Consistent and correctable bias in metagenomic sequencing experiments

Michael R. McLaren[1], Amy D. Willis[2], and Benjamin J. Callahan[1,3,*]

[1]Department of Population Health and Pathobiology, North Carolina State University, Raleigh, NC 27607
[2]Department of Biostatistics, University of Washington, Seattle, WA 98195
[3]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695
[*]Correspondence should be addressed to benjamin.j.callahan@gmail.com

Abstract

Measurements of biological communities by marker-gene and metagenomic sequencing are biased: The measured relative abundances of taxa or their genes are systematically distorted from their true values because each step in the experimental workflow preferentially detects some taxa over others. Bias can lead to qualitatively incorrect conclusions and makes measurements from different protocols quantitatively incomparable. A rigorous understanding of bias is therefore essential. Here we propose, test, and apply a simple mathematical model of how bias distorts marker-gene and metagenomics measurements: Bias multiplies the true relative abundances within each sample by taxon- and protocol-specific factors that describe the different efficiencies with which taxa are detected by the workflow. Critically, these factors are consistent across samples with different compositions, allowing bias to be estimated and corrected. We validate this model in 16S rRNA gene and shotgun metagenomics data from bacterial communities with defined compositions. We use it to reason about the effects of bias on downstream statistical analyses, finding that analyses based on taxon ratios are less sensitive to bias than analyses based on taxon proportions. Finally, we demonstrate how this model can be used to quantify bias from samples of defined composition, partition bias into steps such as DNA extraction and PCR amplification, and to correct biased measurements. Our model improves on previous models by providing a better fit to experimental data and by providing a composition-independent approach to analyzing, measuring, and correcting bias.

Introduction

Marker-gene and metagenomic sequencing (jointly, MGS) have transformed the study of biological communities. Extracting and sequencing total DNA from a community can identify thousands of taxa along with their genes and potential functions, while sequencing a phylogenetic marker gene (e.g., 16S rRNA) can quantify taxon abundances (Li 2015; Quince et al. 2017). MGS measurements of microbial communities are yielding fundamental new insights into the structure and dynamics of microbial ecosystems and the roles of microbes as drivers of host and ecosystem health (Callahan et al. 2017; Graham et al. 2016; Knight and Sogin 2017; Lehman et al. 2015; Zeevi et al. 2015). Applications of MGS, often under the alternative terms eDNA sequencing or metabarcoding, increasingly extend beyond microbes to the measurement and monitoring of plants, insects, and vertebrates (Bell et al. 2018; Krehenwinkel et al. 2017; Thomas et al. 2016). MGS methods are now being adopted in fields ranging from food safety (Cocolin et al. 2018) to wastewater remediation (Rosso et al. 2018) to forensics (Metcalf et al. 2017) along with biology and medicine. Unfortunately, however, the community compositions measured by MGS are wrong.

MGS measurements are *biased*: The measured relative abundances of the taxa and genes in the sample are systematically distorted from their true values (Brooks 2016; Sinha et al. 2017). Bias arises because each step in an experimental MGS workflow preferentially detects (i.e., preserves, extracts, amplifies, sequences, or bioinformatically identifies) some taxa over others (Brooks 2016; Hugerth and Andersson 2017; Pollock et al. 2018). For example, bacterial species differ in how easily they are lyzed and therefore how much DNA they yield during DNA extraction (Costea et al. 2017; Morgan et al. 2010), and they differ in their number of 16S rRNA gene copies and thus how much PCR product we expect to obtain per cell (Kembel et al. 2012). Most sources of bias are protocol-dependent: Different PCR primers preferentially amplify different sets of taxa (Sipos et al. 2007), different extraction protocols can produce 10-fold or greater differences in the measured proportion of a taxon from the same sample (Costea et al. 2017), and almost every choice in an MGS experiment has been implicated as contributing to bias (Hugerth and Andersson 2017; Pollock et al. 2018; Sinha et al. 2017). Every MGS experiment is biased to some degree, and measurements from different protocols are quantitatively incomparable (Gibbons et al. 2018; Hiergeist et al. 2016; Mallick et al. 2017; Nayfach and Pollard 2016; Sinha et al. 2017).

The biases of MGS protocols and the error those biases introduce remain unknown. Thus we do not know whether the measured taxonomic or gene compositions derived from MGS are accurate, or to what extent the biological conclusions derived from them are valid. It is common to assume that conclusions drawn from measurements using the same protocol are robust to MGS bias. But simulated examples have shown that bias can lead to qualitatively incorrect conclusions about which taxa dominate different samples (Edgar 2017; Kembel et al. 2012), which ecosystems are more similar (Kembel et al. 2012), and which taxa are associated with a given disease (Brooks 2016). Furthermore, variation in bias limits our ability to make the direct comparisons between results from different experiments that are central to the scientific process. It has been suggested that these issues would be circumvented if bias were the same in every experiment, leading to a number of efforts to define and promulgate standardized MGS protocols (Costea et al. 2017; Gilbert et al. 2014). However, methodological standardization has several limitations. For example, it can be overly

restrictive given the variety of ecosystems and biological questions where MGS methods are applied as well as the continual advance in technology, and unmeasured technical variability can introduce experiment-specific biases into nominally standardized methods (Yeh et al. 2018). More important, standardized protocols remain biased and thus still do not provide accurate measurements of the underlying communities.

Current attempts to counter bias are limited and of unknown efficacy because of our poor understanding of how bias across the full experimental workflow distorts MGS measurements. Hundreds of published studies compare MGS measurements of defined samples to their expected composition in an effort to characterize the bias of the given protocol (many cited in Hugerth and Andersson (2017) and Pollock et al. (2018)). But this approach has limited value so long as we do not know how the error observed in one sample translates to differently composed samples. If we understood how bias acts across samples we might be able to estimate the effect of bias from measurements of samples of defined composition and use those estimates to calibrate measurements of samples of interest to their true values (Hardwick et al. 2017; Thomas et al. 2016). Alternatively, natural communities measured by multiple experiments could be used to calibrate measurements between experiments using different protocols. A quantitative understanding of how bias distorts MGS measurements would also elucidate how statistical analyses and diagnostics are affected by bias and suggest more robust alternatives.

Here we propose and test a mathematical model of how bias distorts taxonomic compositions measured by MGS from their true values. In our model, bias manifests as a multiplication of the true relative abundances by taxon- and protocol-specific factors that are constant across samples of varying compositions. We validate key components of this model, including that bias acts independently on each taxon in a sample, in 16S rRNA gene and shotgun metagenomic sequencing data from bacterial communities of defined composition. We use our proposed model to quantify bias, to partition bias into steps such as DNA extraction and PCR amplification, and to reason about the effects of bias on downstream statistical analyses. Finally, we demonstrate how this model can be used to correct biased MGS measurements when suitable controls are available.

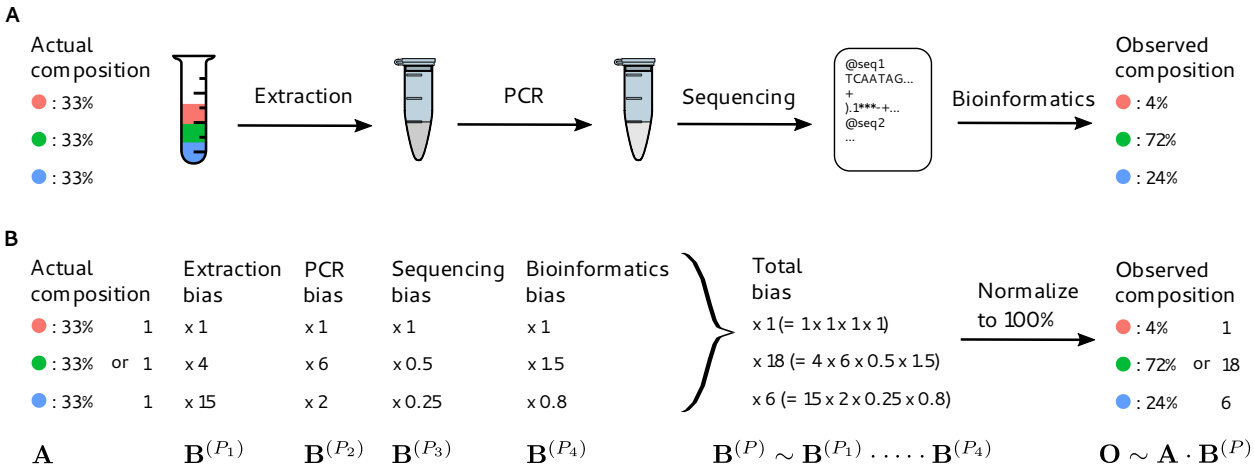

Figure 1: **Bias arises throughout an MGS workflow, creating systematic error between the observed and actual compositions.** Panel A illustrates a hypothetical measurement of an even mixture of three taxa (denoted by the colors red, green, and blue), in which differential detection among taxa biases the observed composition. Panel B illustrates our mathematical model of bias, in which bias multiplies across steps to create the total systematic error in the measured composition.

Results

A mathematical model of MGS bias

Consider a marker-gene or metagenomic sequencing (MGS) experiment as a multi-step transformation that takes as input biological material and provides as output the taxonomic profile corresponding to each sample, i.e. the set of detected taxa and their associated relative abundances (Figure 1A). Each step introduces systematic and random errors that cumulatively lead to error in the measured taxonomic profiles. *Bias* is a particular, ubiquitous form of systematic error that arises from the different efficiencies with which various taxa are detected (i.e., preserved, extracted, amplified, sequenced, or bioinformatically identified and quantified) at each step.

Many bias mechanisms act are thought to act multiplicatively on the taxon abundances, at least to first approximation. For instance, the DNA concentration of a taxon after DNA extraction equals its initial cell concentration multiplied by its DNA yield per cell. This per-cell yield indicates the efficiency of extraction for the taxon, which is is expected to depend on factors such as genome size and the structure of the taxon's cell wall (Morgan et al. 2010). Therefore, we expect extraction efficiencies to vary among taxa, but be approximately constant for any specific taxon across samples treated with the same protocol. Other multiplicative sources of bias include variation in PCR binding and amplification efficiencies (Edgar 2017; Polz and Cavanaugh 1998; Suzuki and Giovannoni 1996; Wagner et al. 1994) and in marker-gene copy number (Kembel et al. 2012).

Inspired by these observations, we propose that at every step in an MGS experiment, the output abundances of individual taxa differ from the input abundances by taxon-specific multiplicative factors (Figure 1), which we refer to as the taxon efficiencies in that step. Taxon efficiencies are determined by the interaction between the experimental protocol and the biological/chemical/physical/informatic state of each taxon in that step, and are there-

fore independent of the composition of the sample. Typical MGS experiments only measure relative abundances, and the change in the *relative abundances* during a step depend only on the *relative efficiencies*. This yields the following mathematical model of bias (Figure 1): **The relative abundances measured in an MGS experiment are equal to the input relative abundances multiplied by taxon-specific but composition-independent factors (the relative efficiencies) at every step.**

The mathematical accounting of bias is simplified by the use of *compositional vectors*: vectors for which only the ratios among elements carry meaning. The relative abundances and relative efficiencies can be described as compositional vectors with K non-negative elements, where K is the number of possible taxa. Two vectors $\mathbf{X}$ and $\mathbf{Y}$ are *compositionally equivalent*, denoted $\mathbf{X} \sim \mathbf{Y}$, if $\mathbf{X} = a\mathbf{Y}$ for some positive constant a (Boogaart et al. 2013) because the ratios among the elements of $\mathbf{X}$ and $\mathbf{Y}$ are the same: $X_i/X_j = aY_i/aY_j$. A compositional vector $\mathbf{X}$ of relative abundances can be converted to proportions, which we denote $\Pr(\mathbf{X})$, by dividing the taxon abundances by their sum, $\Pr(\mathbf{X}) = \mathbf{X}/\sum_i X_i$, without changing its meaning in terms of the ratios among taxa. For example, the vector of observed proportions in Figure 1 of (4%, 72%, 24%) is compositionally equivalent to the vector $(1, 18, 6)$ obtained by dividing all abundances by that of the first (red) taxon.

For a given sample, let $\mathbf{A}$ be the vector of actual relative abundances, $\mathbf{O}$ be the vector of observed (measured) relative abundances, and A_i the actual relative abundance of taxon i. Similarly, let $\mathbf{B}^{(P_l)}$ be the vector of the relative efficiencies of each taxon at step l in protocol P (see Methods for details). Our model of bias can be stated mathematically as

$$\mathbf{O} \sim \mathbf{A} \cdot \mathbf{B}^{(P_1)} \cdot \mathbf{B}^{(P_2)} \cdot \cdots \cdot \mathbf{B}^{(P_L)}, \tag{1}$$

where $\cdot$ denotes element-wise multiplication of two vectors. We define the bias of protocol P by the product over all steps, $\mathbf{B}^{(P)} \sim \mathbf{B}^{(P_1)} \cdot \mathbf{B}^{(P_2)} \cdot \cdots \cdot \mathbf{B}^{(P_L)}$. The observed composition is then simply the actual composition multiplied by the protocol bias,

$$\mathbf{O} \sim \mathbf{A} \cdot \mathbf{B}^{(P)}. \tag{2}$$

When considering samples measured by the same protocol, we will drop the superscript P and simply refer to the total protocol bias as $\mathbf{B}$.

From Equation (2) we see that the ratio between the observed relative abundances of any two taxa i and j is

$$\frac{O_i}{O_j} = \frac{A_i B_i}{A_j B_j}. \tag{3}$$

and the observed proportion of taxon i is

$$\Pr(\mathbf{O})_i = \frac{O_i}{\sum_{j=1}^{K} O_j} = \frac{\Pr(\mathbf{A})_i B_i}{\sum_{j=1}^{K} \Pr(\mathbf{A})_j B_j}. \tag{4}$$

The denominator, $\sum_{j=1}^{K} \Pr(\mathbf{A})_j B_j$, is the *sample mean efficiency*—the average efficiency of the sampled individuals.

The systematic error in the measured composition under our model is $\mathbf{O}/\mathbf{A} \sim \mathbf{B}$, where / denotes element-wise division and is referred to as the compositional difference (Aitchison

1992). The compositional difference unites the experimental notion of bias—variation in the efficiencies with which different taxa are detected—with the statistical notion of bias—the difference between the expected value of an estimate and the true value—with the understanding we are considering the compositional difference rather than the conventional Euclidean difference between compositions.

Properties and implications of the model

Bias is fully described by the relative efficiencies of the total workflow: The bias of individual steps only influences the measurement error through their product, $\mathbf{B}^{(P)}$. Consequently, knowledge of the total protocol bias is sufficient to determine how bias affects the measured taxonomic profiles even if the biases of the individual protocol steps (the $\mathbf{B}^{(P_l)}$) remain unknown. The bias $\mathbf{B}^{(P)}$ has just K-1 parameters, denoting the relative efficiencies with which the K taxa of interest are detected by the protocol as a whole, and fully describes the effect of bias on measurements of those K taxa in all samples.

Systematic error in taxon ratios, but not in taxon proportions, is independent of sample composition: The fold-error in the observed ratios of the abundances of taxon i and taxon j relative to the actual ratio in their abundances is $(O_i/O_j)/(A_i/A_j) = B_i/B_j$ (see Equation (3)). This error depends only on the ratio between the total protocol efficiencies of taxon i and taxon j, and is independent of the rest of the sample. Critically, this means that the systematic error in taxon ratios caused by bias will remain the same in samples of varying composition.

In contrast, the error in the proportion of a taxon depends on the sample composition. The fold-error in the observed proportion of taxon i relative to its actual proportion is

$$\frac{\Pr(\mathbf{O})_i}{\Pr(\mathbf{A})_i} = \frac{B_i}{\sum_{j=1}^{K} \Pr(\mathbf{A})_j B_j} \tag{5}$$

This error depends on the sample mean efficiency $\sum_j \Pr(\mathbf{A})_j B_j$ and thus depends on the proportions of all the other taxa in the sample. Intuitively, bias leads to over-detection of taxa that are more easily detected than average in the given sample. As a result, the same taxon can be over-detected in samples dominated by low-efficiency taxa and under-detected in samples dominated by high-efficiency taxa.

To illustrate, we consider the hypothetical measurement of a second community sample (Sample S_2 in Figure 2) alongside that of the even sample from Figure 1 (Sample S_1 in Figure 2). The dominance of the low-efficiency red taxon in Sample S_2 substantially lowers its sample mean efficiency compared to the even-mixture Sample S_1, changing the fold-error in all taxon proportions. In particular, the blue taxon changes from having a lower-than-average efficiency in Sample S_1 to a higher-than-average efficiency in Sample S_2. As a result, its observed proportion is lower than its actual proportion in Sample S_1, but higher than its actual proportion in Sample S_2! Yet the fold-error in the ratios among taxa is identical in both samples and equal to the bias (Figure 2, bottom row).

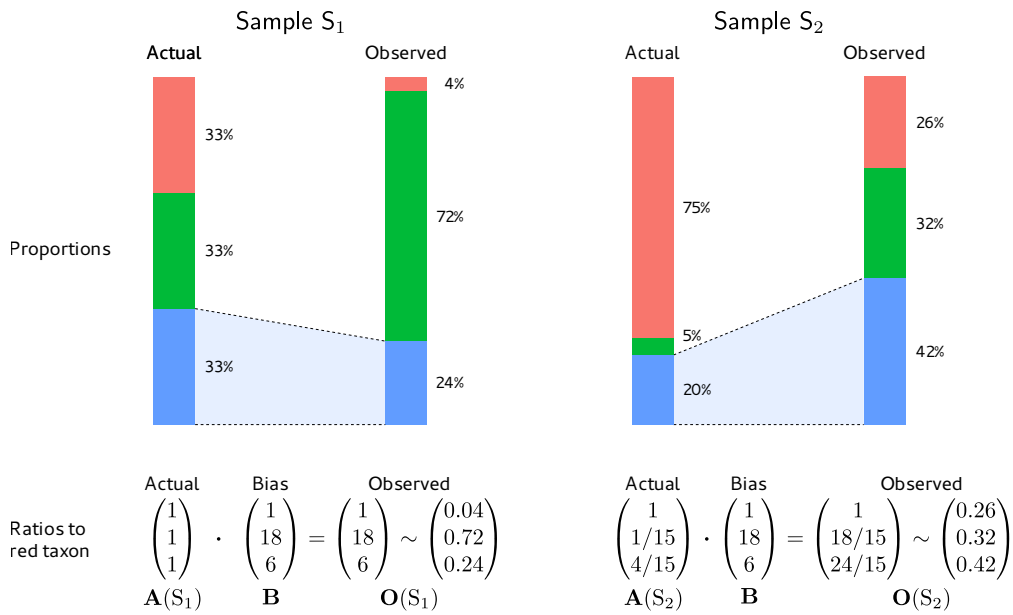

Figure 2: **Consistent multiplicative bias causes systematic error in taxon ratios, but not taxon proportions, that is independent of sample composition.** The even community from Figure 1 and a second uneven community containing the same three taxa in different proportions are measured by a common MGS protocol. Measurements of both samples are subject to the same bias, but the magnitude and direction of error in the taxon proportions depends on the underlying composition (top row). In contrast, when the relative abundances and bias are both viewed as ratios to a fixed taxon (here, the red taxon), the consistent action of bias across samples is apparent (bottom row).

Analyses based on fold-changes in taxon ratios are insensitive to bias, while analyses based on taxon proportions can give spurious results: Although it is widely understood that bias distorts individual community profiles, it is often thought to effectively "cancel out" when analyzing the differences between samples that have been measured by the same protocol. Unfortunately, simulating measurement under our model easily provides examples where common analyses give qualitatively incorrect results. In Figure 2, for example, bias causes the red-dominated Sample S_2 to appear to have a more even distribution of taxa than the actually-even Sample S_1. As a result, any analysis of alpha diversity that incorporates evenness (e.g., the Shannon or Inverse Simpson indices) will incorrectly conclude that Sample S_2 is more diverse. The previous section provides a general explanation as to why, for many analyses, bias does not simply cancel: The underlying statistics are functions of the individual taxon proportions, the error of which varies inconsistently across samples. Consequently, proportion-based analyses can lead to qualitatively incorrect conclusions. As a further example, the actual proportion of the blue taxon decreases from Sample S_1 to Sample S_2 in Figure 2, but the measured proportion increases!

In contrast, the fold-error in taxon ratios is independent of sample composition, and fold-changes in taxon ratios across samples are insensitive to bias. Consider the fold-change in the ratio of a pair of taxa i and j between two samples we call s and t. Following Equation (3), the observed change is

$$\frac{O_i(s)}{O_j(s)} \Big/ \frac{O_i(t)}{O_j(t)} = \frac{A_i(s)B_i}{A_j(s)B_j} \Big/ \frac{A_i(t)B_i}{A_j(t)B_j} = \frac{A_i(s)}{A_j(s)} \Big/ \frac{A_i(t)}{A_j(t)}, \tag{6}$$

and thus equals the true change. That is, *the fold-change in taxon ratios between samples is invariant to bias*. More generally, the compositional difference between samples is invariant to multiplication by a fixed vector (Aitchison 1992) and thus to bias,

$$\mathbf{O}(s) \,/\, \mathbf{O}(t) \sim (\mathbf{A}(s) \cdot \mathbf{B}) \,/\, (\mathbf{A}(t) \cdot \mathbf{B}) \sim \mathbf{A}(s) \,/\, \mathbf{A}(t). \tag{7}$$

Returning to the samples in Figure 2, the actual and observed ratios of the green to the red taxon both change by the same factor of 1/15 from Sample S_1 to Sample S_2, and the actual and observed compositional difference between samples is $(1, 1/15, 4/15)$. Equation (7) shows that any analysis that depends only on the compositional differences between samples will be invariant to bias under our model.

The systematic difference between measurements from different protocols is given by the difference in their biases: Consider protocol P with bias $\mathbf{B}^{(P)}$ and reference protocol R with bias $\mathbf{B}^{(R)}$. If both protocols measure the same sample with actual composition $\mathbf{A}$, the compositional difference between their measurements is

$$\mathbf{O}^{(P)}/\mathbf{O}^{(R)} \sim \mathbf{A} \cdot \mathbf{B}^{(P)}/(\mathbf{A} \cdot \mathbf{B}^{(R)}) = \mathbf{B}^{(P)}/\mathbf{B}^{(R)}. \tag{8}$$

The actual composition drops from the equations and the difference in their measurements is simply the compositional difference in the biases of each protocol, which we refer to as the differential bias $\mathbf{B}^{(P/R)} \equiv \mathbf{B}^{(P)}/\mathbf{B}^{(R)}$ of protocol P relative to the reference R. Measurements on common samples are related to one another by $\mathbf{O}^{(P)} \sim \mathbf{O}^{(R)} \cdot \mathbf{B}^{(P/R)}$, independent of the actual composition of the sample. Usefully, differential bias is mathematically equivalent to bias if we consider the "reference" compositions measured by protocol R as the truth.

Estimates of bias from control samples can be used to correct measurements of other samples: The consistency of bias across samples makes it possible to estimate bias from samples of known composition, referred to as *calibration controls*, and to use that estimate $\hat{\mathbf{B}}$ to *calibrate*, or remove the bias from, measurements of other samples with unknown compositions. A point estimate of the bias of K taxa present with known relative abundances in control sample c is given by the compositional difference between the observed and actual compositions, $\hat{\mathbf{B}} \sim \mathbf{O}(c)/\mathbf{A}(c)$. We have also developed a more general point-estimation procedure that combines measurements from multiple controls with at least partially overlapping sets of taxa into a single estimate of bias across all taxa present in any control by maximizing the explained compositional error in the control measurements (Methods).

Once bias has been estimated for a set of taxa, it can be used to calibrate the relative abundances of those taxa in an unknown sample. Letting $\mathbf{O}$ denote the measured relative abundances for these taxa, the estimate $\hat{\mathbf{A}}$ of the actual relative abundances is

$$\hat{\mathbf{A}} \sim \mathbf{O}/\hat{\mathbf{B}}. \tag{9}$$

That is, the calibrated abundances are found by compositionally subtracting the estimated bias from the original measurement. Through its use of compositional vectors, Equation (9) automatically accounts for differences in composition between the controls and the target

sample. Calibrated estimates of the true taxon proportions are obtained by normalizing the elements of $\hat{\mathbf{A}}$ to sum to 1.

An alternative form of calibration we call *reference calibration* can be performed using control samples whose true composition is unknown but that have been measured by a reference protocol R. Estimation and calibration proceed as before but with the control composition $\mathbf{A}(c)$ replaced by the reference measurement $\mathbf{O}^{(R)}(c)$. In this case, the calibrated composition is an estimate of the measurement we would expect if the target sample had been measured by the reference protocol.

Testing the model with mock communities

We tested our model of bias in data from two studies, Brooks et al. (2015) and Costea et al. (2017), that evaluated the bias of marker-gene and shotgun metagenomic sequencing, respectively, using mock microbial communities in samples of varying composition.

Marker-gene sequencing of even mixtures of various bacterial species

Brooks et al. (2015) generated taxonomic profiles from 71 samples of 58 unique mock communities by amplicon sequencing the V1-V3 region of the 16S rRNA gene. Each unique mock community consisted of an even mixture of between two and seven bacterial species. Each sample was measured in three experiments employing a common experimental workflow, but beginning from different starting points: even mixtures of cells, of extracted DNA, and of PCR product. The authors reported large systematic errors in the taxon proportions measured from the cell and DNA mixtures, which they explained in part by a highly parameterized linear model with many interaction terms. Here we re-analyze the data from this study (Methods and SI R markdown) in order to evaluate our model of bias and its performance relative to alternatives.

The proportions measured from the cell-mixture mock communities differed greatly from the expected even proportions of each species (Figure 3A). The ratios between pairs of species also diverged sharply from the ratio of 1 expected in these even mixtures (Figure 3D). However, and as predicted by our model (see Properties and Implications), the error in the ratios was consistent across samples (Figure 3D) while the error in the proportions varied dramatically in both magnitude and direction (Figure 3C).

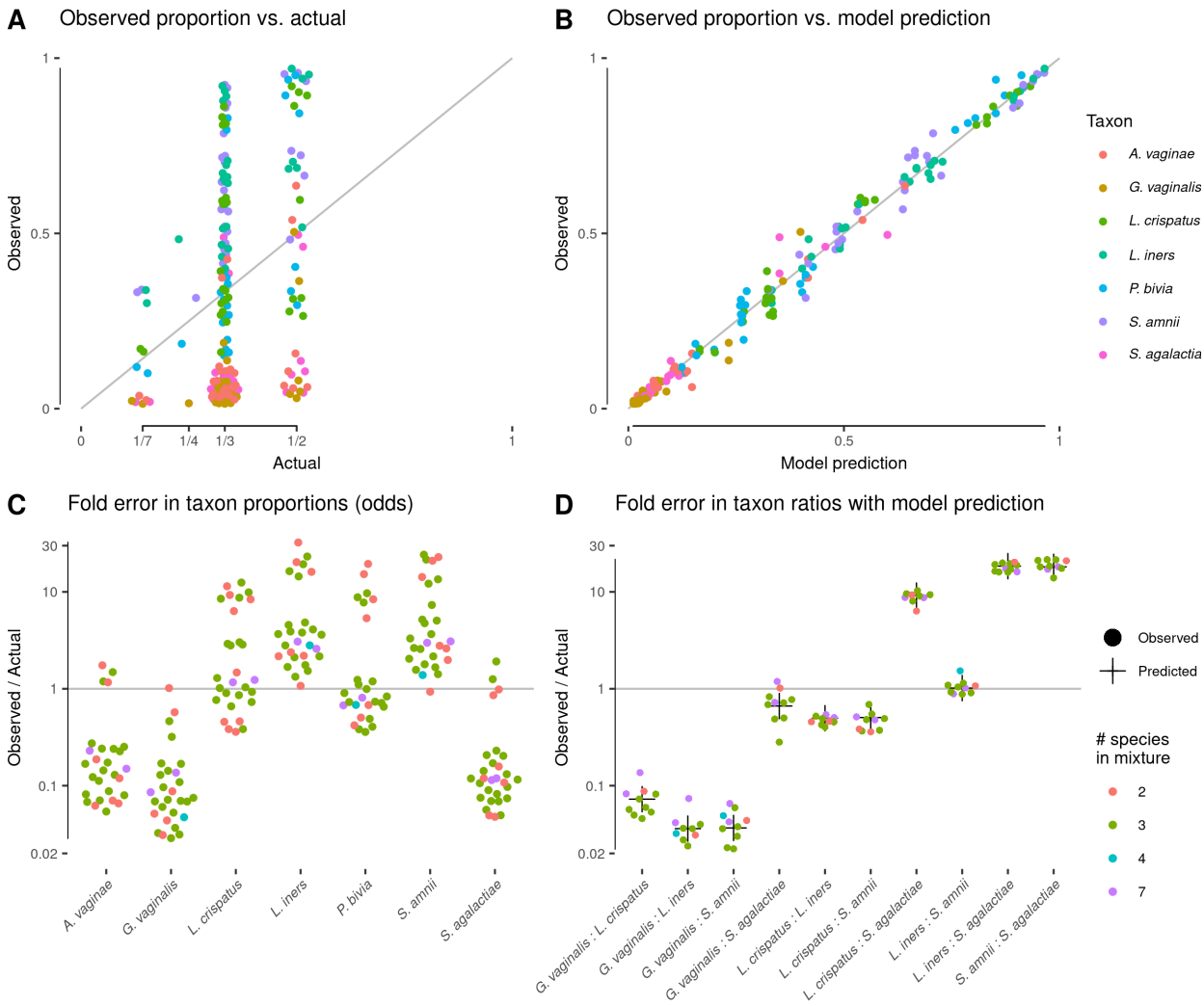

Figure 3: **Our model of bias explains the systematic error observed in the Brooks et al. (2015) cell-mixture experiment.** The top row compares the observed proportions of individual species to the actual proportions (Panel A) and to those predicted by our fitted bias model (Panel B). Panel A shows significant error across all species and mixture types that is almost entirely removed once bias is accounted for in Panel B. Panel C shows the observed error in proportions of individual species, while panel D shows the error in the ratios of pairs of species for five of the seven species. The ratio predicted by the fitted model is given by the black cross in panel D. As predicted by our model, the error in individual proportions (panel C) depends highly on sample composition, while the error in ratios (panel D) does not.

Our model explained almost all of the error in the measured compositions of the cell mixtures. We estimated bias from all samples by a simple point-estimation procedure (Methods; Table 1). We then used the estimated bias to predict the observed compositions from the expected even mixtures using Equation (2). The measured pairwise ratios closely matched the ratios predicted by our model, i.e. the ratios of the biases (or efficiencies) of the two species (black crosses in Figure 3D). The proportions predicted from the fitted model reduced the mean-squared error by 98.8% (SI R markdown) and closely matched the observed proportions (Figure 3B).

The DNA and PCR-product experiments confirmed that our model can also effectively

describe partial MGS workflows. The compositions measured from DNA mixtures were affected by large systematic errors that were well explained by our model, while the systematic error in compositions measured from the PCR mixtures was small compared to the random errors (SI Figure 6). Notably, the bias in the PCR experiment substantially differed from the bias in the cell-mixture experiment (Table 1) which also includes DNA extraction, suggesting that PCR and DNA extraction are both large, independent sources of bias acting in accordance with our model.

Our model better explains the data from the cell and DNA mixtures than previously proposed models, while employing a small number (6) of parameters. A common approach to modeling bias has been to model the observed proportion of a taxon as a linear function of its actual proportion. However, such models cannot explain that the observed proportion of a given taxon can be higher than or lower than its actual proportion in samples of different composition (e.g. *L. crispatus* in Figure 3C), while such behavior is a straightforward consequence of our model. Brooks et al. (2015) attempted to overcome this limitation by adding second and third-order interaction terms between taxa. Their model was able to account for the variation in proportional error but at the cost of vastly increased model complexity and a tendency to overfit and poorly predict in differently composed samples because the parameter estimates are sensitive to the specific compositions on which the model was trained (SI R markdown).

Metagenomic sequencing of fecal samples with a spike-in mock community

In their Phase III experiment, Costea et al. (2017) used shotgun metagenomic sequencing to generate taxonomic profiles from eight fecal specimens to which a fixed amount of a cellular mock community (the spike-in) had been added. A control specimen (M) containing only the ten-species mock community was also characterized. DNA was extracted from each specimen using three distinct DNA extraction protocols (H, Q, and W), and shotgun sequencing was performed on the extracted samples by a common protocol.

There was substantial variation across specimens and DNA extraction protocols in the measured composition of the fecal taxa, and in the proportion of sequencing reads derived from the spike-in mock community (Figure 4A). Protocol W generated substantially more sequencing reads from the mock community than did H or Q, likely contributing to the lower variance of the relative abundances of the mock taxa measured by Protocol W (Figure 4B). In this experiment, the composition of the fecal taxa varied among specimens, but the subcomposition comprising the mock taxa was constant. This was reflected in the consistency of the measured relative abundances within the mock taxa for a given protocol. However, those measured relative abundances varied substantially between protocols, and included qualitative differences such as certain mock taxa being enriched by one protocol and diminished by another (Figure 4B). These results are consistent with a unique bias associated with each protocol acting in accordance to our model. The consistency of bias across specimens, including the mock-only specimen, shows that it is independent of the overall sample compositions that included varying fecal taxa in varying proportions.

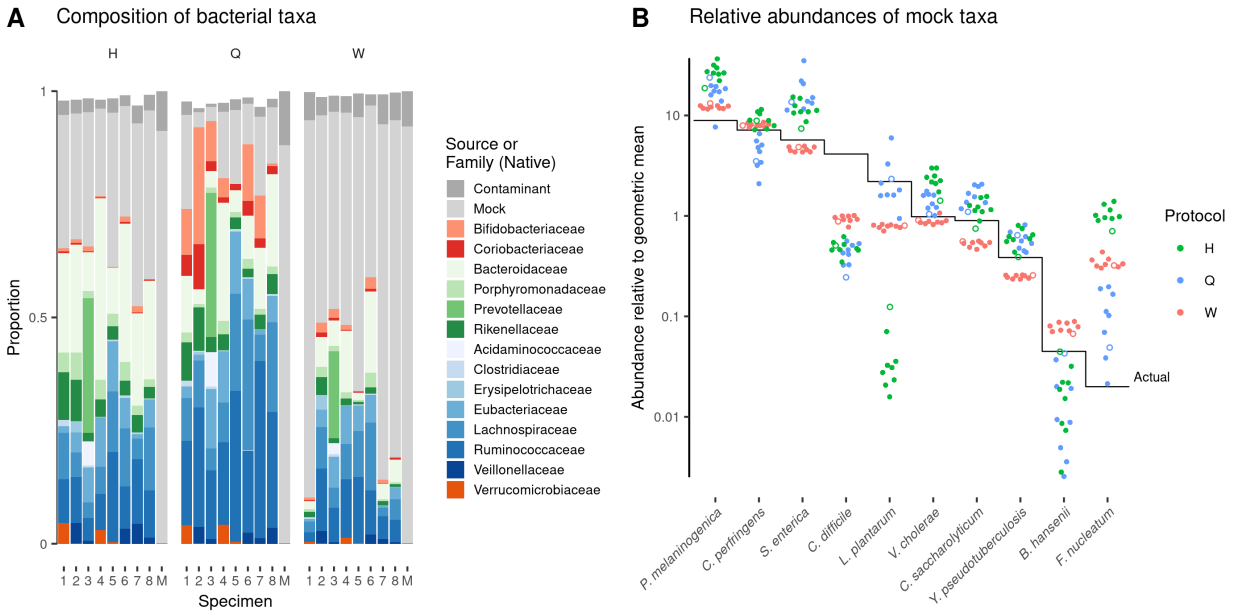

Figure 4: **Bias of the mock spike-in taxa is consistent in the Costea et al. (2017) shotgun dataset despite variation among samples in mock proportion and in composition of native taxa.** Panel A shows the variation in bacterial composition across protocols and specimens (labels 1 through 8 denote fecal specimens and M the mock-only specimen). Composition is colored by source (spike-in Mock, spike-in Contaminant, and Native fecal bacteria) and by taxonomy (native bacterial families with proportion at least 0.02 in at least one sample. Colors are grouped by phylum. Red: *Actinobacteria*, Green: *Bacteroidetes*, Blue: *Firmicutes*, Orange: *Verrucomicrobia*). Panel B shows the relative abundance of each mock taxon across samples (points) against the actual composition (black line); abundances are divided by the geometric mean of all mock taxa in the sample.

Applications of the model

Calibration

Under our model bias is consistent across samples, and can therefore be estimated from control sample(s) of known composition (Methods), and that estimate used to calibrate the measured compositions of unknown samples towards their true composition using Equation (9). In the Brooks et al. (2015) cell mixtures, we estimated bias from two samples containing all seven species and calibrated the the measurements of the other 69 samples based on that estimate, thereby reducing the mean-squared error of the proportions in the calibrated samples by 92.6% and the average Bray-Curtis dissimilarity between the actual and observed compositions from 0.35 to 0.08 (SI R markdown). In the Costea et al. (2017) dataset, the measured composition of the spike-in mock community deviated from the truth in protocol-specific fashion (Figure 5, top row). We estimated the bias of each protocol on the mock taxa in three randomly chosen specimens and calibrated samples from the remaining specimens based on those estimates, thereby removing systematic error and greatly increasing the accuracy of the measurements (Figure 5, middle row).

Differential bias between experiments (Equation (8)) can be estimated from samples common to each experiment, even if the actual composition of the common samples is unknown. Differential bias can then be used to calibrate measurements from various experiments to

those of a chosen reference protocol, thereby making measured compositions from different experiments quantitatively comparable even if their fidelity to the true compositions remains unclear. We illustrate calibration to a reference protocol using the multi-protocol design of the Costea et al. (2017) experiment (Figure 5). We defined the measurements by Protocol W as the reference composition that was then used in place of the actual composition in our calibration procedure. This greatly reduced the systematic differences between measurements from different protocols, without necessarily improving the accuracy compared to the actual composition (Figure 5, bottom row).

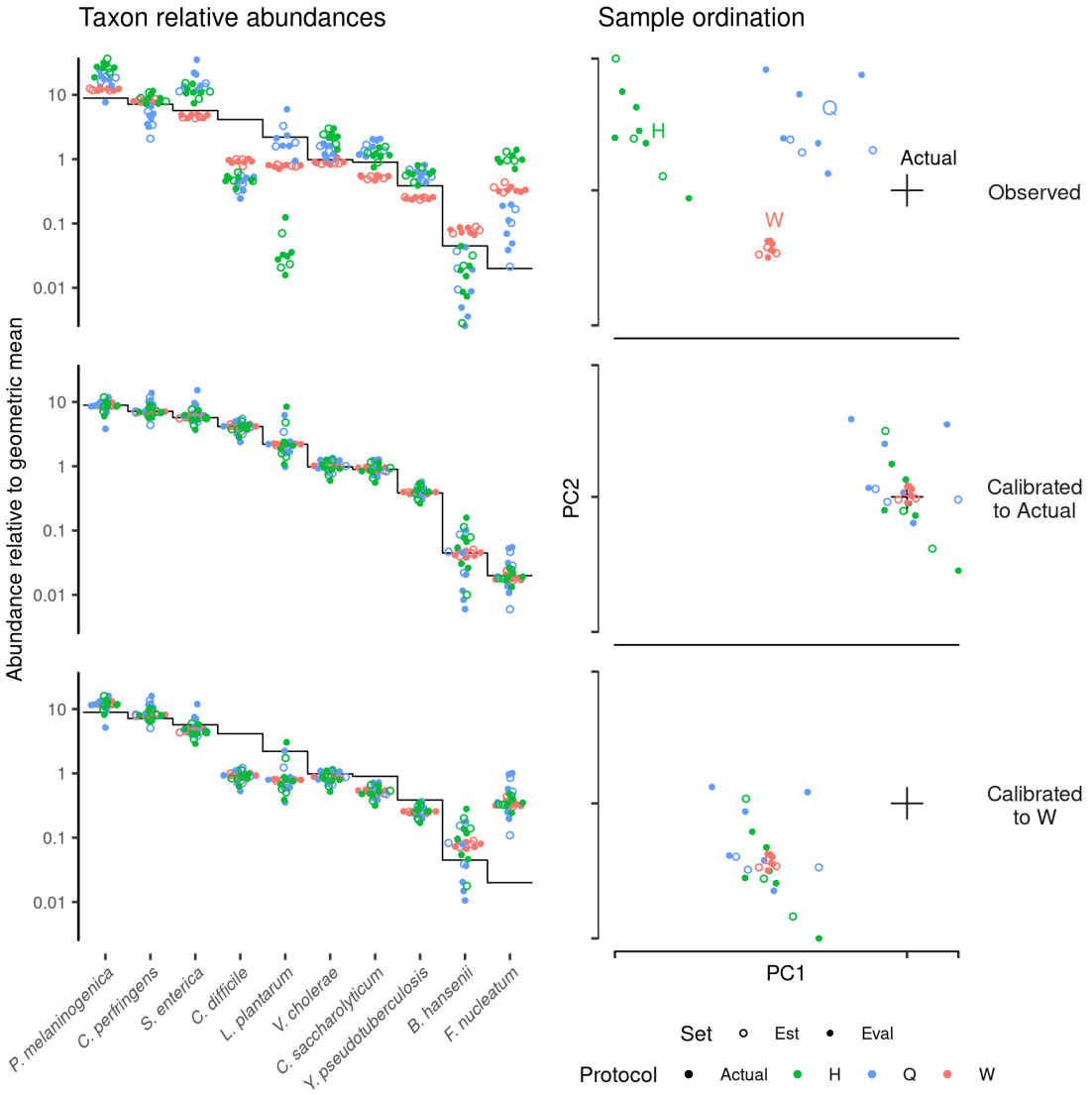

Figure 5: **Calibration can remove bias and make MGS measurements from different protocols quantitatively comparable.** For the sub-community defined by the mock spike-in of the Costea et al. (2017) dataset, we estimated bias from three specimens (the estimation set "Est") and used the estimate to calibrate all specimens. The left column shows taxon relative abundances as in Figure 4B and the right column shows the first two principal components from a compositional principle-components analysis (Gloor et al. 2017). The top row shows the measurements before calibration; the middle, after calibration to the actual composition; and the bottom, after calibration to Protocol W.

Bias measurement as a means of evaluating and improving protocols

Under our model, the compositional vector of relative efficiencies completely describes the effect of bias in samples of any composition, and thus is the correct way to measure and evaluate bias.

One way to use measurements of bias is to select less-biased MGS protocols. For that purpose, we suggest that the magnitude of bias be summarized with ratio-based statistics such as the geometric range and the geometric standard deviation of the relative efficiencies. In the Costea et al. (2017) data, these statistics indicate that Protocol H is substantially more biased than Protocols Q and W, as is evident from the far greater spread among the relative efficiencies of Protocol H (Table 2).

We can also use vectors of relative efficiencies calculated from variations of an MGS protocol to quantify the bias attributable to specific parts of the workflow. In the Brooks et al. (2015) study, the same MGS workflow was run from different starting points: cells, extracted DNA, and PCR product. This design allows bias of parts of the workflow to be quantified by comparing the biases resulting from different starting points (Table 3; Methods). In particular, we estimated the bias of DNA extraction as the compositional difference between the bias in the cell mixtures and the bias in the DNA mixtures, and we estimated the bias of PCR by the compositional difference in bias of the PCR-product mixtures and the DNA mixtures (Table 1). For these taxa and workflow, DNA extraction was the largest single source of bias and divides the taxa into poorly and easily extracted groups (Table 1, SI Figure 8A). PCR bias was also substantial and reinforced the variation in extraction bias within groups (SI Figure 8 B and C). PCR and extraction bias acted in opposite directions for some taxa, such as *L. crispatus* and *P. bivia*, leading to more moderate total efficiencies; and acted in the same direction for others, leading to the lowest and highest total efficiencies for *G. vaginalis* and for *L. iners*, respectively.

In the Costea et al. (2017) study, the same MGS workflow with different DNA extraction protocols was used to measure a common set of samples. This study design allowed us to measure the bias of each individual workflow on the spike-in mock community (Table 2), and also allowed us to measure the difference in the bias, or differential bias, attributable to the DNA extraction step (this includes bias during extraction and possible downstream effects; Methods). Protocol H's bias is substantially larger than its differential bias to the other two protocols (Figure 5 A and B). This observation suggests that some components of bias are shared between protocols, although we are unable to say whether from extraction or a shared step such as library preparation.

We illustrate the use of quantifying bias in terms of the vector of relative efficiencies by considering the hypotheses that PCR of a 16S region yields product in proportion to 16S copy number (CN) in the Brooks et al. (2015) data. Following Brooks et al. (2015), we estimated 16S CN for each species from NCBI RefSeq genome annotations (Methods; SI Table 3). Copy number ranged from 1 to 7 among these species, leading to a substantial expected CN bias. However, we found that CN was not associated with total bias or the estimated bias from PCR (SI Figures 8 and 9) and that CN correction increased error in the measured profiles of both cell and DNA mixtures (SI Figure 7). These results indicate that CN variation is just one component of PCR bias, which itself is just one component of total bias, and thus even perfect correction of CN bias will not ameliorate total bias in

marker-gene sequencing experiments with protocols similar to that in Brooks et al. (2015).

	Mixtures			Steps	
Taxon	Cells	DNA	PCR prod.	Extraction	PCR
Gardnerella vaginalis	0.17	0.4	0.83	0.42	0.48
Streptococcus agalactiae	0.25	2	0.9	0.12	2.2
Atopobium vaginae	0.3	1	1	0.28	1
Prevotella bivia	1.7	0.4	0.91	4.3	0.44
Lactobacillus crispatus	2.3	0.54	0.93	4.3	0.58
Sneathia amnii	4.5	2.4	1.3	1.9	1.9
Lactobacillus iners	4.6	2.3	1.2	2	1.9
Gm. range	28	6.1	1.5	35	5.1
Gm. std. dev.	4.2	2.3	1.2	4.1	2

Table 1: **Estimated bias in the mixture experiments of Brooks et al. (2015).** The bias from the three mixture experiments was used to estimate the bias from the individual steps of DNA extraction and PCR (Methods). The total protocol bias is given by the bias in the Cells mixtures, and the bias from sequencing and bioinformatics by the PCR mixtures. The estimated Extraction bias equals the compositional difference between the Cells and DNA mixtures and the estimated PCR bias equals the compositional difference between the DNA and PCR product mixtures. Each bias vector is shown relative to the (geometric) average taxon.

Taxon	H	Q	W
Prevotella melaninogenica	3	1.8	1.4
Clostridium perfringens	1.3	0.57	1.1
Salmonella enterica	1.9	2.9	0.81
Clostridium difficile	0.12	0.099	0.22
Lactobacillus plantarum	0.016	0.94	0.36
Vibrio cholerae	2.3	1.4	0.9
Clostridium saccharolyticum	1.3	1.7	0.58
Yersinia pseudotuberculosis	1.5	1.5	0.64
Blautia hansenii	0.33	0.24	1.7
Fusobacterium nucleatum	51	4.2	17
Gm. range	3300	42	76
Gm. std. dev.	8.5	3.1	3.2

Table 2: **Estimated bias of mock taxa for the three protocols (H, Q, and W) in the Costea et al. (2017) experiment.** Each bias vector is shown relative to the (geometric) average taxon. Taxa are ordered by their actual abundance (Figure 4B).

Discussion

The lack of a rigorous understanding of how bias distorts marker-gene and metagenomic sequencing (jointly, MGS) measurements stands in the way of accurate and reproducible community-composition measurements. Previous analyses of bias in MGS experiments have largely relied on descriptive statistical models (Bell et al. 2018; Brooks et al. 2015; Krehenwinkel et al. 2017; Sinha et al. 2017) whose parameters cannot be identified with biophysical quantities that one might expect to apply to differently composed samples. Failure to develop more mechanistic models may have stemmed from the seeming hopelessness of accounting for the many verified sources of bias. Here we proposed a mathematical model of bias in MGS experiments as a set of taxon-specific factors (the relative efficiencies) that multiply the true relative abundances to produce the measured relative abundances. Our model was inspired by the observation that many sources of bias, such as differences in DNA extraction efficiency (Morgan et al. 2010), PCR primer binding and amplification efficiency (Edgar 2017; Polz and Cavanaugh 1998; Suzuki and Giovannoni 1996; Wagner et al. 1994), and marker-gene copy number (Kembel et al. 2012), are thought to act multiplicatively, and hence so could their cumulative effect. The parameters in our model (the relative efficiencies) have biophysical interpretations as the relative yield per unit of input for each taxon, for individual steps or for the workflow overall. Our hypothesis that the relative efficiencies are consistent across samples is grounded in existing understanding of individual bias mechanisms, and was supported by marker-gene and shotgun-metagenomic sequencing measurements of mock bacterial communities with varying composition. We further showed how our model could be used to measure, understand, and correct bias.

We found bias to be independent of sample composition only after accounting for the compositional nature of MGS measurements. Bias appeared inconsistent when viewed in terms of taxon proportions—for example the measured proportion of *L. crispatus* was both higher and lower than its true value in different samples (Figure 3C). However, these apparent inconsistencies did not reflect inconsistency in the action of bias, but instead were a consequence of the compositional nature of MGS data. A limited number of sequencing reads are generated from each sample, so if one taxa is enriched by bias then other taxa must be correspondingly diminished. Therefore, a taxon's proportional over- or under-representation depends not on its absolute detection efficiency but on its efficiency relative to other taxa in the sample—*L. crispatus* increased in proportion when it was detected more efficiently than the average in the sample, and decreased otherwise. This compositional effect explains why linear models of bias based on proportions (e.g. Bell et al. 2018; Krehenwinkel et al. 2017) fail to extrapolate to differently-composed samples (SI R markdown), even when complex between-taxon interaction terms are introduced. In contrast, when compositionality was accounted for it became clear that relative efficiencies were consistent across differently composed samples. Bias had the same effect in each sample when we divided out the effect of compositionality by considering ratios of taxa (Figure 3D), and when we fully modeled the normalization involved in constructing proportions we found that a single set of relative efficiencies explained the observed proportions in every sample (Figure 3B).

A quantitative model allows the sensitivity of downstream analyses to bias to be rigorously evaluated. Consider the often unstated assumption that analyses of the differences between samples measured in the same experiment should be robust to bias, because each

sample is biased in the same way. We can formally evaluate this assumption in the simple numerical example shown in Figure 2: Sample S_1 has higher Shannon diversity than Sample S_2 but the measured diversity of Sample S_1 is lower than Sample S_2, and Sample S_1 has a higher proportion of the blue taxon than Sample S_2 but is observed to have a lower proportion of the blue taxon, despite the same bias distorting each sample. Whether such qualitative errors are likely can be investigated by simulating our model with empirical distributions of bias and community compositions. For example, consider the *Bacteroidetes*:*Firmicutes* ratio, a repeatedly-proposed diagnostic of gut health (Finucane et al. 2014; Ley et al. 2006). The range of relative efficiencies within the *Firmicutes* indicates that the *Bacteroidetes*:*Firmicutes* ratio measured by the metagenomic sequencing protocols evaluated in Costea et al. (2017) can differ from the true ratio by very little, or by as much as 100-fold, depending on which Firmicutes species is dominant in the sample!

If bias acts as a consistent multiplication of the relative abundances as we propose, then analyses of MGS data based on ratios of taxa could largely remove the effect of bias, reduce spurious results, and make results from different experiments more comparable. The key insight is that the fold change between samples in the ratios between taxa is invariant to consistent multiplicative bias, because the relative efficiencies divide out (Equations (6) and (7)). To be clear, this bias-invariance property only holds when considering samples biased in the same way, so ratio-based analyses still must be conducted within experiments sharing a common MGS protocol. But by controlling for study-specific bias, these analyses may give results that are more concordant across studies than analyses that are sensitive to bias. One ready source of such methods is the field of Compositional Data Analysis (CoDA), which deals almost entirely with ratio-based analyses (Aitchison 1986; Gloor et al. 2017). In fact, our model of bias is equivalent to what is referred to as a compositional perturbation in the CoDA field, and many CoDA methods are designed to be invariant to compositional perturbations (Aitchison 2003; Boogaart et al. 2013) and, therefore, would be invariant to multiplicative bias. CoDA methods are being increasingly used to analyze MGS data, but to date this has been motivated by the need to account for the compositionality of MGS data. The possibility that such methods could also reduce or remove the effect of bias has not been widely appreciated.

Studies investigating bias and/or optimizing protocols should evaluate the systematic errors introduced by bias in a way that accounts for the compositional nature of MGS data. Most previous studies of bias quantified bias with taxon proportions (e.g., Bell et al. 2018; Brooks et al. 2015; Krehenwinkel et al. 2017) or proportion-based summary statistics such as Bray-Curtis dissimilarities (Sinha et al. 2017) or differences in Shannon diversity (Song et al. 2016). Proportion-based measurements do not consistently measure bias in differently composed samples, and thus are difficult to interpret and can mislead researchers attempting to reduce the effect of bias on their experiments. The adoption of compositionally aware analytical methods to study bias may lead to insights that generalize beyond the specific sample compositions considered in these studies. In particular, quantification of bias in the form of the "bias vector" of relative efficiencies we proposed here has a natural biological interpretation as the relative yields of each taxon and can be naturally decomposed into an element-wise product of bias vectors for each step in a workflow, allowing for granular investigation of MGS protocol choices.

Our results suggest that calibration could become a practical approach to improving

the accuracy of MGS analyses and diagnostics. If bias is composition-independent, then it can be estimated from one or more control samples and those estimates used to correct the relative abundances in target samples. Our results overcomes major limitations in recent attempts at MGS calibration (Bell et al. 2018; Brooks et al. 2015; Krehenwinkel et al. 2017; Thomas et al. 2016) by indicating how to obtain a compositionally independent estimate of bias for many taxa from a small number of control samples. Intriguingly, the differential bias between protocols behaves in the same manner as the bias of an individual protocol under our model. This property opens the possibility of calibration based on a reference protocol's measurements of control samples even if their true composition is not known. Reference calibration does not give the abundances in terms of biologically tangible units like cell concentration, but can make measurements from differently biased experiments quantitatively comparable, allowing diagnostic criteria to be applied outside of the lab in which they were defined. Reference calibration may sidestep the practical challenges of creating defined cellular mixtures of many taxa by using natural samples (or aggregations of natural samples) as calibration controls that would then contain the full range of taxa naturally present.

Limitations and Next Steps

We found bias to act multiplicatively in accordance with our model in two mock-community experiments; however, many sources of bias may deviate from multiplicativity under non-ideal conditions. For example, it has been observed that the efficiency of a target sequence is altered by saturation during PCR amplification when the target is either rare or highly abundant (Gonzalez et al. 2012; Suzuki and Giovannoni 1996). In a non-microbial marker-gene experiment, Thomas et al. (2016) found a strong and consistent saturation effect that may have been caused by such a mechanism. The opposite of saturation, where taxa have lower efficiencies when rarer in the sample, may occur if low-abundance taxa are culled by the minimum-abundance thresholds used by taxonomic profilers such as MetaPhlAn2 (Truong et al. 2015). Such deviations from multiplicativity may be eliminated through protocol design (e.g., to avoid PCR saturation Polz and Cavanaugh (1998)) or accounted for with extensions allowing for deterministic and random variation in efficiencies.

Even when bias does act multiplicatively on individual microbial species, that multiplicativity will not hold for aggregates of species that vary in their efficiencies (SI R markdown). Consider again the *Bacteroidetes*:*Firmicutes* ratio diagnostic. We know that within the *Firmicutes* phylum there is tremendous phenotypic variation, and that variation manifested itself in the Costea et al. (2017) study as order-of-magnitude differences between the relative efficiencies of various *Firmicutes* species. As a result, consistent multiplicative bias could dramatically enrich or diminish the relative abundance of the *Firmicutes* phylum depending on which *Firmicutes* species were present. Unfortunately, the potential for bias to act inconsistently on aggregates of taxa cannot be entirely circumvented by eschewing aggregation in our analyses because even the fundamental units we derive from MGS data effectively aggregate variation at some level (McLaren and Callahan 2018). In the bacterial context, the traditional 97% ribosomal OTU groups variation at roughly the genus level (Yarza et al. 2014), short-read ribosomal ASVs and common shotgun profilers group variation at roughly the species level (Edgar 2018; Hillmann et al. 2018), and metagenome assembly combines

strains too similar to be separated by the assembler or by the subsequent binning (Dick 2018). This does not make bias an intractable problem, but it does mean that additional work is needed to understand the phylogenetic scales over which bias significantly varies. Methods to control bias will need to operate on taxonomic objects with commensurate or finer resolution than that variation in the bias phenotype (McLaren and Callahan 2018).

Cellular phenotype and matrix chemistry can vary substantially between samples of different types, and could affect the action of bias. For example, the efficiency of DNA extraction for a given taxon could be changed by differences in storage methods that impact the integrity of the cellular membrane (Morgan et al. 2010), and inhibitors in soil samples are known to influence the efficiency of PCR (Schrader et al. 2012). In the Costea et al. (2017) experiment, protocols differentially extracted the mock and native gut taxa (Figure 4A), which may have been due to physiological differences between the lab-grown cells and the preserved fecal cells. Potential phenotype and matrix effects are particularly relevant for calibration applications, as accurate calibration will require that bias measured in the controls is representative of the bias in the target samples.

The development of effective calibration methodologies will be limited by our ability to develop control samples that cover the range of taxa present in target communities. The bias estimation procedure we developed here is limited to those taxa present in the controls, and thus only allows for partial calibration of the subcomposition of the target samples consisting of those taxa. However, even partial calibration is useful when controls with key taxa in the target community are available. For example, vaginal microbiome samples from a patient over several clinical visits analyzed by Brooks et al. (2015) were mostly comprised of the seven species in their mock mixtures. It may be possible to effectively augment the range of taxa with credible bias estimates beyond those included in control samples by using phylogenetic inference methods to predict the bias of related taxa (Goberna and Verdú 2016). It will be easier to broadly cover the taxa in a given environment when calibrating to a reference protocol, because any sample measured by the reference protocol can be used as a reference calibration control, including samples from the environment of interest.

The model of bias we explored in this paper treats bias and measurement error more generally as deterministic. However, in practice, we observe variability in the error of taxon ratios around the mean (Figure 3D; Figure 5D). We developed a simple point-estimation procedure for estimating bias from random observations (Methods); however, robust estimation and calibration may require a statistical model that accounts for the variance in bias among samples along with the expected bias. Modeling both the expected bias and the variance in bias using a statistical model is therefore warranted. A statistical model would facilitate the construction of confidence intervals for the calibrated taxon proportions in a sample, which is not possible using our deterministic model. Some of the challenges associated with building a statistical model for bias include modeling the presence of taxa thought to be absent from the community (but observed due to contamination or index switching (Eisenhofer et al. 2018)) and the absence of taxa known to be present (Yeh et al. 2018), and accounting for the noise associated with the count nature of sequencing data. Our finding that multiplicative error in taxon ratios provide a parsimonious model for bias paves the way for the development of a statistical model, and we leave the development of such a model for future work.

Conclusion

We suggest a simple yet profound change in how we view MGS measurements. Currently, researchers tend to either 1) take MGS measurements as telling us only about presence and absence of detectable taxa, 2) hope that bias in the measurements of individual samples will somehow cancel out when analyzing differences between samples within a given experiment, or 3) pretend bias doesn't exist. We propose a new view in which the measured relative abundances within an experiment are biased by unknown—but constant—multiplicative factors. When bias acts consistently in this manner it can be accounted for through the use of bias-insensitive analyses, or corrected by a calibration procedure. Our results lay a foundation for the rigorous understanding of bias in marker-gene and metagenomic sequencing measurements that is required for accurate and reproducible research using MGS methods and for the development of reliable MGS diagnostics and interventions.

Methods

Model details

Qualitative or downstream effects of a step: In the most general formulation of our model first presented in the Results, we allow for the possibility that the protocol choice at a step influences the qualitative, as well as the quantitative, properties of the taxa. For instance, DNA extraction protocols yield differently fragmented DNA (Costea et al. 2017), which may affect the bias of downstream affects such as PCR and sequencing. Therefore, the bias at a step in general will depend on the protocol for that step as well as for all previous steps. To make this dependence explicit, let $P_l \mid P_1 \dots P_{l-1}$ denote the l-th step of protocol P performed after steps 1 through $l-1$ of protocol P. The bias of protocol P is then

$$\mathbf{B}^{(P)} = \mathbf{B}^{(P_1)} \cdot \mathbf{B}^{(P_2|P_1)} \cdot \mathbf{B}^{(P_3|P_1P_2)} \cdot \dots \cdot \mathbf{B}^{(P_L|P_1P_2\cdots P_{L-1})}. \tag{10}$$

Our core assumption is that the bias at each step is fixed within the context of the total protocol P, but independent of the composition of the sample. All properties and implications described in Section [Ref] to bias or differential bias of the total protocols are valid under this condition in the presence of such qualitative or downstream effects.

Downstream effects are important to consider, however, when attempting to estimate the bias of an individual step or protocol choice. Consider using differential bias to estimate the effect of DNA extraction protocol in the Costea et al. (2017) experiment, where the protocols differ only at the extraction step. The differential bias between Protocols H and W, $\mathbf{B}^{(H/W)}$, includes both the direct effect of the differential bias at the extraction step as well as the indirect downstream effects the extraction protocol has on later steps. Here we cannot distinguish the two, as estimating the size of indirect effects requires experiments where protocols at different steps are varied in a combinatorial fashion.

A different but related issue arises in our estimation of bias from different steps in the Brooks et al. (2015) experiment. There, DNA extraction from pure cultures to create the DNA mixtures differed from extraction in the community samples in the cell mixtures. In order to estimate the bias during extraction, we must assume that there is no qualitative

difference in the DNA, and so the bias of PCR, sequencing, etc. will be the same whether starting from the cell or the DNA mixtures. In other words, we must assume that the bias at a step P_l is only a property of the individual step.

Bias estimation

The geometry of compositional vectors

We develop some concepts of the so-called Aitchison geometry for compositional vectors that are useful for estimating bias (Aitchison 2003; Boogaart et al. 2013). The Aitchison geometry defines a normed vector space for compositional vectors of a given length K. In the Aitchison geometry, element-wise multiplication and division take the place of addition and subtraction in the standard Euclidean geometry, and the distance between vectors is determined by the fold-differences in their element ratios (as opposed to absolute differences in elements under Euclidean geometry) (Aitchison 2003). The norm of a composition $\mathbf{X}$ is given by

$$\|\mathbf{X}\| = \sqrt{\frac{1}{K}\sum_{i<j}\left[\ln\frac{X_i}{X_j}\right]^2}. \tag{11}$$

The distance between two compositions $\mathbf{X}$ and $\mathbf{Y}$ is given by the norm of their compositional difference, $\|\mathbf{X}/\mathbf{Y}\|$. Thus under our deterministic model the norm of the protocol bias equals the distance between the observed and actual compositions, $\|\mathbf{B}\| = \|\mathbf{O}/\mathbf{A}\|$. The Aitchison norm is often equivalently defined as

$$\|\mathbf{X}\| = \sqrt{\sum_{i=1}^{K}\left[\ln\frac{X_i}{g(\mathbf{X})}\right]^2} \tag{12}$$

where $g(\mathbf{X})$ is the geometric mean of the elements of $\mathbf{X}$ and $\ln\frac{X_i}{g(\mathbf{X})}$ is the i-th element of the centered-log-ratio (clr) transform of $\mathbf{X}$ (Boogaart et al. 2013). Equation (12) shows that the Aitchison norm is the standard Euclidean norm of $\operatorname{clr}\mathbf{X}$, and we use this form to compute the norm.

The mean of a set $\{X(s) \mid s \in S\}$ of compositional vectors in the Aitchison geometry, called the *sample center* (Boogaart et al. 2013), is given by the element-wise geometric mean,

$$\operatorname{cen}\{\mathbf{X}(s)\} \sim \left(\left[\prod_{s\in S} X_1(s)\right]^{1/|S|}, \ldots, \left[\prod_{s\in S} X_K(s)\right]^{1/|S|}\right), \tag{13}$$

where $|S|$ is the number of elements in the set S. The center of $\{\mathbf{X}(s)\}$ is often defined as the exponential of the Euclidean sample mean of $\{\operatorname{clr}\mathbf{X}(s)\}$ (Boogaart et al. 2013), which is compositionally equivalent to the definition in (13). The sample center is the vector $\mathbf{Y}$ that minimizes the sum of squared distances from the set S to itself, $\sum_s \|\mathbf{X}(s)/\mathbf{Y}\|^2$, where distance is computed with respect to the Aitchison norm [Ref]. It is also linear with respect to compositional addition and subtraction (i.e., element-wise multiplication and division),

$$\operatorname{cen}\{\mathbf{X}(s)\cdot\mathbf{Y}(s)\} = \operatorname{cen}\{\mathbf{X}(s)\}\cdot\operatorname{cen}\{\mathbf{Y}(s)\}.$$

Point estimate of the average bias in a set of samples

In contrast to the deterministic model of Equation (2), the measurement error we observe in our mock communities is random, since 1) the actual composition generally differs from the expected composition due to experimental error during sample construction; 2) bias will vary slightly among samples within an experiment due to unintentional variation in sample handling; and 3) the read counts we observe are subject to random sampling during sequencing. We describe a procedure for combining multiple control observations into a single estimate of the expected bias.

For sample s, let $\mathbf{A}(s)$ denote the sample composition and $\mathbf{O}(s)$ be the observed composition. Under our model, the expected value of $\mathbf{O}(s)$ is $\mathbf{A}(s) \cdot \mathbf{B}$. However, in practice, each sample varies from its expectation by a compositional error term which we call $\epsilon(s)$. Therefore, for all samples,

$$\mathbf{O}(s) \sim \mathbf{A}(s) \cdot \mathbf{B} \cdot \epsilon(s), \tag{14}$$

In other words, we decompose the error $\mathbf{O}(s)/\mathbf{A}(s)$ into a deterministic component (equal to the expected bias $\mathbf{B}$) and a random component $\epsilon(s)$ that accounts for all additional sources of variation. In the case where all samples contain all K taxa, we propose estimating $\mathbf{B}$ as a compositional vector that minimises the sum of squared compositional errors over all samples:

$$\hat{\mathbf{B}} = \arg\min_{B} \sum_{s} \|\mathbf{O}(s)/(\mathbf{A}(s) \cdot \mathbf{B})\|^2. \tag{15}$$

The least-squares estimate defined in (15) is simple to calculate using sample centers and is

$$\begin{aligned} \hat{\mathbf{B}} &\sim \mathrm{cen}\{\mathbf{O}(s)/\mathbf{A}(s)\} \\ &\sim \mathrm{cen}\{\mathbf{O}(s)\}/\,\mathrm{cen}\{\mathbf{A}(s)\}. \end{aligned} \tag{16}$$

However, when some elements of $\mathbf{A}(s)$ are zero (i.e., not all taxa are contained in all samples), some care is required to ensure that the compositional norm in (15) remains well-defined. In this case, we can still minimize the residual sum of squared norms over an appropriate subcomposition, which we now define. Let $\hat{\epsilon}(s) \sim \mathbf{O}(s)/(\mathbf{A}(s) \cdot \mathbf{B})$, and $K(s)$ be the set of taxa in sample s. The residual sum of squares over the subcompositions is defined as

$$\mathrm{RSS} = \sum_{s} \frac{1}{|K(s)|} \sum_{\substack{i<j \\ i,j \in K(s)}} \left[\ln \frac{\hat{\epsilon}_i(s)}{\hat{\epsilon}_j(s)} \right]^2, \tag{17}$$

where $\hat{\epsilon}_i(s)$ is the ith element of $\hat{\epsilon}(s)$ and $|K(s)|$ is the number of taxa in sample s. Our bias estimate $\hat{\mathbf{B}}$ is the compositional vector value of that minimizes this RSS. The bias is defined only up to a constant multiplied by all elements, so to find a unique solution we arbitrarily set the efficiency of the last taxon to 1, minimize with respect to the remaining taxa, and geometrically center the result by dividing the efficiencies by their geometric mean. This solution is unique provided that there is sufficient taxonomic overlap among the samples. In other cases, there will still be infinite solutions as bias cannot be estimated between all taxa.

To see why, consider two taxa, i and j. Given at least one control sample containing both taxa, we can estimate the efficiency of i relative to j, or B_i/B_j. We can also estimate B_i/B_j if there is a third taxon k such that one control contains i and k and another contains j and k, since $B_i/B_j = (B_i/B_k)(B_k/B_j)$. We can similarly imagine longer chains (e.g., i with k, k with l, l with j), and the estimation procedure we have defined allows estimation of B_i/B_j so long as there is any such chain of pairs connecting i and j. Grouping all taxa that co-occur or our connected through such a chain forms non-overlapping groups for which bias can be estimated within but not between groups. The optimized solution can still be used as long as the relative values between groups are properly interpreted as meaningless.

Differential bias: Differential bias to a reference protocol R can be estimated without knowing the actual compositions by replacing the actual compositions $A(s)$ in the above with the reference measurement $\mathbf{O}^{(R)}(s)$. If all samples contain all K taxa, we can simply compute

$$\hat{\mathbf{B}}^{(P/R)} \sim \text{cen}\{\mathbf{O}^{(P)}(s)/\mathbf{O}^{(R)}(s)\}. \tag{18}$$

Otherwise, we choose $\hat{\mathbf{B}}^{(P/R)}$ to minimize

$$\sum_s \frac{1}{|K(s)|} \sum_{\substack{i<j \\ i,j \in K(s)}} \left[\ln \frac{O_i^{(P)}/[O_i^{(R)} \cdot \hat{B}_i^{(P/R)}]}{O_j^{(P)}/[O_j^{(R)} \cdot \hat{B}_j^{(P/R)}]} \right]^2. \tag{19}$$

Bioinformatics and statistical analysis

Data and code availability

All code used to obtain or create the taxonomic profiles and perform all analyses and figures is available at `https://github.com/mikemc/2019-bias-manuscript`. Analysis and visualization is performed using the R software environment and the "tidyverse" suite of R data analysis packages and contained in R-markdown documents, which can be executed to generate all numerical results, tables, and figures. Versions that have already been "knit" into html documents are included.

Brooks et al. (2015) experiment

We used taxonomic profiles generated in the original study and provided as supplemental information. Specifically, we used the code provided in Additional File 7 of Brooks et al. (2015) with sample information and read assignments in Additional Files 2, 10, and 11 to build a table of amplicon sequences assigned to each of the seven mock species in each sample. Brooks et al. (2015) used a classification method and 16S reference database designed for species-level classification of vaginally associated taxa from V1-V3 region amplicons (Fettweis et al. 2012). Reads were assigned to species in the database according to a 97% sequence identity threshold, resulting in 93.5% of reads assigned and for which the vast majority (99.98%) were assigned to the seven mock species. We discarded the small fraction (0.0002%) of reads assigned to other taxa. Most samples were assigned a small fraction of their reads

from mock taxa not expected to be in the sample. These out-of-sample species generally had much lower frequency than expected species, suggesting they were the result of cross-sample contamination rather than mislabeling or misconstruction of the samples. We therefore removed these reads before evaluating and estimating bias.

We took the actual composition of each sample to be an even mixture of the species in the sample, in units of cell concentration, DNA concentration, or PCR-product concentration. Brooks et al. (2015) constructed the cell mixtures to be even mixtures based on CFUs (a proxy for cell concentration); the DNA mixtures based on DNA concentration; and the PCR mixtures based on volume from amplification of a fixed weight of DNA. Extraction and PCR protocols differed somewhat when using pure cultures to create the DNA and PCR product mixtures than when applied to communities in the cell experiment. Thus the DNA and PCR product in the second and third experiments may differ qualitatively from that in the cell mixture experiments, which could in principle affect the bias of downstream steps.

Following Brooks et al. (2015), we estimated the 16S copy number and genome size for the seven mock strains from available genome annotations for the given species in NCBI. In particular, we used annotated reference and representative genomes in RefSeq release 86 as collated by the GTDB (Parks et al. 2018). Current RefSeq annotations report 16S copy numbers for two species, *P. bivia* and 3 for *S. amnii*, that differ from those reported in Brooks et al. (2015) Table 5. The bias expected from copy-number variation for the cell mixtures is simply the compositional vector of copy numbers (number of 16S copies per genome); for the DNA mixtures, the relative efficiency of each taxon is give by its copy number divided by genome size (to give 16S copies per base pair); and for the PCR-product mixtures, is the compositional identity vector $(1, \ldots, 1)$ (i.e., no bias). Denoting the estimated CN bias as $\hat{\mathbf{B}}^{(\mathrm{CN})}$ for the given experiment, the CN-corrected proportions are $\Pr[\mathbf{O}/\hat{\mathbf{B}}^{(\mathrm{CN})}]$.

For each mixture experiment, we estimated bias by minimizing the residual squared compositional error as described above in "Bias estimation." We then used these estimates to partition our estimate of the total protocol bias into three steps—1) DNA extraction, 2) PCR amplification, and 3) sequencing and bioinformatics—under the simplifying assumption that the bias of shared steps are the same across experiments. We assume the bias of the cell mixture experiments is $\mathbf{B}^{(\mathrm{Cells})} = \mathbf{B}^{(P_1)} \cdot \mathbf{B}^{(P_2)} \cdot \mathbf{B}^{(P_3)}$, of the DNA mixtures is $\mathbf{B}^{(\mathrm{DNA})} = \mathbf{B}^{(P_2)} \cdot \mathbf{B}^{(P_3)}$, and of the PCR-product mixtures is $\mathbf{B}^{(\mathrm{PCR\ product})} = \mathbf{B}^{(P_3)}$. We therefore estimate the extraction bias as $\hat{\mathbf{B}}^{(P_1)} = \hat{\mathbf{B}}^{(\mathrm{Cells})}/\hat{\mathbf{B}}^{(\mathrm{DNA})}$, the PCR bias as $\hat{\mathbf{B}}^{(P_2)} = \hat{\mathbf{B}}^{(\mathrm{PCR\ product})}/\hat{\mathbf{B}}^{(\mathrm{DNA})}$, and the sequencing and bioinformatics bias as $\hat{\mathbf{B}}^{(P_3)} = \hat{\mathbf{B}}^{(\mathrm{PCR\ product})}$.

Costea et al. (2017) Phase III experiment

We downloaded raw sequencing reads for the Phase III experiment from ENA study accession PRJEB14847 and generated taxonomic profiles using MetaPhlAn2 (Truong et al. 2015) under default settings. We restricted our subsequent analysis to bacterial taxa and MetaPhlAn2's estimated species proportions. All mock species were detected in every sample except for the rarest species, *Blautia hansenii*, in three samples from Protocol Q. Our approach to visualizing and estimating bias requires all mock taxa to have non-zero observed abundance. We therefore added a small abundance to each taxon equal to the minimum proportion of any detected species in the sample (serving as an estimate of the detection limit for MetaPhlAn2 to call a species as present).

Costea et al. (2017) reported *Escherichia coli* as a contaminant of the spike-in based on profiles, based on its presence in the taxonomic profiles obtained from the mOTU software in the mock-only samples from each lab. We used the mock-only samples to determine which Metaphlan2 taxonomic identifiers corresponded to the spike-in contaminant and found that the identifiers "Shigella_flexneri" and "Escherichia_unclassified" captured the vast majority of non-mock abundance in the mock-only samples, with "Escherichia_coli" and various other identifiers each receiving much smaller amounts. We classified all three *Shigella/Escherichia* identifiers as "Contaminants" in our evaluation of overall community composition across specimens.

We obtained the expected mock community composition from the supplementary files of Costea et al. (2017). Following Costea et al. (2017), we took the actual cell concentration of the mock taxon in the spike-in to be given by FACS measurements, taking the mean for the majority of taxa where there were two replicate measurements. The FACS measurements provided in the SI of Costea et al. (2017) indicate a different abundance ranking for three taxa (*V. cholerae*, *C. saccharolyticum*, and *Y. pseudotuberculosis*) from that shown in Figure 6 of Costea et al. (2017), indicating a mislabeling of taxa in either the provided mock-composition table or their figure. We take the table to be the correct composition. A mislabeling would change the specific bias values we estimate for these taxa but would not qualitatively affect our results.

We estimated the bias of each protocol and the differential bias between protocols as described in "Bias estimation". For calibration, we randomly chose three specimens to use as the "estimation set" to estimate bias, and then calibrated all samples using Equation (9). To visualize bias in Figure 4B and the left column of Figure 5, we normalized the observed and actual relative abundances within each sample by the geometric mean over all taxa. This normalization equates the observed error (fold-difference between the observed and actual abundance) to the bias estimates shown in Table 2.

Acknowledgements

We thank Glen Satten and David Clausen for valuable discussions.

References

Aitchison, J. (1986). *The statistical analysis of compositional data.*

Aitchison, J. (2003). "A Concise Guide to Compositional Data Analysis". In: *2nd Compositional Data Analysis Workshop; Girona, Italy.*

Aitchison, J. (1992). "On criteria for measures of compositional difference". In: *Mathematical Geology* 24.4, pp. 365–379.

Bell, K. L., K. S. Burgess, J. C. Botsch, E. K. Dobbs, T. D. Read, et al. (2018). "Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures". In: *Molecular Ecology.*

Boogaart, K. G. van den, R. Tolosana-Delgado, K. G. van den Boogaart, and R. Tolosana-Delgado (2013). *Analyzing Compositional Data with R.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Brooks, J. P. (2016). "Challenges for case-control studies with microbiome data." In: *Annals of epidemiology* 26.5, 336–341.e1.

Brooks, J. P., D. J. Edwards, M. D. Harwich, M. C. Rivera, J. M. Fettweis, et al. (2015). "The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies". In: *BMC Microbiology* 15.1, p. 66.

Callahan, B. J., D. B. DiGiulio, D. S. A. Goltsman, C. L. Sun, E. K. Costello, et al. (2017). "Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.37, pp. 9966–9971.

Cocolin, L., M. Mataragas, F. Bourdichon, A. Doulgeraki, M.-F. Pilet, et al. (2018). "Next generation microbiological risk assessment meta-omics: The next need for integration". In: *International Journal of Food Microbiology* 287, pp. 10–17.

Costea, P. I., G. Zeller, S. Sunagawa, E. Pelletier, A. Alberti, et al. (2017). "Towards standards for human fecal sample processing in metagenomic studies". In: *Nature Biotechnology* 35.11, pp. 1069–1076.

Dick, G. (2018). *Genomic Approaches in Earth and Environmental Sciences*. Chichester, UK: John Wiley & Sons, Ltd.

Edgar, R. C. (2017). "UNBIAS: An attempt to correct abundance bias in 16S sequencing, with limited success". In: *bioRxiv*, p. 124149.

Edgar, R. C. (2018). "Updating the 97% identity threshold for 16S ribosomal RNA OTUs". In: *Bioinformatics*. Ed. by A. Valencia.

Eisenhofer, R., J. J. Minich, C. Marotz, A. Cooper, R. Knight, et al. (2018). "Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations". In: *Trends in Microbiology*.

Fettweis, J. M., M. G. Serrano, N. U. Sheth, C. M. Mayer, A. L. Glascock, et al. (2012). "Species-level classification of the vaginal microbiome". In: *BMC Genomics* 13.8, S17.

Finucane, M. M., T. J. Sharpton, T. J. Laurent, and K. S. Pollard (2014). "A Taxonomic Signature of Obesity in the Microbiome? Getting to the Guts of the Matter". In: *PLoS ONE* 9.1. Ed. by M. M. Heimesaat, e84689.

Gibbons, S. M., C. Duvallet, and E. J. Alm (2018). "Correcting for batch effects in case-control microbiome studies". In: *PLOS Computational Biology* 14.4. Ed. by M. Langille, e1006102.

Gilbert, J. A., J. K. Jansson, and R. Knight (2014). "The Earth Microbiome project: successes and aspirations". In: *BMC Biology* 12.1, p. 69.

Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue (2017). "Microbiome Datasets Are Compositional: And This Is Not Optional". In: *Frontiers in Microbiology* 8, p. 2224.

Goberna, M. and M. Verdú (2016). "Predicting microbial traits with phylogenies". In: *The ISME Journal* 10, pp. 959–967.

Gonzalez, J. M., M. C. Portillo, P. Belda-Ferre, and A. Mira (2012). "Amplification by PCR Artificially Reduces the Proportion of the Rare Biosphere in Microbial Communities". In: *PLoS ONE* 7.1. Ed. by J. A. Gilbert, e29973.

Graham, E. B., J. E. Knelman, A. Schindlbacher, S. Siciliano, M. Breulmann, et al. (2016). "Microbes as Engines of Ecosystem Function: When Does Community Structure Enhance Predictions of Ecosystem Processes?" In: *Frontiers in Microbiology* 7, p. 214.

Hardwick, S. A., I. W. Deveson, and T. R. Mercer (2017). "Reference standards for next-generation sequencing". In: *Nature Reviews Genetics* 18.8, pp. 473–484.

Hiergeist, A., U. Reischl, and A. Gessner (2016). "Multicenter quality assessment of 16S ribosomal DNA-sequencing for microbiome analyses reveals high inter-center variability". In: *International Journal of Medical Microbiology* 306.5, pp. 334–342.

Hillmann, B., G. A. Al-Ghalith, R. R. Shields-Cutler, Q. Zhu, D. M. Gohl, et al. (2018). "Evaluating the Information Content of Shallow Shotgun Metagenomics". In: *mSystems* 3.6. Ed. by J. F. Rawls, e00069–18.

Hugerth, L. W. and A. F. Andersson (2017). "Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing". In: *Frontiers in Microbiology* 8, p. 1561.

Kembel, S. W., M. Wu, J. A. Eisen, and J. L. Green (2012). "Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance". In: *PLoS Computational Biology* 8.10. Ed. by C. von Mering, e1002743.

Knight, R. and M. L. Sogin (2017). "The Microbiome and Human Biology". In: *Annual Review of Genomics and Human Genetics* 18, pp. 3.1–3.22.

Krehenwinkel, H., M. Wolf, J. Y. Lim, A. J. Rominger, W. B. Simison, et al. (2017). "Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding". In: *Scientific Reports* 7.1, p. 17668.

Lehman, R. M., H. P. Collins, J. M. F. Johnson, V. Acosta-Martinez, C. A. Cambardella, et al. (2015). "Soil biology for resilient, healthy soil". In: *Journal of Soil and Water Conservation* 70.1, 12A–18A.

Ley, R. E., P. J. Turnbaugh, S. Klein, and J. I. Gordon (2006). "Human gut microbes associated with obesity". In: *Nature* 444.7122, pp. 1022–1023.

Li, H. (2015). "Microbiome, metagenomics, and high-dimensional compositional data analysis". In: *Annual Review of Statistics and Its Application* 2, pp. 73–94.

Mallick, H., S. Ma, E. A. Franzosa, T. Vatanen, X. C. Morgan, et al. (2017). "Experimental design and quantitative analysis of microbial community multiomics". In: *Genome Biology* 18.1, p. 228.

McLaren, M. R. and B. J. Callahan (2018). "In Nature, There Is Only Diversity". In: *mBio* 9.1, e02149–17.

Metcalf, J. L., Z. Z. Xu, A. Bouslimani, P. Dorrestein, D. O. Carter, et al. (2017). "Microbiome Tools for Forensic Science". In: *Trends in Biotechnology* 35.9, pp. 814–823.

Morgan, J. L., A. E. Darling, and J. A. Eisen (2010). "Metagenomic Sequencing of an In Vitro-Simulated Microbial Community". In: *PLoS ONE* 5.4. Ed. by F. Rodriguez-Valera, e10209.

Nayfach, S. and K. S. Pollard (2016). "Toward Accurate and Quantitative Comparative Metagenomics". In: *Cell* 166.5, pp. 1103–1116.

Parks, D. H., M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, et al. (2018). "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life". In: *Nature Biotechnology* 36, pp. 996–1004.

Pollock, J., L. Glendinning, T. Wisedchanwet, and M. Watson (2018). "The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies". In: *Applied and Environmental Microbiology* 84.7. Ed. by S.-J. Liu, e02627–17.

Polz, M. F. and C. M. Cavanaugh (1998). "Bias in template-to-product ratios in multitemplate PCR." In: *Applied and environmental microbiology* 64.10, pp. 3724–30.

Quince, C., A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata (2017). "Shotgun metagenomics, from sampling to analysis". In: *Nature biotechnology* 35.9, p. 833.

Rosso, G. E., J. A. Muday, and J. F. Curran (2018). "Tools for Metagenomic Analysis at Wastewater Treatment Plants:Application to a Foaming Episode". In: *Water Environment Research* 90.3, pp. 258–268.

Schrader, C., A. Schielke, L. Ellerbroek, and R. Johne (2012). "PCR inhibitors - occurrence, properties and removal". In: *Journal of Applied Microbiology* 113.5, pp. 1014–1026.

Sinha, R., G. Abu-Ali, E. Vogtmann, A. A. Fodor, B. Ren, et al. (2017). "Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium". In: *Nature Biotechnology* 35, pp. 1077–1086.

Sipos, R., A. J. Székely, M. Palatinszky, S. Révész, K. Márialigeti, et al. (2007). "Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis". In: *FEMS Microbiology Ecology* 60.2, pp. 341–350.

Song, S. J., A. Amir, J. L. Metcalf, K. R. Amato, Z. Z. Xu, et al. (2016). "Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies." In: *mSystems* 1.3, e00021–16.

Suzuki, M. T. and S. J. Giovannoni (1996). "Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR". In: *Applied and Environmental Microbiology* 62.2, pp. 625–630.

Thomas, A. C., B. E. Deagle, J. P. Eveson, C. H. Harsch, and A. W. Trites (2016). "Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material". In: *Molecular Ecology Resources* 16.3, pp. 714–726.

Truong, D. T., E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, et al. (2015). "MetaPhlAn2 for enhanced metagenomic taxonomic profiling". In: *Nature Methods* 12.10, pp. 902–903.

Wagner, A., N. Blackstone, P. Cartwright, M. Dick, B. Misof, et al. (1994). "Surveys of Gene Families Using Polymerase Chain Reaction: PCR Selection and PCR Drift". In: *Systematic Biology* 43.2, pp. 250–261.

Yarza, P., P. Yilmaz, E. Pruesse, F. O. Glöckner, W. Ludwig, et al. (2014). "Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences". In: *Nature Reviews Microbiology* 12.9, pp. 635–645.

Yeh, Y.-C., D. M. Needham, E. T. Sieradzki, and J. A. Fuhrman (2018). "Taxon Disappearance from Microbiome Analysis Reinforces the Value of Mock Communities as a Standard in Every Sequencing Run". In: *mSystems* 3.3. Ed. by J. G. Caporaso, e00023–18.

Zeevi, D., T. Korem, N. Zmora, D. Israeli, D. Rothschild, et al. (2015). "Personalized Nutrition by Prediction of Glycemic Responses". In: *Cell* 163.5, pp. 1079–1094.

Supplemental figures

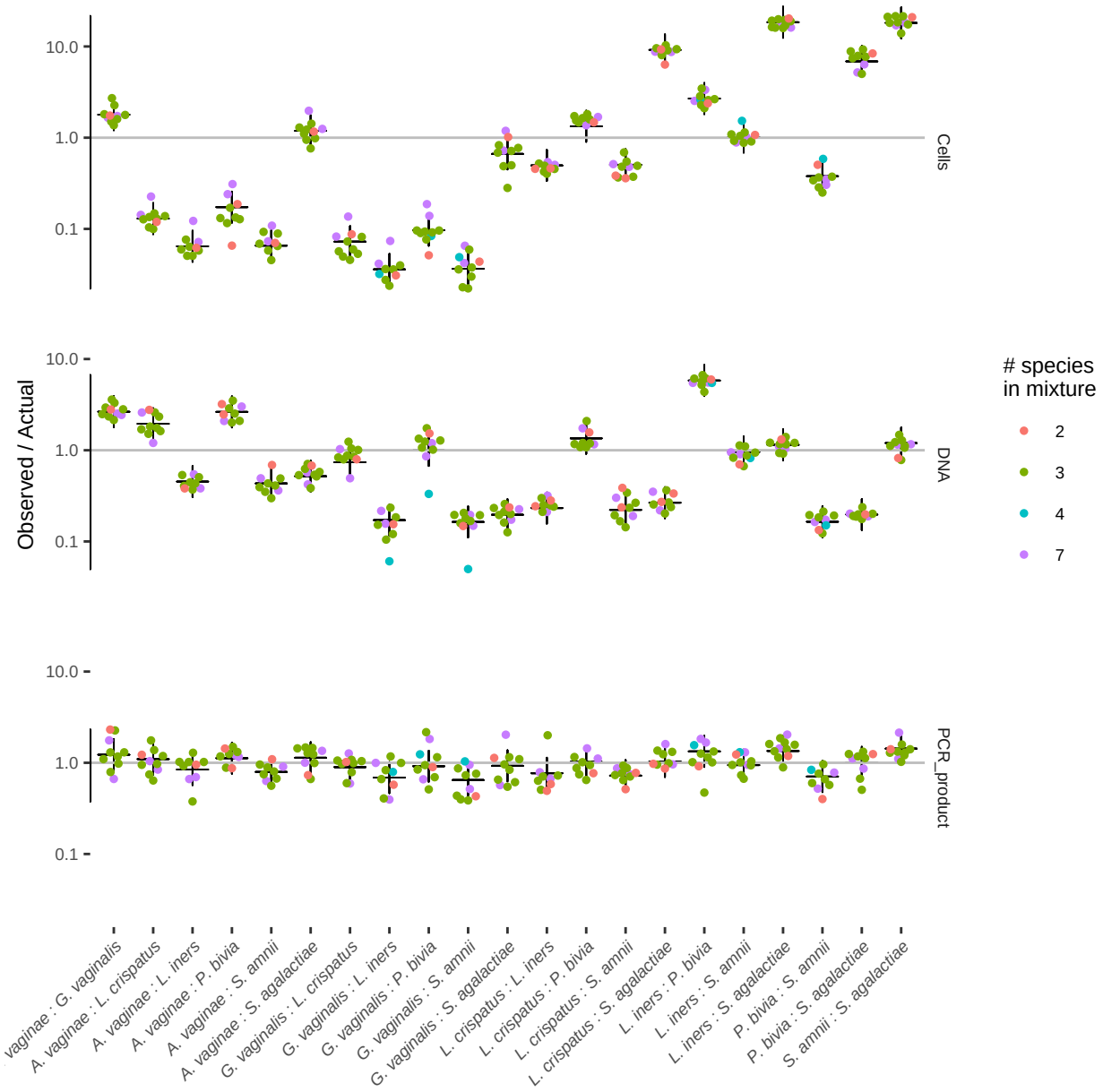

Figure 6: The observed error in species ratios (colored dots) against the fitted model prediction (black cross) for the three mixture experiments of Brooks et al. (2015).

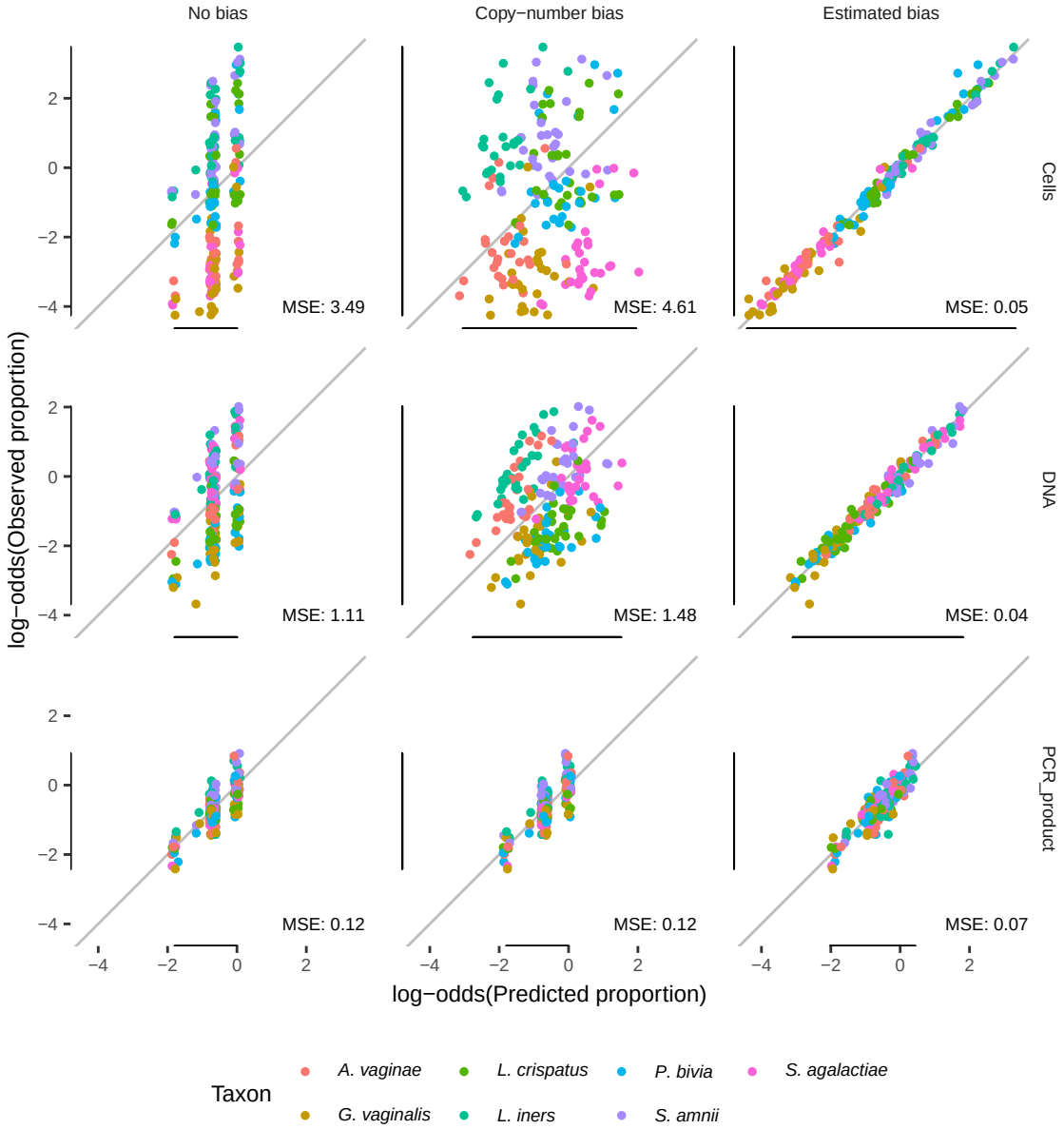

Figure 7: Comparison of observed proportions with three types of expected proportions—the actual proportions, the CN-biased proportions, and the proportions predicted by the fitted bias model—for the three mixture experiments of Brooks et al. (2015). Proportions p are transformed to log-odds ($\ln[p/(1-p)]$) to avoid compressing errors near $p = 0$ and $p = 1$, and average mean-squared error (MSE) of the log-odds are shown.

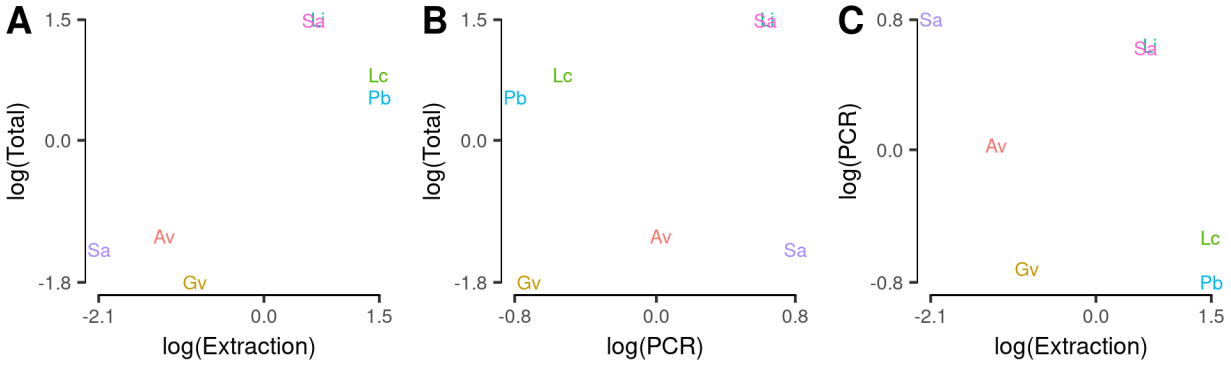

Figure 8: Total bias is driven by extraction, which splits the taxa into low- and high-extraction groups (A). Variation within low- and high-extraction groups is due to extraction (A) and PCR (B). PCR and extraction bias appear to be uncorrelated (C).

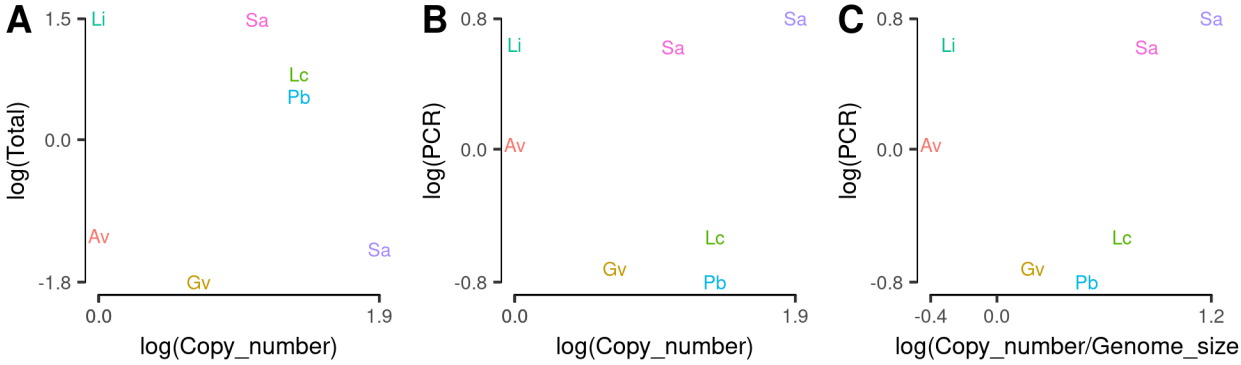

Figure 9: Copy number poorly predicts predict total bias (A) and PCR bias (B,C).

Supplemental tables

Taxon	Genome size (Mbp)	Copy number
Atopobium vaginae	1.44	1
Gardnerella vaginalis	1.64	2
Lactobacillus crispatus	2.04	4
Lactobacillus iners	1.28	1
*Prevotella bivia**	2.52	4
*Sneathia amnii**	1.33	3
Streptococcus agalactiae	2.16	7

Table 3: Genome statistics from NCBI RefSeq genomes for the species in the Brooks et al. (2015) experiment (Methods). *Denotes strains with copy numbers different from that reported in Brooks et al. (2015).