# Inferring true changes in microbial abundance from taxonomically-biased microbiome measurements

Michael R. McLaren[*] Jacob T. Nearing[†] Amy D. Willis[‡]
Karen G. Lloyd[§] Benjamin J. Callahan[¶]

2022-01-16

## Contents

**Preface** **2**

**1 Introduction** **2**

**2 How taxonomic bias affects abundance measurements** **5**
2.1 Relative abundance (proportions and ratios) . . . . . . . . . . 5
2.2 Absolute abundance . . . . . . . . . . 6

**3 Taxonomic bias can cause errors in differential-abundance results** **9**
3.1 Fold changes between a pair of samples . . . . . . . . . . 9
3.2 Regression analysis of many samples . . . . . . . . . . 10

**4 Case studies** **12**
4.1 Foliar fungi experiment . . . . . . . . . . 12
4.2 Vaginal microbiomes of pregnant women . . . . . . . . . . 15
4.3 Microbial growth in marine sediments . . . . . . . . . . 17
4.4 Summary and conclusions . . . . . . . . . . 18

**5 Potential solutions** **19**
5.1 Use ratio-based relative-abundance analyses . . . . . . . . . . 19
5.2 Calibration using community controls . . . . . . . . . . 20
5.3 Calibration using reference species . . . . . . . . . . 20
5.4 Choosing complementary total-density and metagenomics measurements . . . 22
5.5 Bias-sensitivity analysis . . . . . . . . . . 23
5.6 Bias-aware meta-analysis . . . . . . . . . . 23

**6 Conclusion** **24**

**Appendix** **24**

**A Model details** **24**
A.1 MGS measurement . . . . . . . . . . 24

[*] North Carolina State University; send correspondence to m.mclaren42@gmail.com
[†] Dalhousie University
[‡] University of Washington
[§] University of Tennessee
[¶] North Carolina State University

A.2 Taxonomic bias in total-density measurement . . . . . . . . . . . . . . . . . . . 25
A.3 Using reference species for total-density normalization . . . . . . . . . . . . . 27
A.4 Linearity of extraction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28

**B Review of experimental methods for obtaining absolute densities 29**
B.1 Measurement of total cell density . . . . . . . . . . . . . . . . . . . . . . . . . 29
B.2 Measurement of total DNA density post extraction . . . . . . . . . . . . . . . 29
B.3 Equivolumetric protocol . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30
B.4 Housekeeping species . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30
B.5 Spike-ins . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
B.6 Targeted measurements . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

**C Error in estimated regression coefficients and standard errors 32**

**D Supplemental figures 33**

**References 33**

# Preface

```
#> working directory clean
```

*This manuscript was rendered from commit 414f6e78cc9b29c8236c8c3e996b68cd4d733b86.*

**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. It is licensed under a CC BY 4.0 License. See the Zenodo record for how to cite the latest version.

# 1 Introduction

One of the most basic questions we can ask about microbial communities is: How do different microbial taxa vary in abundance—across space, time, and host or environmental conditions? Marker-gene and shotgun metagenomic sequencing (jointly, MGS) can be used to measure the abundances of thousands of species simultaneously, making it possible to ask this question on a community-wide scale. In these *differential-abundance (DA) analyses*, the change in abundance of a microbial taxon across samples or conditions is used to infer ecological dynamics or find microbes that are associated with specific host diseases or environmental conditions. Standard MGS measurements lose information about total microbial density and so are typically used to analyze the abundances of taxa relative to each other. But new methods are increasingly used to provide absolute information, making it possible to analyze changes in absolute cell density. In its various forms, DA analysis remain one of, if not the most, common forms of analyses applied to MGS data to elucidate the inner workings of microbiomes and their relationships to host and environmental health.

Yet these DA analysis are built on a fundamentally flawed foundation. MGS measurements are *taxonomically biased*: Microbial species vary dramatically (e.g. 10-1000X) in how efficiently they are measured—that is, converted from cells into taxonomically classified sequencing reads—by a given MGS protocol (McLaren, Willis, and Callahan (2019)). This bias arises from variation in how species respond to each step in an MGS protocol, from sample collection to bioinformatic classification. Although often associated with features specific to marker-gene sequencing—the variation among species in marker copy numbers and in primer-binding and amplification efficiencies—the existence of large variation in DNA extraction efficiencies and in the ability to correctly classify reads make taxonomic bias a universal feature of both marker-gene and shotgun measurements. As a result of taxonomic bias, MGS measurements provide inaccurate representations of actual community composition and tend to differ across protocols, studies, and even experimental batches (Yeh et al. (2018), McLaren, Willis, and Callahan (2019)). These errors have been found in some cases to supersede sizable biological differences (e.g. Lozupone et al. (2013)) and have may have contributed to failed replications of prominent findings in the human microbiome literature, such as the associations of Bacteroides and Firmicutes in stool with obesity (Finucane et al. (2014)) and the associations of certain species in the vagina of pregnant women with preterm birth (Callahan et al. (2017)).

The extent to which taxonomic bias has impacted the DA results in the scientific literature is unknown. The typical approach taken to counter taxonomic bias is to standardize the measurement protocol used within a given study, with the (often tacit) assumption being that samples measured by the same protocol will be affected by bias in the same way and so the measured differences between samples will be unaffected. For example, if taxonomic bias were to cause the measured proportion of a species to consistently be 10X too high, we would still be able to accurately infer the fold change in its proportion across samples (Kevorkian et al. (2018), Lloyd et al. (2020)). Unfortunately, mathematical arguments and analysis of experiments with artificially constructed ('mock') communities demonstrate that this assumption is not always warranted: Consistent taxonomic bias can lead to variable fold errors in species' proportions (Figure 1, McLaren, Willis, and Callahan (2019)). These varying errors can lead to spurious conclusions for how the proportion of a taxon varies across samples, even in the direction of change (for example, causing a taxon that decreases appear to increase) (McLaren, Willis, and Callahan (2019)). Yet McLaren, Willis, and Callahan (2019) also found that certain types of DA analysis—those based on fold changes in the ratios among species—where robust to bias. The implications of these findings for DA analysis of absolute abundances and for the joint analysis of variation of many species across many samples, as typically done in microbiome association testing, have yet to be investigated.

Here we use a combination of theoretical analysis, simulation, and re-analysis of published experiments to consider when and why taxonomic bias in MGS measurements leads to spurious results in DA analysis of relative and absolute abundance. Our analysis clarifies how the received wisdom that taxonomic bias does not affect the analysis of change across samples is only partially correct and can give a false sense of security in the accuracy of DA results. Yet we also present several potential solutions—methods for quantifying, correcting, or otherwise accounting for the effect of taxonomic bias in DA analyses that can be deployed today with only modest changes to existing experimental and analytical workflows. Over time, application of these methods to past and future experiments will provide crucial quantitative information about the conditions under which taxonomic bias creates spurious results for various DA methodologies. Collectively, these methods and insights may provide practical solutions to taxonomic bias in DA analysis and the confidence that is necessary to codify the statistical findings of microbiome studies into readily-translatable scientific knowledge.

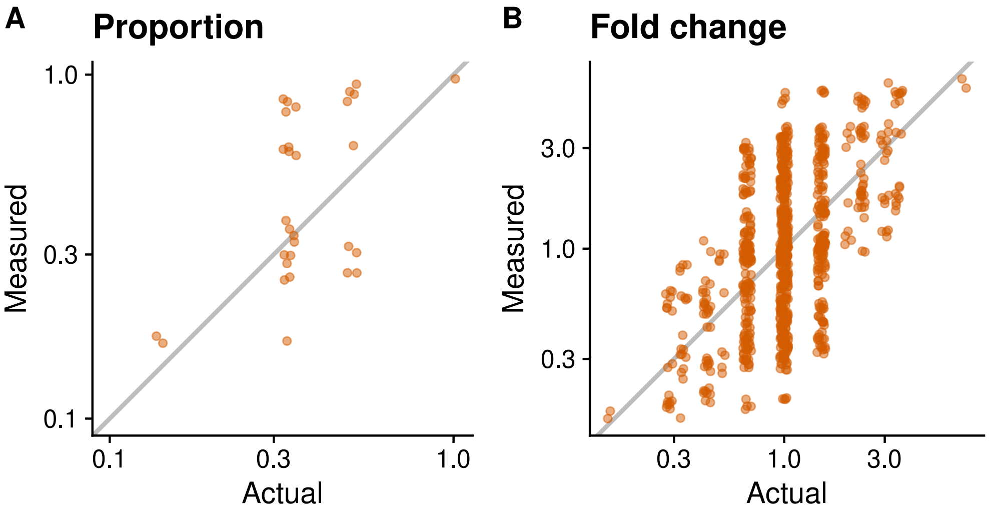

Figure 1: **Mock community experiments show that taxonomic bias can distort the measured fold change in an individual species' proportion across samples.** The figure shows the measured vs. actual proportions for a single bacterial species, *Lactobacillus crispatus*, in a set of bacterial cellular mock communities, and the resulting fold changes between community samples. The inconsistent error in the measured proportions of individual samples (Panel A) leads to inaccurate measurements of fold changes (Panel B). Mock communities were constructed and measured with 16S sequencing by Brooks et al. (2015). The data was re-analyzed by McLaren, Willis, and Callahan (2019), who showed that despite the inconsistency of the errors in Panel A, taxonomic bias acted consistently across samples.

# 2 How taxonomic bias affects abundance measurements

We begin by considering how taxonomic bias affects the relative- and absolute-abundance measurements which serve as the input to microbiome DA analyses.

Our primary tool for understanding the impact of taxonomic bias on MGS measurement is the theoretical model of MGS measurement developed and empirically validated by McLaren, Willis, and Callahan (2019). This model is the simplest that respects the multiplicative nature of taxonomic bias and the *compositional* nature of MGS measurements, in which the total read count for a sample is unrelated to its total cell number or density (Gloor et al. (2017)). We consider a set of microbiome samples measured by a specific MGS protocol that extracts, sequences, and bioinformatically analyzes to taxonomically assign reads to a set of microbial species $S$. Various forms of taxonomic assignment are possible; for simplicity, we suppose that reads are assigned to the species level, with reads that cannot be uniquely assigned being discarded. We ignore the possibility that reads are incorrectly assigned to the wrong species or sample. Unless otherwise stated, we treat the sequencing measurement as deterministic, ignoring the 'random' variation in read counts that arise from the sampling of sequencing reads and other aspects of the MGS process.

In our model, the assigned read count of a species $i$ in a sample $a$ equals its abundance multiplied by species-specific and sample-specific factors,

$$\text{reads}_i(a) = \text{abun}_i(a) \quad \cdot \quad \underbrace{\text{efficiency}_i}_{\substack{\text{species specific,}\\ \text{sample independent}}} \quad \cdot \quad \underbrace{\text{effort}(a)}_{\substack{\text{species independent,}\\ \text{sample specific}}} \quad . \tag{1}$$

The species-specific factor, efficiency$_i$, equals the *relative measurement efficiency* (or simply *efficiency*) of the species—how much more easily that species is measured (converted from cells to assigned reads) relative to a arbitrary fixed reference species (McLaren, Willis, and Callahan (2019)). We assume the efficiencies of particular species are consistent across samples. The variation in efficiency among species corresponds to the taxonomic bias of the MGS protocol. The sample-specific factor, effort$(a)$, we call the *sequencing effort* for that sample; it captures the variation in the total number of assigned reads due to experimental features such as library normalization and total sequencing-run output. Equation (1) implies that the total number of assigned reads in sample $a$ equals

$$\text{reads}_S(a) = \text{abun}_S(a) \cdot \text{efficiency}_S(a) \cdot \text{effort}(a), \tag{2}$$

where $\text{abun}_S(a) \equiv \sum_{j \in S} \text{abun}_j(a)$ is the total abundance of species in $S$ and

$$\text{efficiency}_S(a) \equiv \frac{\sum_{j \in S} \text{abun}_j(a) \cdot \text{efficiency}_j}{\text{abun}_S(a)} \tag{3}$$

is the *mean efficiency* over all species in $S$.

## 2.1 Relative abundance (proportions and ratios)

We distinguish between two types of species-level *relative abundances* within a sample. The *proportion* of species $i$ in sample $a$ equals to its abundance relative to the total abundance of all species in $S$,

$$\text{prop}_i(a) \equiv \frac{\text{abun}_i(a)}{\text{abun}_S(a)}. \tag{4}$$

The *ratio* between two species $i$ and $j$ equals the abundance of $i$ relative to that of $j$,

$$\text{ratio}_{i/j}(a) = \frac{\text{abun}_i(a)}{\text{abun}_j(a)}. \tag{5}$$

Proportions and ratios each form the basis for popular DA methods; ratio-based methods are commonly referred to as Compositional Data Analysis (CoDA) methods.

Taxonomic bias has a different effect on species proportions versus species ratios. The proportion of a species is typically measured by its proportion of assigned reads,

$$\widehat{\text{prop}}_i(a) = \frac{\text{reads}_i(a)}{\text{reads}_S(a)}. \tag{6}$$

We can rewrite the right-hand side (using Equations (1), (2), and (6)) to find

$$\widehat{\text{prop}}_i(a) = \text{prop}_i(a) \cdot \underbrace{\frac{\text{efficiency}_i}{\text{efficiency}_S(a)}}_{\substack{\text{variable}\\ \text{fold error}}}. \tag{7}$$

Taxonomic bias creates a multiplicative error in the species' proportion equal to its efficiency relative to the mean efficiency in the sample. Consequently, the species's proportion is measured as too high in samples that are dominated by species with lower efficiencies, and measured as too low in samples that are dominated by species with higher efficiencies. This phenomenon is illustrated in two hypothetical communities in Figure 2. Species 3 has an efficiency of 6; it is under-measured in Sample 1, which has a mean efficiency of 8.33, but over-measured in Sample 2, which has a mean efficiency of 3.15.

The measured ratio between species $i$ and $j$ is given by the ratio of their read counts,

$$\widehat{\text{ratio}}_{i/j}(a) = \frac{\text{reads}_i(a)}{\text{reads}_j(a)}. \tag{8}$$

From Equations (1) and (8), it follows that

$$\widehat{\text{ratio}}_{i/j}(a) = \text{ratio}_{i/j}(a) \cdot \underbrace{\frac{\text{efficiency}_i}{\text{efficiency}_j}}_{\substack{\text{constant}\\ \text{fold error}}}. \tag{9}$$

Taxonomic bias creates a multiplicative measurement error in the ratio that is equal to the ratio in their efficiencies; the error is therefore constant across samples. For instance, in Figure 2, the ratio of Species 3 (with an efficiency of 6) to Species 1 (with an efficiency of 1) is over-estimated by a factor of 6 in both communities despite their varying compositions.

The fact that the multiplicative error in proportions varies while that in ratios is constant forms the basis for understanding the effect of bias on different methods for measuring absolute abundances from MGS and the extent to which bias cancels in different DA analyses.

## 2.2 Absolute abundance

Researchers often would like to know how the *absolute abundance* of a species changes. In this context, *absolute* means not relative to other species; typically, the abundance of interest is relative to non-microbial entities such the volume, mass, or amount of host DNA in the sample. A wide range of experimental methods have been used to convert the relative abundances from MGS measurements into measurements of absolute abundances. Yet there has so far been little consideration as to how the resulting measurements are affected by taxonomic bias in the MGS measurements or in the supplemental measurements, such as flow cytometry and qPCR, that they often involve. Here and in Appendix B we describe how various methods for measuring absolute abundance are predicted to be affected by taxonomic bias, with an eye towards determining whether the measurement error is variable or constant.

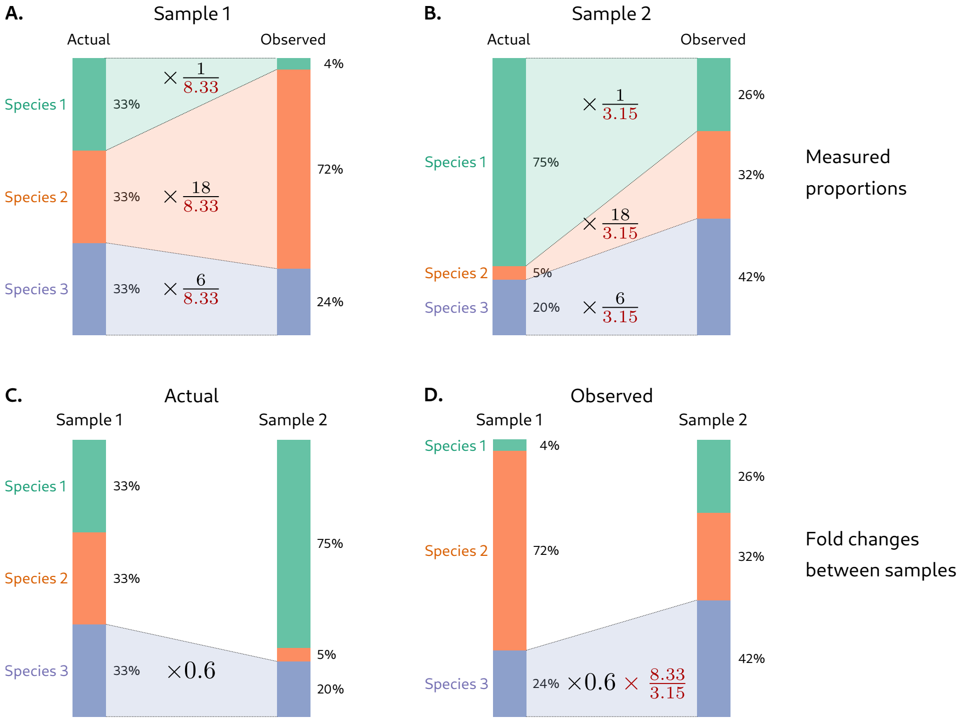

Figure 2: **Taxonomic bias creates sample-dependent multiplicative errors in species proportions, which can lead to inaccurate fold changes between samples.** Top row: Error in proportions measured by MGS in two hypothetical microbiome samples that contain different relative abundances of three species. Bottom row: Error in the measured fold-change in the third species that is derived from these measurements. Species' proportions may be measured as too high or too low depending on sample composition. For instance, Species 3 has an efficiency of 6 and is under-measured in Sample 1 (which has a mean efficiency of 8.33) but over-measured in Sample 2 (which has a mean efficiency of 3.15).

Our discussion supposes the 'abundance' of interest is cell number, though it applies to other quantities, such as biomass and genome copy number, that may be more relevant in different biological contexts.

We consider two general classes of methods: Those in which absolute-abundance information is derived from measurement of the aggregate abundance of the total community, and those in which it drives from targeted measurement (or prior knowledge) of one or more particular species.

**Normalization by total abundance:** Measurements of species absolute abundances are often obtained by making a (non-MGS) measurement of total community abundance, equating it with the aggregate abundance of the species $S$ measured by MGS, and multiplying this total by the proportions from MGS,

$$\widehat{\text{abun}}_i(a) = \widehat{\text{prop}}_i(a) \cdot \widehat{\text{abun}}_S(a). \tag{10}$$

This measurement is affected by taxonomic bias in the MGS measurement as well as in the total-abundance measurement. For example, measurement of total community abundance by 16S qPCR is affected by variation among species in extraction efficiency, 16S copy number, and PCR binding and amplification bias. If the species-level efficiencies of the total-abundance measurement are constant across samples and we neglect other sources, then we can express the total-abundance measurement as

$$\begin{align} \widehat{\text{abun}}_S(a) &= \sum_{i \in S} \text{abun}_i(a) \cdot \text{efficiency}_i^{\text{tot}} \tag{11} \\ &= \text{abun}_S(a) \cdot \text{efficiency}_S^{\text{tot}}(a), \tag{12} \end{align}$$

where $\text{efficiency}_i^{\text{tot}}(a)$ is the absolute measurement efficiency of species $i$ for the total-abundance measurement and $\text{efficiency}_S^{\text{tot}}(a)$ is the mean efficiency of the total-abundance measurement in the sample. The equations (11) and (7) for the error in total-abundance and proportion measurements imply that the species abundance measurement (10) has error given by

$$\widehat{\text{abun}}_i(a) = \text{abun}_i(a) \cdot \frac{\text{efficiency}_i \cdot \text{efficiency}_S^{\text{tot}}(a)}{\text{efficiency}_S(a)}. \tag{13}$$

Equation (13) indicates that the multiplicative error in the measured absolute abundance of a species equals its MGS efficiency relative to the mean MGS efficiency in the sample, multiplied by the mean efficiency of the total measurement. As in the case of proportions (Equation (7)), the error depends on sample composition through the two mean efficiency terms and so may vary across samples. On the other hand, if the mean efficiency of the total-abundance measurement mirrors that of the MGS measurement, the two can offset and lead to a relatively stable error. We discuss how this possibility might apply to real experimental workflows below.

**Normalization by one or more reference species:** Suppose we had a measurement of the (absolute) abundance of a *reference species* $r$. In the absence of taxonomic bias, all species are expected to have the same ratio of reads to abundance in a sample (Equation (1)). Thus the abundance-to-reads ratio for species $r$ can serve as a conversion factor allowing us to obtain the abundance of an arbitrary species $i$,

$$\widehat{\text{abun}}_i(a) = \text{reads}_i(a) \cdot \frac{\widehat{\text{abun}}_r(a)}{\text{reads}_r(a)}. \tag{14}$$

The abundance of one or more reference species can be directly measured using targeted measurement methods like species-specific qPCR and used with Equation (14) to obtain absolute abundance for all species. To our knowledge, this approach has not previously been suggested. Instead, reference-based measurement has been used in the context of spike-in experiments or normalization to a species (such as the host) which is treated as having a constant abundance. In a spike-in experiment, the abundance of reference species are added in known (up to experimental error) abundances to each sample. When normalizing to a species that is treated as having a constant but unknown abundance, we set $\widehat{\text{abun}}_r(a)$ in Equation (14) equal to 1 and interpret the abundances as having fixed but unknown units, which is sufficient for multiplicative DA analysis.

The error in the abundance measurement (14) due to taxonomic bias in the MGS measurement is given by

$$\widehat{\text{abun}}_i(a) = \text{abun}_i(a) \cdot \frac{\text{efficiency}_i}{\text{efficiency}_r} \cdot \frac{\text{fold error in}}{\widehat{\text{abun}}_r(a)} . \tag{15}$$

The constant error in the measured ratio of species $i$ to $r$ (see Equation (9)) propagates to the abundance measurement. There will generally also be systematic error in the abundance of the reference species; however, if the systematic fold error is constant across samples, so will be that of the abundances of other species.

Spike-ins are instead sometimes used to measure absolute abundances using Equation (10). In this case, an intermediate step is taken in which the total community abundance is first measured by the ratio of non-spike-in to spike-in reads. If done correctly, this calculation yields results that are identical to directly applying Equation (14) (Appendix A.3).

**Difference between the two approaches:** Reference-species normalization yields constant fold errors because it is based on species-level read counts and abundances and we assume that efficiencies are constant at the species level. In contrast, total-abundance normalization is based on aggregates of species (for the calculation of proportion and the total-abundance measurement) and so depends on mean efficiencies, which can vary across samples.

# 3 Taxonomic bias can cause errors in differential-abundance results

How do the errors caused by taxonomic bias in the abundances measured for individual samples impact DA analysis? Although there are many ways to quantify the changes in abundance that form the basis of a DA analysis, we focus on DA analyses of the (log) fold changes in proportions, ratios, and absolute abundance; these multiplicative difference measures are common (though not ubiquitous) in extant DA analyses and have more direct ecological interpretations (via the processes of exponential growth and decay) than non-multiplicative measures.

## 3.1 Fold changes between a pair of samples

The building blocks of a DA analysis are the fold changes (FCs) in abundance between individual pairs of samples.

The impact of bias on the measured FCs in species proportions and ratios follows directly from the results of Section 2 for the error in individual-sample measurements. From Equation (7), it follows that the measured FC in the proportion of species $i$ from sample $a$ to sample $b$

is

$$\underbrace{\frac{\widehat{\text{prop}}_i(b)}{\widehat{\text{prop}}_i(a)}}_{\text{measured FC}} = \frac{\text{prop}_i(b) \cdot \cancel{\text{efficiency}_i}/\text{efficiency}_S(b)}{\text{prop}_i(a) \cdot \cancel{\text{efficiency}_i}/\text{efficiency}_S(a)} \tag{16}$$

$$= \underbrace{\frac{\text{prop}_i(b)}{\text{prop}_i(a)}}_{\text{actual FC}} \cdot \underbrace{\left[\frac{\text{efficiency}_S(b)}{\text{efficiency}_S(a)}\right]^{-1}}_{\text{fold error}}. \tag{17}$$

The sample-independent efficiency factor cancels, but the sample-dependent mean efficiency does not, leaving an error equal to the inverse of the change in the mean efficiency. In contrast, when we use Equation (9) to compute the error in the FC in the ratio between two species $i$ and $j$, we find that the constant error $\text{efficiency}_i/\text{efficiency}_j$ exactly cancels, so that the measured FCs remain accurate regardless of whether the mean efficiency varies.

Figure 2 (bottom row) illustrates these two different behaviors for the error in the FCs of proportions versus ratios when the mean efficiency varies between samples. Here the mean efficiency decreases by a factor of 2.6 (FC of 0.4X) from Sample 1 to Sample 2, which causes the FC of the proportion of each species to be measured as 2.6X larger than its true value. Though the fold error for all species is the same, the implications depend on the actual FC and correspond to three distinct types of error: an increase in magnitude, a decrease in magnitude, and a change in direction; we refer to the latter as a *sign error* since in reference to the change in sign in the corresponding log fold change (LFC). We can see each type of error in Figure 2. For Species 1, which increases and thus moves in the opposite direction of the mean efficiency, we see an increase in magnitude of the measured FC (actual FC: 2.3X, measured FC: 6.5X). For Species 2, which decreases and thus moves in the same direction as the mean efficiency but by a larger factor, we see an decrease in magnitude (actual FC: 0.15X, measured FC: 0.44X). For Species 2, which decreases by a smaller factor than the mean efficiency, we see a change in direction (actual FC: 0.6X, measured FC: 1.8X), such that the species actually appears to increase (a sign error). In contrast, the fold error in Equation (9) completely cancels when we divide the ratio measured for one sample $a$ by another sample $b$.

These differences in the FCs of proportions versus ratios are mirrored when we compare FCs in absolute abundance made with normalization to a total-abundance measurement versus a reference-species measurement. FCs computed from total-abundance normalization are subject to error when the ratio of the two mean efficiencies (that for the MGS measurement and for the total-abundance measurement vary across samples). In contrast, FCs computed from reference-species normalization are not subject to error, so long as the (fold) error in the assumed or measured abundance of the reference species is constant.

## 3.2 Regression analysis of many samples

DA analysis of multiple samples can be framed as a regression problem in which we analyze the relationship between a microbial *response variable*, such as the log abundance of species $i$, with one or more *covariates*, such as the pH of the sampled environment or whether the sample is from a healthy or sick person. The simplest regression analysis uses the simple linear regression model,

$$\log \widehat{\text{abun}}_i(a) = \alpha_i + \beta_i x(a) + \varepsilon_i(a), \tag{18}$$

where $x$ is a continuous covariate (e.g. pH) or a binary covariate (e.g. $x = 1$ for treated patients and $x = 0$ for controls), $\alpha_i$ and $\beta_i$ are regression coefficients, and $\varepsilon_i(a)$ is a mean-zero random variable that reflects the residual (unexplained) variation in the response. In a DA analysis, we are interested in the slope coefficient $\beta$, which describes how the species' abundance changes with $x$. We are generally uninterested in the intercept coefficient, $\alpha$, which captures differences in the baseline abundance among species and—it is usually hoped—study- and species-specific systematic error such as that created by taxonomic bias.

How does taxonomic bias in fact impact estimates of the slope coefficient, $\beta$? Appendix C derives the effect of bias on the estimated coefficients for the simple linear regression model under estimation via Ordinary Least Squares (OLS) or Maximum Likelihood Estimation (MLE). The result has a simple interpretation that provides intuition for what to expect in more complicated generalized linear regression models used in popular DA methods. Consider the regression (18) fit to the measured log abundance of species $i$ across all samples. The measurement error in the response variable,

$$\text{error}_i(a) \equiv \log \widehat{\text{abun}}_i(a) - \log \text{abun}_i(a) = \log \frac{\widehat{\text{abun}}_i(a)}{\text{abun}_i(a)}, \tag{19}$$

can be thought of as having three components:

- Error that is constant across samples affects the intercept coefficient, $\alpha_i$.
- Error that varies systematically with $x$ affects the slope coefficient, $\beta_i$.
- Error that varies and is uncorrelated with $x$ affects the residual variation, $\varepsilon_i(s)$.

We saw in Section 2 that the error from taxonomic bias has a species-specific, sample-independent component that is constant across samples; this component causes a species-specific systematic error in the estimated intercept, $\alpha_i$. The error also has a sample-specific, species-independent component. The portion that is correlated with the covariate $x$ creates a systematic error in the slope coefficient $\beta_i$ that is the same for every species and opposite in sign to how the error varies with $x$; the portion that is uncorrelated with $x$ affects the precision (standard errors) of the slope estimates in a species-specific manor.

Figure 3 illustrates with a simulated data for the case of log abundance measured using the total-abundance normalization method (Equation (10)), assuming that total community abundance is measured perfectly accurately. In this case, the variable component of the error is due entirely to variation in the log mean efficiency of the MGS measurement. Variation in the log mean efficiency that is associated with the covariate $x$ creates a systematic error in the estimated slope $\hat{\beta}$ equal to the negative of the (scaled) covariance of log mean efficiency with $x$. The absolute error is the same for all species; however, its relative value depends on the magnitude of the covariance of the log mean efficiency with $x$ relative to that of the response (here, $\log \text{abun}_i$) with $x$ or, equivalently, the relative magnitudes of their slopes. As in the case of fold changes between pairs of samples, the net effect can be decreases in magnitude (Species 9, 10, and 1 in Figure 3), changes in sign (Species 5), or increases in magnitude (remaining species) depending on these relative values. Variation in the log mean efficiency that is uncorrelated with $x$ does not systematically distort $\hat{\beta}$ but does affect its precision, typically leading to increased standard errors as the variation in log mean efficiency effectively acts as an additional source of noise in measured abundance (Figure 3 D). The exception is for species whose residual variation is strongly positively correlated with that of log mean efficiency (here, Species 9), which can appear to have less random variation and receive standard errors that are too small. Decreased magnitudes and increased standard errors can both cause associations to be missed that would otherwise have been detected (Species 10 and 1), while increased magnitudes can turn weak or statistically insignificant associations into strong and statistically significant ones (Species 7, 6 and 4).

Overall, we find that variation in measurement error created by variation in the mean efficiency of the MGS measurement can negatively affect certain DA analyses by creating systematic errors in the slope coefficient estimates and/or reducing their precision. Similar effects occur for regressions of log proportions and of log abundances derived by total-abundance normalization when the total abundance is accurately measured; however, as noted in Section 2, total-abundance measurements that are taxonomically biased in a complementary manner can mitigate these effects. Regression analyses of log-ratios and log abundances derived from reference-species normalization remain unaffected—the errors are constant across samples and so only impact the intercept coefficients.

# 4 Case studies

To better understand the potential impact of taxonomic bias on DA analysis in practice, we conducted several case studies spanning a range of biological scenarios and sequencing technologies.

## 4.1 Foliar fungi experiment

Measurement of control communities along with the primary experimental samples can enable researchers to directly measure and remove the effect of bias prior to or concurrent with downstream DA analysis (McLaren, Willis, and Callahan (2019)). Gnotobiotic community experiments are well suited to this form of *calibration via community controls* since, unlike most natural ecosystems, it is possible to assemble 'mock communities' containing all species in known relative abundances.

Leopold and Busby (2020) are the first to use mock-based calibration in a gnotobiotic microbiome experiment. In a study of host-commensal-pathogen interactions, Leopold and Busby (2020) inoculated plants with 8 commensal fungal species and subsequently exposed plants to a fungal pathogen. To investigate the joint effect of colonization order and host genotype, plants varied in source location and in which commensal colonized first. The authors used ITS amplicon sequencing to measure communities before and after pathogen infection, along with mock communities with quantified genome concentrations of DNA from the 9 species. These DNA mock communities allowed the authors to directly measure the bias due to the sequencing workflow following DNA extraction and correct for it when analyzing the primary experimental samples following the method of McLaren, Willis, and Callahan (2019). Although extraction is not accounted for, considerable bias might still be expected due to preferential PCR binding and amplification and the substantial ribosomal copy-number-variation (CNV) in fungi (Lofgren et al. (2019)). The authors found a 13X difference between the most and least efficiently measured commensal, while the pathogen was measured 40X more efficiently than the least efficiently measured commensal. We reanalyzed this experiment to understand the impact that bias might have had on the study's analyses were it not accounted for.

Leopold and Busby (2020) used the ITS sequencing profiles of the pre-infection samples to understand the impact of host genotype and an experimental treatment—which species was allowed to colonize first—on the commensal community composition prior to infection. Two DA analyses were conducted to address these questions. First, the authors used linear regression of log species proportion (via negative-binomial regression to account for random read-sampling error) to ask whether host genotype, treatment, and their interaction significantly impacted overall community composition. Second, the authors quantified the strength of priority effects—the advantage of being allowed to colonize first—in different host genotypes by the LFC in the proportion of a species when allowed to colonize first versus later with the majority of commensals.

Since these analysis are both based on proportions, bias could in principle impact their results. We re-ran both analyses with and without bias correction, finding nearly identical results regardless of whether bias is accounted for[1]. To understand why, we compared the multiplicative variation taxa proportions with that of the mean efficiency. The proportions of each taxon varied substantially across samples, much more so than the mean efficiency,

[1] The two sets of analyses can be seen at the following links: negative-binomial regression analysis; priority-effects analysis

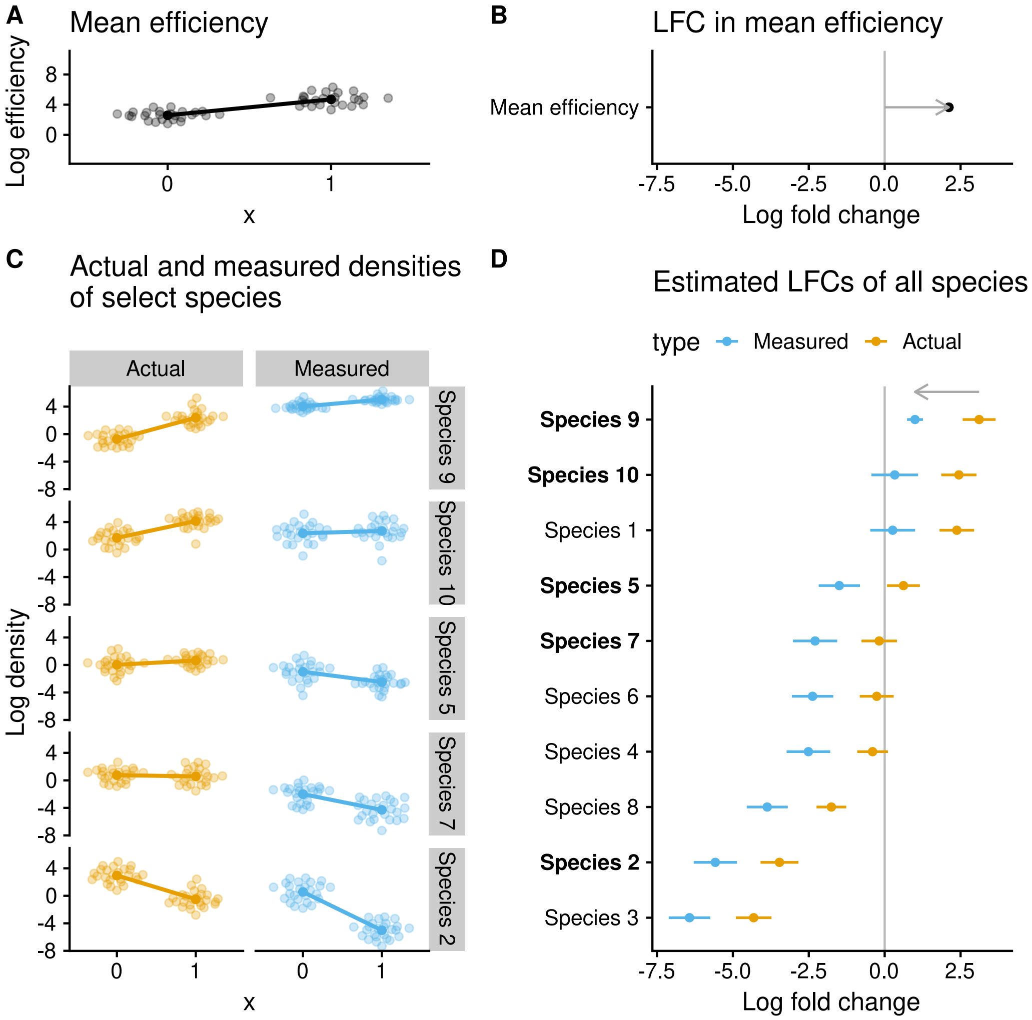

Figure 3: **Taxonomic bias distorts multi-sample differential abundance inference when the mean efficiency of samples is associated with the covariate of interest.** This figure shows the results of a regression analysis of simulated microbiomes consisting of 50 samples and 10 species from two environmental conditions indexed by $x = 0$ and $x = 1$. In this simulation, the species with the largest efficiency (Species 9) also has the largest positive LFC, which drives the positive association of the log mean efficiency with the condition (shown in Panels A and B). This positive LFC in the log mean efficiency induces a systematic negative shift in the estimated LFCs of all species (Panels C and D). Panel D shows the mean LFC (points) with 95% confidence intervals (CIs), for each species estimated from either the actual or the measured densities. The error (difference in LFC estimates on measured and actual) equals the negative LFC of the mean efficiency (shown in Panel B).

which had a geometric range of 1.62X and a geometric standard deviation of just 1.05X (SI Figure 6). Why does the mean efficiency vary so little despite substantial multiplicative variation in the efficiency among taxa and in the taxa proportions across samples? This answer relates to the fact that the mean efficiency is a weighted *arithmetic* average of species proportions and so is insensitive to large multiplicative variation in low-proportion species. The pre-infection samples in the Leopold and Busby (2020) study are always dominated by the three species with the highest efficiencies, whose efficiencies vary only by 3X (and just 1.5X between the two most dominant). The lowest efficiency species, despite having large variation in their proportion on the log scale, are always rare (proportions $\ll 1$, even after bias correction) and hence don't significantly affect the mean efficiency. Because the log mean efficiency varies much less than the log proportions for each taxon, there is little room for it to distort the results of DA analyses based on variation in log proportions.

We additionally considered the potential for bias to impact analysis of how commensal taxa responded to infection. In particular, it is interesting to consider whether any commensals are able to increase in absolute concentration in response to infection. The experiment performed by Leopold and Busby (2020) does not allow absolute-abundance measurement; nevertheless, the results of the previous sections allow us to consider the impact that bias would have on an analysis of absolute genome concentration measured using the total-abundance normalization method (Equation (10)). It is first useful to consider how bias impacts the estimated change in log proportion of each commensal following infection. The pathogen is absent in pre-infection samples but tends to dominate the community post-infection, raising the mean efficiency of post-infection samples (SI Figure 7). Across different host genotypes, the average increase in mean efficiency ranges from 2.5X to 5.2X. We used simple linear regression to measure the average change in log proportion of each commensal for each host genotype, with and without bias correction. Regardless of whether bias is accounted for, the log proportion of each commensal decreased, which is unsurprising given the pathogen's growth and the sum-to-one constraint of species proportions. Without bias correction, however, the LFCs were lower by an amount corresponding to the inverse change in log mean efficiency. The magnitude errors created were substantial for a large fraction of commensal-host pairs; in several cases, the LFC estimates without bias correction were clearly negative, whereas the estimates with bias correction were close to 0.

Now, consider the impact this shift in mean efficiency might have on a regression analysis with log absolute genome concentration as the response, if measured using the total-abundance normalization method (Equation (10)). We consider two scenarios. *Scenario 1:* Total genome concentration in each sample is perfectly known and used to measure the concentrations of individual taxa via Equation (10). In this case, bias in the MGS measurements would create absolute errors in the LFCs for genome concentration identical to those for proportions. The scientific error, however, may be much worse. Since the sum-to-one constraint of proportions no longer applies, the LFCs will necessarily be larger (less negative), such that the absolute error caused by the change in mean efficiency could plausibly create substantial biological errors where some taxa that persist or increase in abundance instead appear to decrease. *Scenario 2:* Total ITS concentration is measured by qPCR with the same primers and PCR protocol used for ITS sequencing, and used without any correction to measure individual taxa concentration with the total-abundance normalization method. In this case, we should expect the increase in mean efficiency to have little to no impact on LFCs, since it would be offset by the increase in mean efficiency of the total-concentration measurement. Moreover, since the total ITS concentration and ITS sequencing are both performed on the extracted DNA, we can also expect the effect of taxonomic bias caused by DNA sequencing to be offset. Thus, the systematic shift in mean efficiency following infection may or may not significantly impact the LFC estimates depending on how the total-abundance measurement is made prior to normalization.

## 4.2 Vaginal microbiomes of pregnant women

The vaginal microbiome during pregnancy has been a source of intensive study due to its apparent connection with the health of both mother and her developing child. Many MGS studies have found associations of specific microbial species and community characteristics with rates of urinary tract and sexually-transmitted infections, bacterial vaginosis (BV), and preterm birth. Yet these associations vary across studies of different populations and using different MGS methods. DA analyses of the vaginal microbiome are commonly based on proportions, creating an opportunity for taxonomic bias to impact results. A number of studies have experimentally demonstrated substantial taxonomic bias among MGS protocols and individual steps (such as extraction and PCR amplification) in vaginal samples or *in vitro* samples of vaginally-associated species (Yuan et al. (2012),Brooks et al. (2015),Gill et al. (2016),Graspeuntner et al. (2018)). Bias has been proposed as a potential explanation for discrepancies across studies (Callahan et al. (2017), Others?), but there has so far been little quantitative analysis of this possibility. Here we use empirical bias measurements from control samples to investigate the role of bias in proportion-based DA analyses of vaginal microbiomes from a recent large-cohort study of pregnant women.

As part of the Multi-Omic Microbiome Study: Pregnancy Initiative (MOMS-PI) study, Fettweis et al. (2019) collected longitudinal samples from over 1500 pregnant women, including nearly 600 that were measured by amplicon sequencing of the 16S V1-V3 hypervariable region to yield species-level bacterial taxonomic profiles. Taxonomic bias of this MGS protocol was previously investigated by Brooks et al. (2015) and McLaren, Willis, and Callahan (2019), using measurements by Brooks et al. (2015) of cellular mock communities of seven common, clinically-relevant vaginal bacterial species. Of these, *Lactobacillus iners* had the highest efficiency, which was nearly 30X larger than that of the species with the lowest efficiency, *Gardnerella vaginalis*. A second *Lactobacillus* species, *L. crispatus*, had an efficiency that was approximately 2X less than *L. iners* and 15X greater than *G. vaginalis*. These species, along with the unculturable *Lachnospiraceae BVAB1*, are the most common top (most-abundant) species (SI analysis) and can reach high proportions in individual samples, indicating that shifts between them might drive large changes in the mean efficiency which might in turn impact DA results.

We sought to assess this possibility using a joint analysis of the control measurements from Brooks et al. (2015) and the microbiome measurements from the MOMS-PI study. To obtain the taxonomic bias among all species identified in the MOMS-PI measurements, we used the taxonomic relationships with the seven control species to impute the efficiencies for the remaining species. We used these imputed efficiencies to calibrate (correct the effect of bias in) the MOMS-PI measurements, examine variation in the mean efficiency varied across samples, and compare the results of a DA analysis with and without bias correction.

The mean efficiency varies substantially across vaginal samples (Figure 4. This variation appears to be primarily driven by variation in which species is most abundant, samples in which a *Lactobacillus* species is most abundant (after calibration) typically have an efficiency that is 3-20X greater than samples in which *G. vaginalis* is most abundant. Shifts between *Lactobacillus*-dominance and *Gardnerella*-dominance are common in between-women comparisons and occasionally occur between consecutive visits in individual women (SI Figure 9). These shifts typically result in substantial fold changes in mean efficiency (SI Figure 9) and can cause spurious fold changes in the trajectories of lower-abundance species (SI Figure 10).

We next considered whether the observed variation in mean efficiency could cause systematic error in a DA analysis. In particular, we hypothesized that DA analysis of species proportions versus a covariate that is associated with *Lactobacillus* and/or *Gardnerella* would be particularly prone to spurious results. Patient metadata was not available due to privacy restrictions; we therefore sought a clinically relevant covariate to use in a regression analysis

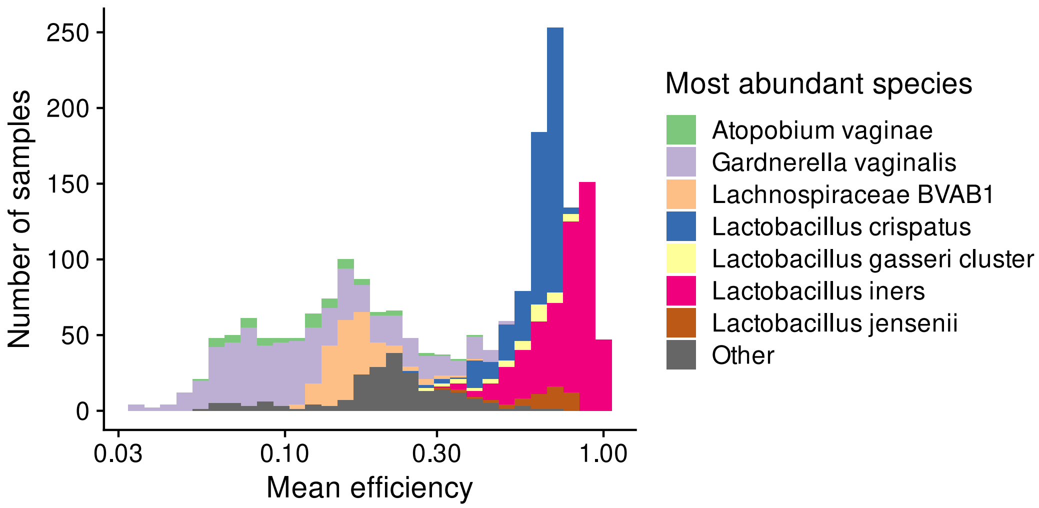

Figure 4: **The mean efficiency in vaginal samples from the MOMS-PI study varies with the most abundant species.**

that could be determined directly from the microbiome profiles and that we *a priori* expected to be associated with the proportions of *Lactobacillus* and *Gardnerella*. Alpha diversity metrics such as species richness, the Shannon index, and the (Inverse) Simpson index have been repeatedly found to be strongly positively associated with bacterial vaginosis (BV; Srinivasan et al. (2012), Cartwright et al. (2018)). Cartwright et al. (2018) found that an observed Simpson Diversity Index of 0.82 (corresponding to an order-2 effective number of species of 5.6) classified high diversity samples as BV positive with a sensitivity of 100% and specificity of 85.1%. In addition, it is commonly observed that samples from women with and without BV that are dominated by *Lactobacillus spp.* tend to have higher diversity, whereas samples dominated by *Gardnerella* tend to have lower diversity. We therefore chose to perform a DA analysis of species proportion versus alpha diversity, hypothesizing that *Lactobacillus* and *Gardnerella* would drive a negative association of mean efficiency with diversity and thereby distort DA estimates for all species. We split samples into low, medium, and high diversity groups based on Shannon diversity in observed (uncalibrated) microbiome profiles. We then estimated the LFC in proportion from low- to high-diversity samples for a 30 common species by simple linear regression, using calibrated (bias-corrected) and observed (uncorrected) microbiome profiles. As expected, the mean efficiency was higher in low-diversity samples due to a larger fraction of samples dominated by *Lactobacillus* and a lower fraction with samples dominated by *Gardnerella*. The decline in mean efficiency in the high diversity group caused a concomitant increase in the LFCs of all species when bias was not accounted for. This error due to bias resulted in serious magnitude errors in nearly all species and sign errors in over one third, with errors of both types seen for several clinically relevant species.

#### 4.2.1 Notes

- It might be cleaner story-wise to redo the diversity partitioning using the threshold identified by Cartwright et al. (2018)
- For the DA analysis of proportions vs. diversity, I tried other types of DA analysis methods (Gamma-Poisson regression and regression on ranks, similar to Wilcoxon test) to understand how the impact of bias differs across methods. In each case bias has a big effect but the species whose estimates are most affected varies. I would need to revise and verify this analysis before including its results. One reason for doing so

would be to show that our qualitative finding about how bias distorts results applies beyond the basic linear regression model.

- Decreases in mean efficiency during transitions from *Lactobacillus* to *Gardnerella* dominance can be expected to be even more extreme for commonly-used vaginal microbiome primers that fail to amplify *Gardnerella*.

## 4.3 Microbial growth in marine sediments

Our route for mean efficiency to become associated with the covariate is if the covariate quantifies a biological process that preferentially selects for a microbial trait that also tends to increase or decrease measurement efficiency. We illustrate this potential mechanism with a study of microbial growth in marine sediments.

The surface layers of marine sediments harbor diverse and abundant microbiomes. Total cell density and species richness decrease with depth as resources are consumed in older, deeper sediment layers; however, some taxa are able to grow and increase in density as they are slowly buried. Lloyd et al. (2020) performed a systematic assessment of growth rates of bacterial and archaeal taxa over a depth of 10 cm (corresponding to ~40 years of burial time) in sediment of the White Oak River estuary. To estimate growth rate, the authors first measured absolute cell density of microbial taxa using the total-abundance normalization method (Equation (10)), with taxa proportions measured with 16S amplicon sequencing and total community density measured by directly counting cells using epifluorescence microscopy (CARD-FISH). The authors refer to these absolute densities as FRAxC measurements, for 'fraction of 16S reads times total cell counts'. The FRAxC measurements were used to infer growth rate from the slope of a simple linear regression of log cell density against burial time over the first 3 cm below the bioirrigation layer (corresponding to $\sim 8$ years of burial). To validate the inference of positive growers, the authors compared the growth rates from FRAxC-based inference from two sediment cores, qPCR measurements in these cores for a few reference taxa, and FRAxC-based inference in two replicate laboratory incubation experiments.

Taxonomic bias could lead to systematic error in FRAxC-derived growth rates if sample mean efficiency tends to systematically vary with burial time (or equivalently, depth). One possibility is that microbes with tougher cell walls 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. In this case, we would expect the relative abundance of tougher species to increase with depth and hence the mean extraction efficiency to decrease, which in turn would lead to inflated growth-rate estimates for all taxa; a possible end result would be that taxa that decay sufficiently slowly would be mistakenly inferred to have positive growth.

We can test the hypothesis that systematic variation in log mean efficiency with depth distorts inferred growth rates using the qPCR measurements of two clades also collected by Lloyd et al. (2020). In Section B, we argue that taxonomic bias in targeted qPCR measurements is expected to create a roughly constant fold error and yield fold changes in absolute density that are relatively unaffected by changes in mean efficiency. Thus comparing qPCR to FRAxC growth rates on the same reference taxa allows us to estimate the systematic error in FRAxC growth rates. Because the systematic error in regression slopes due to bias is the same for each taxon (Section 3), these comparisons allow us to draw conclusions about the accuracy of FRAxC growth rates for all taxa. The first soil core included qPCR measurements of a single archaeal clade, *Bathyarchaeota*, for which growth rates by qPCR and FRAxC were nearly identical (doubling rates of 0.099/yr by FRAxC and 0.097/yr by qPCR). The second soil core included qPCR measurements of *Bathyarchaeota* and a second clade, *Thermoprofundales*/MBG-D. In this core, FRAxC and qPCR growth rates differed more substantially, with growth rates from FRAxC being larger by 0.012/yr for *Bathyarchaeota* (0.112/yr by FRAxC and 0.1/yr by qPCR) and by 0.086/yr for *Thermoprofundales*/MBG-D (0.294/yr by FRAxC and 0.208/yr by qPCR). A low number of experimental samples and noise in both the FRAxC and qPCR measurements place significant uncertainty in these measurements; however, the fact that FRAxC-derived growth rates are larger than qPCR-derived rates in all three cases is consistent with our hypothesis that mean efficiency decreases with depth in a manner that systematically biases FRAxC-derived rates to higher values. The differences in growth rate are small in absolute terms; however, the maximum observed difference of 0.086/yr suggests an error large enough to impact results for some taxa classified as positive growers, whose FRAxC growth rates ranged between 0.04/yr and 0.5/yr. Overall, the comparison between FRAxC and qPCR measurements gives support to the study conclusions, but suggest that species at the lower end of this range of positive FRAxC-derived rates may in fact be merely persisting or even slowly declining in abundance.

## 4.4 Summary and conclusions

The impact of bias can depend on protocol, biological system, and type of DA analysis being done. Though these case studies span a highly limited range of possibilities, when combined with the theoretical results of section X help suggest some general conclusions about how and when bias will impact DA analyses based on fold changes in proportions.

First, the error caused by consistent taxonomic bias *can mostly cancel* in cross-sample comparisons and so not impact DA analyses of fold changes in species proportions. Our theoretical results from Section 3 indicate that when the mean efficiency is roughly constant across samples, the error in proportions cancels in fold change calculations and is absorbed by the intercept term in regression models. We observed this scenario in the analysis of pre-infection fungal microbiome samples of Leopold and Busby (2020) and when analyzing the trajectories of the vaginal microbiomes of women when the dominant species remained constant[2], and a stable mean efficiency within marine soil cores is plausible and consistent with (though unable to be fully determined by) the different growth rate estimates in the Lloyd et al. (2020) experiment. Section 3 showed that absolute DA analysis using total-abundance normalization are susceptible to errors as with DA analysis of proportions; yet our analysis of two (hypothetical) approaches to analyzing absolute changes in response to infection in the fungal microbiome experiment indicates that the increasingly common approach of pairing marker-gene sequencing with qPCR of the total marker density can largely mitigate this effect, since here what matters is the variation in the ratio of mean efficiencies of the MGS measurement to the total-density measurement and this ratio may remain roughly constant if bias is largely shared by the two measurements.

Yet in other cases, the mean efficiency can vary substantially and create substantial error in fold changes between pairs of samples. We saw examples when comparing fungal microbiomes pre- and post-infection in the foliar-fungi experiment and comparing vaginal microbiomes with different dominant species in the MOMS-PI experiment. The impact of these errors on results of DA regression analysis across many samples depends on whether the mean efficiency varies systematically with the covariate of interest. In these two examples, systematic variation of the mean efficiency arose as high-efficiency species tended to dominate samples in one of the sample conditions (post-infection foliar samples or low-diversity vaginal samples). Thus one way for significant error in DA regression results to arise is when a small number of (or just one) species that have particularly high or low mean efficiencies, form a large proportion of the community in a substantial fraction of samples, and are associated with the covariate of interest.

For many experimental systems, however, there may not be any one species that frequently forms a large fraction of the community. In these systems, systematic variation of the mean efficiency can still arise through the collective change of many species that are associated with the covariate of interest. We described a hypothesized scenario in the marine-sediment

[2] This point could use more support in the MOMSPI case study

case study in which increased burial time (the covariate) selects for lysis-resistant (and so lower efficiency) species. As another example, treatment with a certain antibiotic might selectively kill easier-to-lyze Gram negative species (or harder-to-lyse Gram positives) and thereby decreasing (or increasing) the mean efficiency in samples collected from a host post-treatment. Such situations can arise whenever there is a microbial trait that is associated with both measurement efficiency and the biological processes of interest. An example besides cell-wall structure is ribosomal copy number, which increases a species' efficiency in ribosomal amplicon measurements and is positively associated with metabolic rate. A more generic mechanism by which mean-efficiency associations might arise is from the evolutionary relationships among species. Species with more recent common ancestry are expected to be more similar across a wide range of heritable traits, which include traits that affect measurement efficiency (such as cell wall structure, genome size, ribosomal copy number, PCR binding sequence) as well as traits that affect the biological processes under study. For example, differences in the phylum-level composition of samples from two conditions might be driven by many species that show phylum-level conservation in a relevant biological trait. If these species also show phylum-level conservation in (potentially different) traits that affect measurement efficiency, an association of mean efficiency with the condition can arise.

So long as the mean efficiency does not vary systematically with the regression covariates, the error in fold changes serve primarily to reduce the precision and power of regression analyses without systematically distorting estimates and so does not necessarily lead to invalid inferences. But the reduction in precision could be substantial, particularly given the small sample sizes common in many microbiome studies, and thereby substantially limit the ability to draw meaningful conclusions from a study[3].

These observations provide reasons to both worry and hope. It seems likely that in many experiments the mean efficiency is consistent or is at least not associated with the covariate, so that DA inferences remain valid. Yet it is not obvious *a priori* in which studies this condition holds, and there are plausible mechanisms that can create problematic associations of the mean efficiency even in ecosystems with high species diversity. Thus while we should not discount the large set of existing DA results, we should seek ways to better assess the robustness of results from previous studies and to measure and correct the error caused by bias in future ones.

# 5 Potential solutions

The theoretical results from Section 3 suggest a number of potential methods to avoid, correct, or otherwise mitigate the errors created by taxonomic bias.

## 5.1 Use ratio-based relative-abundance analyses

Since bias creates a consistent fold error in species ratios, a natural approach to countering bias when analyzing relative abundances is to use DA methods that are based on fold changes in ratios. A variety of such methods have been developed for microbiome DA analysis that build on earlier work from the field of Compositional Data Analysis (CoDA). These methods analyze log-fold changes in the ratios among a pair of species or, more generally, some product of exponential power of multiple species, which likewise are bias-invariant. Besides a greater robustness to taxonomic bias, such ratio-based analyses have other advantages. First, analysis of ratios avoids other standard criticisms leveled at proportion-based analysis—namely that the change in proportion of a species is affected by change in the density of all other species and the set of species used in the denominator of the proportion calculation (Gloor et al.

[3] I suspect this non-systematic variation in the mean efficiency is the more typical situation, but we currently lack an example of it in the case studies. We can find a semi-real example by taking real microbiome data from a case-control-type design and simulating a strong degree of bias. I did this with the HMP2 IBD study, but thinnk it might be better to find a different example where there are some clear DA results.

(2017)). Second, ratios may be more informative than absolute density in some cases, such as for swabs or other sample types for which the absolute density in the sample may be an unreliable indicator of the absolute density of the sampled ecosystem.

Ratios come with their limitations, however. First, there are a vast number of species ratios (and derivatives) to potentially consider, none of which provide direct information about absolute densities, which can make choosing and interpreting a ratio-based DA analysis challenging. Second, ratio measurements are impacted by noise in both the numerator and denominator and so can be noisier than proportion measurements, particularly when the read counts in the denominator are small. A related problem is that, due to the discrete nature of biological organisms and the reads that are generated during sequencing, species in the denominator are commonly observed to have 0 reads; the assumptions that made by different methods about the meaning of these 0 counts can have a major impact on DA results. Finally, bias invariance at the level of ratios among species does not extend to ratios of additive aggregates of species, such as between bacterial phyla, unless further conditions are met (McLaren, Willis, and Callahan (2019)).

## 5.2 Calibration using community controls

Measurement of community calibration controls along with the primary experimental samples can enable researchers to directly measure and remove the effect of bias prior to or concurrent with downstream DA analysis. Community calibration controls are samples whose species identities and relative abundances are known either by construction or by characterization with a chosen 'gold standard' or *reference protocol* (McLaren, Willis, and Callahan (2019)). MGS measurement of one or more control communities can be used to directly measure the relative efficiencies among the superset of species in the controls (McLaren, Willis, and Callahan (2019)). The measured relative efficiencies can then be used to calibrate (remove the effect of bias from) the relative abundances (ratios) of the control species in a set of primary (non-control) experimental samples that were measured with the same protocol (ideally in the same experiment and sequencing run) (McLaren, Willis, and Callahan (2019)). Calibration can be extended to species not in the controls using statistical methods to impute (predict) the unobserved efficiencies using phylogenetic relatedness, genetic characteristics (such as 16S copy number), and/or phenotypic properties (such as cell-wall structure).

The calibrated compositions can be used for arbitrary downstream microbiome analyses, including relative and absolute DA analysis. To demonstrate the potential for calibration to improve fold-change measurements, we estimated bias from one sample in the mock community data from Brooks et al. (2015) that contained all 7 species and used it to calibrate the compositions of all samples (Figure 5). We then estimated species' densities using the uncalibrated and calibrated proportions using total-abundance normalization, treating the total density as known and fixed. Visual inspection shows that the bias estimated from just a single control community with all species can be sufficient to greatly reduce the effect of taxonomic bias on the measured fold change in species' proportions and densities (Figure 5). To demonstrate a practical application, we used the bias estimated from the full set of mock communities as the basis for calibrating the vaginal community time series measurements from the MOMS-PI study (Fettweis et al. (2019)) that were made using the same 16S sequencing protocol. We imputed the efficiencies of other species using taxonomic relationships (Methods). Calibration then allowed us to explore the impact of bias on the measured trajectories of species proportions within women over the course of pregnancy as described in Section 4.

## 5.3 Calibration using reference species

Obtaining control communities that span the taxa of interest is often not practical. Moreover, the bias that is measured from controls may differ from that in the primary samples due

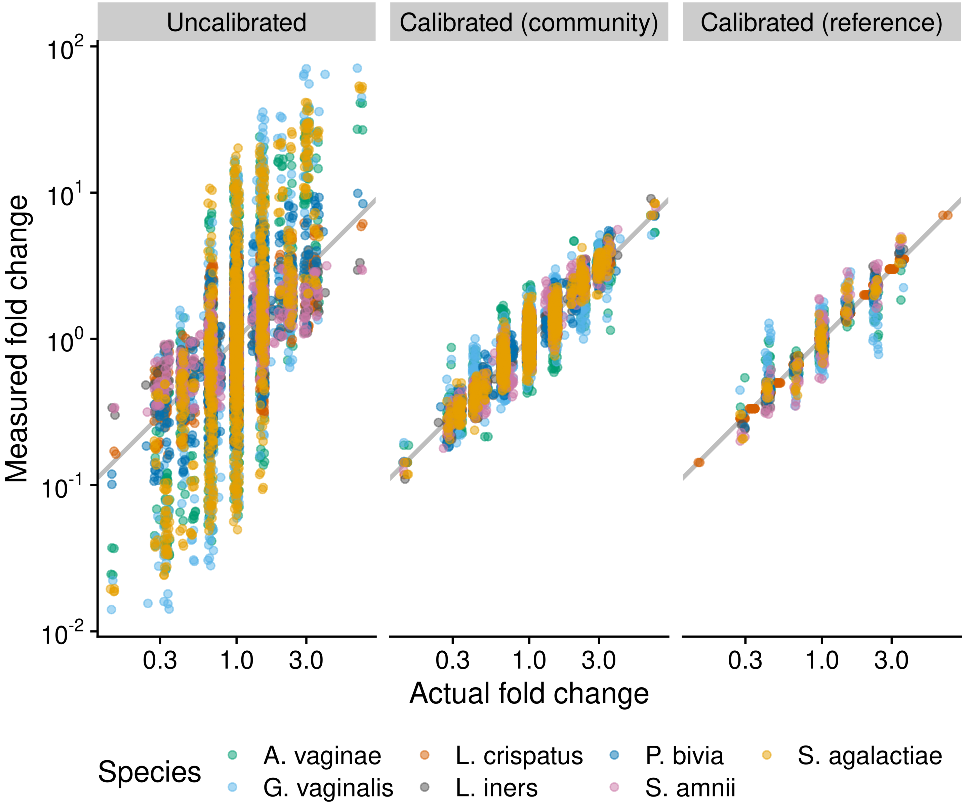

Figure 5: **Fold changes can be calibrated using community controls or reference species.** The figure compares the performance of three methods for measuring fold changes in absolute cell density in cellular mock communities of 7 vaginal species, which were constructed and measured via 16S sequencing by Brooks et al. (2015). The 'Uncalibrated' fold changes are derived directly from uncalibrated individual abundance measurements, which equal the product of the species' proportion by the total density (which here is be known to be constant by construction). The 'Calibrated (community)' measurements are computed from abundance measurements where the proportions are first corrected for the taxonomic bias that was estimated from a single sample that contained all 7 species. The 'Calibrated (reference)' measurements are computed from abundances measured with the reference-species method, with *Lactobacillus crispatus* used as the reference; that is, the true abundance of *L. crispatus* is treated as known and used to infer the abundance of the remaining 6 species. Only samples that contain *L. crispatus* are included.

to differences in cell state or sample preparation. These problems may be overcome for the purpose of species-level absolute DA analysis by using the reference-species approach to absolute-density measurement that was described in Section 3. This *reference-species calibration* allows one to simultaneously account for changes in total density and the sample mean efficiency, obtaining accurate fold changes in absolute density from taxonomically biased, relative metagenomics measurements.

Recall that densities estimated by the reference-species approach have a constant fold error, provided that the fold error in the reference species is also constant (Equation (15)). Three categories of reference species are reviewed for this purpose in Appendix B, including why they might be expected to produce constant fold errors: housekeeping species assumed to have constant abundance, spike-in species added at known or constant abundance, and species whose abundance has been measured with a targeted method like species-specific qPCR.

To demonstrate the ability for reference calibration to improve fold changes, we treated the abundance of one species (*Lactobacillus crispatus*) in the Brooks et al. (2015) mock community data as known, and used it to calibrate the abundances of all species. Doing so improved the resulting measurements of fold changes in density for all species (Figure 5).

## 5.4 Choosing complementary total-density and metagenomics measurements

A variety of methods for measuring total community density have been paired with either amplicon or shotgun sequencing data for the purpose of conducting absolute DA analysis using the total-density approach described in Section 3. Each of these methods is subject to its own taxonomic bias: Cells from different species are likely to contribute more or less efficiently to the total abundance measurement. Consider the popular method of assessing bacterial abundance with total-16S qPCR measurement. Even an ideal qPCR protocol that perfectly amplifies all species remains a biased measurement of cell density, since the total 16S copies in the extracted DNA contributed by each species will be proportional to its (species-specific) lysis efficiency and 16S copy-number. 16S qPCR is commonly paired with 16S amplicon sequencing, with which these sources of bias are shared, perhaps along with variation in primer binding and amplification efficiency. Methods like using flow cytometry that directly measures cell density lacks these biases but likely have their own that are more likely to be orthogonal to the 16S sequencing measurement. By extending the analysis of Section 3 to include consistent taxonomic bias in the total-density measurement, we find that pairings of total-density and sequencing measurements that share large sources of taxonomic bias can lead to an offsetting of errors that reduces the error in fold-change measurement and DA inference (Appendix A.2). This finding suggests that methods of total-density measurement that are more accurate as measures of total cell density may actually perform worse than less accurate methods for the purposes of absolute DA inference.

We illustrate with a hypothetical example of two vaginal communities of equal density, but which are dominated either by *Lactobacillus iners* or *Gardnerella vaginalis*. Based on our earlier analysis in McLaren, Willis, and Callahan (2019), we estimate that *L. iners* yields 10X more 16S copies per cell than *G. vaginalis* in extracted DNA. Ignoring other sources of bias, we expect relative efficiencies of *L. iners*:*G. vaginalis* of 10 for both qPCR and 16S sequencing measurements. Consequently, the *G. vaginalis* community will yield a substantially lower qPCR measurement than the *L. iners* community and we would incorrectly infer a decrease in total abundance. At the same time, this decrease in the mean efficiency of the sequencing measurement would cause artificially high fold changes in the proportions of all species (Section 3). When using qPCR and 16S sequencing measurements for community-based density estimation, these two opposing errors cancel, yielding accurate fold changes in species densities.

More generally, changes in the mean efficiency of the total-density measurement can offset those in the sequencing measurement when comparing log density of species across samples. If the (log) efficiencies of the total and sequencing measurement are positively correlated across species, then the (log) mean efficiencies will tend to be positively correlated across samples, leading to a reduction in error in DA analysis relative to total-density methods whose efficiencies are uncorrelated with those of the sequencing measurement. These observations suggest that qPCR of a marker gene may be the ideal pairing for amplicon sequencing measurements, despite being a poor measure of changes in total cell density. Similarly, bulk DNA quantification may be an ideal pairing for shotgun sequencing, making it possible to account for variation in lysis efficiency and genome length. These pairings even make it possible to account for error caused by variation among samples in the fraction of unclassified reads, which can form a major fraction of amplicon and shotgun data. Importantly, maximizing the offsetting of errors requires thoughtful choices during bioinformatic analysis, perhaps eschewing the filtering and normalization steps used in many software packages and workflows (Appendix A.2).

## 5.5 Bias-sensitivity analysis

For experiments that have been conducted without calibration controls, it is still possible to investigate the likelihood that DA results can be explained by taxonomic bias by performing a computational bias-sensitivity analysis.

A bias-sensitivity analysis examines how sensitive the results of a given DA analysis are to the working assumption about the bias of a given protocol. This working assumption is typically is that there is no bias, but could also be bias measured in a previous experiment or predicted from taxonomic features like 16S copy number. A straightforward approach to bias-sensitivity analysis is to use computer simulation to re-analyze the MGS data across a set of simulated taxonomic biases. First, multiple sets of efficiency vectors are randomly generated, each representing a possible quantification of taxonomic bias for the study. Second, each efficiency vector is used to calibrate the original MGS measurements; each set of calibrated measurements represents the true community compositions under the given hypothesized bias. Finally, the DA analysis is re-run on each set of calibrated measurements and the distribution of results is compared graphically or with summary statistics. A graphical summary of such an analysis being applied to results from the corncob frequentist DA- analysis tool ‘corncob’ are shown in SI Figure 12. The simulated efficiency vectors may be generated to different hypotheses about bias in the given system, such as its overall magnitude and correlation among phylogenetically-related species. Thus the utility of such simulations will increase the more we learn about the properties of taxonomic bias in different systems.

An alternative approach to bias-sensitivity analysis would be to directly include the (unknown) taxonomic bias of the protocol in the statistical model used for DA analysis, as described for bias-aware meta-analysis below. Further development of tools and workflows for performing bias sensitivity analyses could be a valuable way to assay and improve the reliability of microbiome results—for differential abundance and microbiome analyses more generally.

## 5.6 Bias-aware meta-analysis

So far we have considered analysis of samples subject to the same taxonomic bias; however, meta-analysis of microbiome samples measured across multiple studies must often contend with a diversity of protocols and hence taxonomic biases in samples from different studies. This inconsistency in taxonomic bias across studies poses a problem even for ratio-based analyses. A potential solution is to perform a bias-aware meta-analysis that that explicitly accounts for the possibility of distinct taxonomic bias across studies.

Parametric microbiome meta-analysis models include sets of species-specific latent parameters to model “batch effects”—amorphous differences in the measurements of each study. By configuring these models so that these latent parameters correspond to study-specific species efficiencies, we can improve their biological interpretability and gain the ability to directly assess whether disagreements in DA results across studies can be explained by differences in taxonomic bias. To the extent that taxonomic bias truly is consistent within studies, meta-analyses that directly account for it can have greater statistical power to extract the truly consistent and inconsistent microbial patterns across studies.

# 6 Conclusion

*This section is a placeholder for a short (2-3 paragraph) conclusion. We are planning to keep 'discussion' to the earlier sections and appendix, at least for initial submission.*

# A Model details

This appendix provides a more detailed description of our deterministic model of the microbiome measurement process, which it uses to derive some of the additional claims of the main text.

## A.1 MGS measurement

This section expands the description of the model in the main text to more explicitly lay out our assumptions and to consider the effect of variation in measurement efficiencies within a taxonomic group on its MGS measurement.

A taxon can be any group of organisms. For any taxon $I$ and sample $a$, we define the absolute efficiency of that taxon and that sample as the number of reads assigned to that taxon divided by the number of cells of that taxon in the given sample. We assume that these assigned reads really do come from that particular taxon and sample (though this assumption is often violated in practice).

The efficiency of a taxon $I$ is the average number of reads produced by each of its constituent cells. This number varies due to

1. Genetic variation among cells
2. Non-genetic variation among cells
3. Randomness in the experimental process

Variation encoded in the genome of cells determines properties that affect its efficiency, such as how easy it is to lyse, how easily PCR primers will bind to their target locus, how many copies of a marker gene are in the genome, and whether its sequencing reads can be taxonomically assigned. This variation created by genetically encoded traits is what we refer to as 'taxonomic bias' and is the primary focus of this article. In principle, even a single nucleotide change (e.g. in a primer-binding site) could affect the efficiency of a cell. Nevertheless, cells will more similar genomes will tend to have more similar genetically-encoded efficiencies. For the purposes of our analysis, we suppose that cells within the same species have approximately identical genetically-encoded efficiencies. The impact of violations in this assumption can be understood using the approach we describe below for taxa above the species level. More generally, our results can be read as applying to the finest taxonomic unit that is reported by the given MGS method.

Genetically identical cells can differ in cell state such that their efficiencies vary significantly. An extreme example is the effect of sporulation: Spores are generally harder to lyse than vegetative cells, and VanInsberghe et al. (2020) found that the efficiency of a common DNA extraction protocol was 800X on vegetative cells vs spores of a strain of *Clostridium difficile*. Physiological differences might also arise simply due where they are in the cell cycle. To the extent that the cells of a given genotype tend to have a similar distribution of cell states across samples, the efficiency of the genotype will remain stable. We assume this stability throughout our analysis. Violations in this assumption can again be understood using an approach similar to that we use to considering taxa above the species level.

The final number of reads also varies due to idiosyncratic events that we typically think of (and model) as 'random', such as the precise handling of a sample during pipetting or loading of a sequencing flow cell. Such variation may or may not be dwarfed by the 'random' variation in the sample composition (i.e., the variation not explained by the covariates in a regression analysis). For simplicity, we ignore this random variation when considering the impact of taxonomic bias on MGS measurement. The effect of random variation in regression analysis is addressed in Appendix C and the effect of random counting error is discussed in Section 4.

The absolute efficiency of a taxon can also vary simply because more effort is put towards sequencing that sample, either intentionally or unintentionally. The sequencing effort might depend on the taxonomic composition and bias of a sample. For example, a sample with higher output DNA will often be diluted more prior to amplification or sequencing. The critical assumption we make regarding sequencing effort is that it affects all species in the sample equally and hence doesn't affect their relative abundances.

With these assumptions in place, we can partition the multiplicative difference between the reads assigned to a taxon and its density in the source sample into a taxon-specific factor and a sample-specific factor. There is an extra degree of freedom, which we settle by equating the taxon-specific factors with relative measurement efficiencies and arbitrarily choose the relative efficiency of a particular species to equal 1.

Our model for the read counts of a species $i$ in sample $a$ can thus be expressed mathematically as

$$\text{reads}_i(a) = \text{density}_i(a) \cdot \text{efficiency}_i \cdot \text{effort}(a). \tag{20}$$

Note that the species' efficiencies are properties of the species $i$ but not the sample $a$. Similarly, let $\text{density}_I(a)$, $\text{reads}_I(a)$, and $\text{efficiency}_I(a)$ denote the density, read count, and relative efficiency of an arbitrary group of species $I$ in sample $a$. The read count for taxon $I$ in sample $a$ can be written

$$\text{reads}_I(a) = \text{density}_I(a) \cdot \text{efficiency}_I(a) \cdot \text{effort}(a). \tag{21}$$

The efficiency of taxon $I$ in sample $a$ equals the weighted average of its constituent species,

$$\text{efficiency}_I(a) \equiv \frac{\sum_{i \in I} \text{density}_i(a) \cdot \text{efficiency}_i}{\text{density}_I(a)}. \tag{22}$$

The ratio between the read counts of two taxa $I$ and $J$ is

$$\widehat{\text{ratio}}_{I/J}(a) = \frac{\text{density}_I(a)}{\text{density}_J(a)} \cdot \frac{\text{efficiency}_I(a)}{\text{efficiency}_J(a)}. \tag{23}$$

Equation (7) for the error in a species' proportion and Equation (9) for the error in the ratio between two species both follow directly from this more general expression.

## A.2 Taxonomic bias in total-density measurement

To understand the impact of taxonomic bias in the total-density measurement on community normalization, we model error in total-density measurements similarly to that for MGS measurements. Let $\widehat{\text{density}}_S(a)$ be the measurement of the density of all species $S$. We suppose that this number is the sum of the (unobserved) contributions from each species. The contribution of a species $i$ is proportional to its actual density $\text{density}_i(a)$ and its *absolute efficiency for the total-density measurement* $\text{efficiency}_i^{\text{tot}}(a)$. The contributions may also be multiplied by a common sample-specific error factor to account for non-linearity of extraction and/or random error, which we ignore for now. The measured total density therefore equals

$$\widehat{\text{density}}_S(a) = \sum_{i \in S} \text{density}_i(a) \cdot \text{efficiency}_i^{\text{tot}} \tag{24}$$

$$= \text{density}_S(a) \cdot \text{efficiency}_S^{\text{tot}}(a) \tag{25}$$

where

$$\text{efficiency}_S^{\text{tot}}(a) \equiv \frac{\sum_{i \in S} \text{density}_j(a) \cdot \text{efficiency}_i^{\text{tot}}}{\text{density}_S(a)} \tag{26}$$

is the mean efficiency in the sample with respect to the total-density measurement.

The error in the total density depends on the species composition through the mean efficiency of the sample; hence spurious changes in measured density can occur simply due to shifts from high- to low-efficiency species (or vice versa).

How does this error affect the measured densities of individual species when the total is used for normalization? Substituting the mean efficiency of the total measurement for the error term in Equation (13) gives

$$\widehat{\text{density}}_i(a) = \text{density}_i(a) \cdot \frac{\text{efficiency}_i \cdot \text{efficiency}_S^{\text{tot}}(a)}{\text{efficiency}_S(a)}. \tag{27}$$

The mean efficiency of the total measurement appears in the numerator, while the mean efficiency of the MGS measurement is in the denominator. To the extent that these two are positively associated across samples (specifically, their logarithms are positively linearly correlated), then the variation in one will tend to be offset by the variation in the other. In this case the (geometric) variation in the ratio will be reduced, and the fold error in the species density measurement will tend to be more consistent than if the total density measurement was taxonomically unbiased (i.e. $\text{efficiency}_S^{\text{tot}}(a) = 1$ for all samples). The error in species densities for individual samples may not be reduced; but the error in fold changes and regression analysis across samples will be.

The extent to which the log mean efficiency of the MGS measurement and the total measurement are correlated depends on the species efficiencies as well as the changes in species composition across samples, making it difficult to make general statements about this effect. We can get a sense for the effect by considering the correlation between log efficiencies across species between the two measurement types. Suppose that distribution over species of log efficiencies of the MGS measurement has variance $\sigma^2$, the log efficiencies for the total density measurement has variance $\sigma_{\text{tot}}^2$, and the correlation between the two is $\rho$. The variance in the difference $\log \text{efficiency}_i^{\text{tot}} - \log \text{efficiency}_i$ is thus $\sigma^2 + \sigma_{\text{tot}}^2 - \rho\sigma\sigma_{\text{tot}}$. All else equal, the smaller this variance, the smaller we expect the variation in the error factor $\text{efficiency}_S^{\text{tot}}(a)/\text{efficiency}_S(a)$.

**Unclassified reads:** It is common that a given MGS protocol generates sequencing reads from particular taxa that the bioinformatics portion entirely discards, due to an inability to taxonomically assign the reads to the chosen taxonomic units. For instance, 16S data that is analyzed with closed-reference OTU assignment and shotgun data analyzed by mapping to a database of reference genomes will be unable to assign sequences that are less than a chosen similarity cutoff (e.g. 97%) from the reference sequences. These taxa may form an appreciable fraction of the community and thus lead to a large fraction of the total sequenced reads for a given sample may remain unassigned.

So far, we have defined the efficiency of the MGS protocol in terms of taxonomically classified reads and so have treated such taxa as having an MGS efficiency of 0. If such species are included in the total density measurement, then we have a mismatch between the two measurement types. However, we may be able to remove this mismatch by altering the equation we use to form species density measurements. In particular, we can compute the proportion of the focal species from the ratio of its assigned reads to the total sequenced reads, rather than to the total set of assigned reads. This approach raises additional questions around how to treat reads from assignable species that are filtered during quality control steps.

Suppose that the set of species that can be classified by our bioinformatics pipeline, $S$, is a proper subset of a larger set of species $S'$ that are present and yield sequencing reads in our sample. Suppose that there is no additional taxonomic bias in our MGS measurement; that is, all species in $S$ have an efficiency of 1, and the species in $S' \setminus S$ have an efficiency of 0. Moreover, suppose that all reads from the $S$ species are classified; that is, there is no loss of reads due to routine QC filtering. Further suppose that we are capable of perfectly measuring the total density of all cells from the full set of species $S'$; that is, $\text{efficiency}_i^{\text{tot}} = 1$ for all species $i \in S'$.

In this case, we can overcome the mismatch between MGS and total measurement when we use them to measure species densities, by using the total reads rather than the assigned reads to compute the species proportions. That is, instead of

$$\widehat{\text{prop}}_i(a) = \frac{\text{reads}_i(a)}{\text{reads}_S(a)}, \tag{28}$$

we take

$$\begin{aligned} \widehat{\text{prop}}_i(a) &= \frac{\text{reads}_i(a)}{\text{total reads}(a)} &(29)\\ &= \frac{\text{reads}_i(a)}{\text{reads}_{S'}(a)}. &(30) \end{aligned}$$

By accounting for the contribution of unknown species when computing proportions, we are able to resolve the mismatch and obtain accurate species densities.

## A.3 Using reference species for total-density normalization

Constant reference species are sometimes used to measure total density of $S$ by the ratio of $S$ reads to $R$ reads. For example, a study of *Arabidopsis* microbiomes used the ratio of bacterial to host reads in shotgun sequencing as a proxy for total bacterial density, which they then used for total-community normalization of 16S amplicon sequencing measurements (Karasov et al. (2020), Regalado et al. (2020)). Chng et al. (2020) similarly used the ratio of bacterial to host or diet reads in shotgun sequencing of mouse fecal samples as a proxy for total bacterial density (though they did not use this measurement for community normalization). Smets et al. (2016) similarly used the ratio of non-spike-in to spike-in reads to estimate total density.

What is the impact of taxonomic bias on these total density estimates and the species densities derived from them? The measured density of $S$ is

$$\begin{aligned} \widehat{\text{density}}_S(a) &= \frac{\text{reads}_S(a)}{\text{reads}_R(a)} &(31)\\ &= \frac{\text{density}_S(a)}{\text{density}_R} \cdot \frac{\text{efficiency}_S(a)}{\text{efficiency}_R} &(32)\\ &= \text{density}_S(a) \cdot \text{efficiency}_S(a) \cdot \text{constant}. &(33) \end{aligned}$$

Hence variation in the mean efficiency among the species $S$ across samples creates a variable fold error in the density measurement, similar to direct measurements of total density.

The impact of this variable error for species density measurement via Equation (10) depends on whether the species proportions are derived from the same or a different MGS measurement.

If proportions from the same MGS measurement are used, then the measured species densities are

$$\widehat{\text{density}}_i(a) = \frac{\text{reads}_i(a)}{\text{reads}_S(a)} \cdot \widehat{\text{density}}_S(a) \tag{34}$$

$$= \frac{\text{reads}_i(a)}{\text{reads}_R(a)}. \tag{35}$$

where the second line follows by substitution of the previous equation. The variable error in the total density cancels with that in the proportion, yielding a constant fold error. In fact, the measurement we obtain is exactly the same as that from reference normalization (Equation (14)).

The situation differs when the total density and community composition are estimated using different sequencing and/or bioinformatic methods, as in the plant microbiome example above. In this case, the mean efficiency of the total density measurement will not equal that in the proportions, and the more general case of total-density normalization discussed in Appendix A.2 applies.

Although we only consider constant reference species, similar behavior occurs if for varying reference species assayed by targeted measurements.

## A.4 Linearity of extraction

Some popular absolute-abundance methods use post-extraction DNA (or RNA) density as a proxy for cell density in the original sample; in particular, using florescence-based total DNA quantification or using qPCR or ddPCR to quantify a marker-gene. In addition, some have suggested normalization of species to spike-ins of extraneous DNA added after extraction. Both of these approaches presuppose that DNA extraction is linear in the sense that the DNA concentration is proportional to the cell concentration in the original sample (possibly after correction by known dilution or concentration factors). Deviations from linearity can occur due to random and systematic variation in DNA yields. Here we consider the impact of these deviations on species density measurements.

One source of systematic variation in yield is taxonomic bias. We already explored the impact of taxonomic bias in extraction.

NOTE: This might not be a correct way to talk about this; should perhaps make a more explicit model to verify this argument.

To understand the impact of non-linearity after bias has been accounted for, let $C(a)$ be a sample-specific factor that accounts for DNA extraction non-linearity after controlling for bias. From Equation (13), we have the more general equation

$$\widehat{\text{density}}_i(a) = \text{density}_i(a) \cdot \frac{\text{efficiency}_i \cdot \text{efficiency}_S^{\text{tot}}(a)}{\text{efficiency}_S(a)} \cdot C(a). \tag{36}$$

$C(a)$ might fluctuate randomly among samples, or might vary systematically — for example, if DNA yield saturates at high concentrations then $C(a)$ will be smaller in higher-concentration samples. If DNA extraction is not linear, then measures of cell density might give more accurate fold changes even if they share less bias with the MGS measurement.

Post-extraction spike-ins and targeted measurements face a similar problem. These perfectly control for bias in fold change estimation *if* DNA extraction is linear. However, if not then the fold error in $\widehat{\text{density}}_r(a)$ for a species with a targeted measurement will vary with $C(a)$

and cause error in fold changes. The same problem applies to DNA spike-ins added after DNA extraction (although different accounting is needed since $\text{density}_r(a)$ no longer has meaning).

# B Review of experimental methods for obtaining absolute densities

There are many experimental techniques to be able to add absolute-density information to MGS measurements. Here we review the experimental techniques; the next section considers the implications for systematic error.

*NOTE: Right now I don't consistently address why various targeted methods might be expected to produce constant fold errors. In revision, seek to connect each method with the relevant theory.*

## B.1 Measurement of total cell density

Total cell density in the original sample can be directly measured by cell counting, either via microscopy (Kevorkian et al. (2018), Lloyd et al. (2020)) or flow cytometry (Props et al. (2017), Vandeputte et al. (2017)). Total cell density or biomass can also be measured via properties assumed to be proportional to cell density, such as fluorescence (as in fluorescence spectroscopy, Wang et al. (2021)), components of microbial cell membranes (as in PLFA analysis, Smets et al. (2016)), and the rate of microbial respiration (as in SIR method, Smets et al. (2016)). So far, it is primarily the cell counting methods that have been used for species-density measurement (rather than simply for the total density), by multiplying the estimated total density by the MGS proportions.

## B.2 Measurement of total DNA density post extraction

It is also common to use density of bulk DNA or a marker gene as a proxy for total community density. Marker-gene density is typically measured with qPCR or ddPCR using 'universal primers' that target the marker gene of interest (typically the 16S gene for bacterial microbiome experiments) (Tettamanti Boshier et al. (2020), Jian et al. (2020), Galazzo et al. (2020)). Bulk DNA density can be measured using fluorescence-based DNA quantification assays (Contijoch et al. (2019); Korpela et al. (2018)). In either case, the DNA density is measured after DNA extraction and so is affected by taxonomic bias in the extraction process, such as variation in lysis efficiency among species. Other sources of bias that affect the DNA density measurement include variation in marker-gene copy number (for marker-gene density) and variation in genome size (for bulk DNA density).

The measured DNA density is typically used as a direct proxy for cell density in the original sample. In particular, it is assumed that a doubling of cell density in the original sample leads to a doubling of DNA density in the extraction (possibly after adjustment for known dilution factors). This *linearity assumption* may be violated for several reasons. First, because of taxonomic bias. For example, samples dominated by easy-to-lyse species will give more DNA per cell than samples dominated by hard-to-lyse species. Second, systematic non-linearity may occur in the DNA yield as a function of input, even if species composition is held fixed. For example, DNA yield may saturate at high sample inputs. Third, the DNA yield may vary apparently randomly due to subtle differences in sample chemistry or handling during the experiment.

## B.3 Equivolumetric protocol

A large part of the reason that there is not a direct correspondence between total density in the sample and total reads sequenced is that MGS experiments are typically intentionally designed to yield a similar number of sequencing reads from each sample, regardless of total density. Cruz, Christoff, and Oliveira (2021) propose instead designing the MGS experiment so as to make total reads proportional total density. The 'equivolumetric protocol' they develop represents a first attempt in this direction. In their protocol, total reads is a saturating function of total density; this function can be measured with a calibration experiment, and the calibration curve used to predict the total density in the source sample. This total density estimate is then used to scale the read counts to estimate species densities in a manner equivalent to the total-community density method.

## B.4 Housekeeping species

We use *housekeeping species* (by analogy with housekeeping genes used for normalization in RNAseq experiments) to refer to species whose density is assumed to be constant, either in the MGS sample or in the source ecosystem it is derived from.

Housekeeping species can sometimes be identified from prior scientific knowledge. Several studies that have employed shotgun sequencing of host-associated microbiomes have use the plant or animal host for this purpose. A study of *Arabidopsis* microbiomes used the ratio of bacterial to host reads in shotgun sequencing as a proxy for total bacterial density, which they then used for total-community normalization of 16S amplicon sequencing measurements (Karasov et al. (2020), Regalado et al. (2020)). Chng et al. (2020) similarly used the ratio of bacterial to host reads in shotgun sequencing of mouse fecal samples as a proxy for total bacterial density (though they did not use this measurement for community normalization). They also use reads from dietary plants for the same purpose. Wallace et al. (2021) used shotgun sequencing to study the virome of *Drosophila*, and normalized virus reads to *Drosophila* reads to measure viral abundance per fly. Organelle reads can also be used. Diener et al. (2021) use mitochondria reads in 16S sequencing of mouse fecal pellets to assess total microbial load, though in a qualitative fashion (as mitochondrial reads were only non-zero at very low bacterial densities induced by antibiotics).

In some cases, there may also be microbes or viruses thought to have stable densities. A recent example is that the most abundant DNA virus in human feces, crAssphage, and the most abundant RNA virus in human feces, Pepper Mild Mottle Virus, have been treated as stable reference species in wastewater monitoring for SARS-CoV-2. Although primarily used in the context of qPCR measurements, these viruses could also be used as references in RNA and DNA shotgun sequencing experiments.

Housekeeping species may not be known *a priori*; to address this case, several methods have been put forward to computationally identify unchanging microbes species directly from MGS measurements. These studies are often focused on mammalian gut bacterial communities. It is perhaps unreasonable to expect bacterial species to be unchanging across hosts, but weaker assumptions can be made to develop normalization methods for the MGS measurements with a similar spirit to reference-based normalization. These studies have instead developed normalization methods based on assumptions such as that most species do not change between any pair of samples (David et al. (2014)) or that the mean (log) abundance between two sample conditions is unchanged for at least some species (Mandal et al. (2015), Kumar et al. (2018)).

When housekeeping species are sequenced along with the primary MGS measurement, they can be used to obtain species densities via reference-normalization (Equation (14). In this case, the only relevant taxonomic bias is that of the primary MGS measurement; if it is constant then it will cancel in fold-change calculations. Because the density of the housekeeping species is unknown, we can consider either that the density of focal species has a constant error, or is in units of the housekeeping species. Non-constant error might arise if the species is treated as constant when it is in fact not.

Housekeeping species have also been used to estimate total community density by $\text{reads}_S/\text{reads}_R$. This estimate has been used to study variation in total density across samples, or for total-density normalization.

## B.5 Spike-ins

Spike-in methods differ by the biological spike-in material: Cellular spike-ins can be added prior to DNA extraction, and DNA spike-ins can be added prior to or following DNA extraction. In either case, a variety of methods can be used to actually leverage the spike-ins for absolute density analysis.

**Cellular spike-ins:** Cellular spike-ins are added to the sample prior to DNA extraction. Some sample processing has typically occurred prior to spiking, for the purposes of storage (e.g. a freeze thaw cycle) and homogenization. We should expect there to be taxonomic bias between the spike-in species and those naturally in the sample due to genetic differences and because of physiological differences induced by the sample processing prior to spiking and the experimental procedure used to grow and prepare the spike-in cells. Our analysis acknowledges this bias, but assumes that it is consistent across samples. The nominal density of the spike-in species added to each sample is subject to random and systematic fold error; but systematic fold error that is shared across samples will induce a constant fold error in species densities and so not impact DA analysis. For instance, if source stock is actually 1.5X higher concentration than thought, the true spike-in concentration will be 1.5X greater than nominal in all samples and not pose a problem for accurate DA inference beyond leading to a greater than intended sequencing effort being expended on the spike-in.

Example studies include Stämmler et al. (2016), Ji et al. (2019), and Rao et al. (2021).

**DNA spike-ins:** Another possibility is to add DNA spike-ins, derived from natural or artificial sequence. DNA spike-ins can be added the samples before DNA extraction (e.g. Smets et al. (2016), Tkacz, Hortala, and Poole (2018), Zemb et al. (2020)) or after DNA extraction (e.g. Hardwick et al. (2018), Tkacz, Hortala, and Poole (2018)). Adding spike-ins prior to extraction is thought to be preferable as it makes it possible to detect and correct for variation in DNA extraction yield among samples (Tkacz, Hortala, and Poole (2018), Zemb et al. (2020), Harrison et al. (2021)). Below, we consider the distinction between pre- and post-extraction spike-ins in the light of taxonomic bias coupled with other sources of variation in extraction efficiency.

**How spike-ins are used:** Like housekeeping species, spike-ins have been used in a variety of ways to analyze absolute abundances. Let $R$ (for reference) be the spike-in species and $S$ be the native species. Smets et al. (2016) used the ratio of $S$ to $R$ reads as an estimate of total density, which they were interested in for its own sake, though one could imagine then also using this total density estimate in community normalization (Equation (10)). Zemb et al. (2020) used the spike-ins to measure total density from the ratio of $S$ to $R$ qPCR abundance estimates and then used Equation (10). Stämmler et al. (2016) and others used the ratio-based method of Equation (14).

## B.6 Targeted measurements

A variety of methods exist for targeted measurement of absolute density of a specific species (or higher-order taxon). The most common approach is to use qPCR or ddPCR to measure the concentration of a marker gene in the extracted DNA, using primers scoped to the target taxon. This approach is therefore subject to sources of taxonomic bias including extraction, marker-gene copy number, and primer-binding. It is also subject to non-species-specific variation in extraction yields unless these are otherwise controlled for. It is also possible to directly measure cell density using methods. Some species can be measured by CFU counting after plating on selective media (REFs), and ddPCR has been used to direct measure cells (Dreo et al. (2014), Morella et al. (2018)) and viruses (Pavšič, Žel, and Milavec (2016), Morella et al. (2018)) without first performing an extraction. Species-specific florescent probes also make it possible to measure individual species via microscopy or flow cytometry (REFs).

TODO: Argue that these methods may yield constant fold errors.

# C Error in estimated regression coefficients and standard errors

Results from statistics on regression under measurement error can help us understand the effect of bias on DA regression analysis.

Consider Equation (18) for the simple linear regression of log density of a species $i$ on covariate $x$. Let $y$ stand for the actual log density of a focal species $i$, $z$ stand for the log density that we've measured, and $d = z - y$ equal the difference between the two, which here is the log efficiency of species $i$ minus the log mean efficiency of the sample. Let $s_{xy}$ denote the sample covariance between variables $x$ and $y$, $s_x^2 = s_{xx}$ and $s_x$ denote the sample variance and standard deviation, and $r_{xy} = s_{xy}/(s_x s_y)$ denote the sample correlation.

The ordinary least squares (OLS) and maximum likelihood (MLE) estimates of the slope of $z$ equals the sample covariance of $z$ and $x$ divided by the sample variance in $x$ or, equivalently, the sample correlation of $z$ and $x$ multiplied by the ratio of their sample standard deviations,

$$\hat{\beta}_z = \frac{s_{zx}}{s_x^2} = r_{zx} \cdot \frac{s_z}{s_x}. \tag{37}$$

From the (bi)linearity of sample covariances it follows that

$$\hat{\beta}_z = \frac{s_{yx}}{s_x^2} + \frac{s_{dx}}{s_x^2} = \frac{r_{yx} s_y}{s_x} + \frac{r_{dx} s_d}{s_x} = \hat{\beta}_y + \hat{\beta}_d, \tag{38}$$

where $\hat{\beta}_y$ and $\hat{\beta}_d$ denote the slope estimates for $y$ and $d$ (were these values to be known). The absolute error in the estimate $\hat{\beta}_z$ of $\hat{\beta}_y$ is therefore $\hat{\beta}_d$; it is large in a practical sense when $\hat{\beta}_d$ is large (in absolute value) compared to $\hat{\beta}_y$, which corresponds to the covariance of $d$ with $x$ being large compared to the covariance of $y$ with $x$.

In our case, the covariance of $d$ equals the negative of the covariance of log mean efficiency with $x$. The absolute error in $\hat{\beta}$ equals the negative covariance of log mean efficiency scaled by the variance in $x$. This absolute error is the same for all species; however, its practical significance varies depending on its magnitude relative to that of the slope of the actual log densities. For species that covary with $x$ more strongly than the log mean efficiency, the error will be relatively small. This situation might occur either because the mean efficiency varies relatively little across samples or because its variation is relatively less correlated with $x$ compared to the log density of the focal species.

We can similarly understand the impact of measurement error on the precision of our slope estimates. The OLS and MLE estimated standard error in $\hat{\beta}$ are both approximately

$$\hat{\text{se}}(\hat{\beta}) \approx \frac{s_{\hat{\varepsilon}}}{s_x \sqrt{n}}, \tag{39}$$

where $s_{\hat{\varepsilon}}$ is the sample standard deviation of the residuals (Wasserman (2004) Chapter 13). The sample residuals of $z$, $y$, and $d$ have a similar relationship to the regression coefficient estimates,

$$
\begin{aligned}
\hat{\varepsilon}_z &\equiv z - \hat{\beta}_z x && (40)\\
&= (y + d) - (\hat{\beta}_y + \hat{\beta}_d)x && (41)\\
&= (y - \hat{\beta}_y x) + (d - \hat{\beta}_d x) && (42)\\
&= \hat{\varepsilon}_y + \hat{\varepsilon}_d. && (43)
\end{aligned}
$$

(Note, here I've omitted the subscript indicating the dependence on the sample.) It follows that the sample variances of the residuals of $z$, $y$, and $d$ are related through

$$s^2_{\hat{\varepsilon}_z} = s^2_{\hat{\varepsilon}_y + \hat{\varepsilon}_d} = s^2_{\hat{\varepsilon}_y} + s^2_{\hat{\varepsilon}_d} + 2s_{\hat{\varepsilon}_y \hat{\varepsilon}_d}. \quad (44)$$

The standard deviation of the $z$ residuals is increased above that of the $y$ residuals when the $d$ residuals are either uncorrelated or positively correlated with the $y$ residuals, but may be decreased when the $y$ and $d$ residuals are negatively correlated.

In our case, the $d$ residuals equal the negative residuals of the log mean efficiency. It is plausible that for most species, their residual variation will have a small covariance with log mean efficiency and the net effect of variation in the mean efficiency will be to increase the estimated standard errors, as occurs with most species in Figure 3. However, high-efficiency species that vary substantially in proportion across samples may be strongly positively correlated with log mean efficiency such that the estimated standard errors decrease, as we see with Species 9 in Figure 3.

# D Supplemental figures

# 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.* BioMed Central. https://doi.org/10.1186/s12866-015-0351-6.

Callahan, Benjamin J, Daniel B DiGiulio, Daniela S Aliaga Goltsman, Christine L Sun, Elizabeth K Costello, Pratheepa Jeganathan, Joseph R Biggio, et al. 2017. "Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women." *Proc. Natl. Acad. Sci. U. S. A.* 114 (37): 9966–71. https://doi.org/10.1073/pnas.1705899114.

Cartwright, Charles P., Amanda J. Pherson, Ayla B. Harris, Matthew S. Clancey, and Melinda B. Nye. 2018. "Multicenter study establishing the clinical validity of a nucleic-acid amplification–based assay for the diagnosis of bacterial vaginosis." *Diagn. Microbiol. Infect. Dis.* 92 (3): 173–78. https://doi.org/10.1016/j.diagmicrobio.2018.05.022.

Chng, Kern Rei, Tarini Shankar Ghosh, Yi Han Tan, Tannistha Nandi, Ivor Russel Lee, Amanda Hui Qi Ng, Chenhao Li, et al. 2020. "Metagenome-wide association analysis identifies microbial determinants of post-antibiotic ecological recovery in the gut." *Nat. Ecol. Evol.* 4 (9): 1256–67. https://doi.org/10.1038/s41559-020-1236-0.

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.

Cruz, Giuliano Netto Flores, Ana Paula Christoff, and Luiz Felipe Valter de Oliveira. 2021. "Equivolumetric Protocol Generates Library Sizes Proportional to Total Microbial Load in 16S Amplicon Sequencing." *Front. Microbiol.* 12 (February): 1–16. https://doi.org/10.3389/fmicb.2021.638231.

David, Lawrence A, Arne C Materna, Jonathan Friedman, Maria I Campos-Baptista, Matthew C Blackburn, Allison Perrotta, Susan E Erdman, and Eric J Alm. 2014. "Host

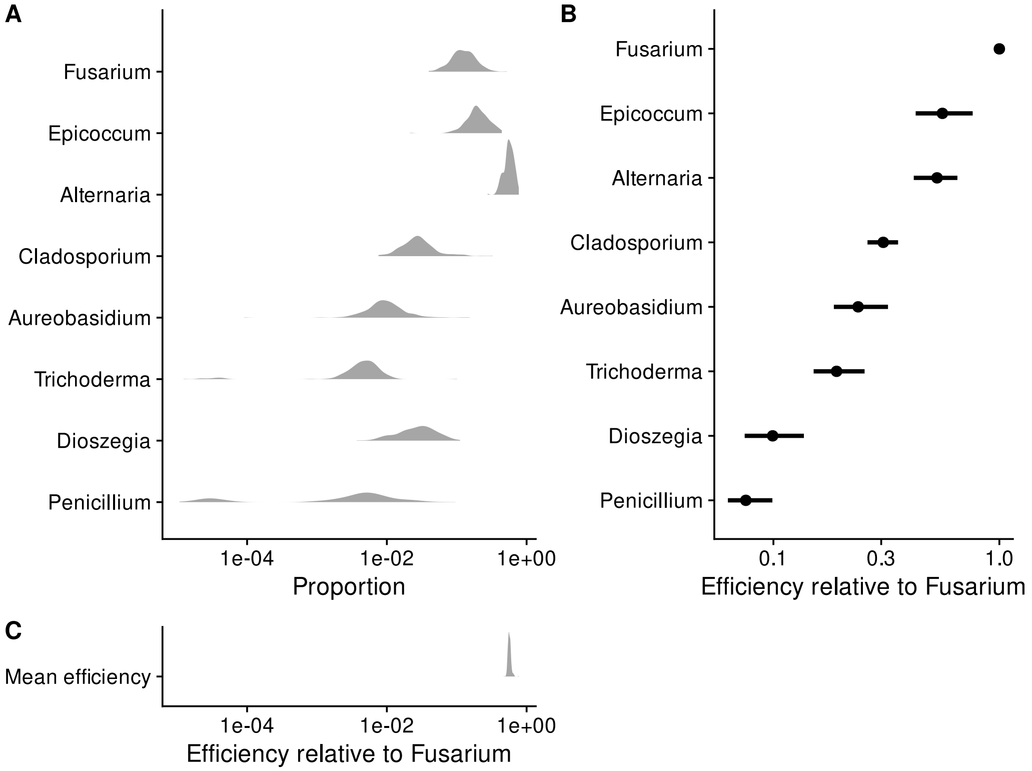

Figure 6: **In the pre-infection samples from Leopold and Busby (2020), multiplicative variation in taxa proportions is much larger than that in the mean efficiency.** Panel A shows the distribution of the proportions of each commensal isolate (denoted by its genus) across all samples collected prior to pathogen inoculation; Panel C shows the distribution of the (estimated) sample mean efficiency across these same samples on the same scale; and Panel B shows the efficiency of each taxon estimated from DNA mock communities as point estimates and 90% bootstrap percentile confidence intervals. Efficiencies are shown relative to the most efficiently measured taxon (*Fusarium*).

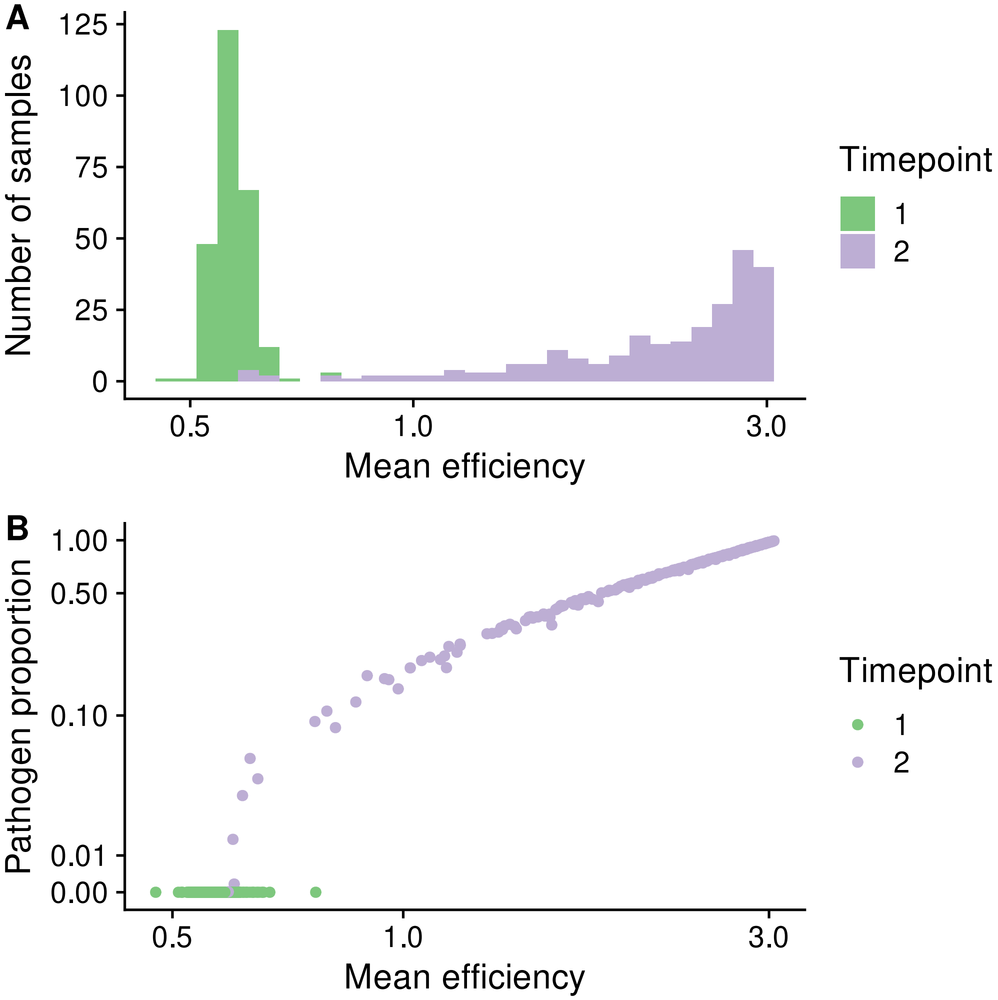

Figure 7: **The mean efficiency tends to increase after infection due to the high proportion of the pathogen.**

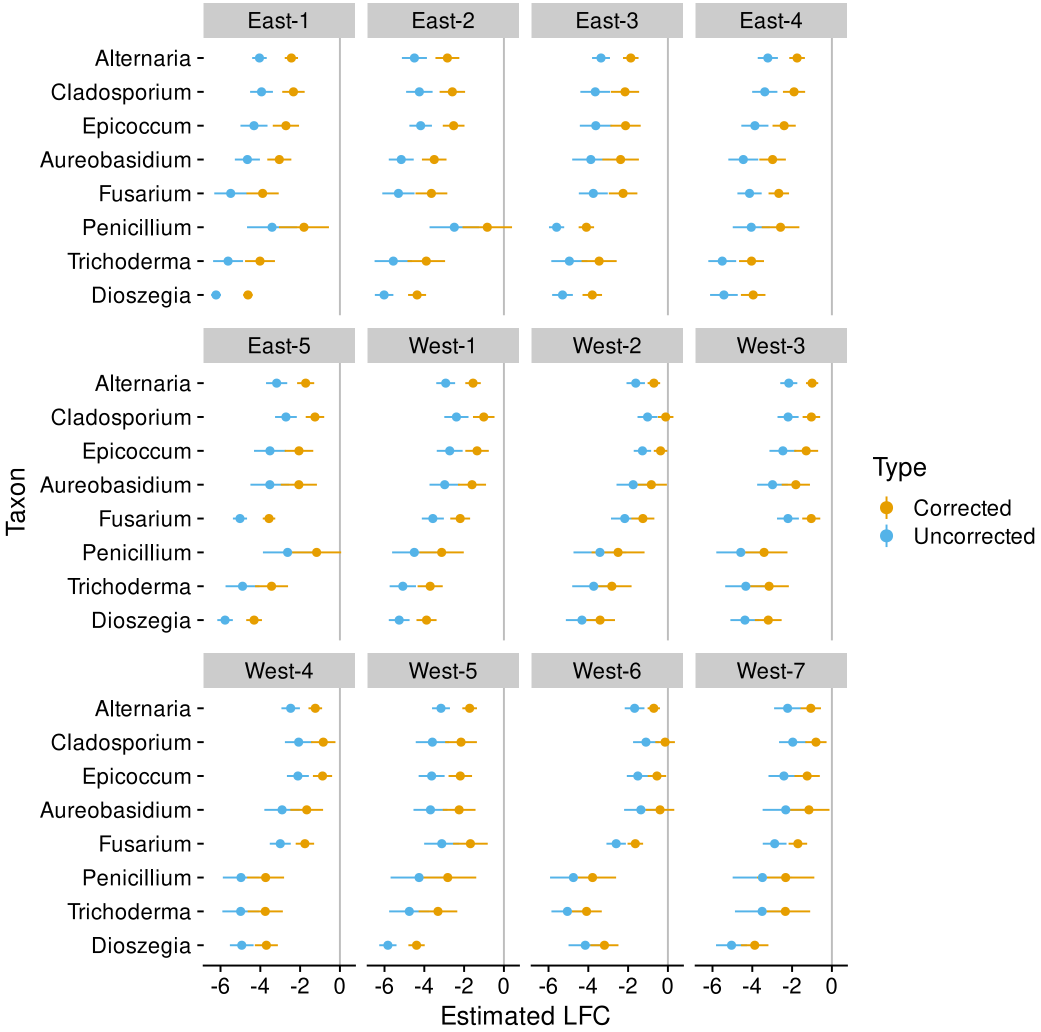

Figure 8: **Bias correction increases the estimated increase in log proportion in response to infection for commensal taxa across all host genotypes.** Shown are the estimated log fold change (LFC) and 95% confidence intervals from simple linear regression of log (base e) proportion against experimental timepoint for commensal taxa. Negative values indicate that the proportion of the taxon decreased on average in response to infection, which we expect due to an increase in pathogen abundance and the sum-to-one constraint of proportions. Bias leads to artificially low estimates, as the increased pathogen proportion drives an increase in mean efficiency.

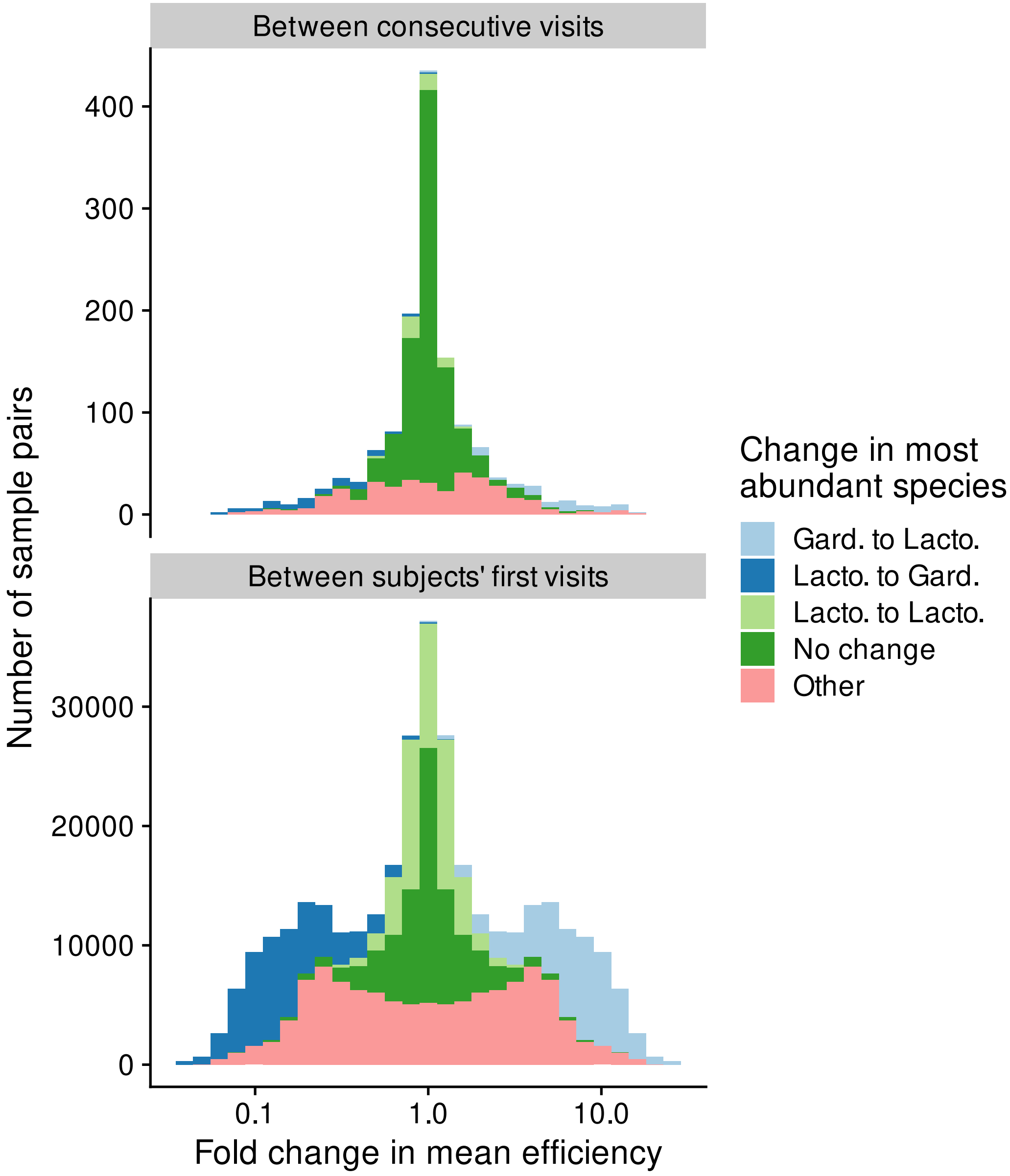

Figure 9: **Fold changes in the mean efficiency within and between women in the MOMS-PI study.**

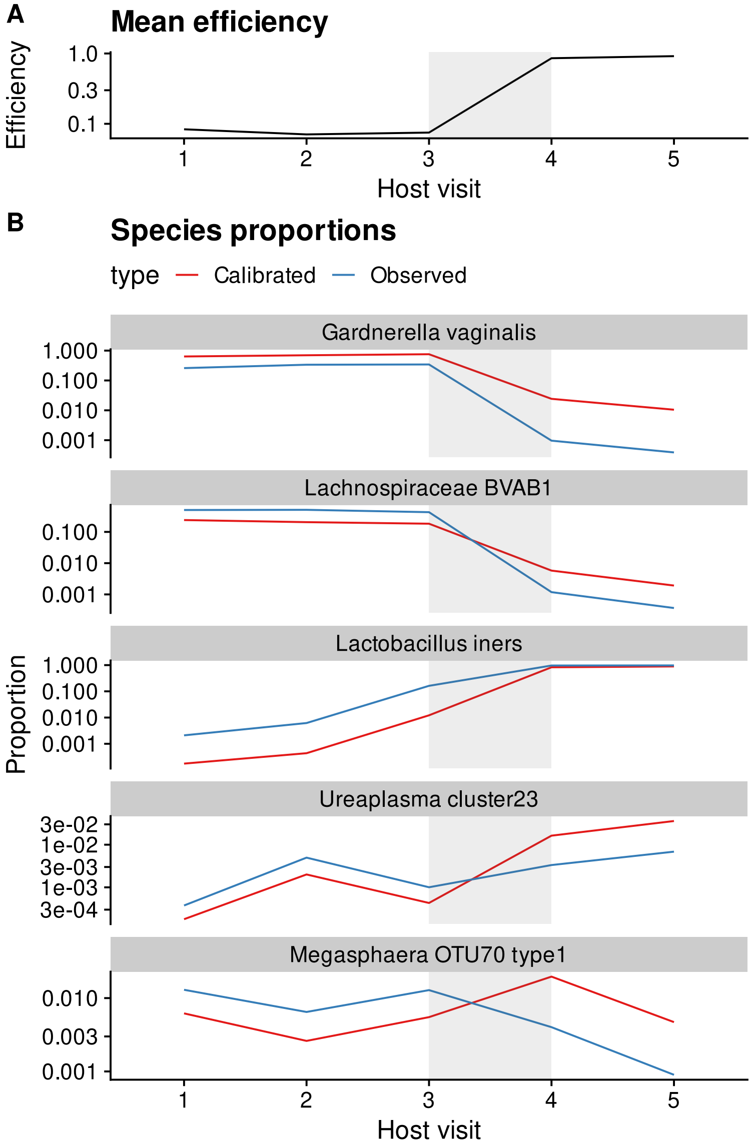

Figure 10: **In vaginal microbiome measurements, shifts between *Lactobacillus* and *Gardnerella* dominance can drive spurious fold changes in other, lower-abundance species.** The figure shows species proportions and mean efficiency trajectories over consecutive clinical visits for a subject in the MOMS-PI study whose microbiome samples showed substantial variation in mean efficiency. The subject's samples are dominated by *Gardnerella vaginalis* and *Lachnospiraceae BVAB1* during the first three visits before transitioning to being dominated by *Lactobacillus iners* between visits 3 and 4. This transition drives a sharp increase in the mean efficiency, which significantly distorts the fold changes in the observed (uncalibrated) microbiome measurements for species with less dramatic fold changes. Two exemplar species are shown to illustrate the magnitude (*Ureaplasma cluster 23*) and sign (*Megasphaera OTU70 type1*) errors that can arise in this situation.

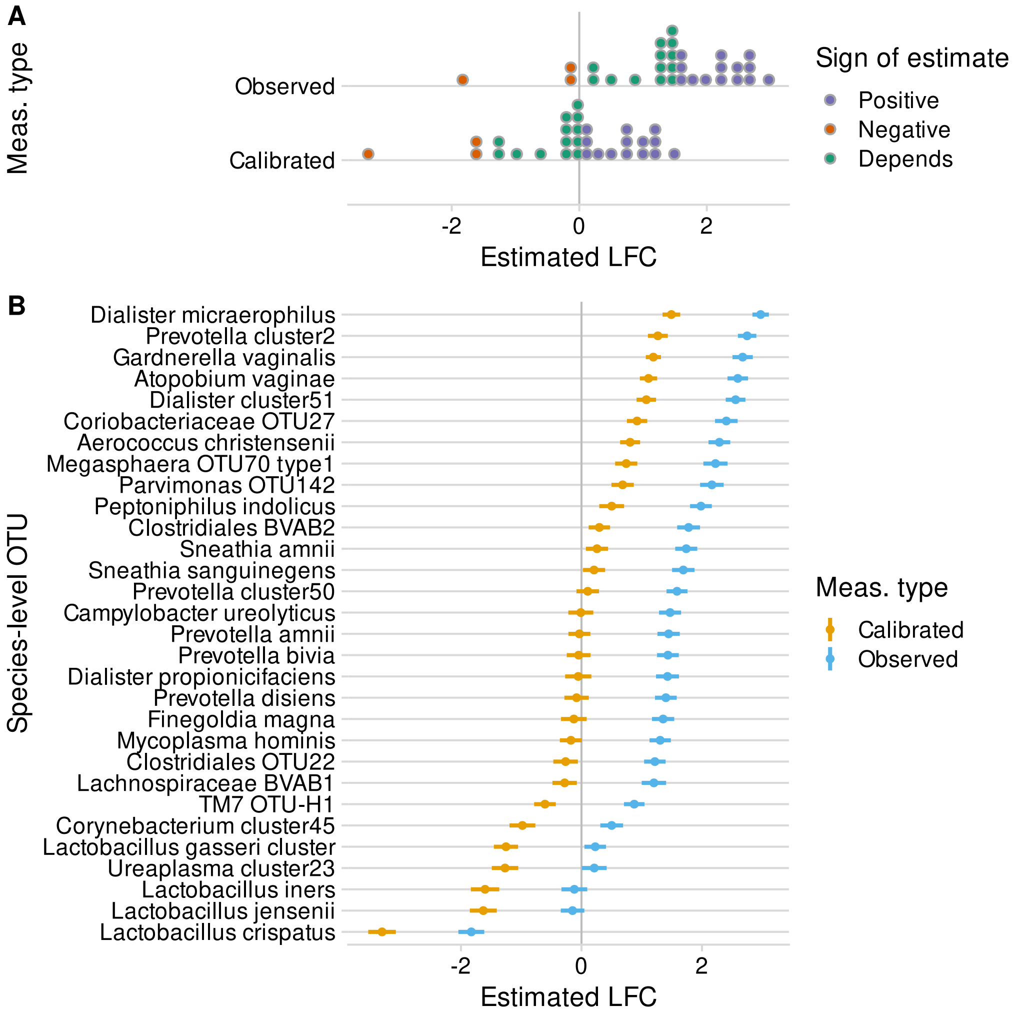

Figure 11: **Bias distorts log fold changes (LFCs) in species proportions in a regression analysis of vaginal microbiome samples from the MOMS-PI study.** Samples were split into low, medium, and high diversity groups based on Shannon diversity in observed (uncalibrated) microbiome profiles. The LFC in proportion from low- to high-diversity samples was estimated for 30 common species by simple linear regression, using calibrated (bias-corrected) and observed (uncorrected) microbiome profiles following a simple zero-replacement procedure. Panel A shows the distribution of point estimates; Panel B shows the point estimates and 95% confidence intervals for each species. The difference between the calibrated and observed estimate for each species equals the negative LFC in mean efficiency.

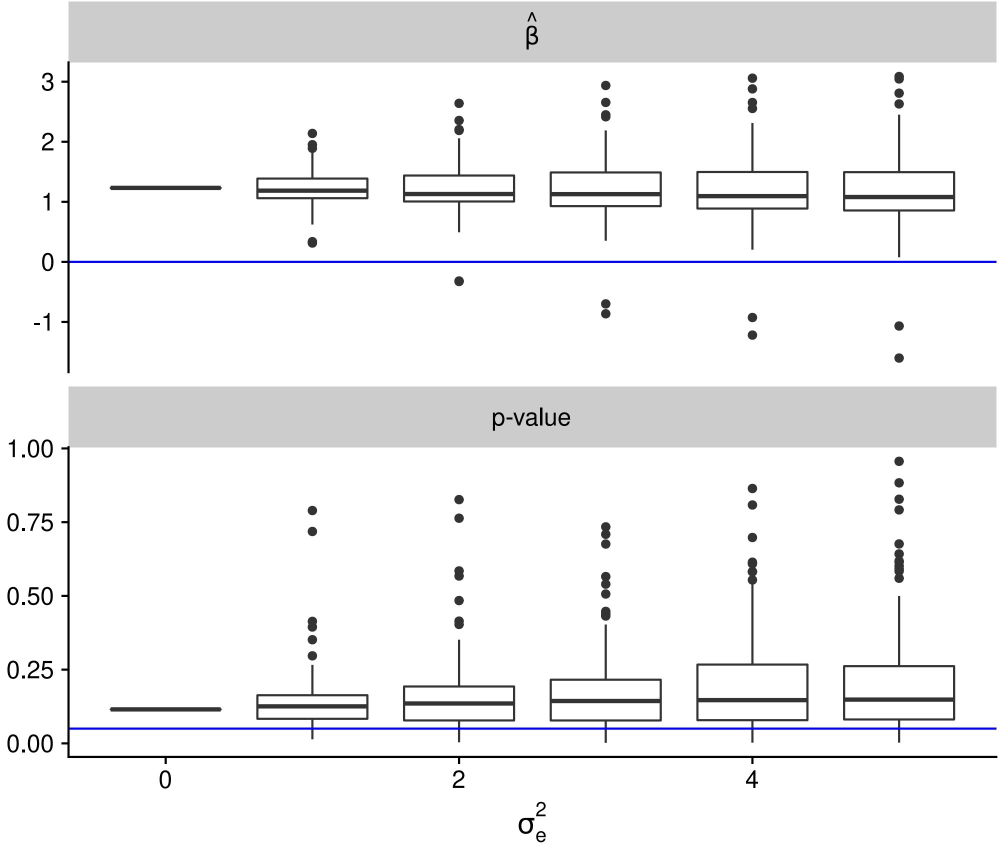

Figure 12: **A bias-sensivity analysis can be performed to examine how sensitive the results of a DA analysis are to assumptions about taxonomic bias in community measurements.** The figure shows the results of a bias-sensitivity analysis used to study the effect of bias on the association of *Gardnerella vaginalis* and preterm birth that was investigated by Callahan et al. (2017). 100 random efficiency vectors were drawn at 6 different bias strengths (quantified by the variance in log efficiency, $\sigma_e^2$). Each efficiency vector was used to calibrate the MGS profiles and perform a DA association test of *G. vaginalis* versus the host's preterm birth outcome; regression coefficients $\hat{\beta}$ indicate the increase of average logit proportion of *G. vaginalis* in women who experienced preterm birth.

lifestyle affects human microbiota on daily timescales." *Genome Biol.* 15 (7): R89. https://doi.org/10.1186/gb-2014-15-7-r89.

Diener, Christian, Anna C. H. Hoge, Sean M. Kearney, Ulrike Kusebauch, Sushmita Patwardhan, Robert L. Moritz, Susan E. Erdman, and Sean M. Gibbons. 2021. "Non-responder phenotype reveals apparent microbiome-wide antibiotic tolerance in the murine gut." *Commun. Biol.* 4 (1). https://doi.org/10.1038/s42003-021-01841-8.

Dreo, Tanja, Manca Pirc, Živa Ramšak, Jernej Pavšič, Mojca Milavec, Jana Žel, and Kristina Gruden. 2014. "Optimising droplet digital PCR analysis approaches for detection and quantification of bacteria: a case study of fire blight and potato brown rot." *Anal. Bioanal. Chem.* 406 (26): 6513–28. https://doi.org/10.1007/s00216-014-8084-1.

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.

Finucane, Mariel M., Thomas J. Sharpton, Timothy J. Laurent, and Katherine S. Pollard. 2014. "A Taxonomic Signature of Obesity in the Microbiome? Getting to the Guts of the Matter." Edited by Markus M. Heimesaat. *PLoS One* 9 (1): e84689. https://doi.org/10.1371/journal.pone.0084689.

Galazzo, Gianluca, Niels van Best, Birke J. Benedikter, Kevin Janssen, Liene Bervoets, Christel Driessen, Melissa Oomen, et al. 2020. "How to Count Our Microbes? The Effect of Different Quantitative Microbiome Profiling Approaches." *Front. Cell. Infect. Microbiol.* 10 (August). https://doi.org/10.3389/fcimb.2020.00403.

Gill, Christina, Janneke H. H. M. van de Wijgert, Frances Blow, and Alistair C. Darby. 2016. "Evaluation of Lysis Methods for the Extraction of Bacterial DNA for Analysis of the Vaginal Microbiota." Edited by Peter E. Larsen. *PLoS One* 11 (9): e0163148. https://doi.org/10.1371/journal.pone.0163148.

Gloor, Gregory B., Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. 2017. "Microbiome Datasets Are Compositional: And This Is Not Optional." *Front. Microbiol.* 8 (November): 2224. https://doi.org/10.3389/fmicb.2017.02224.

Graspeuntner, Simon, Nathalie Loeper, Sven Künzel, John F. Baines, and Jan Rupp. 2018. "Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract." *Sci. Rep.* 8 (1): 9678. https://doi.org/10.1038/s41598-018-27757-8.

Hardwick, Simon A., Wendy Y. Chen, Ted Wong, Bindu S. Kanakamedala, Ira W. Deveson, Sarah E. Ongley, Nadia S. Santini, et al. 2018. "Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis." *Nat. Commun.* 9 (1): 3096. https://doi.org/10.1038/s41467-018-05555-0.

Harrison, Joshua G., W. John Calder, Bryan Shuman, and C. Alex Buerkle. 2021. "The quest for absolute abundance: The use of internal standards for DNA-based community ecology." *Mol. Ecol. Resour.* 21 (1): 30–43. https://doi.org/10.1111/1755-0998.13247.

Ji, Brian W., Ravi U. Sheth, Purushottam D. Dixit, Yiming Huang, Andrew Kaufman, Harris H. Wang, and Dennis Vitkup. 2019. "Quantifying spatiotemporal variability and noise in absolute microbiota abundances using replicate sampling." *Nat. Methods* 16 (8): 731–36. https://doi.org/10.1038/s41592-019-0467-y.

Jian, Ching, Panu Luukkonen, Hannele Yki-Järvinen, Anne Salonen, and Katri Korpela. 2020. "Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling." Edited by Ivone Vaz-Moreira. *PLoS One* 15 (1): e0227285. https://doi.org/10.1371/journal.pone.0227285.

Karasov, Talia L., Manuela Neumann, Alejandra Duque-Jaramillo, Sonja Kersten, Ilja Bezrukov, Birgit Schröppel, Efthymia Symeonidi, et al. 2020. "The relationship between microbial population size and disease in the Arabidopsis thaliana phyllosphere." *bioRxiv.* https://doi.org/10.1101/828814.

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.

Korpela, Katri, Elin W. Blakstad, Sissel J. Moltu, Kenneth Strømmen, Britt Nakstad, Arild E. Rønnestad, Kristin Brække, Per O. Iversen, Christian A. Drevon, and Willem de Vos. 2018. "Intestinal microbiota development and gestational age in preterm neonates." *Sci. Rep.* 8 (1): 1–9. https://doi.org/10.1038/s41598-018-20827-x.

Kumar, M. Senthil, Eric V. Slud, Kwame Okrah, Stephanie C. Hicks, Sridhar Hannenhalli, and Héctor Corrada Bravo. 2018. "Analysis and correction of compositional bias in sparse sequencing count data." *BMC Genomics* 19 (1): 799. https://doi.org/10.1186/s12864-018-5160-5.

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.

Lofgren, Lotus A., Jessie K. Uehling, Sara Branco, Thomas D. Bruns, Francis Martin, and Peter G. Kennedy. 2019. "Genome-based estimates of fungal rDNA copy number variation across phylogenetic scales and ecological lifestyles." *Mol. Ecol.* 28 (4): 721–30. https://doi.org/10.1111/mec.14995.

Lozupone, Catherine A, Jesse Stombaugh, Antonio Gonzalez, Gail Ackermann, Janet K Jansson, Jeffrey I Gordon, Doug Wendel, Yoshiki Va, and Rob Knight. 2013. "Meta-analyses of studies of the human microbiota." *Genome Res.*, 1704–14. https://doi.org/10.1101/gr.151803.112.

Mandal, Siddhartha, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. 2015. "Analysis of composition of microbiomes: a novel method for studying microbial composition." *Microb. Ecol. Heal. Dis.* 26 (1): 27663. https://doi.org/10.3402/mehd.v26.27663.

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.

Morella, Norma M., Shangyang Christopher Yang, Catherine A. Hernandez, and Britt Koskella. 2018. "Rapid quantification of bacteriophages and their bacterial hosts in vitro and in vivo using droplet digital PCR." *J. Virol. Methods* 259 (May): 18–24. https://doi.org/10.1016/j.jviromet.2018.05.007.

Pavšič, Jernej, Jana Žel, and Mojca Milavec. 2016. "Digital PCR for direct quantification of viruses without DNA extraction." *Anal. Bioanal. Chem.* 408 (1): 67–75. https://doi.org/10.1007/s00216-015-9109-0.

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.

Rao, Chitong, Katharine Z. Coyte, Wayne Bainter, Raif S. Geha, Camilia R. Martin, and Seth Rakoff-Nahoum. 2021. "Multi-kingdom ecological drivers of microbiota assembly in preterm infants." *Nature* 591 (7851): 633–38. https://doi.org/10.1038/s41586-021-03241-8.

Regalado, Julian, Derek S. Lundberg, Oliver Deusch, Sonja Kersten, Talia Karasov, Karin Poersch, Gautam Shirsekar, and Detlef Weigel. 2020. "Combining whole-genome shotgun sequencing and rRNA gene amplicon analyses to improve detection of microbe–microbe interaction networks in plant leaves." *ISME J.*, May, 823492. https://doi.org/10.1038/s41396-020-0665-8.

Smets, Wenke, Jonathan W. Leff, Mark A. Bradford, Rebecca L. McCulley, Sarah Lebeer, and Noah Fierer. 2016. "A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing." *Soil Biol.*

*Biochem.* 96: 145–51. https://doi.org/10.1016/j.soilbio.2016.02.003.

Srinivasan, Sujatha, Noah G. Hoffman, Martin T. Morgan, Frederick A. Matsen, Tina L. Fiedler, Robert W. Hall, Frederick J. Ross, et al. 2012. "Bacterial Communities in Women with Bacterial Vaginosis: High Resolution Phylogenetic Analyses Reveal Relationships of Microbiota to Clinical Criteria." Edited by Adam J. Ratner. *PLoS One* 7 (6): e37818. https://doi.org/10.1371/journal.pone.0037818.

Stämmler, Frank, Joachim Gläsner, Andreas Hiergeist, Ernst Holler, Daniela Weber, Peter J. Oefner, André Gessner, and Rainer Spang. 2016. "Adjusting microbiome profiles for differences in microbial load by spike-in bacteria." *Microbiome* 4 (1): 28. https://doi.org/10.1186/s40168-016-0175-0.

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.

Tkacz, Andrzej, Marion Hortala, and Philip S. Poole. 2018. "Absolute quantitation of microbiota abundance in environmental samples." *Microbiome* 6 (1): 110. https://doi.org/10.1186/s40168-018-0491-7.

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.

VanInsberghe, David, Joseph A. Elsherbini, Bernard Varian, Theofilos Poutahidis, Susan Erdman, and Martin F. Polz. 2020. "Diarrhoeal events can trigger long-term Clostridium difficile colonization with recurrent blooms." *Nat. Microbiol.* 5 (4): 642–50. https://doi.org/10.1038/s41564-020-0668-2.

Wallace, Megan A., Kelsey A. Coffman, Clément Gilbert, Sanjana Ravindran, Gregory F. Albery, Jessica Abbott, Eliza Argyridou, et al. 2021. "The discovery, distribution, and diversity of DNA viruses associated with Drosophila melanogaster in Europe." *Virus Evol.* 7 (1): 1–23. https://doi.org/10.1093/ve/veab031.

Wang, Xiaofan, Samantha Howe, Feilong Deng, and Jiangchao Zhao. 2021. "Current Applications of Absolute Bacterial Quantification in Microbiome Studies and Decision-Making Regarding Different Biological Questions." *Microorganisms* 9 (9): 1797. https://doi.org/10.3390/microorganisms9091797.

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.

Yuan, Sanqing, Dora B. Cohen, Jacques Ravel, Zaid Abdo, and Larry J. Forney. 2012. "Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome." Edited by Jack Anthony Gilbert. *PLoS One* 7 (3): e33865. https://doi.org/10.1371/journal.pone.0033865.

Zemb, Olivier, Caroline S Achard, Jerome Hamelin, Marie-Léa De Almeida, Béatrice Gabinaud, Laurent Cauquil, Lisanne M. G. Verschuren, and Jean-jacques Godon. 2020. "Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard." *Microbiologyopen* 9 (3): 1–21. https://doi.org/10.1002/mbo3.977.