Computational and Structural Biotechnology Journal 20 (2022) 2259–2267

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Method article

Beta-diversity distance matrices for microbiome sample size and power calculations — How to obtain good estimates

By IMPACTT investigators [1]

ARTICLE INFO

Article history:
Received 29 October 2021
Received in revised form 13 April 2022
Accepted 22 April 2022
Available online 27 April 2022

Keywords:
Beta-diversity
Sample size calculations
Power calculations
Simulation
Microbiome
Jaccard distance
Community structure

ABSTRACT

In microbiome studies, researchers often wish to compare the taxa count distributions between groups of samples. Commonly used corresponding methods of analysis are built on examining distance matrices, where distances describe the beta-diversity between samples. Analyses then compare the distribution of distances within groups to the distributions between groups. However, when performing *a priori* sample size or power calculations for such study designs, appropriate within and between group distance distributions can be challenging to obtain. When available, pilot study data, or data from prior studies of similar design should provide realistic distance estimates. However, when these are not available, distances can be extracted from available studies where one can assume similar beta-diversity. Alternatively, distances can be generated by simulation methods. Here, we describe and illustrate these three strategies for obtaining realistic distance matrices. For simulation methods, we illustrate the procedures required starting from existing benchmark data, as well as how to simulate directly from population assumptions. Using data from the American Gut project, we provide tables of observed distances for use by researchers planning their own studies, as well as R codes for generating similar matrices in other datasets. Furthermore, for simulated data, we compare methods, provide R codes, and demonstrate how challenging it is to obtain realistic distance distributions without any benchmark data. This code and illustrative distance tables are provided by the IMPACTT Consortium as a resource to the microbiome research community.

1. Introduction

More than a decade after the Human Microbiome Project (HMP) [1], large-scale microbiome studies have been initiated worldwide to characterize the large diversity of microbial populations living in and on humans. This is being achieved by next generation sequencing technologies to genetically identify all microbiota present in a biosample. This was a newly developed method at the time of the HMP and one that continues to offer advantages over culture-based methods. In fact, the ability to identify all community members in human microbiota at a site of interest was the 'promised land' of large-scale sequencing technologies [2]. The microbiome research field has also embraced principal component or coordinate clustering methods to show visually and intuitively in a figure, how whole microbial communities cluster by a health outcome or an environmental determinant. As many times microbiota clusters overlap visually, these figures are often accompanied by analysis methods such as PERMANOVA [3] to test for statistically significant differences in cluster composition. However, it is rare to find study plans that carefully consider whether they have good power for detection of meaningful differences, particularly for complex community composition questions. However, this may be coming shortly: reporting guidelines have been recently developed for microbiome studies (STORMS: https://www.stormsmicrobiome.org/ [4], STROBE-metagenomics [5]) have been recently developed for microbiome studies and recommend examining study power carefully.

In an overview paper by Casals-Pascual et al [6] on sample size requirements for microbiome studies, an example was provided on how to test for desired differences in microbial beta-diversity, a metric that characterizes the whole microbiome community [7]. It assumed a normal distribution of microbial beta-diversity but often this is not the case. Appreciating this challenge, Kelly *et al.* developed a series of equations to determine sample size for beta-diversity measures, [8], built on the permutation-based

[1] The members of the IMPACTT investigators Collaboration are listed in Appendix A at the end of the article.

https://doi.org/10.1016/j.csbj.2022.04.032

analysis method PERMANOVA [3]. However, use of their equations requires researchers to find literature reference values for variances and desired differences in beta-diversity. Although several alternative analysis methods for detecting global microbiome diversity have been proposed [9–11], to perform sample size and power calculations using these methods, realistic estimates of beta diversity are still required in advance. Recently, Gail et al. (2021) [12] compared the power of several of these methods; they also mention that beta-diversity measures are required when estimating sample size and power. This presents microbiome researchers with another challenge. 'Actual values' for the beta-diversity metric by groupings of interest are not often reported in microbiome papers. Although obtaining data from a small pilot study would be ideal [12], is not always feasible. We -- several researchers from IMPACTT (Integrated Microbiome Platforms for Advancing Causation Testing and Translation; https://www.impactt-microbiome.ca/) – describe in detail, including code, how to obtain values for microbial beta-diversity variances and helpful group differences from existing literature, to be used in sample size calculations. When such data are missing or not reported, we also show how to use simulation to create ranges for these beta-diversity values.

2. Distance metrics

Sample size calculations for beta-diversity require measures of the "distance" between samples, where distance captures the dissimilarity between two microbiome profiles. Many different definitions can be used to describe the similarity or differences between vectors of microbiome counts. Perhaps the most common distance metric is the Unifrac distance (weighted or unweighted), which measures the phylogenetic dissimilarity of samples by accounting for the evolutionary tree estimated from sequence similarity [13]. The unweighted UniFrac distance uses only estimated species presence/absence information, e.g. the molecularly-defined operational taxonomic unit (OTU), and counts the fraction of branch length unique to each sample, while the weighted UniFrac distance weights the branch length with the OTU abundance difference. The unweighted Unifrac distance has been shown to be more sensitive to outliers, i.e. OTUs with significantly different abundance from other OTUs, or to OTUs with abundance near detection limits [14].

Non-phylogenetic distance metrics that are commonly used for beta-diversity include the Jaccard (weighted and unweighted) and Bray-Curtis distances. The unweighted Jaccard distance for microbiome is defined as one minus the ratio of the intersection to the union of OTU presence/absence sets of two samples [15]. The weighted Jaccard distance considers abundance information, and a distance between two samples j and k is defined as:

$$d_{jk} = 1 - \frac{\sum_i min(x_{ik}, x_{ij})}{\sum_i max(x_{ik}, x_{ij})},$$

where i indexes different OTUs.

A third common choice, chosen as the default metric in the popular vegan R package [16] is the Bray-Curtis distance [17] defined as:

$$d_{jk} = \frac{\sum_i |x_{ij} - x_{ik}|}{\sum_i |x_{ij} + x_{ik}|}.$$

The choice of distance metric in a microbiome study depends on the study goals since difference metrics reflect different perspectives on beta-diversity. The Unifrac distance is a logical choice when the evolutionary tree is relevant. Use of weighted metrics that incorporate abundance information will be more important when considering the microbiome as a system, i.e. a set of of OTU community members where all the frequencies are correlated, whereas unweighted metrics may be more appropriate when looking for new mutations or the presence of a few new rare (estimated) species. Furthermore, beta-diversity measures might be chosen to make it easier to assess a specific hypothesis, for example changes in microbiome composition over time. For example, a Bray Curtis distance that uses all the abundance differences may be more sensitive to examine gradients of diversity with location or time [7].

Researchers should consider experimental design (e.g. effect size expectations, abundance levels, the range of anticipated OTUs), data quality (e.g. phylogeny validity, presence of outliers) and priorities (e.g. type I error) to choose an appropriate metric for a specific power analysis [18]. Code for calculation of a Unifrac distance matrix illustrated with data from the American Gut project [19] can be found in Supplement A, Box 1.

3. Estimation of distance matrices for sample size calculations

Having chosen a distance metric, beta-diversity sample size calculations require knowledge of the distributions of this distance measure, both within and between groups [8]. For example, distances between pairs of individuals receiving the same treatment could be compared to distances between pairs of individuals receiving different treatments. Ideally, sample size calculations are then built on a matrix of distances between all pairs of samples, but they can also be built on summary statistics describing the distributions of these distances within and between groups.

The best way to obtain realistic distance matrix estimates for a chosen beta-diversity distance metric in planning a study is from a pilot study or from a publication of a similar study. Assuming the existence of a pilot or a published study involving samples and groups similar to the planned study, between-group and within-group distance distributions can be calculated from those studies and used for inferring effect sizes and conducting power calculations. This approach was taken by Gail et al. (2021) [12]; a small set of pilot data were generated and then used to illustrate power calculations.

However, pilot or similar published data may not always be available. In such cases, there are two main strategies that can be used. Firstly, data from a previous study can be used as a benchmark dataset in simulations that expand or shrink the observed distances to desired values. Alternatively, if no benchmark data can be found, Kelly *et al.* (2015) [8] proposed a simulation strategy that employs rarefaction of randomly generated of OTU counts to achieve desired within-group and between-group distance. In either case, analysis of the simulated data for a range of sample sizes will allow estimation of statistical power.

3.1. Extracting distance metrics from a publication

When microbiota beta-diversity distance measures are reported in a publication, they can be used to inform sample size calculation assumptions. Although pairwise distances are not often reported for each pair of samples compared, we illustrate here, using a study of the breast milk microbiome, how summary statistics can provide sufficient information. Tannock *et al.* (2013) [20] displayed the distributions of between and within group beta-diversity distances for 90 infants, sampled at the age of 2 months, who were fed breast, cow or goat milk. We extracted the means and standard errors of the distances from Fig. 4b of their paper, then converted the standard errors to standard deviations, and we provide the results in Table 1. Both within group and between group distances are shown, and it can be seen that between group distances tend to be slightly larger. These values could then be used in power calculations that are based on t-tests comparing group means. For more

Table 1
Within group and between group beta-diversity means and standard deviations for stool microbiota, for infants fed breast milk, goat milk or cow milk. Data extracted from Tannock, Lawley *et al.* 2013) "CtoB": Cow to Goat. "GtoB": Goat to human Breast. "GtoC": Goat to Cow.

	Breast	Goat	Cow	CtoB	GtoB	GtoC
Mean	0.707	0.734	0.72	0.761	0.755	0.737
SD	0.014	0.01	0.011	0.007	0.007	0.007

than two groups, it may be possible to transform means and standard deviations like those in Table 1 to an ANOVA R^2 estimate, which can then motivate sample size calculations based on an ANOVA-based effect size parameter. We have recently illustrated such calculations in a review issue on microbiome studies [21], and this is well described in the classic text by Cohen [22].

3.2. Simulating distance matrices based on benchmark data

Benchmark data, usually from a single group of samples or individuals, can be used as a starting point to create distance matrices for power calculations. Ideally, benchmark data would be sampled from the same body site as planned for the upcoming study, and the microbiome should be assayed with the same technique. Then, for power calculations, individuals in the benchmark data can be randomly allocated to groups, and pairwise distances between individuals assigned to different groups are multiplied by a scaling factor to vary the between group distances relative to within-group distances.

These simulations and calculations depend on the statistical quantity known as "effect size". Often, for beta-diversity comparisons, the effect size is defined as a function of between and within group sums of squares from an analysis of variance [8]. Let N be the total number of subjects, G be the number of groups and assume an equal number of individuals (n) will be recruited for each group, so that $N = nG$. The effect size, α, is then defined as:

$$\alpha = \frac{SST - SSW(N-1)/(N-G)}{SST + SSW/(N-G)}$$

$$SSW = \frac{1}{n}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} d_{jk}^2 \varepsilon_{jk}$$

$$SST = \frac{1}{N}\sum_{j=1}^{N-1}\sum_{k=j+1}^{N} d_{jk}^2$$

where SSW is the within group sum of squared distances, and SST is the total sum of squared distances between individuals j and k. The quantity $\varepsilon_{jk} = 1$ is an indicator variable, which is 1.0 only when subject j and k are in the same group, and is 0.0 otherwise.

Kelly *et al.* (2015) [8] proposed that simulations starting from benchmark data could be undertaken as follows:

- Bootstrap the benchmark data set and randomly allocate the bootstrapped set of individuals into G groups. Within group distances (SSW) can then be estimated from the distances within these groups of individuals.
- To create larger distances between groups than within groups, a scaling factor, $\sigma > 1.0$, is introduced and set to a value chosen by the researcher. Every bootstrapped distance between individuals in different groups is multiplied by σ, and then the total sum of squares, $SST(\sigma)$, between all pairs of individuals, can be calculated. This procedure is then repeated for a range of chosen values for σ.
- Average effect sizes comparing between group sums of squares to within group sums of squares, for each chosen value of σ, can then be calculated from the bootstrapped datasets:

$$\alpha(\sigma) \approx \frac{SST - SSW\left(\frac{G-1}{(N-G)\sigma} + 1\right)}{SST + SSW\left(\frac{N-G+1}{\sigma} - 1\right)} \quad (1)$$

The effect sizes will increase with larger values of σ. By choosing a fine grid of values for σ, it is possible to find a value corresponding to the average desired effect size α. Also, although we illustrate calculations for two groups in the code in the Supplementary material, the general concept can be extended to several groups by defining more than one σ for different pairs of groups.

- To perform power calculations, the steps above are then repeated many times, generating datasets for analysis, and testing for the ability to detect differences in beta diversity. Details for this last step are provided in the section below on calculating power using beta-diversity measures. The code for simulating distance matrices from benchmark data can be found in Supplement A, Box 2.

3.2.1. Observed beta-diversity distances from the American Gut project

To illustrate beta-diversity-based distances from a public dataset that is frequently analyzed, we took data from the American Gut project (AG) [19] and calculated distances between microbiome samples using weighted and unweighted Unifrac, Jaccard and Bray-Curtis metrics at the phylum level. These distances were calculated for several sampling sites (e.g. feces, tongue, skin, etc.) for pairs of samples either from the same sampling site [within group distances] or from two different sampling sites [between group distances]. The resulting distributions of pair distances are summarized by boxplots in Fig. 1 for Jaccard distances and in Supplement B, Figs. S1-S3, for the three other distance measures (Bray-Curtis, unweighted and weighted Unifrac). Several quantiles as well as the means and standard deviations (SD) for all 4 distance measures are also shown in Supplement Table S1. It can be seen that the distances can have long asymmetric tails.

Inter-sampling site beta-diversity distances, such as seen in the American Gut data, are not likely to be relevant for most sample size calculations, since multiple sites were sampled from the same individual leading to potential dependence between site information. Furthermore, microbiome community structures are known to be extremely distinct between sampling sites. Beta-diversity changes due to an intervention at a single site are likely to be smaller than differences between sites. On the other hand, differences in beta diversity between independent individuals may be larger than between sites sampled from the same individual. Therefore, we also calculated the distances within and between subgroups of the AG data defined by age ($\geq$45 versus <45), sex and asthma diagnosis. Fig. 2 shows these statistics for the Jaccard distance, and Supplement B, Figs. S4-S6, and Tables S2-S4 show similar results for the other distance measures.

As seen for age, sex, and asthma status, the distribution of distances in these data is extremely similar within and between groups. For example, the median of the Jaccard distances is 0.66 for males, females, and between the sexes. In contrast, larger differences are seen between body sampling sites, as expected. For example, the median Jaccard distance within nostril samples is 0.74 but only 0.65 for feces and 0.46 for hair. The median Jaccard distances between sampling sites was 0.66. We also observed that

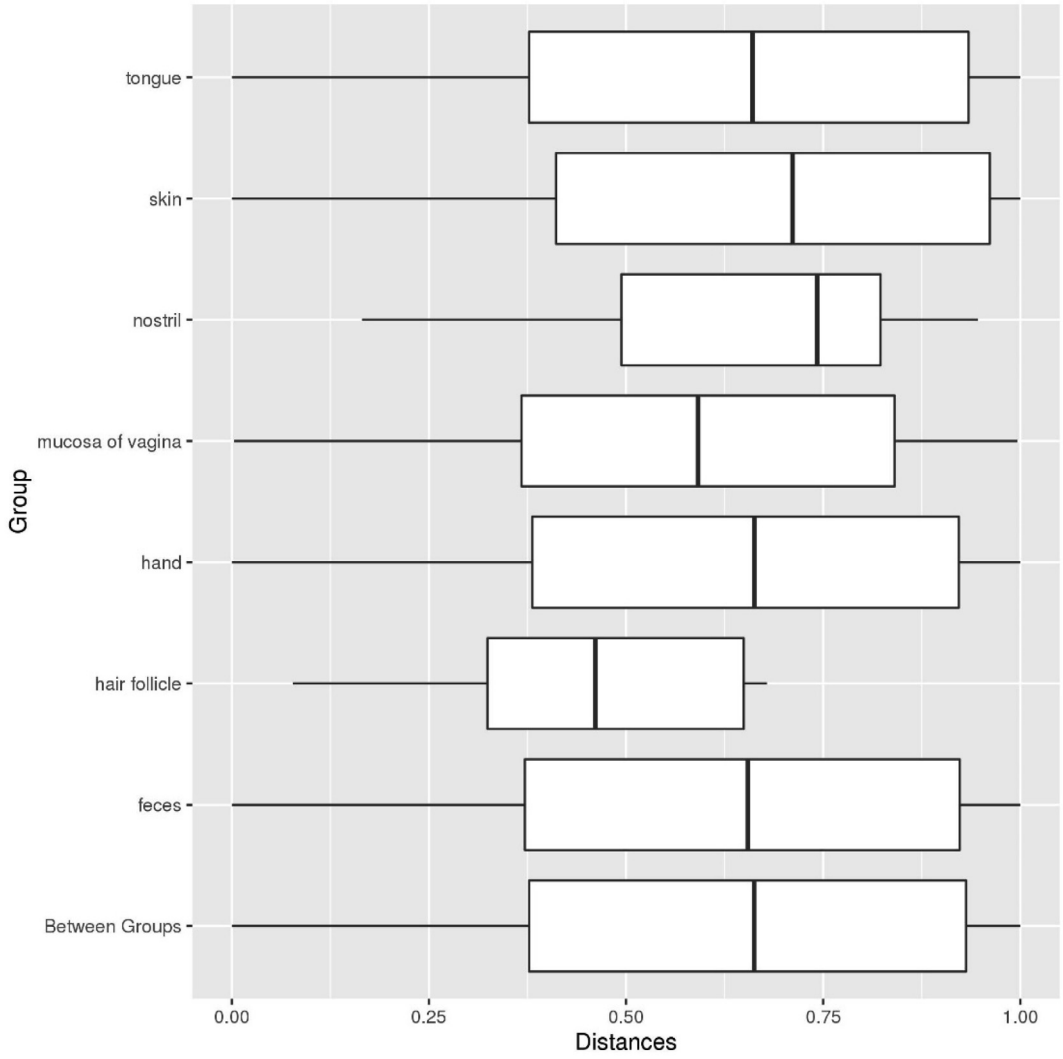

Fig. 1. Boxplots of within- and between-group Jaccard distances from different body sites from the AG data. Samples missing body site information are excluded.

the Unifrac distances (both weighted and unweighted) have smaller variance across the body sites than Bray-Curtis and Jaccard distances (Supplement B Table S1). Some of the body sites included only a small number of samples. Unifrac distance takes the phylogenetic relatedness of the OTUs at these sites into account, which may partially compensate for the small sample size. However, Bray-Curtis and Jaccard distances do not incorporate such assumptions, and these metric-specific attributes may be more affected by sample size.

3.3. Simulating distance matrices without benchmark data

We followed Kelly *et al.* (2015) [8] to illustrate one way to simulate distance matrices when benchmark data are not available. For fixed choices of N, G, and a specific effect size (α) the following procedure can be followed:

- Start by simulating a uniform vector of species or OTU counts for each of a sample size of N individuals. Then perform rarefication (subsampling) by randomly choosing a proportion (the rarefication proportion: P_r) of the counts to keep [23]; these counts are then randomly subsampled for each individual. This step creates pairwise distances that will be considered to be within-group distances, dependent on the proportion P_r. Distances between samples will be larger if fewer sequences are retained, i.e. a smaller value of P_r. This enables the generation of samples with a desired within-group mean distance.
- Then the sets of OTU counts for each of the N individuals are allocated to G groups to match the desired, pre-specified effect sizes. Specifically, a proportion of the OTUs in one of the G groups are assumed to belong to new species, unique to that group; we call this the segregation proportion. By varying the segregation proportion of OTUs that are renamed in this way, the within-group distances across all groups can be preserved, but between-group distances are boosted [8]. The choices of within-group distances and effect sizes then determine the between group distances. These rarefication and segregation steps are implemented in the R package ***micropower***.
- Distance matrices can then be calculated from the rarified data; the code for generating distance matrices with this approach can be found in Supplement A, Box 3.

Importantly, to calculate a Unifrac distance, the simulation method must provide data that corresponds to a valid phylogenetic tree. Frequently, we found that Kelly's method for generating data with rarefication did not provide results where Unifrac distances could be calculated. In contrast, Jaccard distances or Bray-Curtis distances can always be calculated.

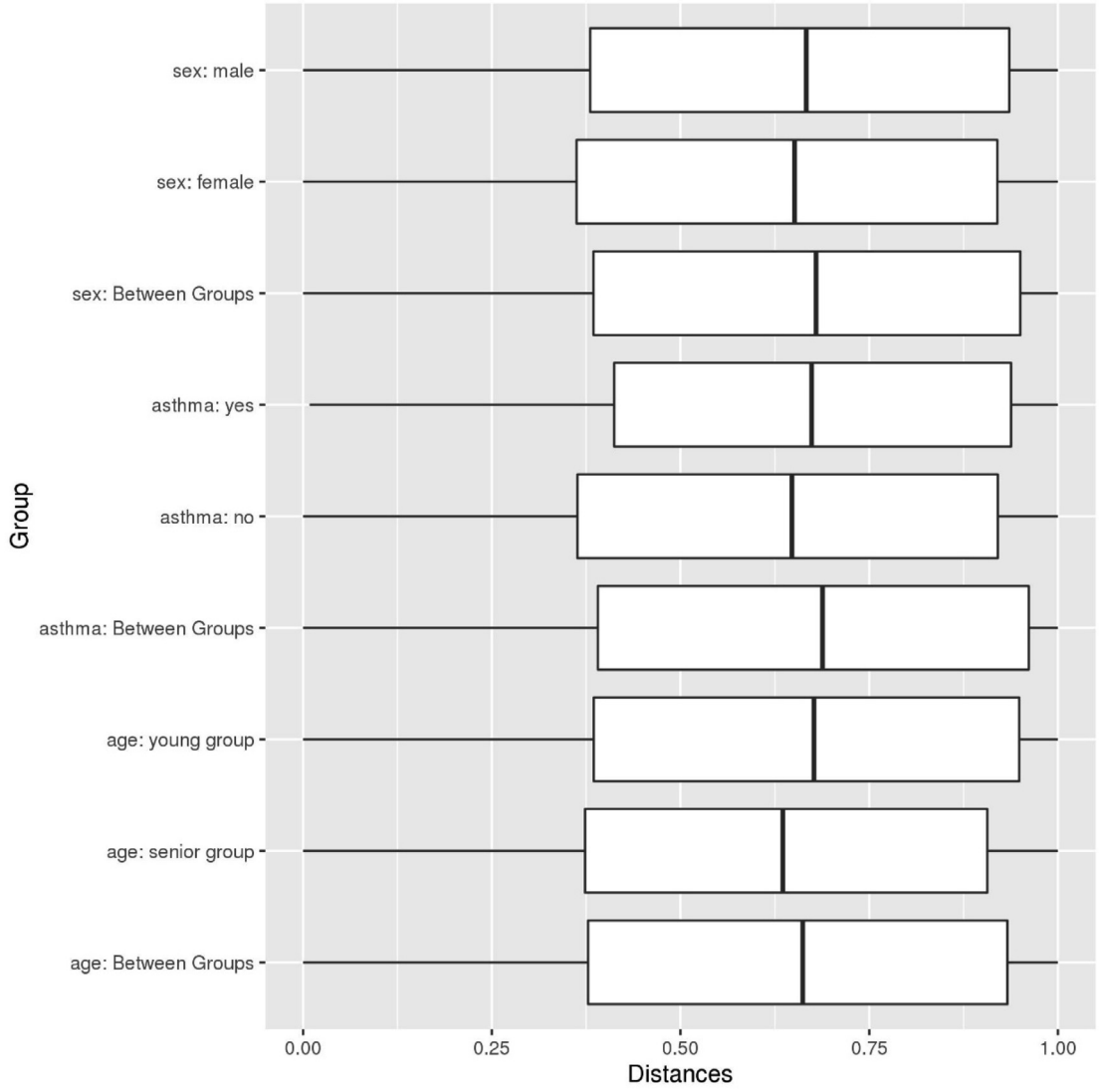

Fig. 2. Within group and between group distance measures for Jaccard distances and subgroupings of the AG data by age, asthma diagnosis and sex. Samples with missing age are excluded from boxplots for age; Samples missing asthma are excluded from boxplots for asthma.

It can be challenging to choose the parameters of the simulation procedure to obtain realistic distances. In Fig. 3, we show a heatmap of how well the simulated data would match the AG distances for groups defined by sex (male vs female) and for different parameter choices. From the AG data, we calculated the weighted Unifrac distances for male and female groups and between the sexes. For the simulations, we assumed two equally sized groups, N = 100, 50 OTUs, and a sequence depth of 50. Then we varied the other simulation parameters for Kelly's method, and for each set of parameters we calculated the deviance between the observed distance distribution and the simulated one. Specifically, the deviance between two distance distributions is defined as $\sum_q \frac{|d_q{}^{obs} - d_q{}^{sim}|}{d_q{}^{obs}}$, where quantile q is taken from the set of quantiles (0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95), $d_q{}^{obs}$ is the qth quantile of the observed distance distribution, and $d_q{}^{sim}$ is the qth quantile of the simulated distances. Finally, the deviances from between and within the male and female groups are calculated and added together to become our evaluation measure for goodness of fit each set of parameters. Results are shown for different sets of parameters in the heatmap. Although Fig. 3 shows two clear bands where deviance is smaller, the trends across the heatmap are not smooth and small changes in either within or between group diversity can lead to abrupt changes in the similarity of the distributions.

In Fig. 4, we show boxplots of the simulated weighted Unifrac beta-diversity estimates corresponding to the best situation in Fig. 3, as well as three other parameter choices that gave less good matches to AG data. It can be seen that the distance distributions vary substantially in magnitude and variability. By selecting the rarefaction parameter of 0.10 and an effect size of 0.18, the deviances were the smallest, indicating best match to the AG sex beta-diversity measures.

In Kelly's study [8], the starting distribution of counts across the OTUs was assumed to be uniform, so that an equal number of sequences is assigned to each OTU. However, this is not a realistic assumption since usually there is a large range between the most common and least common OTU. This assumption can be easily altered to simulate from a non-uniform distribution. In the following section, we show power results when simulations started with non-uniformly distributed OTU counts.

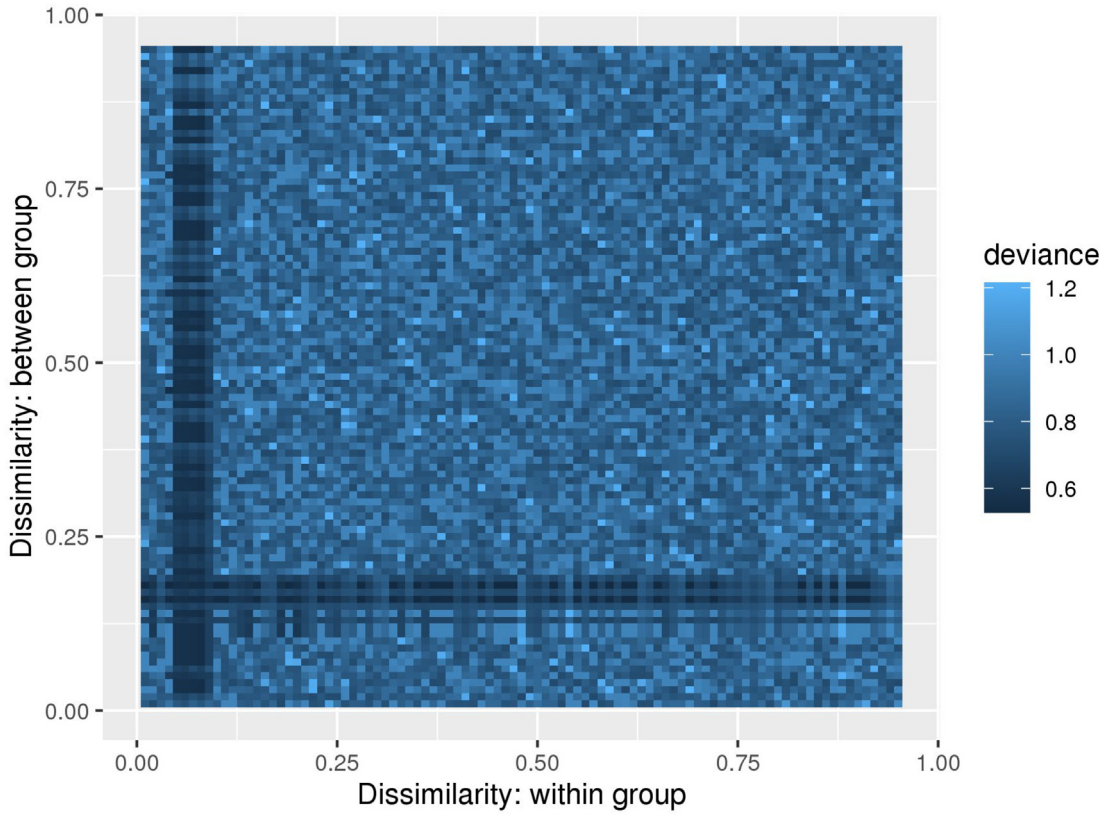

Fig. 3. Heatmap of summed absolute deviances across 7 quantiles between simulated beta-diversity distances and observed distances from AG data for two groups defined by sex. Simulated data are obtained with Kelly's method, varying the simulation parameters in each cell to match observed data most closely.

4. Power calculations for beta-diversity measures

With a method in hand to obtain or generate an appropriate distance matrix, power can be estimated by bootstrap sampling if we follow Kelly's [8] method. This approach samples individuals from the distributions implied by the distance matrix, allocates them into groups with the desired effect size, and then analyzes between- and within-group distances using a method such as PERMANOVA [3]. Performing this process repeatedly enables estimation of the proportion of datasets where the null hypothesis is rejected at a chosen significance threshold. For example, to estimate the statistical power of a study that includes 10 subjects per group, we can randomly select, with replacement, 10 subjects per-group from each simulated distance matrix. If we repeat the selection and analysis, say 100 times, power will be the percentage out of 100 where the p-value is less than the threshold. Hence, by varying these calculations for different sample sizes, one can choose an appropriate study size with good power. Sample code for calculating power using this approach can be found in Supplement A, Box 4.

4.1. Illustration of power calculations with American Gut benchmark data

Using two of the body sites within the AG data (skin and hand) as benchmark data, we illustrate the results of two power calculations in Fig. 5.

There were 169 skin samples and 165 hand samples, which were combined to create a dataset of 334 individuals to calculate pairwise distances. We then sampled from this large distance matrix assuming a desired study sample size of 100 individuals divided into 2 groups (Fig. 5a) or 100 individuals divided into 10 groups (Fig. 5b), and a type 1 error of 0.05. The effect sizes were varied by changing the scaling factor σ from 0.9 to 10. After creating between and within distances, the resulting effect sizes were calculated for each simulated data set, and this was repeated 100 times. Due to stochasticity associated with the bootstrap sampling, a chosen scaling factor leads to distance matrices with similar but not identical effect sizes; the effect sizes obtained ranged from 0.003 to 0.98. Therefore, we calculated power by the proportion of datasets rejected for a given value of the scaling factor, rather than for the effect sizes. We smoothed the results in Fig. 5 so that the general trend can be seen. The power increases with effect size as would be anticipated. With 10 groups, the effect sizes must be larger before good power, of 80% or more, is obtained.

4.2. Illustration of power calculations without benchmark data

Fig. 6 shows the results from power calculations built on Kelly's data simulation method, using unweighted Unifrac distances, assuming the same parameters as for Fig. 5, although the segregation proportion was varied to lead to a range of effect sizes and powers between zero and one.

In Fig. 6a, power is shown assuming a uniform distribution across the OTUs. However, as discussed above, all OTUs are unlikely to be equally common in microbiome samples. Therefore, we adapted the simulation to vary the initial number of counts across the 50 OTUs, following a binomial distribution of size 62, proportion 0.8 (giving a mean of 50), and results are shown in Fig. 2b. Power estimates become more variable with the latter simulation. Hence, these results imply that to ensure good power in planning a study, it would be best to anticipate the variability in the distribution of OTU counts, and to use a larger sample size if possible.

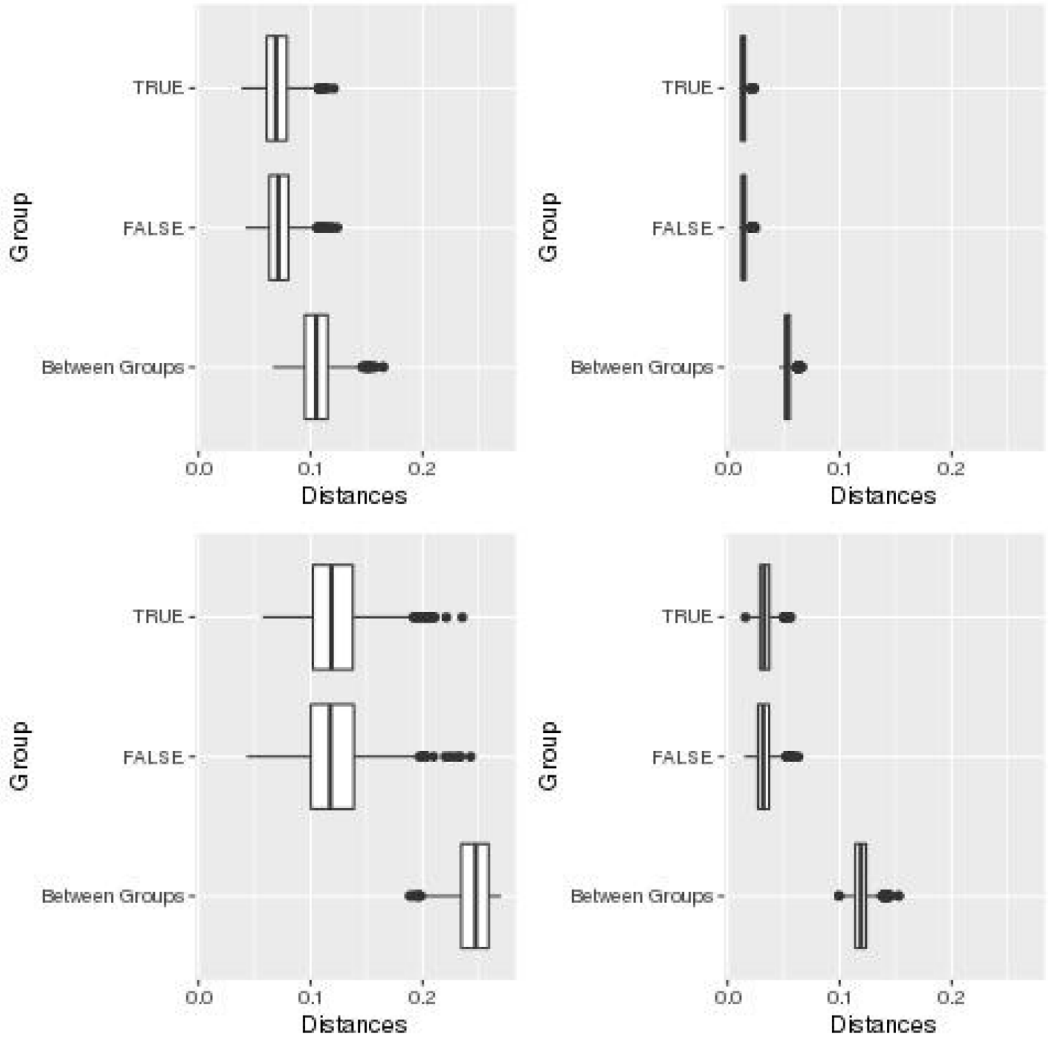

Fig. 4. Distributions of simulated beta-diversity using Kelly's method for four scenarios. Top left: best match of best: best grid on heatmap with low deviance as defined for Fig. 3 (rarefaction 0.10, effect size 0.18). Top right: high deviance between groups (rarefaction 0.07, effect size 0.81). Bottom left: high deviance within groups (rarefaction 0.83, effect size 0.11). Bottom right: cell in heatmap with highest deviance (rarefaction 0.34, effect size 0.50).

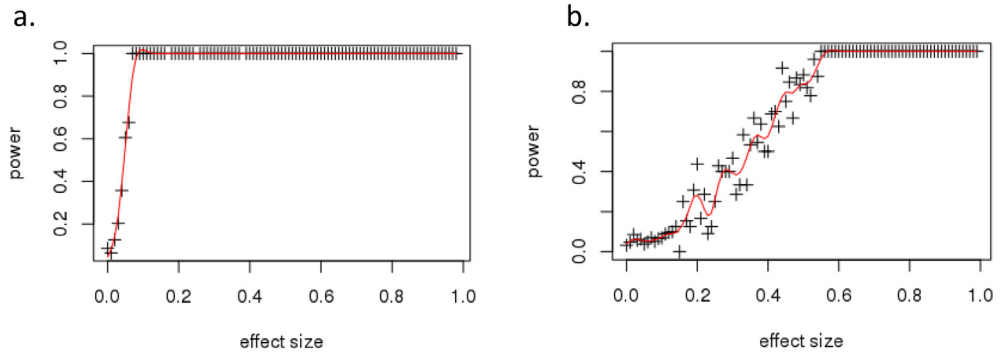

Fig. 5. Power calculations based on benchmark data. We simulated distance matrices based on the observed distance matrix calculated on 169 skin samples and 165 hand samples in the AG dataset. Data were simulated for 100 individuals from 2 groups (a) or 10 groups (b), with a Type I error of 0.05 in both simulations.

Fig. 7 shows simulation results for unweighted Unifrac distances using Kelly's method with 4 groups, assuming an equal distribution of counts across the OTUs to start the simulation. It can be appreciated in Fig. 7 that estimated power in this simulation setup is extremely variable. For instance, when effect sizes are 0.3 or larger, power can be close to one, whereas for many other values, power was close to zero. In fact, we often found it difficult to choose parameter values for Kelly's simulation method that always gave realistic distances, particularly when there were more than 2 groups.

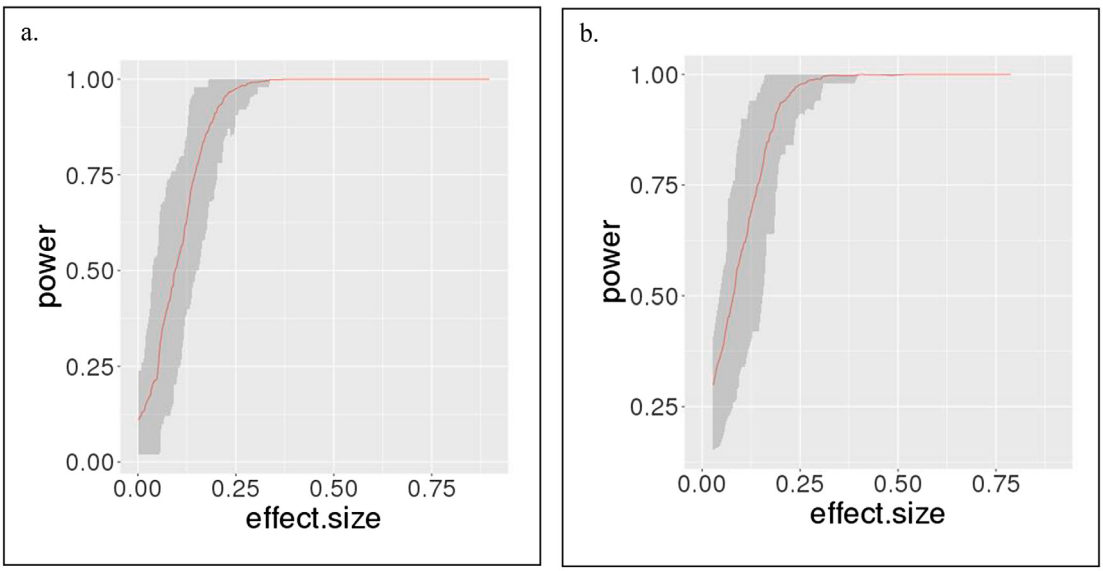

Fig. 6. Power calculations using Kelly's method for unweighted Unifrac distances starting from uniform OTUs (Fig. 6a) and non-uniform (Binomial distributed) OTUs (Fig. 6b). The rarefication proportion (P_r) is 0.03, the segregation proportion is 0.1, N = 100, there are two groups, and 50 OTUs. In panel 2a, there were 50 sequence counts per OTU bin. In panel 2b, counts per bin were generated from Binom (62, 0.8). Power was calculated using the code in Supplement A Box 4 with a type I error of 0.05. The shaded area corresponds to a 95% confidence bank obtained within a bandwidth of size 0.05. Specifically, for a specified value of effect size, α, the simulations rendering estimated effect sizes +/- 0.05 from α were collated, and the 95% CIs of the values of power are shaded accordingly.

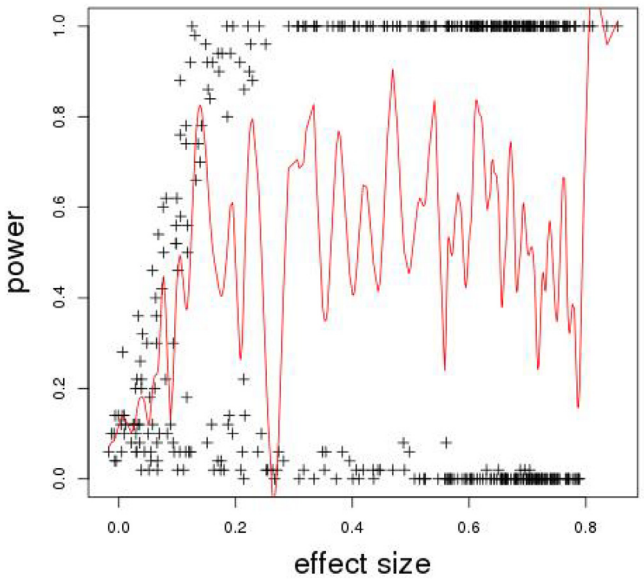

Fig. 7. Power calculations using Kelly's method for unweighted Unifrac distances starting from uniform OTUs with N = 100 from 4 groups. We set the rarefication proportion (P_r) = 0.03, while fixing other parameters as N = 100, 4 groups, 50 OTUs, 50 sequence counts per OTU bin with a type I error of 0.05. The segregation proportion was varied between 0.1 and 1.0, and thence effect sizes ranged from 0 to 0.65. Power calculations were performed using code in Supplement A Box 4. The red curve is a smoothed cubic spline fit to the relationship between effect size and power. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Concluding remarks

Clustering methods for microbial beta-diversity have become popular in human and animal microbiome studies to show how whole microbial communities differ across groups of interest. For most part, these are visual tests and microbiota cluster differences are not tested statistically. Appropriate power calculations, performed before undertaking a study, can reassure the researchers (and their audience) that meaningful differences in microbiome composition are likely to be detected – if they exist – with the study being planned. However, obtaining realistic estimates of beta-diversity distance metrics can represent a significant challenge when performing power or sample size calculations. The goal of this paper was to aid researchers planning human or animal microbiome studies. We provided definitions of commonly-used microbial beta-diversity distance metrics, discussed methods for estimating distance distributions from pilot studies, or from summary statistics in existing publications, and demonstrated how to use inference to calculate sample size and power for one analysis method. We illustrated two simulation approaches for generating realistic distance matrices with or without preliminary (benchmark) data. Finally, using sample sizes that are typical in microbiome studies, we reported the power for detecting beta-diversity differences between groups for a range of assumptions.

The procedures that we have described apply when comparing beta diversity between independent groups of individuals. If a longitudinal study were desired, then one would need estimates of the distribution of intra-individual *changes* in beta-diversity. Analogously, for dependent samples, such as those from the same family, then both within cluster and between cluster beta diversity distributions would be needed for sample size calculations. The approach of Field et al. (2013) [24] can be used to adjust the effect size equation (1) to account for repeated measures on the same individual. However, there is no self-evident way to adapt the simulation procedure of Kelly et al. [8] to create correlated sets of data. This could be an interesting research direction.

Researchers should be cautioned that the distance summary statistics that we showed should be considered as illustrative of the *methods* to calculate distances, rather than illustrative of the distances themselves. American Gut-derived distances include multiple sampling sites taken from the same individuals, and so there could be dependence between the microbiome count distributions. Also, our calculations were shown for the phylum level, and hence will not apply to other levels. Researchers should

thoughtfully select relevant publications or pilot data for use when undertaking sample size calculations.

The choice of the distance metric depends on the objectives and priorities of the power calculation, and the choice of simulation method depends on the: (i) availability of the pilot dataset, and (ii) assumptions researchers want to make. In particular, we found that simulated distances, within- and between-groups need careful scrutiny before conducting power calculations. We found that results from Kelly's method can be unreliable, dependent on assumptions, this particularly when planning comparisons among several groups. Moreover, when planning to study microbiota environments, interventions, or sampling sites different from those examined here, microbial beta-diversity distances provided in this paper—either from simulation or from the American Gut project—may not resemble those intended for study. Nevertheless, by varying assumptions, a range of distributions can be obtained and a study can be designed with conservative assumptions.

Although the beta-diversity distance calculations and power curves presented in this paper are specific to the data choices and modelling assumptions we made, we anticipate that our examples will provide researchers with concrete starting points for choosing distance estimates for their own studies. Throughout this paper, we have provided detailed R code to assist in simulating and analyzing data on beta-diversity distance measures, and in calculating sample size or power. The inclusion of examples of distances that we obtained, despite their limitations, may guide researchers in the generation of their own distance matrices. We encourage microbiome scientists to test and optimize the tools/scripts presented here, and future ones, as microbiome datasets become more plentiful in public repositories.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We gratefully acknowledge platform support from the Canadian Institutes for Health Research, with grant number IMC-161484.

Appendix A

Lai Jiang[1,2], Tahsin Ferdous[1,3], Irina Dinu[1,4], Julie Groizeleau[1,3], Jayne Danska[1,5,6], Kathy D. McCoy[1,3], Marie-Claire Arrieta[1,3], Anita L. Kozyrskyj[*,1,4,7], Celia M.T. Greenwood[*,1,2,8,9].

[1]IMPACTT: Integrated Microbiome Platforms for Advancing Causation Testing & Translation, Canada

[2]Lady Davis Institute for Medical Research, Montreal, QC, Canada

[3]Dept. of Physiology & Pharmacology, Cumming School of Medicine, University of Calgary, AB, Canada

[4]School of Public Health, University of Alberta, Edmonton, AB, Canada

[5]Hospital for Sick Children Research Institute, Toronto, ON, Canada

[6]Dept. of Immunology and Dept. of Medical Biophysics, Faculty of Medicine, University of Toronto, Toronto ON, Canada

[7]Department of Pediatrics, University of Alberta, Edmonton, AB, Canada

[8]Gerald Bronfman Department of Oncology, McGill University, Montreal, QC, Canada

[9]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada

*Authors contributed equally.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.04.032.

References

[1] Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214, doi:10.1038/nature11234 (2012).
[2] Ghannam RB, Techtmann SM. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Comput Struct Biotechnol J 2021;19:1092–107. https://doi.org/10.1016/j.csbj.2021.01.028.
[3] Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecol 2001;26:32–46. https://doi.org/10.1111/j.1442-9993.2001.01070. pp. x.
[4] Mirzayi C et al. Reporting guidelines for human microbiome research: the STORMS checklist. Nat Med 2021;27:1885–92. https://doi.org/10.1038/s41591-021-01552-x.
[5] Bharucha T et al. STROBE-metagenomics: a STROBE extension statement to guide the reporting of metagenomics studies. Lancet Infect Dis 2020;20:e251–60. https://doi.org/10.1016/S1473-3099(20)30199-7.
[6] Casals-Pascual C et al. Microbial diversity in clinical microbiome studies: sample size and statistical power considerations. Gastroenterology 2020;158:1524–8. https://doi.org/10.1053/j.gastro.2019.11.305.
[7] Anderson MJ et al. Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist. Ecol Lett 2011;14:19–28. https://doi.org/10.1111/j.1461-0248.2010.01552.x.
[8] Kelly BJ et al. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics 2015;31:2461–8. https://doi.org/10.1093/bioinformatics/btv183.
[9] Zhao N et al. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. Am J Hum Genet 2015;96:797–807. https://doi.org/10.1016/j.ajhg.2015.04.003.
[10] Wu C, Chen J, Kim J, Pan W. An adaptive association test for microbiome data. Genome Med 2016;8:56. https://doi.org/10.1186/s13073-016-0302-3.
[11] Koh H, Blaser MJ, Li H. A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. Microbiome 2017;5:45. https://doi.org/10.1186/s40168-017-0262-x.
[12] Gail MH, Wan Y, Shi J. Power of microbiome beta-diversity analyses based on standard reference samples. Am J Epidemiol 2021;190:439–47. https://doi.org/10.1093/aje/kwaa204.
[13] Hall BG. Building phylogenetic trees from molecular data with MEGA. Mol Biol Evol 2013;30:1229–35. https://doi.org/10.1093/molbev/mst012.
[14] Fukuyama J, McMurdie PJ, Dethlefsen L, Relman DA, Holmes S. Comparisons of distance methods for combining covariates and abundances in microbiome studies. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. p. 213–24.
[15] Jaccard P. The distribution of the flora in the alpine zone. New Phytol 1912;11:37–50.
[16] Oksanen J. *et al.* Vegan: Community Ecology Package. R package, version 2.5-7. (2020).
[17] Bray JR, Curtis JT. An ordination of the upland forest communities of Southern Wisconsin. Ecol Monogr 1957;27:325–49.
[18] Wong RG, Wu JR, Gloor GB. Expanding the UniFrac Toolbox. PLoS ONE 2016;11:. https://doi.org/10.1371/journal.pone.0161196e0161196.
[19] McDonald, D. *et al.* American Gut: an open platform for citizen science microbiome research. *mSystems* **3**, doi:10.1128/mSystems.00031-18 (2018).
[20] Tannock GW et al. Comparison of the compositions of the stool microbiotas of infants fed goat milk formula, cow milk-based formula, or breast milk. Appl Environ Microbiol 2013;79:3040–8. https://doi.org/10.1128/AEM.03910-12.
[21] Ferdous T. *et al.* The rise to power of the microbiome: power and sample size calculation for microbiome studies. (Submitted).
[22] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Lawrence Erlbaum Associates; 1988.
[23] Hughes JB, Hellmann JJ. The application of rarefaction techniques to molecular inventories of microbial diversity. Methods Enzymol 2005;397:292–308. https://doi.org/10.1016/S0076-6879(05)97017-1.
[24] Field A. Discovering statistics using IBM SPSS statistics. Sage; 2013.