could even be accounted for in this manner. Yet ideally we would still be able to estimate and correct remaining error due to unshared bias or to mechanisms, such as saturation during DNA extraction, that could break the assumed proportionality between qPCR measurements and total abundance.

3.2 We cannot simply ignore taxonomic bias

These examples show that there is potential for bias to distort DA results in both low and high diversity settings. We therefore need methods to control or correct for bias to ensure the validity of future results, as well as methods to probe the robustness of past results. It is possible that spurious results are unlikely in many experimental contexts. By applying such methods across a range of biological and experimental contexts we will deepen our understanding of when bias is unlikely to significantly distort biological findings and when additional measures are needed to mitigate its effects.

4 Potential solutions

4.1 Using control samples to calibrate relative abundances

In principle, control ("mock") communities containing representative taxa from the environment of interest can be used to directly estimate the measurement efficiencies and correct bias in the MGS measurements ("calibration"). Such an approach may be sufficient for synthetic-community experiments, where all taxa are culturable, and for relatively simple natural communities like the vaginal microbiome that are dominated by a small number of culturable taxa. But suitable mock controls may not be feasible for most complex natural ecosystems and require significant effort to develop.

An closely-related alternative to mocks are controls derivedd from natural samples, which provide a way to calibrate measurements from protocols to a reference protocol (McLaren, Willis, and Callahan (2019)). A natural fecal standard is currently being developed by NIST and one has recently been made available commercially by Zymo Research (ZymoBIOMICS Fecal Reference). Ensuring the stability and homogeneity of such control samples can be challenging. Careful testing is also needed to ensure that the preparation and storage of the controls has not significantly affected the taxonomic bias relative to the experimental samples in a given application. As it is feasible to characterize a single standard much more extensively than a typical community sample, we may be able to obtain an estimated composition we feel comfortable treating as the ground truth. Yet even when this is not possible, such natural standards can allow us to reconcile results across studies despite not knowing the truth.

4.2 Ratio-based relative and absolute abundance inference

A comprehensive approach of estimating the efficiency of all taxa is needed for getting calibrated relative abundances of all taxa within individual samples. But what if we only want calibrated DA analysis? That is, we are comfortable with not knowing whether *E. coli* is 10% of our sample if we can confidently determine that it doubled or halved between sample conditions? This problem is easier to solve, and one solution is to use analysis methods based on the ratios among taxa rather than the proportions of individual taxa.

A subset of methods for analyzing differential relative abundance are derived from the field of Compositional Data Analysis (CoDA). The defining feature of CoDA methods is that they are based on fold-changes in ratios among elements (here, the taxa) (REF AITCHISON).

The use of CoDA methods in microbiome analysis has largely been motivated by concerns over the negative correlations between taxa induced by the sum-to-one constraint in taxon proportions. But McLaren, Willis, and Callahan (2019) showed that, due to their property of perturbation invariance (REF AITCHISON), the results of CoDA differential relative abundance analysis are invariant to consistent taxonomic bias. From Equation (1), the observed ratio of a taxon i to a taxon j in a sample s is

$$\frac{\text{observed}_i}{\text{observed}_j} = \frac{\text{actual}_i}{\text{actual}_j} * \frac{\text{efficiency}_i}{\text{efficiency}_j} \tag{4}$$

The error is independent of the sample composition and thus cancels in any function that is a fold-change of this (or any) ratio of taxa between samples. (Perhaps insert example from Figure 1ABC all giving same ratio FCs.)

Might ratio-based analysis also be used to overcome the effects of bias in differential absolute abundance? In fact, some methods for determining absolute abundance are based on ratios among taxa in the MGS measurement instead of proportions. Rather than using a measurement of total abundance, these methods require determining the abundance of one or more reference taxa. To estimate the abundance of a focal taxon i, this approach multiplies the ratio of reads (or proportions) of taxon i to a reference taxon r by a known or estimated abundance of the reference taxon, (while ignoring bias)

$$\text{estimate}_i = \frac{\text{observed}_i}{\text{observed}_r} * \text{abundance}_r \tag{5}$$

So far this approach has been mainly used with spike-in experiments, in which an extraneous taxon is added in a known (and typically constant) abundance to each sample so that it can serve as the reference taxon. Yet the reference taxon could also be a naturally occurring taxon whose absolute abundance we have estimated using a method such as ddPCR directly on cells or (q/dd)PCR on the extracted DNA. Because taxa do not compete (or competition is greatly reduced) in such targeted measurements, any taxonomic bias associated with them is expected to create constant multiplicative error across samples and so not affect fold change estimates [APPENDIX]. APPENDIX describe various theoretical and experimental considerations for both approaches.

4.3 Estimating error with targeted measurements of a small number of taxa

Because targeted absolute-abundance measurements are expected to provide (relatively) accurate estimates of fold changes, they can be used to validate and even correct the DA results derived from a proportion-based absolute DA analysis. We illustrate the basic idea using the Lloyd et al. (2020) experiment described above. Absolute abundances in this study were estimated by multiplying MGS proportions by total abundance measured by cell counting; for comparison, qPCR was used to measure specific microbial taxa. Because the error in fold changes or a regression from the MGS measurements is taxon-independent, the difference between the two methods for these reference taxa informs us about the difference for all taxa. Lloyd et al. (2020) found close agreement for the specific taxa also measured by qPCR (Table 1), suggesting that variation in sample mean efficiency did not significantly distort their results. In principle, a joint statistical analysis of all measurements would allow inferring variation in the sample mean efficiency across samples and obtaining calibrated fold change and regression estimates for all taxa [APPENDIX].

4.4 Computational approaches

What can we do for experiments that have already been conducted without mock controls, targeted control measurements, or spike-ins? At least two purely computational approaches can still be used.

Bias sensitivity analysis: A straightforward and universally available approach is to use computer simulation to determine how the results change under a range of possible sets of efficiencies, which can be randomly generated to reflect certain hypotheses about bias in the given system. The utility of such simulations can increase the more we learn about the magnitude of bias in different systems and the taxonomic and protocol features that determine it. Future work in developing tools and methods for simulating efficiency vectors and performing bias sensitivity analyses could be a valuable way to assay and improve the reliability of microbiome results—not just for differential absolute abundance, but for all microbiome analyses.

Bias-aware meta-analysis: When the goal is to perform a meta-analysis that combines studies that have used different protocols, the unknown measurement efficiencies of each protocol can be explicitly included as parameters of the statistical model that is used. Unknown efficiencies can be included in "compositional" linear modeling frameworks such as ALDEx2, DivNet, and fido simply by adding a protocol-specific term to the linear model of taxon log ratios. Thus such bias-aware meta-analyses are already technically feasible and—if bias is truly consistent within a protocol or study—may provide a more powerful alternative to non-parametric or other meta-analysis methods that do not explicitly model bias.

A Measurement models

Deterministic measurement error models that extends the McLaren, Willis, and Callahan (2019) model of metagenomics measurement to 1) describe absolute as well as relative abundance, 2) include spike-in taxa, and 3) include supplementary measurements of bulk and targeted (absolute) abundance.

A.1 Actual abundances and general notation

(For now) Number of samples = J; Number of taxa = I.

A_{ij} is the actual absolute abundance of taxon i in sample j, in the target units. For variables that hold matrixes or vectors, capital letters denote absolute abundances and lower case letters denote proportions. Thus a_{ij} is the matrix of actual relative abundances in terms of cells as a proportion of the total amount of cells in the sample; $a_{ij} = A_{ij}/A_{Tj}$. Let A_{Tj} (or $A_{\cdot j}$??) be the total abundance in sample j, $A_{Tj} = \sum_i A_{ij}$. For a subset of taxa $Q \subset \{1, \ldots, I\}$, define $A_{Qj} \equiv \sum_{i \in Q} A_{ij}$ to be the total abundance of the taxa in Q.

Unless specified otherwise, A always refers to abundance in the original sample material, not in extracted DNA. I will assume the target units are concentration of cells per unit mass. However, other units (such as biomass per volume) may be relevant.

To consider various addition information that can be used for absolute abundance inference, I let B denote bulk abundance measurements (such as via flow cytometry or broad-range 16S qPCR), S denote the known taxon abundances of a spike-in, and T denote the abundances of a set of taxa made via targeted measurement. I let $R \subset \{1, \ldots, I\}$ denote the set of *reference taxa* whose abundance is known independently from the metagenomics measurement (up to experimental bias factors) because they were spiked at a known density, subjected to a targeted measurement, or have been determined through some other means to have a constant or known abundance across samples.

Bias vectors $B^{(P)}$ are vectors of length I indicating the taxon-specific bias associated with a measurement protocol P. I generally assume these to be sample-independent. An overbar denotes the average value over taxa weighted by actual abundance; i.e., $\bar{B}_j = I^{-1} \sum_i B_i a_{ij}$. An overbar and a subcript set of taxa indicates the mean bias within that subset: $\bar{B}_{Qj} = \sum_{i \in Q} B_i a_{ij} / \sum_{i \in Q} a_{ij}$.

Measurement type	Units	Abundance matrix	Bias vector
Actual abundance	[cells]	A, a $(I \times J)$	-
Metagenomics	counts	M, m $(I \times J)$	$B^{(M)}$
Bul**k** abundance	[cells], [DNA], or [gene copies]	K ($I \times 1$ column vector)	$B^{(K)}$
Targeted abundance	[cells], [DNA], or [gene copies]	T $(I \times J)$	$B^{(T)}$
Spike-in abundance	[cells], [DNA], or [gene copies]	S $(I \times J)$	$B^{(S)}$

Note: S_{ij} and T_{ij} are only defined for the reference taxa $i \in R$.

A.2 Metagenomics measurement

M is an $I \times J$ matrix representing the (marker-gene or) metagenomics measurement; M_{ij} is the sequencing count associated with taxon i in sample j, such that $M_{ij} = 10$ indicates that 10 fragments of DNA were sequenced and assigned to taxon i and sample j. In the case of paired-end sequencing, a single fragment of DNA yields two reads, which may or may not be counted jointly or independently by different algorithms; here I assume that read-pairs are assigned jointly, and will use "read" and "fragment" interchangeably. For simplicity, I ignore the possibility for index switching or other sources of cross-contamination and assume that reads assigned to sample j truly arose from sample j.

Following the notation defined above, I let $M_{Tj} = \sum_i M_{ij}$ be the total count for sample j; I will sometimes refer to this number as the *sequencing depth* or *read depth* of the sample, though in reality the total count will be less than the sample's sequencing depth since often a significant fraction of reads are lost by filtering steps prior to producing the final count matrix. I use $m_{ij} = M_{ij}/M_{Tj}$ to denote the proportion of reads assigned to taxon i in sample j.

I assume that the composition (relative abundances) is given by the deterministic error model presented in McLaren, Willis, and Callahan (2019). In this model, the bias vector $B^{(M)}$ represents the *relative* efficiencies with which various taxa are measured, and so has the same effect on measurement if we rescale it $B \to cB$ for any $c > 0$. To remove this degree of freedom, I will suppose that $B_1^{(M)} = 1$, so that $B_i^{(M)}$ is the efficiency of taxon i relative to the first taxon. We may take the definition of the MWC model as being that, conditional on the total count M_{Tj}, the observed read count of taxon i is proportional to its actual abundance times its efficiency times a taxon-independent factor C_j chosen to make sample j's counts sum to M_{Tj},

$$M_{ij} = A_{ij} B_i^{(M)} C_j, \qquad C_j = \frac{M_{Tj}}{\sum_i A_{ij} B_i^{(M)}} = \frac{M_{Tj}}{A_{Tj} \bar{B}_j^{(M)}}. \tag{6}$$

In this model, the ratios among taxa are distorted by constant factors; for two taxa i and i',

$$\frac{M_{ij}}{M_{i'j}} = \frac{m_{ij}}{m_{i'j}} = \frac{A_{ij}}{A_{i'j}} \cdot \frac{B_i^{(M)}}{B_{i'}^{(M)}} \tag{7}$$

for all samples j. If one or both taxa are absent ($A_i = 0$) or undetectable ($B_i = 0$), both sides of the equation are either 0, ∞, or 0/0 (undefined). The observed proportion of a taxon equals its actual proportion multiplied by its efficiency relative to the sample mean efficiency,

$$m_{ij} = a_{ij} \cdot \frac{B_i}{\bar{B}_j^{(M)}}. \tag{8}$$

A.3 Bulk absolute abundance measurement

Consider estimating the total absolute abundance A_T via an aggregate or broad-range measurement, such as cell counting, flow cytometry, universal 16S qPCR, or total DNA concentration (e.g. Qubit). I let K_j donote the bulk absolute abundance measurement for sample j (mneumonic: "K" for the last letter in "bulk").

Bulk measurements are affected by taxon-specific bias $B^{(K)}$, where $B_i^{(K)}$ indicates how efficiently cells of taxon i contribute to the bulk measurement K. Unlike the metagenomics efficiencies, these bulk measurement efficiencies are absolute numbers. For example, cells of one taxon may be more reliably counted than another; and for broad-range 16S qPCR measurement, we expect taxa to contribute in proportion to how reliably they are lysed and to their 16S copy number. Let $B_i^{(K)}$ denote the absolute efficiency with which a cell of taxon i contributes to the measurement in sample s, which I assume is sample-independent. We can write the estimate as

$$K_j = \sum_i A_{ij} B_i^{(K)} = A_{Tj} \bar{B}_j^{(K)}, \tag{9}$$

where $\bar{B}_j^{(K)}$ is the mean bulk-measurement efficiency in the sample.

A.4 Targeted absolute abundance measurement

Targeted measurement of the absolute abundance of specific taxa can be made via a method such as qPCR, ddPCR, or counting CFUs selective media. These measurements too may be subject to bias; I suppose that the measurement of taxon i is given by

$$T_{ij} = A_{ij} B_i^{(T)}, \tag{10}$$

where $B_i^{(T)}$ is the sample-independent, taxon-specific efficiency of the targeted measurement for taxon i.

In discussing the case of multiple reference taxa, it will be useful to refer to the total targeted and metagenomics abundance and the average targeted and metagenomics efficiencies of the reference taxa. I do so with the notation A_{Rj} and M_{Rj} (total abundances) and $\bar{B}_{Rj}^{(T)}$ and $\bar{B}_{Rj}^{(M)}$ (mean efficiencies) as defined above. Note that

$$T_{Rj} = A_{Rj} \cdot \sum_{r \in R} \frac{A_{rj}}{A_{Rj}} \bar{B}_{Rj}^{(T)} \tag{11}$$

$$= A_{Rj} \bar{B}_{Rj}^{(T)}. \tag{12}$$

A.5 Spike-ins

The matrix S describes the nominal abundances with which spike-in taxa were added to each sample. Spike-in experiments are often designed so that a taxon i is added in the same concentration to all samples, which we can represent by supposing the defined columns of S to be identical. The bias $B^{(S)}$ should be interpreted as error in the quantification of how much spike-in was added, such that

$$S_{ij} = A_{ij} B_i^{(S)}. \tag{13}$$

I assume that this error is constant across samples. The motivation for this error model is as follows: Suppose that we misestimated the proportion of spike-in taxon $s \in I_S$ in our original spike-in stock, by a factor $B_s^{(S)}$; now its true abundance will be off by that factor in any spike-in derived from the starting stock. Unlike the metagenomics efficiencies, these error factors are absolute numbers.

Taxa can be added as cells (pre-extraction) or as DNA (post-extraction). In an application where DNA spike-ins are used to make inferences about changes in absolute cell abundances, we can include the difference between the nominal abundances S (the spiked DNA concentration) and the actual abundances A that are expected due to factors such as DNA extraction and copy-number variation in the bias term $B^{(S)}$, so long as DNA extraction is linear (see section on saturation effects below).

(If we want to allow for different targeted and spike-in taxa, we can use R_S and R_T.)

A.6 Complications

These models are the simplest for these types of experiments that have taxon-specific bias. They therefore are idealizations that may be violated by real experiments in ways that affect our conclusions. Here I list some relevant complications that we will return to in our analysis. For more considerations, see the Discussion of McLaren, Willis, and Callahan (2019).

A.6.1 Abundances units and protocol stages

Still need to find the right way to synthesize these issues.

qPCR measurements directly estimating the concentration of a marker gene in the extracted DNA. Efficiency factors can translate the units from DNA to cell concentration *and* account for the bias imposed by extraction and marker-gene CNV (copy-number variation). Sometimes we may want to refer to the DNA concentrations in the sample after extraction, and in that case I will use the variable D, and (perhaps) use $B^{(T/D)}$ to refer to the bias of the targeted measurement vs the DNA concentrations it was applied to. This can be useful when we are considering qPCR and also DNA spike-ins.

Note that the "actual" units could be chosen by the researcher to be something other than cell density; e.g. biomass, or even 16S density. But for concreteness I'll take to be cell density, as this seems to be what is most typically meant.

A.6.2 Saturation in DNA extraction yields

The equation for bulk measurement reflects an idealized form of bulk measurement in which, for a fixed taxonomic composition, K_j is directly proportional to A_T. However, a more realistic model for qPCR may be that K_j is a saturating function of A_T, due to factors such as enzyme consumption or saturation in the elution step creating saturation in DNA yield during extraction. However, it is not clear what this function should depend on (input biomass, cell concentration, total DNA?) and it is not clear what bias components should be included in determining the saturation, and these are not simply $\bar{B}$ since that includes 16S copy-number which isn't relevant here. Rather than try to accurately model the relationship between K and A in this more complicated scenario, I will use the models $K_j = f(A_{Tj})$ and/or $K_j = f(A_{Tj}\bar{B}_j^{(K)})$ to get some intuition for the effect that a strong saturation effect has on our inferences when it is not accounted for.

Saturation during extraction also implications for targeted PCR measurements but not for measurements that work directly on cells; and for DNA spike-ins, but not cellular spike-ins.

A.6.3 Taxonomic resolution of measurements

I suppose that the "taxa" we consider are well-defined in a special sense, such that they can be equated both across samples within a measurement type and across different measurement types. In other words, I assume that there is a real group of organisms corresponding to "taxon i" with abundance $A_{i\cdot}$ that the various "taxon i" measurements are really measuring, such that we can, for example, directly equate the taxonomic source of the counts $M_{i\cdot}$ and the targeted measurement $T_{i\cdot}$. Achieving this situation in practice requires careful attention

to things like primer design and how the metagenomic taxonomic assignment is done, and even then it may not always be possible.

One important special case where this assumption breaks down is when measurements aggregate one or more of the *atomic taxa* (corresponding to the rows in A), possibly in different ways by different measurement types. This raises at least two sorts of complications. First, constant multplicative bias at the level of atomic taxa does not translate into constant bias at the level of aggregate taxa (McLaren, Willis, and Callahan 2019). Second, measurements may target different aggregates. For example, if our metagenomics protocol provides ASV or species-level estimates but our PCR-based targeted measurements may quantify a genus or family. Often through sufficiently careful bioinformatics we can at least nest the metagenomics taxa within the targeted taxon, but complications remain if efficiency varies within the aggregates.

Maybe:

- Give example of effect of aggregation when efficiencies vary among the aggregated taxa; understand how in terms of the average efficiency within the aggregate.
- Write out algebraically some of the cases I'll consider later.

B Differential relative abundance

This section describes the effects of bias on common approaches to analyzing differential relative abundance. We introduce the critical distinction between analyses based on proportions and analyses based on ratios. Analyses based on proportions can be distorted by bias, whereas analyses of log fold changes (LFCs) in ratios of atomic taxa are invariant to bias. If atomic taxa may be aggregated into synthetic taxa by multiplication (or by adding log abundance), and ratios formed from such aggregates; LFCs for such generalized ratios remain bias invariant. But if atomic taxa are aggregated additively by summing their read counts—as often done to analyze higher-order taxa such as phyla—then bias invariance only applies if the aggregated taxa have the same efficiency. The error induced by bias in (some) proportion-based methods has a simple form that suggests tractable experimental methods by which it may be estimated and corrected, which we discuss in Section **??**. Section C shows that methods for obtaining absolute abundances from MGS data can generally be split into those based on proportions and those based on ratios, so that the proportion-ratio dichotomy also provides a useful framework for understanding analyses of differential absolute abundance.

B.1 Proportion-based analyses

B.1.1 Log proportion

To simplify notation where possible, I will write B for $B^{(M)}$ when only the metagenomics measurements are relevant.

It follows from (8) that the observed fold change in the proportion of taxon i from sample j to j' is

$$\frac{m_{ij'}}{m_{ij}} = \frac{a_{ij'}\cancel{B_i}/\bar{B}_{j'}}{a_{ij}\cancel{B_i}/\bar{B}_j} = \frac{a_{ij'}}{a_{ij}} \cdot \frac{\bar{B}_j}{\bar{B}_{j'}}. \tag{14}$$

The taxon-specific error term (B_i) cancels, leaving a multiplicative error equal to the inverse change in the sample mean efficiency $\bar{B}^{(M)}$. Framed in terms of the log fold change, this equation becomes

$$\underbrace{\log m_{ij'} - \log m_{ij}}_{\text{observed LFC}} = \underbrace{\log a_{ij'} - \log a_{ij}}_{\text{true LFC}} - \underbrace{\left(\log \bar{B}_{j'} - \log \bar{B}_j\right)}_{\text{LFC in mean efficiency}}, \tag{15}$$

so that the additive error is the LFC in mean efficiency between the two samples.

Analyzing fold changes in a regression framework typically entails considering the expectation of the logarithm of the response conditional on the value of a set of covariates X. I use $E[Y \mid X]$ as shorthand for $E[Y \mid X = x]$, the expected value of a random variable Y given a vector of covariate values x. Applying logarithms to (8) gives

$$\log m_i = \log a_i + \log B_i - \log \bar{B} \tag{16}$$

which, after conditioning on the covariates, becomes

$$E[\log m_i \mid X] = E[\log a_i \mid X] + \log B_i - E[\log \bar{B} \mid X]. \tag{17}$$

The taxon-specific term $\log B_i$ creates a constant error that is unproblematic for differential abundance analysis, which is typically ignores the baseline abundance of the taxon. However, the sample-specific error $\log \bar{B}$ can distort the inferred relationship between X and the expected value of the true log proportion if its expected value also varies with X.

TODO: Explain this γ notation for the regression coefficients, or find a more recognizable notation.

Consider the special case of the linear regression

$$\log a_i = X\gamma + \epsilon, \tag{18}$$

where X is a $J \times p$ covariate matrix and γ is a $p \times 1$ vector of coefficients. Suppose we knew the values of each of the terms in (16). Setting each term as the left-hand-side of its own corresponding regression, the least-squares coefficients are related as ((74))

$$\hat{\gamma}(\log m_i) = \hat{\gamma}(\log a_i) + \hat{\gamma}(\log B_i) - \hat{\gamma}(\log \bar{B}). \tag{19}$$

For the simple linear regression,

$$\log a_i = \gamma_0 + \gamma_1 x + \epsilon, \tag{20}$$

with an intercept coefficient γ_0 and a slope coefficient γ_1, the relation (19) implies

$$\hat{\gamma}_0(\log m_i) = \hat{\gamma}_0(\log a_i) + \log B_i - \hat{\gamma}_0(\log \bar{B}) \tag{21}$$

$$\hat{\gamma}_1(\log m_i) = \hat{\gamma}_1(\log a_i) - \hat{\gamma}_1(\log \bar{B}), \tag{22}$$

Since $\log B_i$ is constant, it only affects the intercept. The slope estimate for the metagenomics proportions, $\hat{\gamma}_1(\log m_i)$, is systematically decreased from that of the true proportions, $\hat{\gamma}_1(\log a_i)$, by that for the sample mean efficiency, $\hat{\gamma}_1(\log \bar{B})$.

B.1.2 Log odds (logit)

The proportion of a taxon saturates at 1 as its absolute abundance increases, while the odds, $a_i/(1 - a_i) = A_i/(A_T - A_i)$, continue to increase. For this reason it is often more natural to perform regression on the log odds, or *logit-transformed proportion*, rather than log proportion.

The observed log odds of taxon i is

$$\operatorname{logit} m_{ij} = \log \frac{m_{ij}}{1 - m_{ij}} = \log \frac{M_{ij}}{M_{Tj} - M_{ij}}. \tag{23}$$

To understand the effects of bias on the measured log odds, it is helpful to define a variable $\bar{B}_{-i,j}$ to denote the mean efficiency in sample j among taxa *other than* i,

$$\bar{B}_{-i,j} \equiv \frac{\sum_{k \neq i} B_k A_{kj}}{\sum_{k \neq i} A_{kj}} = \frac{\sum_{k \neq i} B_k a_{kj}}{1 - a_{ij}}. \tag{24}$$

The measured proportion of non-i taxa is

$$1 - m_{ij} = \sum_{k \neq i} m_{kj} \tag{25}$$

$$= \sum_{k \neq i} a_{kj} \cdot \frac{B_k}{\bar{B}_j} \tag{26}$$

$$= (1 - a_{ij}) \cdot \frac{\bar{B}_{-i,j}}{\bar{B}_j}, \tag{27}$$

where the final expression follows from $\sum_{k \neq i} a_{kj} B_k = (1 - a_{ij}) \bar{B}_{-i,j}$. Intuitively, Equation (25) says that the fold error in the proportion of "not-i" equals the mean efficiency of the "not-i" part of the sample relative to the mean efficiency of the sample as a whole. This result means that the measured odds of taxon i is

$$\frac{m_{ij}}{1 - m_{ij}} = \frac{a_{ij}}{1 - a_{ij}} \cdot \frac{B_i}{\bar{B}_{-i,j}}, \tag{28}$$

This formula resembles Equation (7) for the measured ratio of two taxa i and i', but with taxon i' corresponding to all not-i taxa. In this case, however, the efficiency of not-i varies among samples as changes in their relative abundances change $\bar{B}_{-i,j}$. Hence, in contrast to the ratio of two taxa, the error is not consistent across samples.

The error in the logit of taxon i is

$$\operatorname{logit} m_i = \operatorname{logit} a_i + \log B_i - \log \bar{B}; \tag{29}$$

note the log rather than logit operators in the efficiency terms. The corresponding regression equation is

$$E[\operatorname{logit} m_i \mid X] = E[\operatorname{logit} a_i \mid X] + \log B_i - E[\log \bar{B}_{-i} \mid X]. \tag{30}$$

B.2 Ratio-based analyses

The previous section showed that differential abundance analyses based on proportions may be sensitive to bias, due to the dependence of the multiplicative error in measured proportions on the sample mean efficiency (Equation (8)). In contrast, the multiplicative error in the ratios among taxa (Equation (7)) is independent of sample composition, such that a differential abundance analysis based on log ratios is invariant to bias. The simplest such analysis is when "abundance" is equated with the log ratio of two specific taxa; however, any linear combination of log ratios may be used, which includes a variety of so-called Compositional Data Analysis (CoDA) methods that have recently been applied in microbiome DA analysis (see below). McLaren, Willis, and Callahan (2019) described the bias invariance of ratio-based methods (see their Equations 6 and 7). Here we show how this general result applies to specific types of differential relative abundance.

We first consider the ratio a pair of taxa i and i'. It follows from Equation (7) that the measured fold change in the ratio of taxon i to taxon i' from sample j to sample j' is

$$\frac{m_{ij'}/m_{i'j'}}{m_{ij}/m_{i'j}} = \frac{a_{ij'}\cancel{B_i}/a_{i'j'}\cancel{B_{i'}}}{a_{ij}\cancel{B_i}/a_{i'j}\cancel{B_{i'}}} = \frac{a_{ij'}/a_{i'j'}}{a_{ij}/a_{i'j}}. \tag{31}$$

The error in each ratio, $B_i/B_{i'}$, is constant and so cancels completely, leaving the true fold change. To see the effect of bias on regression of log ratios, we note that the error equation (7) implies

$$\log \frac{m_i}{m_{i'}} = \log \frac{a_i}{a_{i'}} + \log \frac{B_i}{B_{i'}}. \tag{32}$$

Taking the conditional expectation given a covariate vector X gives

$$E\left[\log \frac{m_i}{m_{i'}} \mid X\right] = E\left[\log \frac{a_i}{a_{i'}} \mid X\right] + \log \frac{B_i}{B_{i'}}. \tag{33}$$

In contrast to the case of a log proportion (Equation (17)), the effect of bias on the log ratio is simply to create a constant shift and therefore does not affect differential-abundance analysis.

CoDA DA methods often draw upon generalized notion of ratio, in which the numerator and/or denominator consist of a product of powers of multiple taxa. Such a generalized ratio is determined by length-I vectors n and d giving the power of each taxon in the numerator and denominator, with $n_i = 0$ indicating that the taxon does not affect the term. Denoting the actual value of a given generalized ratio in sample j as y_j and the observed as z_j, we have

$$\text{actual:} \quad y_j = \frac{\prod_i A_{ij}^{n_i}}{\prod_i A_{ij}^{d_i}} \qquad\qquad \text{observed:} \quad z_j = \frac{\prod_i M_{ij}^{n_i}}{\prod_i M_{ij}^{d_i}}. \tag{34}$$

Under the MWC bias model, the observed and actual log ratios are related as

$$\log z_j = \log y_j + \log \frac{\prod_i B_{ij}^{n_i}}{\prod_i B_{ij}^{d_i}}. \tag{35}$$

Hence, as with the simple ratio between two taxa, the error due to bias is independent of sample composition and so does not affect the observed variation in $\log z_j$.

Most of the log-ratio transformations that are used in CoDA DA analysis can be understood as particular applications of the generalized ratio (34) and thus lead to bias-invariant DA results under the deterministic MWC model. For example, the additive log-ratio (ALR) transformation consists of computing the log of the ratios obtained by dividing each taxon's abundance by that of a particularly chosen reference taxon r,

$$\operatorname{alr} A_j = \left[\log \frac{A_{ij}}{A_{rj}}, \ldots, \log \frac{A_{Ij}}{A_{rj}}\right], \tag{36}$$

while the centered log-ratio (CLR) transformation instead sets the denominator to the geometric mean $g(A_j) = \left(\prod_i A_{ij}\right)^{1/I}$ of all taxa,

$$\operatorname{clr} A_j = \left[\log \frac{A_{ij}}{g(A_j)}, \ldots, \log \frac{A_{Ij}}{g(A_j)}\right]. \tag{37}$$

More generally, one might choose any set R of reference taxa and use their geometric mean in the denominator; the ALR and CLR transformations then correspond to taking $R = r$ and $R = I$, respectively. Quinn et al. (2019) call this transformation the *multiple additive log-ratio* (MALR) transformation,

$$\operatorname{malr} A_j = \left[\log \frac{A_{ij}}{g(A_{R,j})}, \ldots, \log \frac{A_{Ij}}{g(A_{R,j})}\right]. \tag{38}$$

The actual and observed MALR-transformed abundance associated with a specific numerator taxon i is

$$\operatorname{malr}(M_j)_i = \operatorname{malr}(A_j)_i + \operatorname{malr}(B)_i; \tag{39}$$

hence the error due to bias is constant and does not factor into the inferred changes in a DA analysis. Note: Unlike Quinn et al. (2019) we do not include the "robust centered log-ratio transformation" of Martino et al. (2019) as an example of an MALR; this transformation

chooses a distinct set of reference taxa for each sample and so does not have bias invariance under the MWC model.

Other generalized log-ratios commonly used in microbiome data analysis include *balances*. A balance is defined by two sets of taxa, Q and R, and equals the log of the ratio of the geometric mean of Q taxa to the geometric mean of R taxa, multiplied by a scaling factor,

$$\sqrt{\frac{|Q||R|}{|Q|+|R|}} \log \frac{g(A_{Q,j})}{g(A_{R,j})}. \tag{40}$$

Balances have become a popular tool for identifying biomarkers or groups of taxa associated with an environmental or health condition while operating within the CoDA framework (Washburne et al. (2017), Rivera-Pinto et al. (2018), Quinn and Erb (2020)). The exponential of a balance is an example of the generalized ratio in (34) and so regression of balances is again invariant to bias.

The products in the numerators and denominators in (34) can be seen as a way to multiplicatively aggregate abundances of different taxa, as opposed to the additive aggregation that is commonly used when relative abundances are viewed as proportions. Multiplicative aggregation preserves the property of perturbation invariance that is the source of bias invariance under the MWC model. In contrast, additive aggregation (known as "amalgamation" in the CoDA literature) violates perturbation invariance and can lead to regression results that depend on bias even when they are ostensibly based on the ratios of two (non-atomic) taxa, as we now explain.

B.3 Higher-order taxa formed by additive aggregation

Microbiome researchers often combine lower-order taxa into higher-order taxa prior to conducting a DA analysis. For instance, Amplicon Sequence Variants (ASVs), Operational Taxonomic Units (OTUs), or species-level counts may be aggregated into genus- or family-level counts by simply summing the counts within each group of lower-level taxa. Such aggregation can increase statistical power by reducing noise, sparsity, and the number of tests conducted. It also simplifies the task of interpretation by limiting the number of taxa considered. In addition, some degree of uncontrolled aggregation is inevitable due to the inherent limitations of our sequencing and bioinformatics protocols to distinguish sufficiently similar organisms (McLaren, Willis, and Callahan (2019)), which may be lumped together into an OTU or species-level feature. Finally, though our focus is on analysis of taxa, we note that taxonomic aggregation also occurs (at least implicitly) in gene or function analyses, where counts are combined from different taxa that share the same genes or predicted functions. Hence, to understand the effects of bias on DA analysis as it is actually practiced, we must understand the effects of such taxonomic aggregation on differential abundance.

Consider a set of atomic taxa given by the set $Q \subset \{1, \ldots, I\}$ and let A_{Qj} be the vector of abundances in sample j of the taxa in Q. The actual absolute abundance of the synthetic taxon Q is simply the sum of abundances of its component taxa, $\text{sum}\,(A_{Q,j}) = \sum_{q\in Q} A_{qj}$. Similarly, the actual proportion of Q is given by the sum of proportions, $\text{sum}\,(a_{Q,j}) = \sum_{q\in Q} a_{qj} = \text{sum}\,(A_{Q,j})\,/A_{Tj}$. The abundance of Q in the metagenomics measurement is $\text{sum}\,(M_{Q,j})$, and its proportion $\text{sum}\,(m_{Q,j})$.

Such a synthetic taxon has a consistent, composition-independent measurement efficiency only if its component taxa have equal efficiencies. Let $\bar{B}_{Q,j}$ denote the mean efficiency of the taxa in Q in sample j,

$$\bar{B}_{Q,j} = \frac{\sum_{q\in Q} B_q A_{qj}}{\sum_{q\in Q} A_{qj}}. \tag{41}$$

Unless all the taxa in Q have the same efficiency, the mean efficiency of Q varies with the relative abundances among these taxa. Summing over Equation (8) and performing some rearranging shows that the observed proportion of Q is

$$\text{sum}\left(m_{Q,j}\right) = \frac{1}{\bar{B}_j} \sum_{q \in Q} a_{qj} B_q \sum_{q \in Q} m_{qj} \tag{42}$$

$$= \text{sum}\left(a_{Q,j}\right) \cdot \frac{\bar{B}_{Q,j}}{\bar{B}_j}. \tag{43}$$

In words, the error in the proportion of the aggregate taxon Q is given by the mean efficiency of taxa in Q relative to the sample mean. Hence only if the atomic taxa that make up Q all have the same efficiencies or always appear in the same ratios to each other will the efficiency of Q be constant across samples. This observation was first made in McLaren, Willis, and Callahan (2019) (Discussion and Appendix 1) and is here extended to its general form.

The lack of consistency in the efficiency of the synthetic taxon Q complicates the form of the error in both proportion- and ratio-based DA analyses. The equivalent of the regression equation (17) for the log proportion of Q is

$$E\left[\log \sum_{q \in Q} m_q \mid X\right] = E\left[\log \sum_{q \in Q} a_q \mid X\right] + E\left[\log \bar{B}_Q - \log \bar{B} \mid X\right]; \tag{44}$$

we can no longer move the log efficiency of the focal taxon out of the expectation operator and so must consider the variation in the mean efficiency of Q as well as of the sample as a whole. To consider the effect on ratio-based analyses, we define another synthetic taxon $R \subset \{1, \ldots, I\}$ and consider the ratio of Q to R. The observed ratio,

$$\frac{M_{Q,j}}{M_{R,j}} = \frac{A_{Q,j}}{A_{R,j}} \cdot \frac{\bar{B}_{Q,j}}{\bar{B}_{R,j}}, \tag{45}$$

has an error equal to the ratio in mean efficiency of Q to R, both of which can vary with the relative abundances of the component taxa. As the ratio no longer as a consistent error, it is possible for bias to lead to spurious inferences in the fold change in the ratio.

C Differential absolute abundance

This section considers various ways to estimate absolute abundances of individual taxa for the purposes of estimating log fold changes (LFCs) in absolute abundance, or more generically of performing linear regression on log absolute abundance.

C.1 Using bulk abundance measurements

The absolute abundance of taxon i is estimated by multiplying its metagenomics proportion by the bulk abundance measurement,

$$\hat{A}_{ij} = m_{ij} K_j. \tag{46}$$

The error in this estimate due to the experimental bias in M and K is given by substituting equations (8) and (9) into (46), giving

$$\hat{A}_{ij} = \frac{a_{ij} B_i^{(M)}}{\bar{B}_j^{(M)}} \cdot A_{Tj} \bar{B}_j^{(K)} \tag{47}$$

$$= A_{ij} \cdot \frac{B_i^{(M)} \bar{B}_j^{(K)}}{\bar{B}_j^{(M)}}. \tag{48}$$

The fold error in the estimated abundance of a particular taxon in a particular sample thus equals the metagenomics efficiency of that taxon times the ratio of the sample mean efficiencies of the bulk measurement to the metagenomics measurement. Next I consider the error in the estimated fold change in A_i between two samples j and j'. As in the case of log proportion estimates, the taxon-specific, sample-independent error term ($B_i^{(M)}$) cancels, but the taxon-independent, sample-specific error term ($\bar{B}_j^{(K)}/\bar{B}_j^{(M)}$) does not, giving

$$\frac{\hat{A}_{ij'}}{\hat{A}_{ij}} = \frac{A_{ij'}}{A_{ij}} \cdot \frac{\bar{B}_j^{(M)} \bar{B}_{j'}^{(K)}}{\bar{B}_{j'}^{(M)} \bar{B}_j^{(K)}}. \tag{49}$$

$$= \frac{A_{ij'}}{A_{ij}} \cdot \frac{\bar{B}_{j'}^{(K)}/\bar{B}_{j'}^{(M)}}{\bar{B}_j^{(K)}/\bar{B}_j^{(M)}}. \tag{50}$$

In words, the fold error in the estimated fold change in the absolute abundance of any taxon equals the fold change in the ratio of the mean efficiency of the bulk abundance measurement to that of the metagenomics measurement.

TODO: Add regression result.

C.1.1 Shared bias components

Motivation: Consider an experiment in which the community has been profiled by 16S sequencing and bulk abundance has been estimated by 16S qPCR. The taxon-specific bias associated with each of these measurements may be highly similar; both are measuring 16S copies in the same aliquot of already-extracted DNA, which has already been affected by DNA extraction bias and 16S copy-number variation. If the same primers are used then PCR bias might also be shared. How does having a large amount of bias being shared between the bulk and sequencing measurement affect the error in abundance estimates in individual samples and the fold changes between samples?

In this case, the ratio of mean efficiencies, $\bar{B}_j^{(K)}/\bar{B}_j^{(M)}$, will vary much less across samples than either mean efficiency itself, such that the error term in (49) will be small. Therefore the estimated fold changes may be accurate despite potentially large variation in the metagenomics sample mean efficiency across samples. Note that the estimated absolute abundance in the individual sample, (47), remains inaccurate since the unknown taxon-specific efficiency remains.

TODO: Add derivation, where we formally define $B^{(M)}$ and $B^{(K)}$ in terms of shared and non-shared components, and show that the variation in the ratio of mean efficiencies reduces to that of the non-shared components.

C.2 Using reference taxa

TODO: Consider rewriting to be in terms of the ratio of counts * the reference abundance, to emphasize that this is a ratio-based method.

C.2.1 Taxa with targeted abundance measurements

First consider the case of just a single reference taxon, r, for which we have targeted measurements. We can estimate the absolute abundances of all taxa by scaling the read counts by the ratio of the abundance measurement to read count for the reference taxon,

$$\hat{A}_{ij} = M_{ij} \cdot \frac{T_{rj}}{M_{rj}}. \tag{51}$$

Accounting for systematic error in both the metagenomcs measurement (6) and the targeted measurement (10) gives

$$\hat{A}_{ij} = A_{ij} \cdot \frac{B_i^{(M)} B_r^{(T)}}{B_r^{(M)}}. \tag{52}$$

Since the error in is sample-independent, it cancels out when we use this equation to compute the fold change between two samples,

$$\frac{\hat{A}_{ij'}}{\hat{A}_{ij}} = \frac{\hat{A}_{ij'}}{\hat{A}_{ij}}. \tag{53}$$

Thus our model predicts that normalizing read counts to a reference taxon, while providing systematically-distorted abundance estimates, provides accurate fold-change estimates.

C.2.2 Cellular spike-ins

Spike-ins are a way to create a reference taxon whose abundance is known by design and so doesn't need to be measured. Typically the spike-in is added in a fixed concentration to all samples; however, this is unnecessary, so long as we know the amount it was added. In either case we can use (51) to estimate the abundances of all taxa from the spike-in taxa, substituting the spike-in abundance S_{ir} for T_{ir}, Our model predicts that normalizing read counts to a spike-in taxon gives systematically-distorted abundance estimates but accurate fold-change estimates for the same reason as in the targeted case.

C.2.3 DNA spike-ins

C.2.4 Taxa assumed to have a constant abundance

By this approach, we identify one or more taxa that we assume to have a constant absolute abundance across samples. Inference proceeds as in the targeted and spike-in case, using (51) but with the reference abundance T_r set to an arbitrary constant as we may not know the true abundance. In this case, the effect of bias cancels in fold-change calculations for the same reasons as the targeted and spike-in cases. Note: The reference taxon could be the host.

C.2.5 Multiple reference taxa

Under the deterministic model we consider here, having multiple reference taxa in the same sample is completely redundant - we learn no new information. In reality, having information from multiple taxa from a single sample should allow us to make more precise estimates, provided that we use them in a way so that the effects of bias do not offset any reductions in noise. How might we leverage measurements of multiple reference taxa, $r \in R$, in a way that remains robust to bias?

First consider the situation where we can assume that all reference taxa are in each sample, as would typically be the case for a spike-in. One possibility is to sum the metagenomic and targeted abundances from the different taxa,

$$\hat{A}_{ij} = M_{ij} \cdot \frac{\sum_{r \in R} T_{rj}}{\sum_{r \in R} M_{rj}}. \tag{54}$$

To see this, let $A_{Rj} = \sum_{r \in R} A_{rj}$ be the summed actual abundances of the reference taxa. Similarly define $M_{Rj} = \sum_{r \in R} M_{rj}$ and $T_{Rj} = \sum_{r \in R} T_{rj}$ be the summed read counts and targeted measurements. Now

$$\hat{A}_{ij} = M_{ij} \cdot \frac{T_{Rj}}{M_{Rj}}. \tag{55}$$

Also let $\bar{B}_{Rj}^{(M)} = \sum_{r \in R} B_{rj}^{(M)} A_{rj}/A_{Rj}$ be the mean metagenomics efficiency among the reference taxa and similarly define $\bar{B}_{Rj}^{(T)}$ as the mean targeted efficiency. A little algebra shows that

$$T_{Rj} = A_{Rj} \cdot \sum_{r \in R} \frac{A_{rj}}{A_{Rj}} \bar{B}_{Rj}^{(T)} \tag{56}$$

$$= A_{Rj} \bar{B}_{Rj}^{(T)} \tag{57}$$

while

$$M_{Rj} = A_{Rj} \bar{B}_{Rj}^{(M)} C_j. \tag{58}$$

And since $M_{ij} = A_{ij} B_i^{(M)} C_j$ (6) we can write the estimated (54) abundance of taxon i as

$$\hat{A}_{ij} = A_{ij} \cdot \frac{B_i^{(M)} \bar{B}_{Rj}^{(T)}}{\bar{B}_{Rj}^{(M)}}. \tag{59}$$

The error now depends on relative abundances among the reference taxa through their mean bias values. Therefore, fold changes estimated using this approach are not necessarily robust to bias if the relative abundances among the reference taxa vary. Spike-in experiments are typically designed so that the reference taxa typically have the same relative abundances with respect to each other in all samples, in which case this method would remain robust to bias.

Another approach is to multiply by the geometric mean abundances of the taxa,

$$\hat{A}_{ij} = M_{ij} \cdot \left[\frac{\prod_{r \in R} T_{rj}}{\prod_{r \in R} M_{rj}} \right]^{1/|R|}. \tag{60}$$

The error in this case is

$$\hat{A}_{ij} = A_{ij} B_r^{(M)} C_j \cdot \left[\frac{\prod_{r \in R} A_{rj} B_r^{(T)}}{\prod_{r \in R} A_{rj} B_r^{(M)} C_j} \right]^{1/|R|} \tag{61}$$

$$= A_{ij} \cdot \left[\frac{\prod_{r \in R} B_r^{(T)}}{\prod_{r \in R} B_r^{(M)}} \right]^{1/|R|}. \tag{62}$$

The multiplicative error is now constant and will cancel in fold-change calculations even if the reference taxa vary in their relative abundances.

Thus the geometric-mean method appears to be robust to bias, but as written it requires that all reference taxa are present in all samples, whereas the summation approach is sensitive to bias but only requires at least one reference taxon to be present in the sample.

Now consider the case where different reference taxa are present (above our detection limit) in each sample, as might be the case for naturally occuring reference taxa. We can modify the geometric-mean approach to work (so long as at least one reference is present in each sample) as follows. From the samples where multiple references are present, we are able to measure the differnetial bias $B_r^{(M/T)}/B_{r'}^{(M/T)}$ among the references r, r'. From measurement of reference r', we know the value of $T_{r'j}/M_{r'j}$, which combined with the differential bias tells us what the ratio T_{rj}/M_{rj} should equal, namely

$$\frac{T_{rj}}{M_{rj}} = \frac{T_{r'j}}{M_{r'j}} \cdot \frac{B_r^{(M/T)}}{B_{r'}^{(M/T)}}. \tag{63}$$

Therefore, in samples where r is missing but r' is present, we can replace $\frac{T_{rj}}{M_{rj}}$ in the product with its predicted value. (More generally, we can use the geometric mean of the values predicted by each of the reference taxa that are present).

This imputation procedure is a simple plug-in approach to show that it is possible to use a multiplicative approach to using multiple references even when the references vary in their presence among samples. Real inference should ideally be done with a fully generative statistical model that includes bias as well as noise-generating processes and can be used for maximum likelihood or Bayesian inference. In such models, the imputation might be implicitly handled as part of the likelihood, and more generally the values $\frac{T_{rj}}{M_{rj}}$ for various r can be naturally and automatically weighted (in terms of their influence on $\hat{A}_{ij}$) according to their predicted precision by the model fitting procedure.

C.3 Using an equivolumetric protocol

C.4 Computational methods

D Proofs of regression results

D.1 General regression

Regression analysis can often be framed as seeking the regression function $r(x)$ that describes how the expected value of some response variable Y varies with the a vector of covariates X (Wasserman (2004)),

$$r(x) = E[Y \mid X = x]. \tag{64}$$

For example, the response might be log absolute abundance of a particular taxon. In our case, however, we don't know Y, but rather a measure that is subject to random and systematic error, which I call Z. Let $D = Z - Y$ be the difference between the true response Y and its measurement Z, so that $Z = Y + D$. It follows that

$$E[Z \mid X] = E[Y \mid X] + E[D \mid X]. \tag{65}$$

D.2 Linear least-squares regression

Theorem D.1 (Linearity of regression coefficients)**.** *Consider a scalar response variable with J observations, which I represent by the vector y, that equals a sum of K component response variables $y^{(k)}$ each scaled by a non-zero factor $c^{(k)}$,*

$$y = \sum_{k=1}^{K} c^{(k)} y^{(k)}. \tag{66}$$

(We may need to assume that the $y^{(k)}$ are linearly independent.) Let X be a J-by-p matrix of covariates for the J observations, with linearly independent columns. Let $\hat{\gamma}$ denote the least-squares estimate of the coefficient matrix γ in the linear regression

$$y = X\gamma + \epsilon. \tag{67}$$

Similarly, let $\hat{\gamma}^{(k)}$ denote the least-squares estimates for the K linear regressions

$$y^{(k)} = X\gamma^{(k)} + \epsilon^{(k)}. \tag{68}$$

The least-squares coefficient estimates for y are given by the sum of those of the $y^{(k)}$,

$$\hat{\gamma} = \sum_{k=1}^{K} c^{(k)} \hat{\gamma}^{(k)}. \tag{69}$$

Proof. The Moore-Penrose pseudoinverse of X is $X^+ = (X^T X)^{-1} X^T$. The least squares estimates are given by multiplying the matrix X^+ by the corresponding response vector (Wikipedia), so that $\hat{\gamma} = X^+ y$ and $\hat{\gamma}^{(k)} = X^+ y^{(k)}$ (for $k = 1, \ldots, K$). The result then follows from the stipulation (66) and the linearity of matrix multiplication,

$$\hat{\gamma} = X^+ y = X^+ \sum_{k=1}^{K} c^{(k)} y^{(k)} = \sum_{k=1}^{K} c^{(k)} X^+ y^{(k)} = \sum_{k=1}^{K} c^{(k)} \hat{\gamma}^{(k)}. \tag{70}$$

□

We can use D.1 to describe how measurement error, such as that caused by experimental bias, affects least-squares estimates of regression coefficients.

Theorem D.2 (Error in regression coefficients)**.** *Suppose that y is the response variable we wish to understand, z is our imperfect measurement of y, and $d = z - y$ the difference between the two, so that $z = y + d$. Let each be a vector of length J describing a set of J observations. Let X be a J-by-p matrix of covariates for the J observations, with linearly independent columns. Consider the linear regression equations for y, d, and z,*

$$y = X\gamma^{(y)} + \epsilon^{(y)} \tag{71}$$

$$z = X\gamma^{(z)} + \epsilon^{(z)} \tag{72}$$

$$d = X\gamma^{(d)} + \epsilon^{(d)}. \tag{73}$$

The relationship between the least-squares estimates mirrors that between the variables themselves,

$$\hat{\gamma}^{(z)} = \hat{\gamma}^{(y)} + \hat{\gamma}^{(d)}. \tag{74}$$

Proof. The result follows directly from $z = y + d$ and D.1.

□

Corollary D.1. *For the special case of simple linear regression,*

$$y = \gamma_0^{(y)} + \gamma_1^{(y)} x + \epsilon^{(y)} \tag{75}$$

$$z = \gamma_0^{(z)} + \gamma_1^{(z)} x + \epsilon^{(z)} \tag{76}$$

$$d = \gamma_0^{(d)} + \gamma_1^{(d)} x + \epsilon^{(d)}, \tag{77}$$

the estimates for the intercept and slope coefficients satisfy

$$\hat{\gamma}_0^{(z)} = \hat{\gamma}_0^{(y)} + \hat{\gamma}_0^{(d)} \tag{78}$$

$$\hat{\gamma}_1^{(z)} = \hat{\gamma}_1^{(y)} + \hat{\gamma}_1^{(d)}. \tag{79}$$

Next, aiming to show the forms for the different transformations of interest. It might be useful to cover the case where d is split into a constant and a sample-specific part, since probably everything of interest will follow from that.

What, if any general results should go in the main text?

Also relate the expected values of y and z conditional on X?.

To figure out what I need:

1. extend the below result for log proportions to multiple regression
2. work out the derivation in terms of D.2
3. Write a corollary if it seems to be useful for doing all the cases (e.g. logit etc)

D.2.1 Linear regression of microbiome abundances

First, consider log proportions. In this case, the intercept coefficient is biased upward by $\log B_i$ and the other coefficients are biased downward by the corresponding coefficient of $\log \bar{B}$, based on the error formulas for individual samples.

TODO: write the general result (for multiple regression, assuming that an intercept term is present.

Proposition D.1. *Consider the linear equations*

$$\log a_i = \alpha_0 + \alpha_1 x \tag{80}$$

$$\log \bar{B} = \beta_0 + \beta_1 x \tag{81}$$

$$\log m_i = \gamma_0 + \gamma_1 x, \tag{82}$$

and let $\hat{\alpha}_0$, etc. denote the least-squares estimates for the regression coefficients supposing perfect information (i.e., that the a_i and B_i are known). The coefficients for $\log m_i$ are related to those for $\log a_i$ and $\log \bar{B}$ through the equations

$$\hat{\gamma}_0 = \hat{\alpha}_0 + \log B_i - \hat{\beta}_0 \tag{83}$$

$$\hat{\gamma}_1 = \hat{\alpha}_1 - \hat{\beta}_1. \tag{84}$$

E DNA measurement and spike-ins

This appendix aims to derive expressions for the error in individual samples and figure out whether the error is constant across samples (so that it cancels out LFC analysis). Why difference? Extraction has occured prior to the targeted measurement.

Punchline: Bias will cancel like before if and only if DNA extraction yield is perfectly proportional to input, in a particular sense I should perhaps create a name for. It needs to be proportional to input for a fixed composition; it doesn't have to be proportional across compositions.

E.1 Expanded model and notation

I partition the experiment into extraction and sequencing steps,

$$A_{ij} \xrightarrow[F'_{ij}]{\text{extraction}} A'_{ij} \xrightarrow[F''_{ij}]{\text{sequencing}} A''_{ij} = M_{ij}. \tag{85}$$

where A'_{ij} denotes the absolute abundances after extraction. The abundances A''_{ij} after sequencing are just the counts M_{ij}. I define a matrix F'_{ij} of factors equalling the fold change in abundance during the extraction step, $F'_{ij} = A'_{ij}/A_{ij}$, and similarly let F''_{ij} be the fold changes during the sequencing step, The total fold change is just the product: $F_{ij} = F'_{ij}F''_{ij} = M_{ij}/A_{ij}$.

The MWC model implies that the relative values of F_{ij} to $F_{i'j}$ are independent of j, for both the protocol as a whole and the individual steps. Therefore we can write the factors as product of the taxon-specific bias B_i and a sample-specific scaling factor C_j, as in Section A. I use $'$s to do this for the different steps, for example $F'_{ij} = B'_i C'_j$. The total metagenomics bias and scaling factors thus split into steps as $B_i = B'_i B''_i$ and $C_j = C'_j C''_j$.

TODO: Add Note about units. We could be using bp, ng, genome copies, or marker-gene copies. It shouldn't affect the conclusions, but affects the units and values of the multipliers. Though perhaps it will help to pick ng DNA or genome copies

What is C'_j?

$$C'_j = \frac{A'_{Tj}}{A_{Tj}\bar{B}'_j}. \tag{86}$$

Note that we have so far been defining the metagenomics efficiencies B as being (arbitrarily) the efficiency relative to the first taxon, which affects the scale of C'.

E.2 Targeted DNA measurement

TODO: explain the big picture of what I'm doing here

Let $B_r^{(T)}$ be the conversion factor associated with the targeted measurement of reference taxon r from the extracted DNA, which I assume is sample-independent but can vary by taxon. NOTE: I'm using B for "bias" here but we need to keep in mind that I'm talking about the bias relative to the DNA and not the cellular abundance. In other words, the result of the targeted DNA measurement is

$$T_{rj} = A'_{rj}B_r^{(T)} \tag{87}$$

$$= A_{rj}B'_rC'_jB_r^{(T)}. \tag{88}$$

Substituting the second expression for T_{rj} in (51) gives an expression for the estimate of the abundance of taxon i from the targeted measurement,

$$\hat{A}_{ij} = \frac{M_{ij}}{M_{rj}} \cdot T_{rj} \tag{89}$$

$$= \frac{A_{ij}B_iC_j}{A_{rj}B_rC_j} \cdot A_{rj}B'_rC'_jB_r^{(T)} \tag{90}$$

$$= A_{ij} \cdot \frac{B_iB_r^{(T)}}{B''_r} \cdot C'_j, \tag{91}$$

where I have replaced B_r/B'_r by B''_r in the denominator. Compared to the case of cellular measurement, the efficiency ratio for the reference taxon is just of the post-extraction steps, and we also have an additional term C'_j that could vary across samples.

Question: Is that the same as the mean DNA extraction efficiency varies across samples?

E.3 DNA spike-ins

HERE: Do same thing as above, and get result that

Suppose we add the spike-in taxon r to sample j in amount A'_{rj} with error $B_r^{(T)}$, such that we believe the spike-in's abundance to be T_{rj} in (87). As with the cellular spike-in, the abundance of taxon i can be estimated by (51) (for the targeted cellular method); and as with the case of a targted DNA measurement above, the resulting estimate is given by (89).

E.4 Implications

These results show that, for both methods, the fold variation in the error across samples equals that in the scaling factor C'_j. Ideally, C'_j would be constant, corresponding to the situation where doubling the cell concentration in a sample (while leaving the taxonomic composition unchanged) results in double the DNA yield. In this case, these DNA-based methods would work just as well with respect to bias not affecting fold-change estimates as their cell-based counterparts.

Perhaps even easier to see why this is the "ideal" situation (and perhaps also a common default assumption when thinking about microbiome experiments), is to think in terms of

the factors $F'_{ij} = B'_i C'_j$. Intuitively, we would like the factor increase for a taxon to be the same across samples; this situation corresponds to $C_j = c$ being constant across samples. (The value of c depends on the experimental details but also how we've set $B'_j = 1$.)

NOTE: Also need to consider whether it is true that the relative abundances are biased by a constant amount; and remember this is am important assumption I'm making as well, and we have some evidence that it can break down. But the motivation for how I've done it is that it seems reasonable to worry about saturation in the extraction yields that is independent of taxa once we account for variation in extraction efficiency and genome size that is simply due to the elution step. May need to give this motivation earlier.

References

Brooks, J Paul, David J Edwards, Michael D Harwich, Maria C Rivera, Jennifer M Fettweis, Myrna G Serrano, Robert A Reris, et al. 2015. "The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies." *BMC Microbiol.* 15 (1): 66. https://doi.org/10.1186/s12866-015-0351-6.

Contijoch, Eduardo J, Graham J Britton, Chao Yang, Ilaria Mogno, Zhihua Li, Ruby Ng, Sean R Llewellyn, et al. 2019. "Gut microbiota density influences host physiology and is shaped by host and microbial factors." *Elife* 8 (January). https://doi.org/10.7554/eLife.40553.

Fettweis, Jennifer M., Myrna G. Serrano, Jamie Paul Brooks, David J. Edwards, Philippe H. Girerd, Hardik I. Parikh, Bernice Huang, et al. 2019. "The vaginal microbiome and preterm birth." *Nat. Med.* 25 (6): 1012–21. https://doi.org/10.1038/s41591-019-0450-2.

Kevorkian, Richard, Jordan T Bird, Alexander Shumaker, and Karen G Lloyd. 2018. "Estimating Population Turnover Rates by Relative Quantification Methods Reveals Microbial Dynamics in Marine Sediment." *Appl. Environ. Microbiol.* 84 (1): e01443–17. https://doi.org/10.1128/AEM.01443-17.

Leopold, Devin R, and Posy E Busby. 2020. "Host Genotype and Colonist Arrival Order Jointly Govern Plant Microbiome Composition and Function." *Curr. Biol.* 30 (16): 3260–3266.e5. https://doi.org/10.1016/j.cub.2020.06.011.

Lloyd, Karen G., Jordan T. Bird, Joy Buongiorno, Emily Deas, Richard Kevorkian, Talor Noordhoek, Jacob Rosalsky, and Taylor Roy. 2020. "Evidence for a Growth Zone for Deep-Subsurface Microbial Clades in Near-Surface Anoxic Sediments." *Appl. Environ. Microbiol.* 86 (19): 1–15. https://doi.org/10.1128/AEM.00877-20.

Martino, Cameron, James T. Morton, Clarisse A. Marotz, Luke R. Thompson, Anupriya Tripathi, Rob Knight, and Karsten Zengler. 2019. "A Novel Sparse Compositional Technique Reveals Microbial Perturbations." *mSystems* 4 (1): e00016–19. https://doi.org/10.1128/mSystems.00016-19.

McLaren, Michael R, Amy D Willis, and Benjamin J Callahan. 2019. "Consistent and correctable bias in metagenomic sequencing experiments." *Elife* 8 (September): 46923. https://doi.org/10.7554/eLife.46923.

Props, Ruben, Frederiek-Maarten Kerckhof, Peter Rubbens, Jo De Vrieze, Emma Hernandez Sanabria, Willem Waegeman, Pieter Monsieurs, Frederik Hammes, and Nico Boon. 2017. "Absolute quantification of microbial taxon abundances." *ISME J.* 11 (2): 584–87. https://doi.org/10.1038/ismej.2016.117.

Quinn, Thomas P., and Ionas Erb. 2020. "Interpretable Log Contrasts for the Classification of Health Biomarkers: a New Approach to Balance Selection." *mSystems* 5 (2): 1–11. https://doi.org/10.1128/mSystems.00230-19.

Quinn, Thomas P., Ionas Erb, Greg Gloor, Cedric Notredame, Mark F. Richardson, and Tamsyn M. Crowley. 2019. "A field guide for the compositional analysis of any-omics data." *Gigascience* 8 (9): 1–14. https://doi.org/10.1093/gigascience/giz107.

Rivera-Pinto, J., J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle. 2018. "Balances: a New Perspective for Microbiome Analysis." *mSystems* 3 (4): e00053–18. https://doi.org/10.1128/mSystems.00053-18.

Tettamanti Boshier, Florencia A., Sujatha Srinivasan, Anthony Lopez, Noah G. Hoffman, Sean Proll, David N. Fredricks, and Joshua T. Schiffer. 2020. "Complementing 16S rRNA Gene Amplicon Sequencing with Total Bacterial Load To Infer Absolute Species Concentrations in the Vaginal Microbiome." *mSystems* 5 (2): 1–14. https://doi.org/10.1128/mSystems.00777-19.

Vandeputte, Doris, Gunter Kathagen, Kevin D'hoe, Sara Vieira-Silva, Mireia Valles-Colomer, João Sabino, Jun Wang, et al. 2017. "Quantitative microbiome profiling links gut community variation to microbial load." *Nature* 551 (7681): 507. https://doi.org/10.1038/nature24460.

Vieira-Silva, Sara, João Sabino, Mireia Valles-Colomer, Gwen Falony, Gunter Kathagen, Clara Caenepeel, Isabelle Cleynen, Schalk van der Merwe, Séverine Vermeire, and Jeroen Raes. 2019. "Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses." *Nat. Microbiol.* 4 (11): 1826–31. https://doi.org/10.1038/s41564-019-0483-9.

Washburne, Alex D., Justin D. Silverman, Jonathan W. Leff, Dominic J. Bennett, John L. Darcy, Sayan Mukherjee, Noah Fierer, and Lawrence A. David. 2017. "Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets." *PeerJ* 5 (February): e2969. https://doi.org/10.7717/peerj.2969.

Wasserman, Larry. 2004. *All of Statistics*. Springer Texts in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-21736-9.

Yeh, Yi-Chun, David M. Needham, Ella T. Sieradzki, and Jed A. Fuhrman. 2018. "Taxon Disappearance from Microbiome Analysis Reinforces the Value of Mock Communities as a Standard in Every Sequencing Run." *mSystems* 3 (3): e00023–18. https://doi.org/10.1128/mSystems.00023-18.