The power of microbiome studies: some considerations on which alpha and beta metrics to use and how to report analysis the results

Jannigje Gerdien Kers (annelies.kers@wur.nl)
Wageningen University & Research https://orcid.org/0000-0002-9951-7549
Edoardo Saccenti
Wageningen University and Research: Wageningen University & Research

Research Article

Keywords: Microbiota, power analysis, multivariate analysis, sample size

Posted Date: July 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-698991/v1

The power of microbiome studies: some considerations on which alpha and beta metrics to use and how to report analysis the results

Kers J.G.[1], Saccenti E.[2]*

[1]Laboratory of Microbiology, Wageningen University & Research, Stippeneng 4, 6708 WE, Wageningen, the Netherlands

[2]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708 WE, Wageningen, the Netherlands

Corresponding Author

* Edoardo Saccenti, Laboratory of Systems and Synthetic Biology, Wageningen University, Stippeneng 4, 6708 HB, Wageningen, The Netherlands. Email: edoardo.saccenti@wur.nl; Tel: +31 (0)317 482018; Fax: +31 (0) 3174 83829

ABSTRACT

Since sequencing techniques become less expensive, larger sample sizes are applicable for microbiota studies. The aim of this study is to show how, and to what extent, different diversity metrics and different compositions of the microbiota influence the needed sample size to observed dissimilar groups. Empirical 16S rRNA amplicon sequence data obtained from animal experiments, observational human data, and simulated data was used to perform retrospective power calculations. A wide variation of alpha diversity and beta diversity metrics were used to compare the different microbiota data sets and the effect on the sample size. Our data showed that beta diversity metrics are most sensitive to observe differences compared to alpha diversity metrics. The structure of the data influenced which alpha metrics are most sensitive. Regarding beta diversity, the Bray-Curtis metric is in general most sensitive to observe differences between groups, resulting in lower sample size and potential publication bias. We recommend to perform power calculations and to use multiple diversity metrics as an outcome measure. To improve microbiota studies awareness needs to be raised on the sensitivity and bias for microbiota research outcomes created by the used metrics rather than biological differences. We have seen that different alpha and beta diversity metrics lead to different study power: on the basis of this observation, one could be naturally tempted to try all possible metrics until one or more are found that give a statistically significant test result, *i.e*. *p*-value < α. This way of proceeding is one of the many forms of the so-called *p*-value hacking. To this end, in our opinion, the only way to protect ourselves from (the temptation of) *p*-hacking would be to *publish*, and we stress here the word publish, a statistical plan before experiments are initiated: this practice is customary for clinical trials where a statistical plan describing the endpoints and the corresponding statistical analyses must be disclosed before the start of the study.

Keywords: Microbiota, power analysis, multivariate analysis, sample size

INTRODUCTION

For a few decades now, researchers have left culture-based methods and used molecular technologies, and more recently mostly sequencing-based approaches, to characterize microbial communities within a certain environment, referred to as the microbiome. In humans and animals, the microbiome has an important role in health and disease. For example, animals raised without or fewer microbes showed an underdeveloped immune system and are more susceptible to diseases [1-3]. Microbiome studies have as goal to investigate, characterize and understand the compositional and functional variability of microbiomes. The question "What is different?" between different groups of interests can be translated into an hypothesis testing procedure.

Hypothesis testing rests on the definition and choice of four parameters: *i*) the effect size, *i.e.* the quantification of the outcome of interest (in the simple case the difference between two groups), *ii*) the sample size *n, i.e.* the number of samples (to be) collected; *iii*) the power of tests 1-β, *i.e.* the probability of the test of rejecting the Null hypothesis when actually false and *iv*) the confidence level α, *i.e.* the probability of rejecting the Null hypothesis when actually true.

It is necessary to perform power analysis before performing experiments. This is well acknowledged in all fields of research, however, microbiome studies are challenged with conflicting results [4]. Under-powering and the failure to correct for false positives are among the causes underlying the lack of reproducibility of many biological findings [5, 6]. The power of a test is linked to the probability β of accepting the Null hypothesis when actually false (False negative error or Type II error), and α describes the False positive error or Type I error. Once acceptable error rates α (usually 0.05 or 0.01) and β (usually 0.2, although context-dependent), and the effect that one is interested to assess statistically are chosen, it is possible, at least in principle, to determine the optimal sample size, *i.e.* the number of samples that one need to collect/analyze to obtain, with probability 1-β, a statistically significant result with confidence α.

Given the nature of microbiome data it is possible to quantify differences between groups at two levels: the alpha (within-sample) and beta (between samples) diversity (**Figure 1**). Alpha diversity metrics summarize the structure of a microbial community with respect to its richness (number of taxonomic groups), evenness (distribution of abundances of the

groups), or both [7]. Commonly used alpha metrics are Phylogenetic diversity [8], Observed number of amplicon sequence variants (ASV) [9], Chao1 [10], Simpson [11, 12] and Shannon indices [12, 13]. Beta diversity metrics summarize which samples differ from one another by considering sequence abundances or considering only presence-absence of sequences. Commonly used beta metrics are Bray-Curtis dissimilarity [14], Jaccard [15], unweighted UniFrac [16] and weighted UniFrac [17]. The choice of the diversity metrics affect the subsequent statistical testing and, as a result, how, and to which extent, power analysis can be performed.

Using an alpha diversity metric a single diversity value is obtained for each sample containing measurement over *m* taxa; thus the problem of assessing differences between two (or more) groups can be addressed with a univariate test, like *t*-test, ANOVA or a non-parametric version. The use of a beta diversity metric implies that all *samples* are to be considered simultaneously, and several methods to compare groups of samples measured on $m > 1$ have been proposed like ANOSIM [18] and PERMANOVA [19] to replace classical multivariate tests like the Hotelling T^2 or Multivariate ANOVA, which are in general not applicable because microbiome data does not satisfy basic assumptions. These assumptions include the independence of the sample units, the multivariate normality of errors, homogeneity of variance–covariance matrices among the groups or because the number of variables is larger than the number of samples making it impossible to apply the test [20] [6, 21-23].

While sample size, Type I and Type II error are well defined concepts, the definition of effect size depends on the outcome quantity of interest and how this quantity is mathematically defined. A fundamental step when performing power analysis is to define the effect size: for a simple two-sample *t*-test the effect size can be expressed as the Cohen's δ [24].

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma} \tag{1}$$

where μ_1 and μ_2 are the population means of the two groups and σ^2 is the pooled variance. Since μ_1 and μ_2 are population parameters which are inaccessible and on which we want to perform inference, an *a priori* estimation, or educated guess, is necessary. This can be

accomplished by taking the sample means m_1 and m_2 and pooled variance s from a pilot study or existing data to obtain estimates of the population parameters. A critical aspect that is not sufficiently acknowledged is that the effect size from Equation (1) is sensitive to the particular diversity metric used to the point that sample size calculation can be severely affected.

The aim of this study is to show how, and to what extent, different diversity metrices influence the sample size needed to assess statistical significance of dissimilarities between different microbial communities. Both simulated and empirical 16S rRNA gene amplicon sequence data sets are used to perform retrospective power calculations on microbiota studies. A broad selection of alpha and beta diversity metrics were used to compare the different microbiota data sets. This study generated insight into the sensitivity and bias of certain statistical methods used in microbial ecology on microbiota research outcomes. We conclude with some recommendation for the reporting of power analysis and sample size calculations for microbiome studies.

MATERIAL AND METHODS

Literature search

To support the choice of alpha and beta diversity metrics to consider in our comparison, we performed a literature search on PubMed (www.pubmed.org) with query: (microbiota[Title] OR microbiome[Title]) NOT Review[Publication Type] 2020/01:2020/02 [Date of Publication]). This strategy aimed to include a broad scope of microbiota studies. We limited our search to studies published in English in with free full text.

Alpha diversity metrics

Richness

Richness is the number of taxa, most often defined as operational taxonomic unit (OTU) or amplicon sequence variant (ASV) observed [9]. Where s is the number of observed taxa, calculated as:

$$S_{rich} = \sum_s 1_{\{ps>0\}}. \quad (2)$$

Phylogenetic Diversity

Phylogenetic diversity (PD) is a phylogenetically-weighted measure of richness. Although, the name suggests diversity, it does not take into account the abundance of taxa. The PD is defined as the sum of the lengths of all those branches on the tree that span the members of the set, given the phylogenetic tree spanning *s* taxa [8]:

$$PD = \sum_i b_i \,, \tag{3}$$

where *s* is the number of observed taxa and b_i is the length of the *i*-th branch in the tree. The index *i* runs on all branches.

Chao1

The Chao1 index is an abundance-based non-parametric estimator of taxa richness [10]. Its defined as

$$Chao1 = s + \frac{F_1(F_1-1)}{2(F_2-1)}, \tag{4}$$

where *s* is the number of observed taxa, F_1 and F_2 are the number of OTU/ASV with only one sequence (*i.e.* "singletons") and two sequences (*i.e.* "doubletons"). This metric assumes the number of organisms identified for a taxon to follow a Poisson distribution. Its definition rests on the concept that rare taxa bring most information about the number of missing taxa. This index gives more weight to the low abundances taxa and only the singletons and doubletons are used to estimate the number of missing taxa [10]. This index is particularly useful for data sets skewed towards low-abundance taxa [25, 26]. However, singletons and doubletons are often removed from 16S rRNA amplicon sequence data sets because the difficulty in robustly differentiating singleton errors from real singleton sequences [27, 28].

Shannon index

The Shannon index *H* is an estimator of taxa diversity, combining richness and evenness) [12, 13]. It is defined as

$$H = -\sum_{i=1}^{S} p_i \log(p_i) \,, \tag{5}$$

where *s* is the number of OTU/ASV and p_i is the proportion of the community represented by the *i*-th OTU/ASV. Basically this index is the entropy associated to a given sample and quantifies the uncertainty in predicting the taxa identity of an individual selected at random from the sample. The Shannon's index uses the relative abundances of different taxa, thus diversity depends on both taxa richness and evenness with which organisms are distributed among the different taxa. This index places a greater weight on taxa richness [26].

Simpson index

The Simpson index *D* is an estimator of taxa diversity, combining richness and evenness [11, 12]. It is defined as

$$D = \frac{1}{\sum_{i=1}^{s} p_i^2} \ , \tag{6}$$

where *s* is the number of OTU/ASV and p_i is the proportion of the community represented by the *i*-th OTU/ASV. This index considers taxa evenness more than taxa richness in its measurement [26]; it indicates the taxa dominance and gives the probability of two individuals that belong to the same taxa being randomly chosen. It varies from 0 to 1 and the index increases as the diversity decreases [26].

Metrics for beta diversity

Bray-Curtis dissimilarity

The Bray-Curtis index (BC) [14] measures the compositional dissimilarity between the microbial communities of two samples *i* and *j* based on counts on each samples. It is defined as

$$BC = 1 - \frac{2C_{ij}}{S_i + S_j}, \tag{7}$$

where C_{ij} is the sum of the smallest values for only those taxa in common between the sample *i* and *j*, S_i and S_j are the total number of taxa counted in sample *i* and *j*, respectively. This index ranges between 0 (the two samples share all taxa) and 1 (the two samples do not share any

taxa). It gives more weight to common taxa [29]. The Bray-Curtis dissimilarity is computed pair-wise between all samples.

Jaccard distance

The Jaccard distance (*J*) between two samples *i* and *j* is defined as *J* = 1 – *J(i,j)* where *J*(*i,j*) is the Jaccard index which is defined as

$$J(i,j) = \frac{|i \cap j|}{|i \cup j|}, \quad (8)$$

which is the ratio between the number of members that are common between the two samples and the number of members that are distinct; it is a measure of similarity for the two communities, and ranges between 0 (the communities are different) and 1 (the two communities are identical).

UniFrac distances

Unweighted UniFrac (UF) and weighted UniFrac distances between two samples take into account the phylogenetic tree, and thus phylogenetic distances between community members [17]. In unweighted UniFrac, the distance is calculated as the fraction of the branch length and in weighted UniFrac, branch lengths are weighted by the relative abundance of sequences. The sum of unshared branch lengths is divided by the sum of all tree branch lengths, which results in the fraction of total unshared branch lengths that is defined as

$$\sum_i^n b_i \times \left[\frac{A_i}{A_T} - \frac{B_i}{B_T}\right]. \quad (9)$$

Lozupone et al., [17] defined n as the total number of branches in the tree, b_i as the length of branch *i*, A_i and B_i are the numbers of sequences that descend from branch *i* in communities *A* and *B*, respectively, and A_T and B_T are the total numbers of sequences in communities A and *B*, respectively. In order to control for unequal sampling effort, A_i and B_i are divided by A_T and B_T [17].

Experimental data sets

Chickdata data set: this data set contains 16S rRNA gene amplicon sequence data obtained from a broiler chicken experiment. The data set is described in detail in (Kers et al., 2019[30]); briefly, chickens were raised under three different housing conditions with the same medium chain fatty acid feed intervention. Between those housing conditions, bird management was kept as similar as possible. At the hatchery the chicks were randomly allocated to three different experimental facilities. Data set A contains 70 broilers from a grow-out feed trial facility, data set B contains 70 broilers raised at a floor stable and data set C contains 70 broilers raised in isolators. A feed intervention was used as a tool to create differences in cecal microbiota between broilers within the same housing condition.

HMP data set: this data set was obtained from the Human Microbiome Project (HMP) phase I [31]. It contains 16S rRNA gene amplicon sequence data of 169 stool samples, 150 oral samples, 86 vaginal samples and 69 skin samples. The bodyside microbiomes are all diverse in terms of community membership [31].

In all data sets ASVs were defined as unique sequences. All data was analysed using NG-Tax [32]. Taxonomy was assigned using the SILVA 128 16S rRNA gene reference [33]. An overview of data set characteristics (sample size, the number of ASVs, mean values of different alpha- and beta diversity metrics.) is shown in **Table 1**.

Simulated data set

We simulated two different scenarios described by Simulated data set 1 and 2.

Simulated data set 1, is built starting from 1995 microbial features observed in 169 stool samples from the Human Microbiome Project (HMP), indicated as $\mathbf{X}_1$ in the following. Data sets (name $\mathbf{X}_2$ for sake of simplicity) were created where 2%, 5%, 10%, 25%, 50% and 75% of bacterial features were randomly removed.

Simulated data set 2*,* is a case-control scenario ($\mathbf{X}_1$ controls, $\mathbf{X}_2$ cases, 1995 features and 169 samples each) where 1%, 2%, 5%, 10%, 15% and 20% of bacterial features were differentially abundant. Both simulated data sets have the same phylogenetic structure as the HMP data set. An overview of the characteristics of the simulated data sets is shown in **Table 2**.

Statistical tests for group differences

Univariate statistical analysis

Differences between groups using alpha diversity as determined by using Phylogenetic diversity, Richness (defined as Observed), Chao1, Simpson and Shannon, were assessed using the a Kruskal-Wallis test (setting K = 1000)[34]. A significance threshold α = 0.01 was used in all calculations.

PERMANOVA

Differences between groups using beta diversity as determined by using Bray-Curtis, Jaccard, unweighted UniFrac and weighted UniFrac were assessed using the Permutational Multivariate Analysis of Variance (PERMANOVA) [19]. PERMANOVA is a robust approach to compare groups of samples measured on $m > 1$ variables. It construct ANOVA-like test statistics from a matrix of resemblances (distances, dissimilarities, or similarities) calculated among the sample units, and assesses significance of the observed differences using random permutations of observations among the groups [35]. The Null-hypothesis H_0 tested by PERMANOVA is that the centroids of the groups (in the space of the chosen resemblance measure), are the same for all groups. This test assumes that samples are exchangeable under the Null hypothesis, are independent and have similar multivariate dispersion. The PERMANOVA test statistic is a pseudo ANOVA *F*-ratio:

$$F = \frac{SS_B \backslash (g-1)}{SS_W \backslash (n-g)} \tag{10}$$

where SS_B is the total sum of squares of the (diss)similarities between groups, SS_W is the total sum of squares of the (diss)similarities within groups, g is the number of groups and n is the total number of samples.

The significance of the *F* statistics is calculated by means of permutations (k=9999). The distribution of *F* under the Null hypothesis is generated by permuting g times the sample group labels and recalculating *F* on the permuted data. Significance is expressed as *p*-value calculated as the fraction of permuted *F*-statistics, which are equal to or greater than the pseudo *F*-ratio observed on the original data.

Calculation of the empirical power

From $N_1 \times m$ and $N_2 \times m$ data matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ we random sampled with replacements K $n_1 \times m$ and $n_2 \times m$ data sets and applied a statistical test to assess the difference between $\mathbf{X}_1$ and $\mathbf{X}_2$ (as quantified by any of the alpha and beta metrics) at significance level α =0.01 under the assumption of the Null hypothesis being false. The empirical power of the test is defined as the empirical probability *EPr* of H_0 being rejected calculated as

$$EPr = \frac{\#(H_0 \, rejected | H_0 \, false)}{K} \quad (11)$$

where #() indicate the number of times that H_0 is rejected. For sake of simplicity we consider $n_1 = n_2 = n$ and we varied *n* between X and Y.

Software

All statistical analyses were performed in R version 4.0.2 (R Foundation for Statistical Computing, Austria [36]), using the packages: Phyloseq, Microbiome, and Vegan [37-39]. PERMANOVA was performed using the *adonis* function for the Vegan package. Other power calculations were performed using the G*power software [40] using the "Means: Difference between two independent means (two groups)" as Statistical test and "a priori" and "post hoc" option for the Type of Power analysis. Differentially abundant microbiota profiles were simulated with the microbiomeDASim R package (Williams et al., 2019[41]) using the *gen_norm_microbiome* function. The R scripts can be found on the Github page: https://github.com/mibwurrepo/KersSaccenti-Power.

Data availability

Human Microbiome Project data used in this study is available at (https://github.com/mibwurrepo/Microbial-bioinformatics-introductory-course-Material-2018/tree/master/input_data\). The poultry data is available at https://www.ncbi.nlm.nih.gov/bioproject/ with accession number PRJNA553870. Simulated data sets are available at https://www.systemsbiology.nl/datasets-2/

RESULTS

Motivation example

We begin with a motivational example to show how the choice of the diversity metrics affects the power of a microbiome study, and how the same study may be underpowered if a different metric is used.

Let's suppose we want to plan an experiment to assess whether gut and oral microbial communities are different. A very simple and basic study design would be to collect n_1=n_2 gut and oral samples and compare the alpha diversity between the two conditions (gut *vs* oral) using a two-sample Kruskal-Wallis *t*-test.

We can base our estimation *d* of a very similar effect size δ on data from HMP (**Table 1**). Using four different alpha metrics and Equation (1) we obtained *d* = 1.27 (Phylogenetic diversity), *d* = 0.3621 (Shannon), *d* = 0.58 (Chao1), *d* = 0 (Simpson). These values are markedly different: fixing the power to 0.8 (β = 0.2) and confidence α=0.05, they will lead to dramatically different required total sample size (**Figure 2A**). This clearly indicates that microbiome studies may be severely underpowered depending on which alpha metric was used to compare two (or more) groups.

We also explored the achievable power by fixing the sample size ($n = n_1 + n_2 = 50 + 50 = 100$) and using different effect sizes (**Figure 2B**). Consistently with what observed in **Figure 2A** results vary strongly, providing a clear indication of the risk of underpowering when Shannon diversity is used.

Note that with the use of beta diversity metrics, performing a priori power analysis becomes much more complicated. The classical tools for power analysis cannot be applied since the statistical tools are not parametric: solutions have been proposed in the literature: see, for instance [42-44].

Literature search

Our literature search returned 632 papers matching the search criteria. We selected randomly 100 papers, and of those the material and methods or full text was investigated to obtain an overview of the most frequently used alpha and beta diversity metrics and of the sample sizes used. Of the 100 full text papers, 92(%) papers contained alpha metrics and 83(%) papers contained beta metrics

In 58% of the papers more than one alpha metric was used. In 21% of the papers more than one beta metric was used. An overview of the frequency of the different metrics showed that the Shannon index and Bray-Curtis dissimilarity are the most common metrics (Table 3). There was a wide variance in the used sample size, defined as the smallest number per group: 46(%) papers had a sample size between ≤10 samples, 34(%) papers used between 11-50 samples, 7(%) papers between 51-100 samples, 10(%) papers between 101-1000 samples, and three papers more than >1000 samples.

The power of microbiome studies

As shown in the Motivational example, the choice of a particular alpha (and beta) diversity metric determines the number of samples required to achieve a pre-determined power. Based on this observation we examined two simulated data sets using both alpha and beta diversity metrics to understand the relationship between the sample size, the observed power and the diversity metrics, together with two experimental data sets (Chicken and HMP data sets).

As representatives of testing procedures using alpha and beta diversity measures to compare two groups we selected the Kruskal-Wallis test (for alpha metrics) and PERMANOVA (for beta metrics), respectively. The Kruskal-Wallis test is the non-parametric choice for comparing two groups when the normality assumption does not hold. When comparing two (or more) groups using beta diversity metrics PERMANOVA and ANOSIM (Analysis of Similarity, [18] are popular choices. The two approaches are equally popular (359 hits on Pubmed for PERMANOVA and 341 for ANOSIM), however, Anderson and Walsh [35] showed that while both approaches are sensitive to unbalanced designs and differences in variance within groups, PERMANOVA is a more robust approach: on this ground we based our choice for PERMANOVA.

Power analysis of simulated data sets

For the simulated data sets, the effect size is known a priori and it is expressed as the % of differentially abundant or present/absent microbial features (ASV) (Figure 3). The achievable power for Simulated data set 1 is shown as a function of the sample size (n) for different percentages of present/absent ASV. If 2%, 5% of the ASV are deleted from the data set, none of the alpha diversity metrics was able to capture the difference between data sets X1 and X2, irrespective of the sample size used (Figure 3A-B). When more than 10% of ASV were removed in data set X2 (Figure 3C-E) all measures were somehow able to capture the difference but the

resulting actual power was very different. Overall, Chao1 and observed diversity allowed higher power with the lower sample size (are more sensitive to observe differences), especially in the medium range of differences (10% to 25%, Figure 3C-D), whereas differences are minimal for >25%. Note that in contrast with the Motivation example, here the Phylogenetic diversity was not the metric resulting in the smallest sample size.

The same approach was used across different beta diversity metrics (Figure 4). The Jaccard diversity metric was most sensitive and weighted UniFrac was least sensitive to observe the differences in presence/absence between the data sets (Figure 4B-F). When more than 10% of ASV were removed in data set X2, no difference between data sets was observed by the UniFrac metric (Figure 4C), while with 25% removed, Bray-Curtis and unweighted UniFrac showed a comparable power and sample size (Figure 4D). In this simulated data set, Weighted UniFrac distance needed the highest sample size to observe the difference (Figure 4D-F).

The achievable power for Simulated data set 2 is also shown as function of the sample size for different percentages of differentially abundant ASV (Figure 5). If ≤ 5% of the ASV were differentially abundant in data set X2 as compared to X1, the Simpson metric needed the lowest sample size (is most sensitive) to observe differences between the data (Figure 5A-C). However, if 10% of the ASV were differentially abundant, the Phylogenetic diversity and Chao1 were more sensitive and the Simpson and Shannon metrics less sensitive (Figure 5D). With 15% of the ASV differentially abundant no differences were observed with the Simpson metrics (Figure 5E).

The same approach was used across different beta diversity metrics (Figure 6). Bray-Curtis distance was most sensitive to observe differences, whereas unweighted UniFrac needed the largest sample size. If 2% of the ASV were differentially abundant power of the beta metrics was totally different, for example, a sample size of 15 would result in power: 100 for Bray-Curtis, 50 for Weighted UniFrac, 40 for Jaccard distance, and just 10 with unweighted UniFrac (Figure 6B). However, if 10% of the ASV were differentially abundant, all metrices would result in a power higher than 0.80 (Figure 6D-F).

Power analysis of experimental data sets: Chicken data set

The Shannon index was the most sensitive alpha metric and Chao1 and Phylogenetic diversity were less sensitive metrics to observe a difference between the groups in data set B (Figure 7A). In data set B, Shannon index was also the most sensitive alpha metric but Simpson was

the least sensitive metric (Figure 7B). Unweighted UniFrac distance was the most sensitive beta diversity metric to observe a difference between groups in data set A (Figure 8A). Jaccard distance was the only metric that showed that H3 needed the smallest sample size, indicating that in H3 specific ASV are differentially present between the groups (Figure 8B). Weighted UniFrac was more sensitive compared to unweighted UniFrac to observe a difference between the groups based on their microbial communities (Figure 8C, D). In general, the alpha diversity measures were less sensitive to observe differences between the broilers compared to the beta diversity (Figure 7A-B, Figure 8).

Although no difference in alpha diversity was observed between broilers fed with or without MCFA raised in housing condition 1, the average daily gain and the average daily feed intake were lower in MCFA broilers [30]. Therefore, the difference only observed based on the beta diversity might already be biologically relevant and hence sufficient to draw conclusions in this case. Based on this data set we observed that Shannon is the most sensitive alpha diversity metric to observe differences between groups, resulting in the lowest needed sample size. The sensitivity of the beta diversity, however, was different per data set. Based on this retrospective power calculation two conclusions can be drawn on this study design. First, not enough chickens were sampled to observe a difference in the alpha diversity between broilers fed with or without MCFA raised in data set A. Second, 15 chicken samples instead of 35 samples per group would have resulted in the same conclusion.

Power analysis of experimental data: Human Microbiome Project data set

The samples in this data set were collected from different body sites and are known to have a very distinct origin, and therefore expected to be different in microbial composition. The comparison between different body sites showed wide variation in sample size across different alpha diversities (Figure 7C-F). The difference in sample size was small in the comparison between skin vs oral microbiome samples, all-around ten samples (threshold power 80, 1-β)(Figure 7C). In the skin vs gut microbiome samples, the Simpson and Shannon alpha diversities did not differ, and the phylogenetic diversity was the most sensitive to observe differences (Figure 7D). In contrast, when comparing the gut vs the oral microbiome, the phylogenetic diversity was least sensitive to observe differences, whereas the Shannon and Simpson metrics were different between gut and oral samples (Figure 7E). In the skin vs vaginal microbiome comparison, Simpson and Shannon's alpha diversities were more sensitive

compared to the Observed/Chao1 and phylogenetic diversity (Figure 7F). Based on the different beta diversity metrics, all comparisons between different body sites supported significant differences even when just five samples were compared (data not shown), due to the large difference between communities (Supplementary Figure 1). Therefore the retrospective power calculations were not informative for this data set.

Are microbiome studies underpowered?

Figure 9A shows the distribution of the sample size of the data sets that were analysed, using the Chao1 diversity measure (among others) in 28 of the 100 papers considered in the literature review. The distribution is highly skewed towards 0 with a median of 39 samples per group and a mode of 8 samples. Removing the two outlying studies with >300 samples, resulted in a median of 23 samples.

Even considering that Chao1 was one the best performing measures in both simulated and experimental data sets, these numbers appear to be worryingly low: on experimental data, which have a complicated structure that is impossible to replicate in simulations, it is rarely possible to attain a power of 80% with less than 40 samples per group. Similar considerations hold when PERMANOVA is applied (see Figure 9B), with a median group size of 22 and mode 3.

Reporting of power analysis <<BOX>>

One of the studies we examined in the literary review reported that sample size and power analysis were performed: “Sample sizes were chosen on the basis of pilot experiments and on our experience with similar experiments.” This is commendable but we believe that the way forward is to employ and report in full a standardized summary of sample size calculations performed. The software G*power generates a Protocol for power analysis. For instance for a two group- comparison with a Mann-Withney/Kruskal-Wallis test, one should report, once adapted

t tests – Means: Wilcoxon–Mann–Whitney test (two groups)

Options:	A.R.E. method		
Analysis:	A priori: Compute required sample size		
Input:	Tail(s)	=	One
	Parent distribution	=	Normal
	Effect size d	=	0.5
	Alpha metric	=	Shannon
	α err prob	=	0.05

	Power (1−β err prob)	=	0.8
	Allocation ratio N2/N1	=	1
Output:	Noncentrality parameter δ	=	2.51
	Critical t	=	1.66
	Df	=	99.2
	Sample size group 1	=	53
	Sample size group 2	=	53
	Total sample size	=	106
	Actual power	=	0.803

Together with this, information should be provided as to how effect size was determined, i.e. which pilot data were used and how the effect size was determined.

A similar reporting protocol could be devised if simulations are used in a PERMANOVA setting. Since simulation and /or pilot data must be used in this case, details on the simulations or pilot data should be reported. For instance, using the Chicken data 1 as pilot, one could report the following protocol, taking 100 re-samplings of size 6 to calculate the achievable power:

Test – PERMANOVA

Options:	9999 permutation		
	100 iterations		
Analysis:	Compute achievable power		
Input:	Beta meitcs	=	Bary-Curtis
	α err prob	=	0.05
	Number of groups	=	2
	Number of taxa	=	363
	Sample size group 1	=	6
	Sample size group 2	=	6

Output:

Observed Effect size (average)	ϖ^2	=	0.120682
	Min\Max effect size	=	0.025922\0.3500687
Observed Effect size (average)	f	=	0.2696886
Min\Max effect size		=	0.1319342\ 0.746349
	Numerator df	=	1
	Denominator df	=	10
	Power (1−β err prob)	=	0.97

DISCUSSION

The aim of this study was to assess how, and to what extent, different diversity metrices and compositions of the microbiota influence the needed sample size to observed dissimilar groups. Based on our literature survey we observed that the Shannon and Bray-Curtis metrics are most published metrics. This might because they are often the most sensitive metrics to

observe differences between groups, resulting in a lower sample size. Our results are in line with previous literature that showed that the choice of distance metric may significantly influence the observed results [45].

A well-known phenomenon that can hamper progress in every research field concerns publication biases in reporting mainly positive findings [46]. In microbiota research this might even occur rather unintentionally, by using certain alpha and beta diversity metrics, but it might also be that researchers selectively report only results for the metric that shows significance even when other metrics had been assessed during the analyses. Our results lead to the speculation that many microbiome studies may be underpowered or, conversely, only reporting evidence of very large effects that can be assessed to be statistically significant also with small sample size. However, since effect size and test statistic are not reported, it is impossible to judge on the quality of the results. Not only, this also hampers the use of published studies as pilot studies to perform power analysis and sample size calculations since, as long as data is not de novo re-analysed.

None of the 100 microbiome studies that we have considered reported the effect size. A collaborative project aiming to investigate reproducibility of 100 high-profile psychological studies reported the average effect size observed in the replication studies was approximately half the magnitude of those given in the original studies, leading to a replication success of only 36% [47]. The lack of reported effects makes it impossible to analyse retrospectively microbiome studies, to perform meta-analysis and, more importantly, makes it impossible to check the consistency of the statistical analysis or detect errors.

On the basis of this, reporting of effects and test statistics should be made compulsory in microbiome studies. For the highly used Kruskal-Wallis test, the *H* test statistic is given by:

$$H = \frac{12}{N^2 - 1} \sum_i \frac{R_i^2}{n_i} - 3N + 1 \qquad (13)$$

where N is the total number of samples, n_i is the number of samples in group i and R_i is the total sum of ranks in the i-th group. Note that H is easily obtainable from most software packages.

For Kruskal-Wallis the most common effect is the η^2 which is defined as:

$$\eta^2 = \frac{H - k + 1}{N - k} \tag{14}$$

where H is the value obtained in the Kruskal-Wallis test, k is the number of groups. For instance, for the comparison of the two feedings (feed A and B, see Table 1) from the chicken data set using the Observed alpha diversity one could report:

> *Feed A (n_1= 35) and Feed B (n_2 = 35) samples were compared with Kruskal-Wallis test using the Chao1 metric: H(df) = H(1) = 14.68, P-value 0.0001, η^2 = 0.66, δ = 0.58.*

where *df* indicates the degrees of freedom. Note here that for a two group comparison, the Kruskal-Wallis test is equivalent to the Wilcoxon-Mann-Whitney (WMW) test [48, 49]. For the WMW test the Cohen's δ effect size definition (Equation (1)) also applies [50]. This greatly simplifies power analysis and sample size calculations: we advise to also report δ when two groups are considered.

In addition, performing power analysis for a Kruskal-Wallis is not a simple matter and requires the use of a rather advanced statistical machinery [51, 52]; for instance, such calculations are not included in G*power [40], which is the most complete software for power analysis. Kruskal-Wallis is the non-parametric counterpart of *one*-way ANOVA and as such is used in situations where there are more than two groups. However, whereas power analysis and sample size calculation for a *one*-way ANOVA with more than two groups are "easily" accessible within R or other software packages, this is not the case Kruskal-Wallis testing. We could locate a R package 'MultNonParam' [53] that performs power analysis for the Kruskal-Wallis test with more than two groups, however, it requires the specification of the offsets for the various populations, under the alternative hypothesis. Relating such tools to determine the effect size observed in microbiome data is a matter we believe to be worthy of exploration and brings us back to the problem that statistics and effect size are not easily available for microbiome studies.

The principle of reporting the effect size should also apply when testing is performed using beta diversity metrics, in which case the PERMANOVA pseudo *F*-statistics (see Equation

(10)) and the effect size should be reported. Typical effect measures in ANOVA are the Choen's f^2 and η^2 and the ϖ^2

> *Feed A (n_1= 35) and Feed B (n_2 = 35) samples were compared with PERMANOVA test using the Unweighted Unifrac metric: $F(df_B,df_W) = F(1,68)=6.27$, P-value = 0.0001, $f^2 = 0.092$, 1000 permutations.*

where df_B, df_W indicate the between-groups and within degrees of freedom. These notations follows guidelines of the American Psychological Association which provides standardized formats for the reporting of statistical analysis for statistical procedures [54]. For PERMANOVA the matter complicates considerably: to estimate statistical power and calculate sample size, one must quantify the expected within-group variance and the effect to be expected when comparing two or more groups. A package like micropower [43] in principle allows estimation of PERMANOVA effects quantified by the ϖ^2 value (limited to the UniFrac measure): unfortunately, at the best of our knowledge, the package seems not to be maintained and lacks a proper manual. The original paper presents a table with effects calculated from different studies that could be used as guide, however this metric is not standard. It can be calculated from the PERMANOVA table as

$$\omega^2 = \frac{SS_{effect} - df_{effect}MS_{residual}}{SS_{effect} + MS_{residual}} \quad (15)$$

For the chicken data we observed ϖ^2 values in the range 0.04 – 0.1, depending on the beta metrics, and these are consistent with those reported in Table 1 from [43].

A more common effect measure is the Cohen's f^2-value, i.e. the between group to within group ratio can be easily obtained by the ANOVA table provided by software like the R package Vegan by taking the ratio between the Treatment sum of squares and the Residual sum of squares. The f^2-value is used for power calculation in the ANOVA setting, however, it should not be used to perform sample size calculation for PERMANOVA, not even to obtain a rough indication, since the corresponding F-statistics do not follow an *F* distribution. For instance, when comparing Feed A and B from the first chicken data with PERMANOVA we can derive a Cohen's f^2 = 0.38; if this value is used to perform power calculation for an ANOVA with

two groups with power 80% at α = 0.01 we obtain that 42 sample per group are need. However, comparing with the results in **Figure 8** we see that a 100% power can be obtained with 25 samples per group, no matter which measure is used: this is a clear indication that power analysis for PERMANOVA can be obtained only by mean of simulations. In this light **Figure 8** can be viewed as a priori power calculations using pilot data.

Furthermore, there is a convention, more or less widely accepted, of classifying Cohen's δ effect which we have used in the power calculation for the Kruskal-Wallis into trivial (δ < 0.2) small (δ = 0.2) , medium (δ = 0.5) and large (δ > 0.8). However, this classification is based on what is observed in psychology and does not apply automatically to other fields of research [55]. In microbiome studies the effects may be in same order of magnitude that may be considered large or very large using the standard convention.

In sum, we have seen that different alpha and beta diversity metrics lead to different study power: on the basis of this observation, one could be naturally tempted to try all possible metrics until one or more are found that give a statistically significant test result, *i.e*. *p*-value < α. This way of proceeding is one of the many forms of the so-called *p*-value hacking (*p*-hacking) [56]. *P*-hacking (also called data dredging, significance chasing, significance questing, or selective inference [57]) is the improper use of data (like adding or removing observations) or statistical procedures (like applying many different tests) until a configuration is found that produces a statistically significant result at the desired confidence level [58]. *P*-hacking is an illegitimate practice that promotes unreproducible results, polluting literature and adding to publication bias [59-61].

To this end, in our opinion, the only way to protect ourselves from (the temptation of) *p*-hacking would be to *publish*, and we stress here the word publish, a statistical plan before experiments are initiated: this practice is customary for clinical trials where a statistical plan describing the endpoints and the corresponding statistical analyses must be disclosed before the start of the study and must be adhered to if results are going to be published [62]. This is the only guarantee that data analysis is not manipulated towards artificially inflated significant results. We appreciate that clinical trials are inherently different from microbiome (and other omics) studies which are often exploratory in nature, but as far as statistics is concerned, they are prey of the same traps and pitfalls. It is obvious that such a change in the approach to

microbiome studies requires the concerted cooperation of researchers, journal editors, reviewers and publishers.

FUNDING

This research was funded by NWO Earth and Life Sciences (ALW) and Cargill Animal Nutrition with project number 868.15.020, and the European Union (H2020-SFS-2018-1 project MASTER-818368).

ACKNOWLEDGMENTS

We would like to acknowledge and thank Hauke Smidt for this input into this manuscript.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

FIGURES

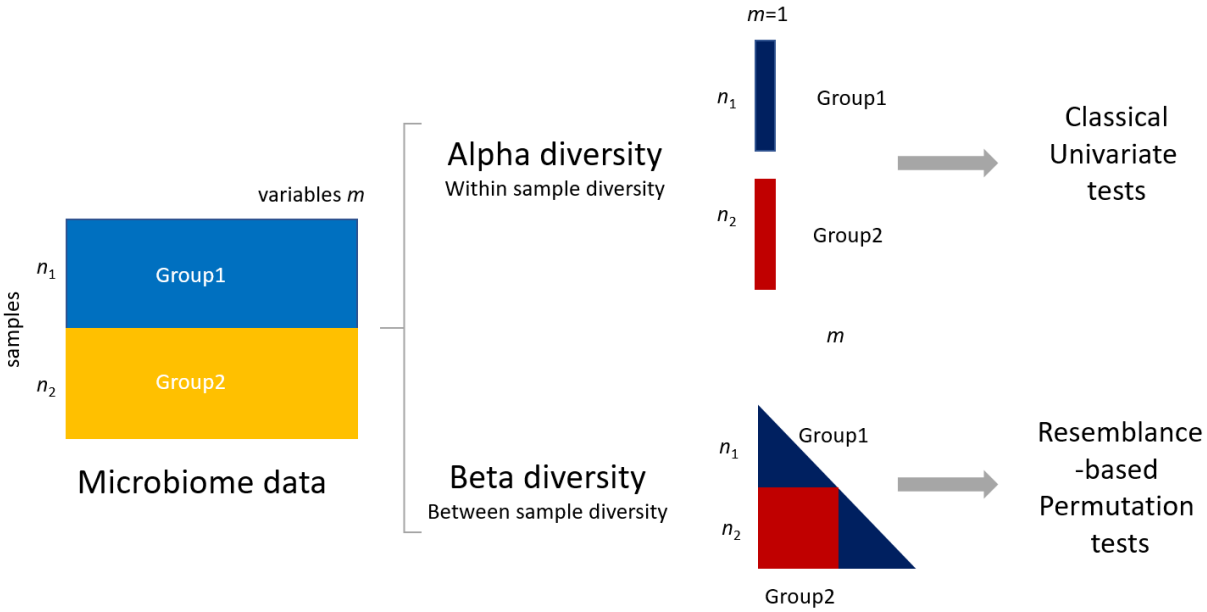


Figure 1

Differences between Group 1 and Group2 in microbiome data can be assessed using either alpha (within sample diversity) or beta (between sample diversity) metrics. The use of alpha metrics allows the use of classical univariate testing, either parametric or non-parametric. The use of beta metrics leads to the use of permutation-based testing approaches like PERMANOVA (see Material and Methods).

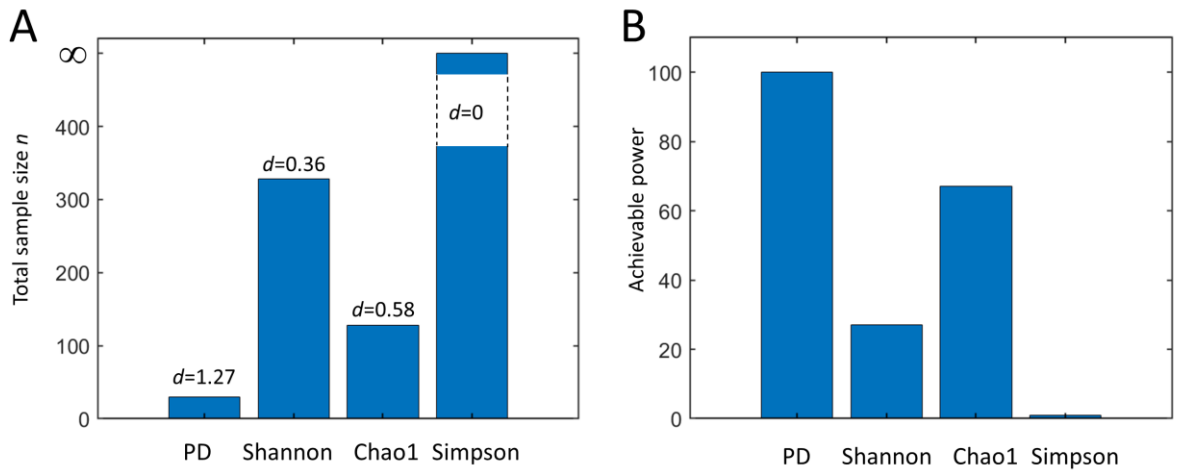


Figure 2

Total sample size (n) required to assess the statistical significance of an effect d using a two-sample Kruskal-Wallis test with a power equal to 0.8 and confidence α = 0.01. B) Achievable power attainable by a Kruskal-Wallis test using a total sample size $n = n_1 + n_2 = 50 + 50 = 100$ using different alpha metrics. Note that for a null effect (d=0) the achievable power coincides with α.

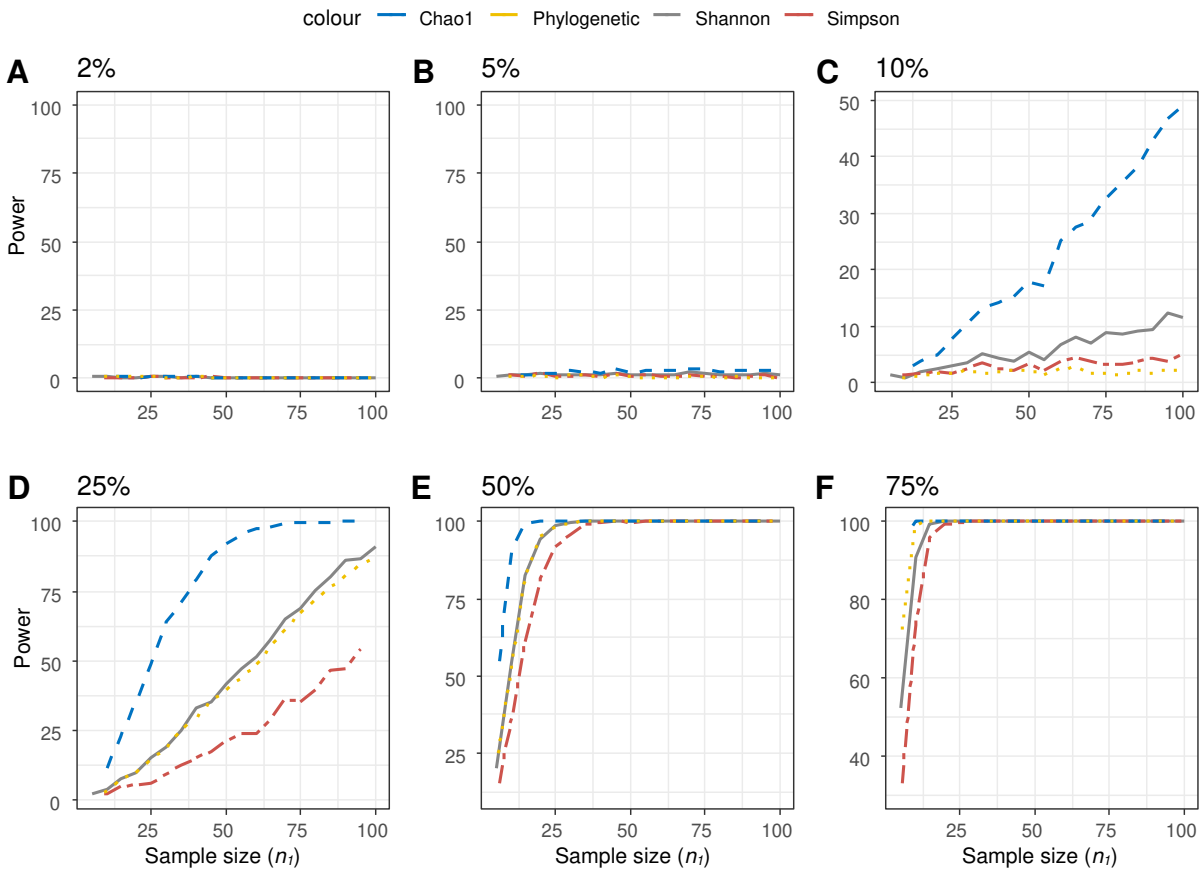


Figure 3

Empirical power for the statistical comparison of two groups $\mathbf{X}_1$ and $\mathbf{X}_2$ (Simulated data set 1). The Empirical power is calculated using Equation (11) as a function of the sample size n_1 of group 1 (with $n_2 = n_1$ and total sample size $n = n_1 + n_2$). Alpha diversity metrics and Kruskal-Wallis test are used to test for differences.

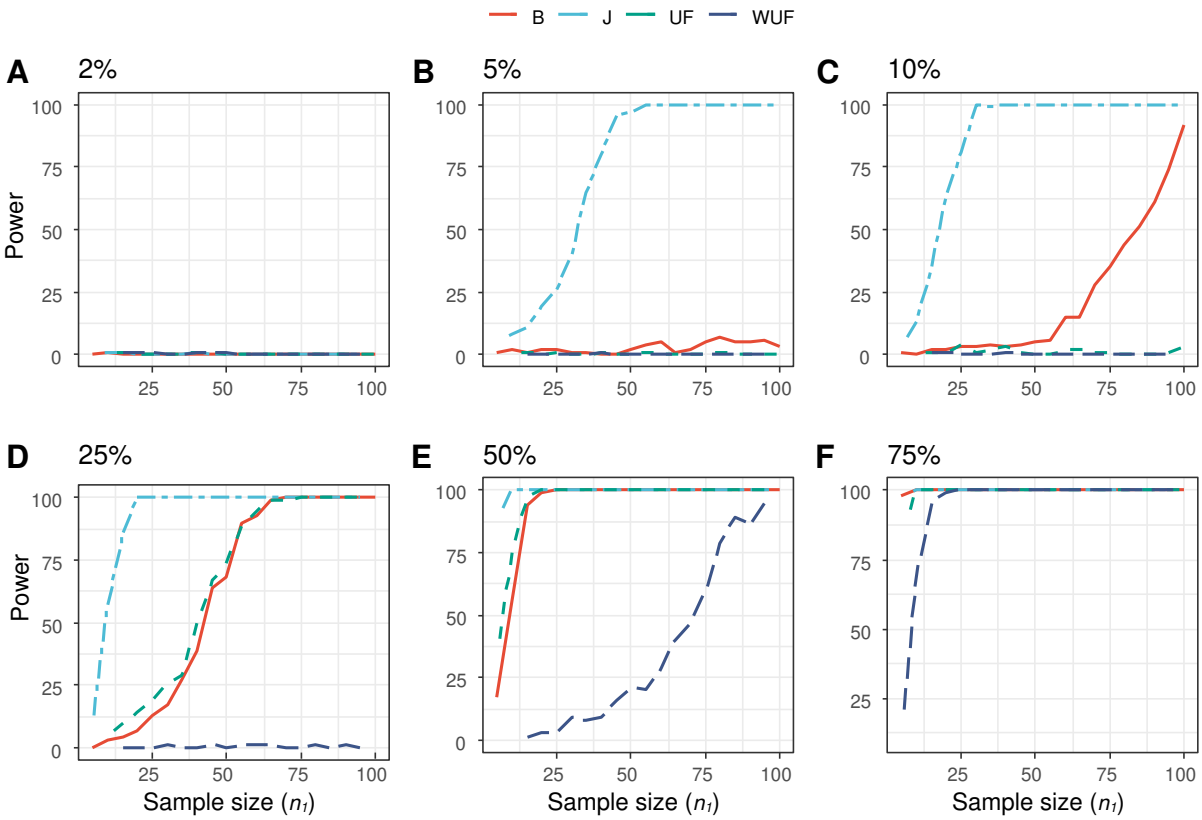


Figure 4

Empirical power for the statistical comparison of two groups $\mathbf{X}_1$ and $\mathbf{X}_2$ (Simulated data set 1). The Empirical power is calculated using Equation (11) as a function of the sample size n_1 of group 1 (with $n_2 = n_1$ and total sample size $n = n_1 + n_2$). Beta diversity metrics and PERMANOVA test are used to test for differences (B = Bray-Curtis, J = Jaccard, UF= unweighted UniFrac, WUF = Weighted UniFrac).

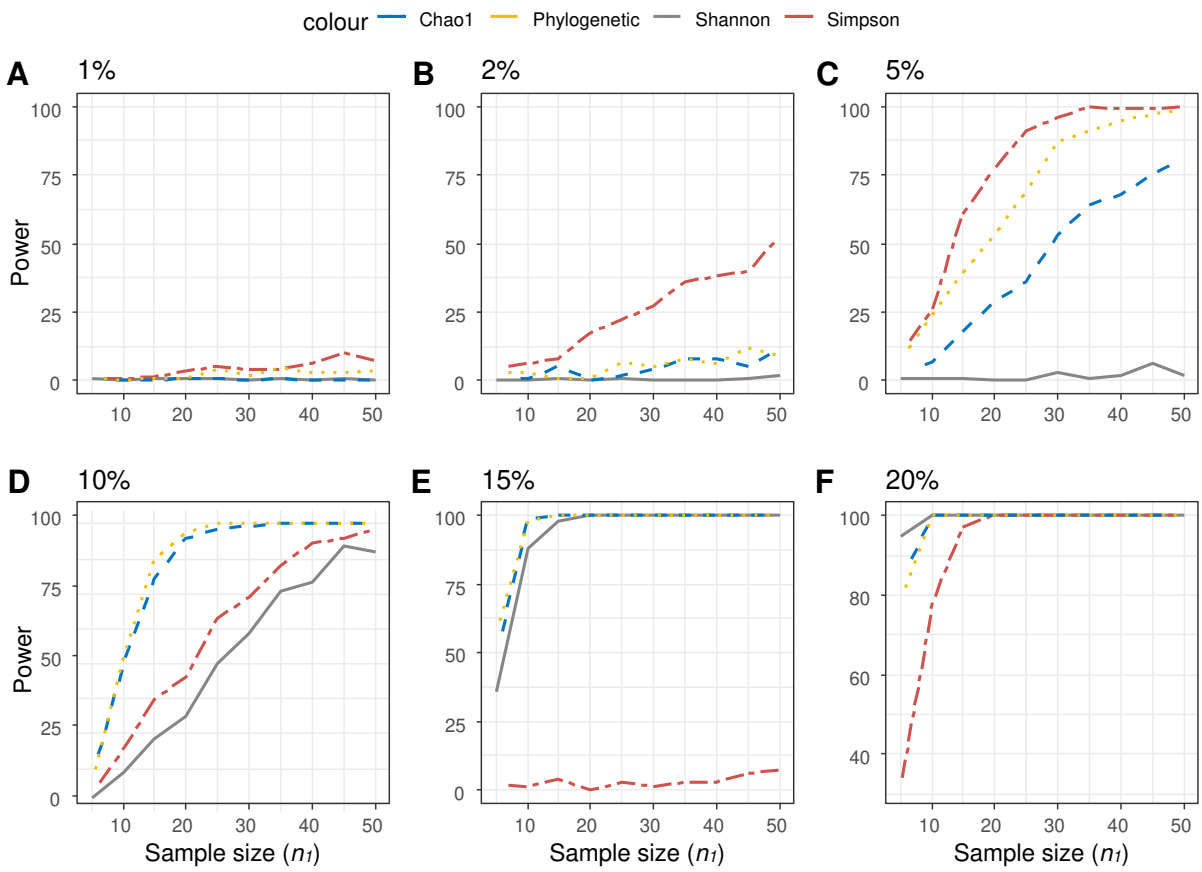


Figure 5

Empirical power for the statistical comparison of two groups $\mathbf{X}_1$ and $\mathbf{X}_2$ (Simulated data set 2). The Empirical power is calculated using Equation (11) as a function of the sample size n_1 of group 1 (with $n_2 = n_1$ and total sample size $n = n_1 + n_2$). Alpha diversity metrics and Kruskal-Wallis test are used to test for differences.

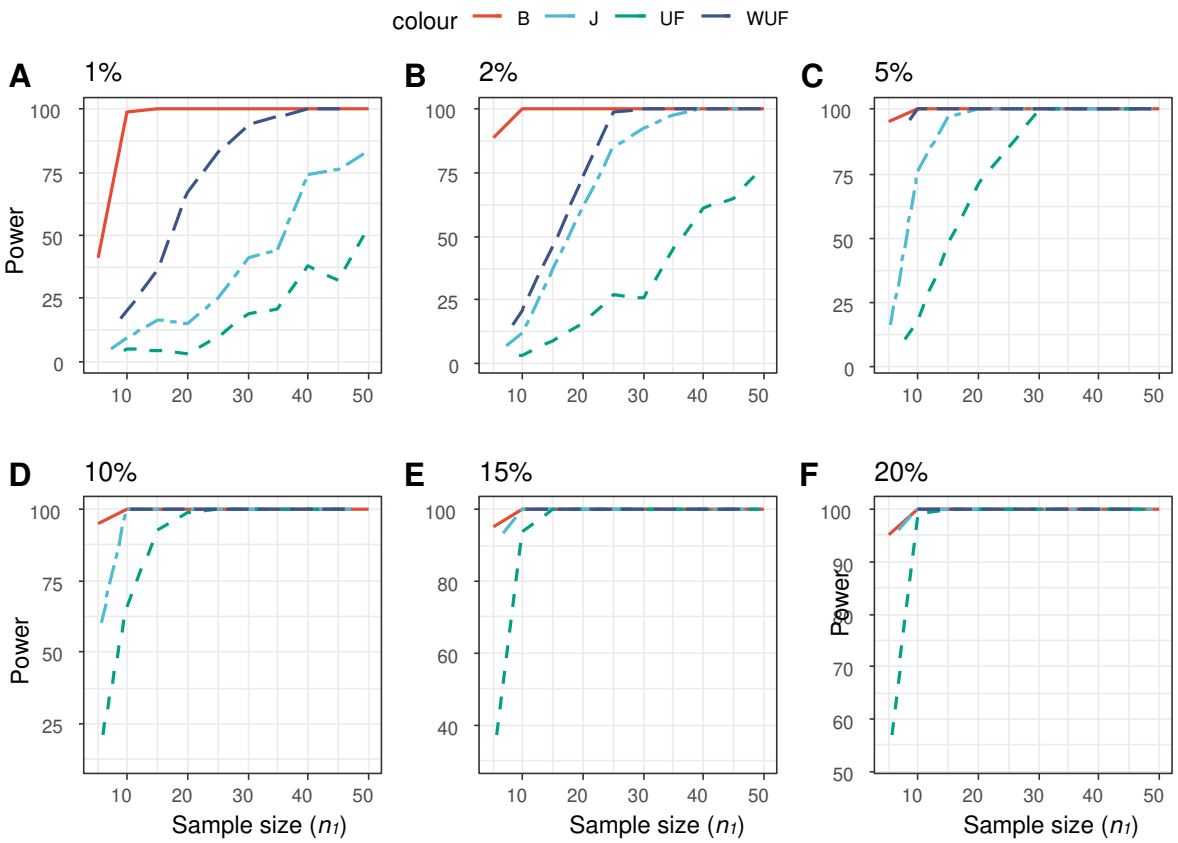


Figure 6

Empirical power for the statistical comparison of two groups $\mathbf{X}_1$ and $\mathbf{X}_2$ (Simulated data set 2). The Empirical power is calculated using Equation (11) as a function of the sample size n_1 of group 1 (with $n_2 = n_1$ and total sample size $n = n_1 + n_2$). Beta diversity metrics and PERMANOVA test are used to test for differences.

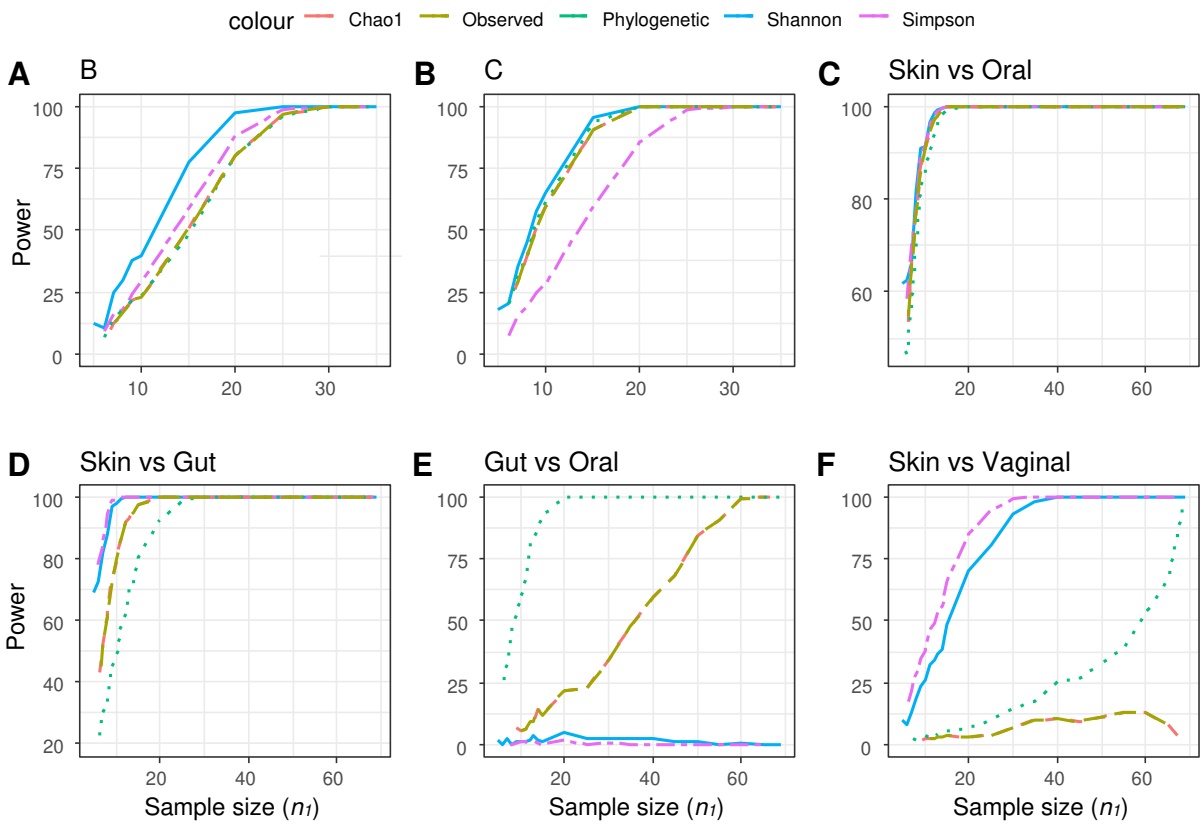


Figure 7

Empirical power for the statistical comparison of the skin, gut and oral microbiome data sets from the Human Microbiome Projects. The Empirical power is calculated using Equation (11). Differences between any two data sets X_1 and X_2 is assessed using alpha diversity metrics using a Kruskal-Wallis test. Empirical power is calculated as a function of the sample size n_1 of group 1 (with $n_2 = n_1$ and total sample size $n = n_1 + n_2$) over k=2000 tests per each sample size.

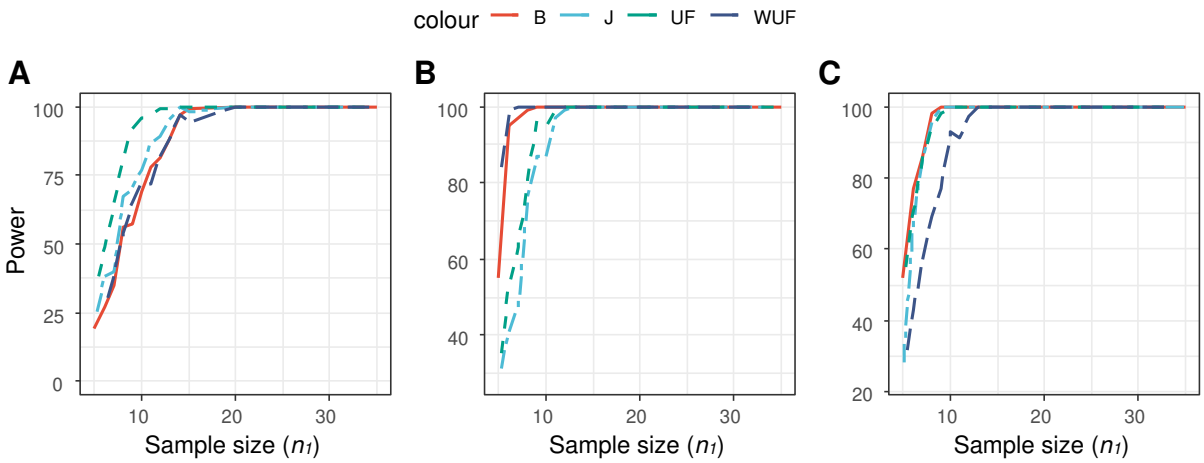


Figure 8

Different beta diversity metrics to estimate different sample sizes (n=100 repetitions). A. Chicken data set A. B. data set B, C data set C.

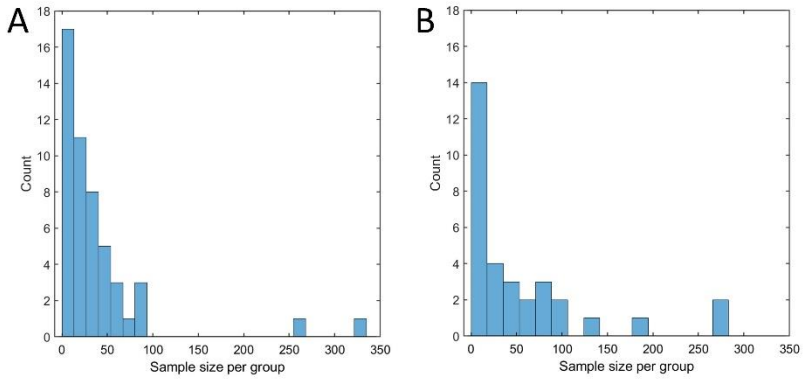


Figure 9

The distribution of the sample size of the data sets that were analysed, using the Chao1 diversity measure (among others) in 28 of the 100 papers considered in the literature review.

TABLES

Table 1

Overview of data set characteristics of the different data sets. The mean alpha and beta diversity, the standard deviation between brackets (A,B). The mean of the beta diversities is also calculated as the mean distance of groups members to the group centroid (C). *n* = sample size. Feed 1, is the intervention without same medium chain fatty acid. PD is the phylogenetic diversity.

A	n	ASV	PD	Shannon	Chao1	Simpson
Chickdata A Feed1	35	780	28.3 (2.9)	4.1 (0.3)	136.3 (20.8)	0.96 (0.02)
Chickdata A Feed2	35	794	28.2 (2.8)	4.1 (0.3)	137.9 (22.0)	0.96 (0.02)
Chickdata B Feed1	35	537	22.6 (3.7)	3.5 (0.5)	109.6 (21.3)	0.90 (0.08)
Chickdata B Feed2	35	588	26.9 (2.7)	4.0 (0.3)	139.5 (20.2)	0.95 (0.02)
Chickdata C Feed1	35	466	17.4 (3.1)	3.2 (0.5)	79.0 (17.7)	0.90 (0.06)
Chickdata C Feed2	35	518	20.2 (2.3)	3.7 (0.3)	98.1 (14.4)	0.95 (0.02)
HMP gut	168	1996	17.3 (3.7)	3.2 (0.6)	70.1 (21.7)	0.9 (0.1)
HMP oral	150	1740	22.6 (4.6)	3.4 (0.5)	82.6 (21.2)	0.9 (0.1)
HMP skin	69	899	12.2 (6.3)	1.7 (0.6)	41.4 (20.7)	0.7 (0.1)
HMP vaginal	86	678	8.6 (3.7)	1.1 (0.6)	31.8 (10.2)	0.4 (0.2)

B	n	ASV	BC	Jaccard	UF	WUF
Chickdata A Feed1	35	780	0.74 (0.09)	0.84 (0.06)	0.40 (0.05)	0.30 (0.08)
Chickdata A Feed2	35	794	0.71 (0.09)	0.83 (0.06)	0.39 (0.05)	0.30 (0.07)
Chickdata B Feed1	35	537	0.59 (0.13)	0.74 (0.10)	0.41 (0.12)	0.27 (0.10)
Chickdata B Feed2	35	588	0.63 (0.12)	0.77 (0.09)	0.35 (0.06)	0.28 (0.09)
Chickdata C Feed1	35	466	0.72 (0.15)	0.83 (0.12)	0.45 (0.12)	0.33 (0.09)
Chickdata C Feed2	35	518	0.69 (0.11)	0.81 (0.08)	0.36 (0.07)	0.29 (0.06)
HMP gut	168	1996	0.80 (0.10)	0.89 (0.07)	0.54 (0.08)	0.39 (0.13)
HMP oral	150	1740	0.70 (0.13)	0.82 (0.09)	0.49 (0.11)	0.33 (0.12)
HMP skin	69	899	0.59 (0.20)	0.72 (0.16)	0.66 (0.10)	0.29 (0.17)
HMP vaginal	86	678	0.71 (0.29)	0.79 (0.24)	0.70 (0.11)	0.21 (0.15)

C	n	ASV	BC	Jaccard	UF	WUF
Chickdata A Feed1	35	780	0.51 (0.06)	0.59 (0.04)	0.28 (0.04)	0.21 (0.06)
Chickdata A Feed2	35	794	0.50 (0.05)	0.58 (0.04)	0.27 (0.03)	0.21 (0.05)
Chickdata B Feed1	35	537	0.41 (0.09)	0.51 (0.07)	0.28 (0.09)	0.19 (0.08)
Chickdata B Feed2	35	588	0.44 (0.07)	0.54 (0.05)	0.24 (0.04)	0.20 (0.06)

Chickdata C Feed1	35	466	0.51 (0.07)	0.58 (0.05)	0.32 (0.08)	0.24 (0.05)
Chickdata C Feed2	35	518	0.48 (0.07)	0.57 (0.05)	0.25 (0.05)	0.20 (0.04)
HMP gut	168	1996	0.57 (0.07)	0.63 (0.04)	0.38 (0.06)	0.27 (0.09)
HMP oral	150	1740	0.49 (0.09)	0.58 (0.06)	0.33 (0.09)	0.23 (0.08)
HMP skin	69	899	0.40 (0.15)	0.50 (0.12)	0.46 (0.06)	0.20 (0.13)
HMP vaginal	86	678	0.51 (0.23)	0.56 (0.18)	0.49 (0.05)	0.14 (0.12)

Table 2

Overview of data set characteristics of the simulated data sets. The mean alpha and beta diversity, the standard deviation between brackets (A,B). n = sample size, PD is the phylogenetic diversity, BC is Bray-Curtis dissimilarity.

A	n	ASV	PD	Shannon	Observed/Chao1	Simpson
Simulation 1 - 1%	169	1975	17.3 (3.7)	3.1 (0.6)	69.6 (21.5)	0.90 (0.1)
Simulation 1 - 2%	169	1955	17.2 (3.7)	3.1 (0.6)	69.2 (21.4)	69.2 (0.1)
Simulation 1 - 5%	169	1895	17.0 (3.6)	3.1 (0.6)	66.2 (20.8)	66.2 (0.1)
Simulation 1 - 10%	169	1795	16.7 (3.6)	3.0 (0.6)	62.7 (19.8)	62.7 (0.1)
Simulation 1 - 25%	169	1496	15.7 (3.4)	2.9 (0.5)	55.3 (17.0)	55.3 (0.1)
Simulation 1 - 50%	169	997	12.8 (2.8)	2.5 (0.5)	35.5 (10.7)	35.5 (0.1)
Simulation 1 - 75%	169	498	8.8 (2.2)	2.1 (0.4)	16.4 (6.1)	16.4 (0.1)
Simulation 2 - 1%	169	1995	106.9 (1.2)	7.1 (0.0)	1385.5 (19.9)	1385.5 (0.0)
Simulation 2 - 2%	169	1995	107.0 (1.2)	7.1 (0.0)	1393.4 (19.2)	1393.4 (0.0)
Simulation 2 - 5%	169	1995	107.7 (1.1)	7.1 (0.0)	1411.5 (18.3)	1411.5 (0.0)
Simulation 2 - 10%	169	1995	108.3 (1.2)	7.1 (0.0)	1439.5 (19.6)	1439.5 (0.0)
Simulation 2 - 15%	169	1995	109.4 (1.1)	7.1 (0.0)	1472.2 (19.4)	1472.2 (0.0)
Simulation 2 - 20%	169	1995	110.3 (1.2)	7.1 (0.0)	1502.8 (17.0)	1502.8 (0.0)

B	n	ASV	BC	Jaccard	UF	WUF
Simulation 1 - 1%	169	1975	0.80 (0.10)	0.89 (0.07)	0.54 (0.08)	0.39 (0.15)
Simulation 1 - 2%	169	1955	0.80 (0.10)	0.89 (0.07)	0.53 (0.08)	0.39 (0.14)
Simulation 1 - 5%	169	1895	0.81 (0.10)	0.89 (0.07)	0.54 (0.08)	0.39 (0.14)
Simulation 1 - 10%	169	1795	0.80 (0.10)	0.88 (0.07)	0.53 (0.08)	0.39 (0.14)
Simulation 1 - 25%	169	1496	0.79 (0.11)	0.86 (0.07)	0.53 (0.08)	0.39 (0.14)
Simulation 1 - 50%	169	997	0.77 (0.13)	0.91 (0.09)	0.53 (0.08)	0.40 (0.15)
Simulation 1 - 75%	169	498	0.84 (0.12)	0.67 (0.08)	0.59 (0.10)	0.50 (0.16)
Simulation 2 - 1%	169	1995	0.50 (0.01)	0.66 (0.01)	0.24 (0.01)	0.06 (0.01)
Simulation 2 - 2%	169	1995	0.49 (0.01)	0.64 (0.01)	0.24 (0.01)	0.06 (0.01)
Simulation 2 - 5%	169	1995	0.47 (0.01)	0.61 (0.01)	0.23 (0.01)	0.05 (0.01)
Simulation 2 - 10%	169	1995	0.43 (0.01)	0.58 (0.01)	0.22 (0.01)	0.05 (0.00)
Simulation 2 - 15%	169	1995	0.40 (0.01)	0.55 (0.01)	0.21 (0.01)	0.05 (0.00)
Simulation 2 - 20%	169	1995	0.38 (0.01)	0.89 (0.01)	0.20 (0.01)	0.04 (0.01)

Table 3

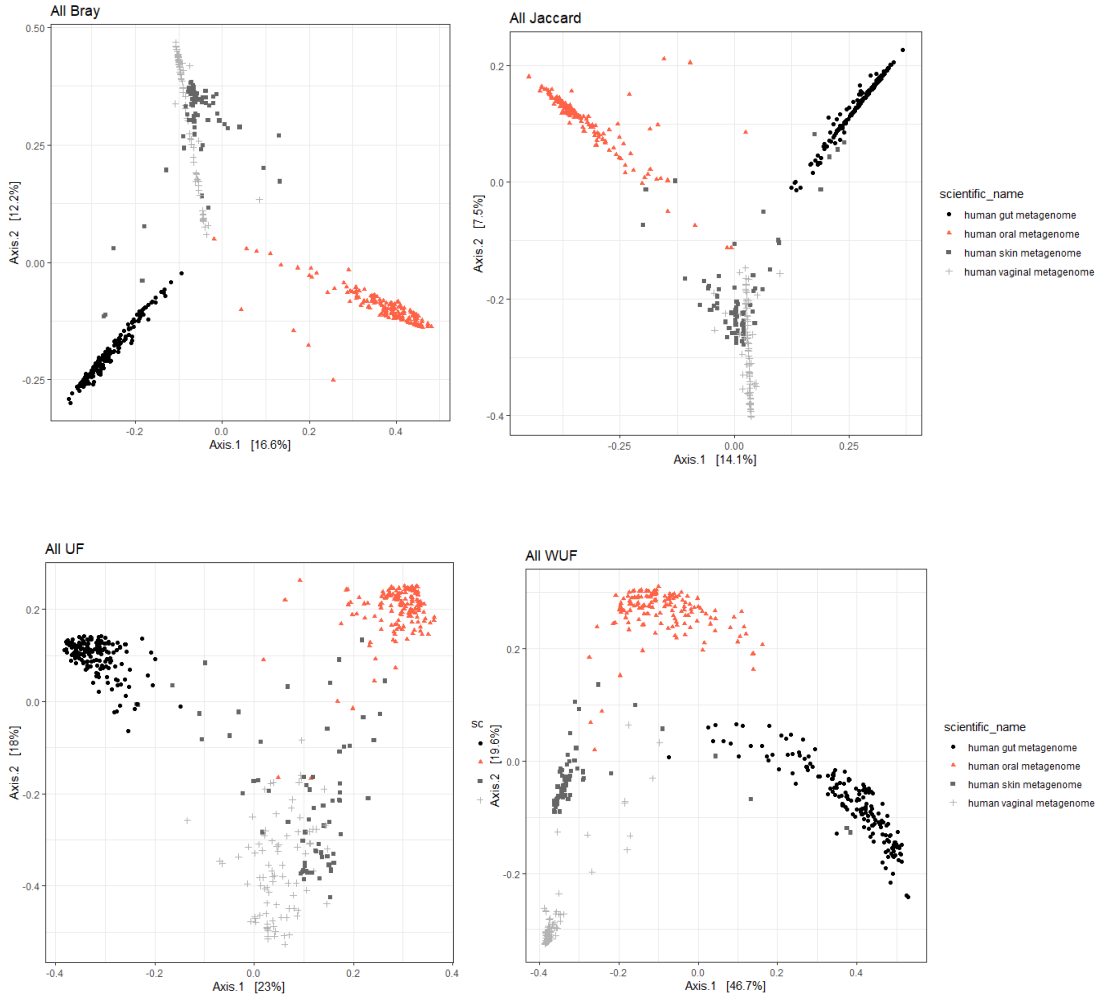
The frequency of the different alpha and beta metrics in published papers with microbiome or microbiota in the title and published between 01/2020 and 02/2020 (n=100, multiple metrics per paper were often used)

Alpha metrics	n	Beta metrics	n
Shannon index	78	Bray-Curtis	41
Chao1	39	Weighted UniFrac	35
Observed OTU/ASV	32	Unweighted UniFrac	21
(Inverse) Simpson	29	Jaccard	4
Phylogenetic	7	Euclidean	3
ACE	5	Jackknifed	2
Coverage	3	Yue and Clayton	2
Pielou	3	Sorensen	1
Sobs	2	Jensen-Shannon	1
Gini-Simpson	1		
Shannon-Wiener	1		

Supplementary figures

Figure S1

Principal coordinate plots (PCoA) based on (a) Bray-Curtis and (b) Jaccard and (c) unweighted UniFrac and (d) Weighted UniFrac distances between the different human sample types. Di erent colors indicate different sample types.



REFERENCES

1. Mulder IE, Schmidt B, Lewis M, Delday M, Stokes CR, Bailey M, et al. Restricting microbial exposure in early life negates the immune benefits associated with gut colonization in environments of high microbial diversity. PLoS One. 2011;6(12):e28279. doi: 10.1371/journal.pone.0028279.
2. Inman CF, Haverson K, Konstantinov SR, Jones PH, Harris C, Smidt H, et al. Rearing environment affects development of the immune system in neonates. Clin Exp Immunol. 2010;160(3):431-9. doi: 10.1111/j.1365-2249.2010.04090.x.
3. Williams SC. Gnotobiotics. Proceedings of the National Academy of Sciences. 2014;111(5):1661-.
4. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410-22. doi: 10.1038/s41579-018-0029-9.
5. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circulation research. 2015;116(1):116-26.
6. Casals-Pascual C, González A, Vázquez-Baeza Y, Song SJ, Jiang L, Knight R. Microbial diversity in clinical microbiome studies: sample size and statistical power considerations. Gastroenterology. 2020;158(6):1524-8. doi:10.1053/j.gastro.2019.11.305
7. Willis AD. Rarefaction, alpha diversity, and statistics. Frontiers in microbiology. 2019;10:2407. doi:10.3389/fmicb.2019.02407
8. Faith DP. The Role of the Phylogenetic Diversity Measure, PD, in Bio-informatics: Getting the Definition Right. Evolutionary Bioinformatics Online. 2006;2:277-83. doi: 10.1177/117693430600200008.
9. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME journal. 2017;11(12):2639-43. doi:10.1038/ismej.2017.119
10. Chao A. Nonparametric estimation of the number of classes in a population. Scandinavian Journal of statistics. 1984:265-70.
11. Simpson EH. Measurement of diversity. nature. 1949;163(4148):688.
12. Lemos LN, Fulthorpe RR, Triplett EW, Roesch LF. Rethinking microbial diversity analysis in the high throughput sequencing era. Journal of microbiological methods. 2011;86(1):42-51. doi: 10.1016/j.mimet.2011.03.01
13. Magurran AE. Measuring biological diversity: John Wiley & Sons; 2013. ISBN: 978-1-118-68792-5
14. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecological Monographs. 1957;27(4):325-49. doi: 10.2307/1942268.
15. Jaccard P. The distribution of the flora in the alpine zone. New phytologist. 1912;11(1):37-50.
16. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology. 2005;71(12):8228-35. doi: 10.1128/AEM.71.12.8228-8235.2005
17. Lozupone CA, Hamady M, Kelley ST, Knight R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol. 2007;73(5):1576-85. Epub 2007/01/16. doi: 10.1128/aem.01996-06.
18. Clarke KR. Non-parametric multivariate analyses of changes in community structure. Australian journal of ecology. 1993;18(1):117-43. doi:
19. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001;26(1):32-46. doi: 10.1111/j.1442-9993.2001.01070.pp.x.
20. Hanson BM, Weinstock GM. The importance of the microbiome in epidemiologic research. Ann Epidemiol. 2016;26(5):301-5. doi: 10.1016/j.annepidem.2016.03.008.

21. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5(1):27. doi:
22. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Frontiers in microbiology. 2017;8:2224. doi:
23. Li CI, Samuels DC, Zhao YY, Shyr Y, Guo Y. Power and sample size calculations for high-throughput sequencing-based experiments. Brief Bioinform. 2017. doi: 10.1093/bib/bbx061.
24. Cohen J. Statistical power analysis for the behavioral sciences: Academic press; 2013.
25. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: statistical approaches to estimating microbial diversity. Applied and environmental microbiology. 2001;67(10):4399-406. doi: 10.1128/AEM.67.10.4399-4406.2001.
26. Kim B-R, Shin J, Guevarra RB, Lee JH, Kim DW, Seol K-H, et al. Deciphering diversity indices for a better understanding of microbial communities. Journal of Microbiology and Biotechnology. 2017;27(12):2089-93. doi: 10.4014/jmb.1709.09027.
27. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nature methods. 2016;13(7):581. doi: 10.1038/nmeth.3869.
28. Allen HK, Bayles DO, Looft T, Trachsel J, Bass BE, Alt DP, et al. Pipeline for amplifying and analyzing amplicons of the V1–V3 region of the 16S rRNA gene. BMC research notes. 2016;9(1):1-6. doi: 10.1186/s13104-016-2172-6
29. Borcard D, Gillet F, Legendre P. Numerical ecology with R: Springer; 2018. doi: 10.1186/s13104-016-2172-6.
30. Kers JG, Velkers FC, Fischer EAJ, Hermes GDA, Lamot DM, Stegeman JA, et al. Take care of the environment: housing conditions affect the interplay of nutritional interventions and intestinal microbiota in broiler chickens. Animal Microbiome. 2019;1(1):10. doi: 10.1186/s42523-019-0009-z.
31. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. nature. 2012;486(7402):207. doi: 10.1038/nature11234.
32. Ramiro-Garcia J, Hermes GDA, Giatsis C, Sipkema D, Zoetendal EG, Schaap PJ, et al. NG-Tax, a highly accurate and validated pipeline for analysis of 16S rRNA amplicons from complex biomes. F1000Research. 2016;5:1791. doi: 10.12688/f1000research.9227.1.
33. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue): D590-6. Epub 2012/11/30. doi: 10.1093/nar/gks1219.
34. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association. 1952;47(260):583-621.
35. Anderson MJ, Walsh DC. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? Ecological monographs. 2013;83(4):557-74. doi: 10.1890/12-2010.1
36. Team R. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0 2008.
37. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8(4):e61217. Epub 2013/05/01. doi: 10.1371/journal.pone.0061217.
38. Lahti L SSea. Tools for microbiome analysis in R. Version 1.5.28. 2017. doi: http://microbiome.github.com/microbiome.

39. Oksanen J, Blanchet F G , Kindt R , Legendre P , O'Hara R B , Simpson G L , Solymos P , Stevens M H H , Wagner H. Vegan: community ecology package. R package version 1.17-4. http://cran r-project org
40. Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods. 2007;39(2):175-91. doi: 10.3758/bf03193146.
41. Williams J, Bravo HC, Tom J, Paulson JN. microbiomeDASim: Simulating longitudinal differential abundance for microbiome data. F1000Research. 2019;8. doi: 10.12688/f1000research.20660.2.
42. La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PloS one. 2012;7(12):e52078. doi:
43. Kelly BJ, Gross R, Bittinger K, Sherrill-Mix S, Lewis JD, Collman RG, et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics. 2015;31(15):2461-8. doi: 10.1093/bioinformatics/btv183.
44. Xia Y, Sun J, Chen D-G. Power and sample size calculations for microbiome data. Statistical analysis of microbiome data with R: Springer; 2018. p. 129-66.
45. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, et al. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS computational biology. 2013;9(1):e1002863. dio: 10.1371/journal.pcbi.1002863
46. Ioannidis JP. Why most published research findings are false. PLos med. 2005;2(8):e124. doi: 10.1371/journal.pmed.0020124
47. Collaboration OS. Estimating the reproducibility of psychological science. Science. 2015;349(6251). doi: 10.1126/science.aac4716
48. Happ M, Bathke AC, Brunner E. Optimal sample size planning for the Wilcoxon-Mann-Whitney test. Statistics in medicine. 2019;38(3):363-75. doi: 10.1002/sim.7983.
49. Hoffman J. Baic Biostatistics for Medical and Biomedical Practitioners. London, UK: Academic Press.
50. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in psychology. doi: 2013;4:863. /10.3389/fpsyg.2013.00863
51. Fan C, Zhang D, Zhang CH. On sample size of the Kruskal–Wallis test with application to a mouse peritoneal cavity study. Biometrics. 2011;67(1):213-24. doi: 10.1111/j.1541-0420.2010.01407.x.
52. Fan C, Zhang D. A note on power and sample size calculations for the Kruskal–Wallis test for ordered categorical data. Journal of biopharmaceutical statistics. 2012;22(6):1162-73. dio: 10.1080/10543406.2011.578313
53. Kolassa JE, Jankowski S. MultNonParam-package R Documentation. 2020.
54. Association AP. Publication manual of the American psychological association: DAR ALMHRER ELADABE; 1994.
55. Saccenti E, Timmerman ME. Approaches to sample size determination for multivariate data: Applications to PCA and PLS-DA of omics data. Journal of Proteome Research. 2016;15(8):2379-93. doi: 0.1021/acs.jproteome.5b01029. Epub 2016 Jul 7
56. Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology. Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011;22(11):1359-66. doi: 10.1177/0956797611417632.
57. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Taylor & Francis; 2016. doi: 10.1080/00031305.2016.1154108

58. Smith GD, Ebrahim S. Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers. British Medical Journal Publishing Group; 2002. doi: 10.1136/bmj.325.7378.1437
59. Raj AT, Patil S, Sarode S, Salameh Z. P-Hacking: a wake-up call for the scientific community. Science and engineering ethics. 2018;24(6):1813-4. doi: 10.1007/s11948-017-9984-1
60. Ioannidis JP. Why most published research findings are false. PLoS medicine. 2005;2(8):e124. doi: 10.1371/journal.pmed.0020124
61. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics. 2014;15(1):1-12. doi: 10.1093/biostatistics/kxt007
62. Gamble C, Krishan A, Stocken D, Lewis S, Juszczak E, Doré C, et al. Guidelines for the content of statistical analysis plans in clinical trials. Jama. 2017;318(23):2337-43. doi: 10.1001/jama.2017.18556. doi: 10.1001/jama.2017.18556