Guidelines to Statistical Analysis of Microbial Composition Data Inferred from Metagenomic Sequencing

Vera Odintsova[1]*, Alexander Tyakht[1,2] and Dmitry Alexeev[1,2]

[1]Federal Research and Clinical Centre of Physical-Chemical Medicine, Malaya Pirogovskaya 1a, Moscow, Russian Federation.
[2]Moscow Institute of Physics and Technology, Institutskiy pereulok 9, Dolgoprudny, Russian Federation.

*Correspondence: vera.odints@gmail.com

https://doi.org/10.21775/cimb.024.017

Abstract

Metagenomics, the application of high-throughput DNA sequencing for surveys of environmental samples, has revolutionized our view on the taxonomic and genetic composition of complex microbial communities. An enormous richness of microbiota keeps unfolding in the context of various fields ranging from biomedicine and food industry to geology. Primary analysis of metagenomic reads allows to infer semi-quantitative data describing the community structure. However, such compositional data possess statistical specific properties that are important to consider during preprocessing, hypothesis testing and interpreting the results of statistical tests. Failure to account for these specifics may lead to essentially wrong conclusions as a result of the survey. Here we present a researcher introduction to the field of metagenomics with the basic properties of microbial compositional data including statistical power and proposed distribution models, perform a review of the publicly available software tools developed specifically for such data and outline the recommendations for the application of the methods.

Introduction

Microbiota, complex communities consisting of microbial species, appear to inhabit literally any environmental niche in the world. Recent advances in molecular genetic techniques allowed the study of microbiota in a cultivation-independent way, leading to the discovery of enormous diversity. One of the most advanced and widely used techniques is metagenomic sequencing: classification and quantification of metagenomic sequences can be used to assess both taxonomic and functional composition of microbiota. However, when a researcher is interested in proper statistical assessment of hypothesis regarding the differences between the samples from different groups or its dependence from multiple factors, quantification of metagenomic reads is just an intermediate step. Generally, the specific features and problems of analysing metagenomic data are universal whether the focus is on microbiota of a subway station, saline lake or host-associated communities. However, here we will illustrate these concepts on the example of human microbiota, as it possesses particular interest to researchers due to its essential role in a biomedical context.

The human body is a habitat for a great amount of microbes (approximately 1–3% of body mass). They play important roles in biological processes that maintain its vital activity. The microbial community consists of thousands of species, and its structure noticeably varies both among subjects and body parts (Levy and Borenstein, 2013; Stein *et al.*, 2013; Tyakht *et al.*, 2013; Human Microbiome Project Consortium, 2012). As cultivation and examination *in vitro* of the majority of microorganisms is difficult, metagenomic analysis is one of the most common ways to study the human-associated microbial community structure and functionality.

An important part of metagenomic analysis is testing hypotheses about the associations between microbiome structure and some factors. Age, diet type, presence of some disease and so on may play the role of such factors in the case of human microbiome research. Factors may take on discrete or continuous values. Such characteristics as sex, body site and stage of the disease represent discrete ones. They allow us to distinguish two or more groups to compare structure of their microbiota. While designing the groups and the inclusion/exclusion criteria for a metagenomic survey, a researcher should make sure that the groups are matched and differ only by the selected features. For example, in a case when the microbiota of healthy subjects and patients with dysbiosis are compared, then there should be no substantial differences in weight, age, sex and other parameters between the groups.

Age, body mass index and drug dosage are the examples of continuous factors. It is impossible to distinguish groups in such case. So the researcher's aim is to obtain the functional dependence of microbiome structure on factor value. A study may aim to explore the association of the microbiome with more than one factor at a time. Multifactor analysis is useful when several characteristics may contribute to the changes in microbial structure and each of them is presumed to have substantial influence. Such analysis attempts to estimate the individual impact of each factor.

A common metagenomic study of association between clinical data and microbiome composition consists of the following stages (Goodrich *et al.*, 2014):

1 formulation of the aims and experiment design (stating the hypothesis, describing the groups of subjects, determining their minimal size, choosing the optimal sequencing technology, targeted sequencing depth and methods of experimental data analysis);
2 collecting the required number of microbial samples;
3 metagenomic sequencing of each sample;
4 taxonomic profiling for each metagenome;
5 statistical analysis of the compositional data.

This review describes basic concepts and models of statistical inference of associations between microbiome structure and factors of interest, with the main focus on human microbiota. It is intended to serve as an assistance in choosing the correct statistical methods for metagenomic analysis realized in R (R Core Team, 2015) and applying them properly (stages 1 and 5).

To state the mathematical formulation of the problem it is necessary to get familiar with the data format inherent for metagenomics. We will first briefly describe the first steps of a metagenomic study from collecting samples to obtaining its taxonomic profile. Next, the statistical formulation of the aim of the research and design of an experiment will be discussed. Then we will focus on the main steps and specifics of metagenomic statistical analysis. The last section contains an overview of the actual and widespread approaches to such analysis implemented in R packages.

Preparing data for statistical analysis

Collection of microbiome samples, sample preparation and sequencing, as well as the primary bioinformatical analysis for taxonomic profiling are complex processes, with each step having its own subtleties. A comprehensive description of them is out of scope of this text and is described elsewhere (for example, in Goodrich *et al.*, 2014). Here we will outline only the essential points.

After collecting the microbiota samples from the individuals under investigation, the samples are stored and transported under low temperature and other conditions to prevent changes in the structure of microbial community before the sample gets to a laboratory. There the DNA is extracted from the samples through a multistage process with the use of special reagents. Then a so-called sequencing library is prepared for each sample. The obtained libraries are subject to DNA sequencing resulting in thousands to tens of millions of short nucleotide sequences (reads) corresponding to the genomes of all microorganisms and viruses present in a sample. There are two main types of sequencing: shotgun sequencing (random sequences for the totality of the genetic material are obtained) or amplicon sequencing (reads belong to a fixed gene of each species, most commonly 16S rRNA gene). The resulting reads are subject to quality filtering. Then taxonomical classification is performed so each read is put into correspondence with the available taxonomically annotated database of microbial genomes or genes. It is also possible to perform a *de novo* analysis without the reference. The result of the taxonomical classification for all reads of a sample is a vector of feature abundances. Each element of this vector reveals the number of reads related to some feature – taxon, gene or gene group. In the context of 16S rRNA sequencing, the common type of feature are referred to as operational taxonomic units (OTUs); as the concept of OTU is not normally used in 'shotgun' sequencing, we will use a microbial species as a feature in the text.

'Shotgun' sequencing allows us to describe the functional structure of microbial community in addition to its taxonomic structure. It shows semiquantitative portrait of genes, gene groups or metabolic paths in the community. Currently 'shotgun' sequencing is a much more costly procedure than the amplicon format. So it is less widespread, albeit it does not give an idea about gene structure of a microbiota. In this paper, without loss of generality, we will describe methods of statistical analysis on the example of 16S rRNA data. The result of the taxonomical classification for all reads of a sample is a vector of feature relative abundance values. Each element of this vector reveals a read count related to some feature (microbial taxon). The microbial communities are compared via the analysis of these vectors.

Statistical properties of metagenomic data

Basic steps

As an exhaustive search across all gut microbiotas is not feasible, the statistical analysis is based on a representative sample from this entire assembly. The researcher is in charge for controlling the validity of the outcome. To identify the changes in microbiota composition associated with certain factors, a researcher postulates and then tests a null hypothesis that the groups do not differ. For instance, that the observed difference in relative abundance of a microbial species between the healthy subjects and patients is just due to chance.

Each feature describing microbial composition (e.g. relative abundance of a certain species) can be considered as a random value realized in each individual metagenome. Then a null hypothesis is postulated stating that the distribution of this random value is independent of the examined factor. Often the differences between distributions are assessed by comparing their mean (or median) values, with null hypothesis stating that the parameter is equal across the groups.

As a means for hypothesis testing, a test statistic is introduced – a numeric function reflecting the degree of similarity between the samples. Then a *P*-value is computed – a probability that such or a more extreme statistics value can be observed assuming the null hypothesis is true. The absolute value of deviation (two-sided test) or deviation in certain direction (one-sided test) may be considered. A one-sided test is convenient for experiments with a strong a priori-supported direction of the factor influence for example, when assessing the deleterious effect of antibiotics on the abundance of sensitive species. One-sided testing allows the detecting of changes with a smaller effect given the same significance level. But more often the direction of influence is not known or the relative abundance of each of the taxa in a community is compared at a time, so, while the abundance of some taxa increases, the abundance of the others has to decrease. In such cases, two-sided test is the right choice.

A small *P*-value means that it is unlikely to obtain two such samples from the same distribution. If *P*-value is lower than a defined threshold then the null hypothesis is rejected. The threshold value (known as a critical value or a significance level of a test) is not fixed absolutely and may vary but it is commonly set to 5% ($P < 0.05$ corresponds to significant differences).

A *P*-value reflects the chance of type I error – rejecting null hypothesis assuming that it is true ('false discovery'). Importantly, 5% significance does not imply that the probability that the null hypothesis is true is 5%. It just means that there is 5% chance to observe these differences when the null hypothesis is true. Moreover, the estimate of significance does not provide information about the scale of the observed differences. A slight increase in levels of a species may be significant but have little impact on the microbial ecology whereas large-scale changes might be assessed as insignificant. Therefore, one should be wary when basing the conclusions of a metagenomic study on *P*-value analysis alone (Baker, 2016; Wasserstein and Lazar, 2016).

In the case of multiple comparisons – for example, when the distribution for each of the hundreds of detected microbial species is compared – it is necessary to control for the number of 'false discoveries'. Assuming the critical *P*-value of 0.05 for each test, we allow 5% probability of making a 'false discovery' – i.e. concluding that the fraction of the taxon is significantly different between the groups while in reality it is not. However, the probability of making a 'false discovery' is much higher: for at least two of the tests, it is $1 - (1 - 0.05) \times ($

1 – 0.05) = 0.0975; in three of the tests it is 0.14; and in four it is 0.4. In order to avoid making such erroneous conclusions, the *P*-values should be adjusted for multiple comparisons via a dedicated routine, e.g. the Bonferroni method (a rather strict method that also decreases the number of 'true discoveries') or a more widely used Benjamini–Hochberg method that controls false discovery rate (FDR) only among 'discoveries'. A FDR of 0.05 means that approximately 5% of significant tests will be 'false discoveries'.

While metagenomic researchers commonly control the rate of type I error using *P*-values, they often fail to take into account the type II error – the error of accepting null hypothesis in the case when it is wrong: for example, of concluding that the level of a bacterial species does not vary significantly between the groups while in fact it does. The respective solution is to use the concept of statistical power. This measure characterizes the quality of detecting the differences between the groups; its value equals to 1 minus the probability of type II error (Table 2.1).

The relation between the *P*-value and statistical power can be illustrated by an example of comparing the abundance of a potential biomarker species between healthy subjects and patients with a specific disease. While the null hypothesis is that the abundance values for the two groups were sampled from the same distribution, an alternative hypothesis is that the samples are different, i.e. the expected levels for the microbiota of control subjects and patients are different. Type I error will occur if the researcher decides that the differences exist while in fact they do not. In such case, a species will be by mistake considered a potential microbial biomarker of the disease thus leading to further useless consumption of resources to examine in detail the species which is ultimately not involved in the disease.

On the other hand, type II error will happen if a researcher does not detect the strong association between the microbial species and the disease – in the case when indeed such connection exists – thus missing an important discovery that might result in understanding of the pathogenesis mechanism and, potentially, development of novel prediction and treatment approaches.

Obviously, a researcher will attempt to reduce the chances of type I and II errors at the same time. However, these rates are related in a way that the decrease of one leads to the increase of the other, and vice versa (Fig. 2.1). So one has to find the reasonable balance between the two measures depending on the aims of the study. One should not be limited only to the control of the significance level, as the low power increases the probability of finding the differences that do not exist, given a fixed *P*-value threshold (Sham and Purcell, 2014). Therefore, improper control of the power leads to low reproducibility of results.

Table 2.1 Relations between type I and II errors in hypothesis testing. The null hypothesis H_0 states that there are no differences between the two distributions, the alternative hypothesis H_1, that the distributions are different

Reality	Statistical conclusion	Error	Probability	Term
H_0	H_0	No error	$1-\alpha$	
H_0	H_1	Type I error	α	Significance
H_1	H_1	No error	$1-\beta$	Power
H_1	H_0	Type II error	β	

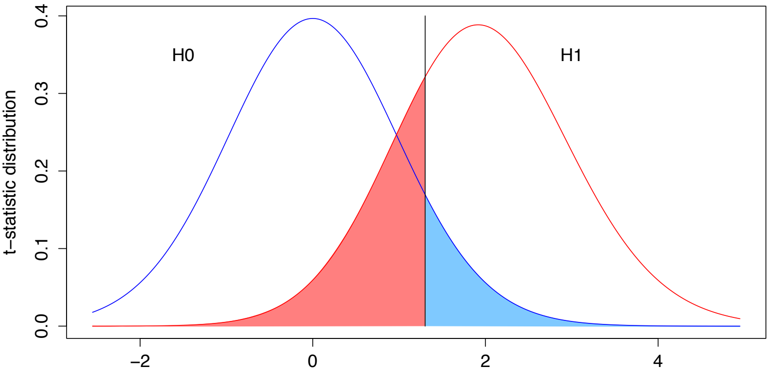

Figure 2.1 Distribution of the test statistics according to null and alternative hypotheses for one-sided Student's test. The black line denotes significance level, the area of the blue region is the probability of a type I error (α) and the area of the red region is the probability of a type II error (β). The illustration is based on the analysis of log-transformed data for healthy subjects and patients with type 2 diabetes (Egshatyan *et al*., 2016).

Statistical power in a metagenomic study is affected by a number of experimental parameters. These include the effect size [measure of observed differences caused by the factor(s)], the DNA sequencing depth, the level of taxonomical description, the choice of the test statistics, the critical value for the type I error probability, the method of multiple comparison correction, and the sample size (Hair *et al.*, 2010; La Rosa *et al.*, 2015). The dependence of power on some of these parameters is presented in Jonsson *et al.* (2016), Kelly *et al.* (2015) and La Rosa *et al.* (2015). The dependence on sample size is especially important for metagenomic studies, as the high cost of sequencing is very restrictive. Therefore, at the stage of experiment planning it is prudent to ensure that the selected sample size is sufficient to detect the underlying dependences but is not redundant and that the sequencing depth allows to capture the minor community members at the desired level of detail. In practice, 80% power threshold is commonly used in biostatistical analysis (La Rosa *et al.*, 2012). Details on the choice of power threshold are described elsewhere (Sham and Purcell, 2014).

In order to determine the required number of samples for obtaining the desired power level, one should proceed from the choice of statistical method, acceptable significance level, taxonomic level and expected effect size. The latter parameter can be determined during a pilot experiment (when a small subgroup of metagenomes is sequenced) or from published data revealing typical effect sizes in similar experiments (such as Kelly *et al.*, 2015; La Rosa *et al.*, 2015). It is worth noting that it can possibly be overestimated in experiments with a small sample size (Button *et al.*, 2013). A researcher should choose the minimal group size allowing to achieve the target level of power. Some of the software packages for statistical analysis of metagenomic data offer functions for power calculations specific for the included statistical methods. These functions should be applied to pilot data. However, if such functions are not available and the pilot sample size is sufficient, a non-parametric method (e.g. permutation-based) should be used to make sure the statistical power is acceptable. When choosing the method of analysis, one should particularly pay attention to the validity of the assumed parameterization, as the wrong choice may lead to overestimate of power or significance leading to a higher rate of type I or II errors that the researcher expects (Hair *et al.*, 2010). The choice of a model will be discussed in detail below.

Specific statistical features of metagenomic data

Metagenomic data possess inherent statistical properties that require a careful approach to the choice of statistical analysis methods (Paulson *et al.*, 2013).

Total read count varies between the samples

The first problem is caused by the technical variability of sequencing data volume per sample: the total sum of reads may be substantially different. Modern DNA sequencers allow the processing of hundreds of samples at a time, but do not guarantee uniform distribution of reads across the metagenomes. Metagenomic data are compositional and direct comparison of the reads that correspond to a feature in two metagenomes does not give a correct representation of the relative abundance of this feature in samples. For example, let 20 reads be classified as the taxon in the first sample and 50 in the other one, while the total reads sum is 1000 in the first sample and 500 in the second one. Then 2% of the first metagenome – a fraction greater than 2% of the second one – correspond to the taxon, although 20 reads are less than 50 reads.

There are several ways of solving the problem. The comparison of them is presented in a paper by McMurdie *et al.* (2014). The first way is equalization of the total sum of reads by a random rarefaction of the data so that the read count is equal across all samples. However, in such case a large fraction of data are being discarded in a wasteful way and the precision of measurement is decreased, especially for metagenomes with highest sequencing depth (McMurdie *et al.*, 2014). As a result, a high rate of type I and II errors is observed. Above all, the random choice of reads decreases the repeatability of an experiment and adds biases (McMurdie *et al.*, 2014).

The second approach is normalizing the data. The most straightforward method is to divide each component of a feature vector by the total read sum for the sample. One disadvantage of such normalization is that the transformed vector loses information about the sequencing depth and thus precision of measurements, and therefore the variance cannot be correctly estimated. This leads to high level of false positives (McMurdie *et al.*, 2014).

More sophisticated methods of normalization methods implemented in packages for metagenomic and differential expression analysis were shown to overcome the problem more successfully (McMurdie *et al.*, 2014).

Dependence of the feature abundance variance on the mean value

Another specific property of metagenomic data is that the normalized variance of taxa differs among taxa and depends on factors whose association with data is investigated. An example of feature variability is presented in Fig. 2.2a. Thus, the property of homoscedasticity (independence of the variance from the factor values) is not fulfilled – however, it is required for correct use of many statistical methods, both parametric and non-parametric. So one should either resort to specialized statistical tests designed for heteroscedastic data or apply a variance-stabilizing data transformation to the data. Common solutions include the square-root transformation, arcsine-square-root- and log-transformation paired with adding a pseudocount (in order to avoid infinite values) to all taxa (Jonsson *et al.*, 2016;

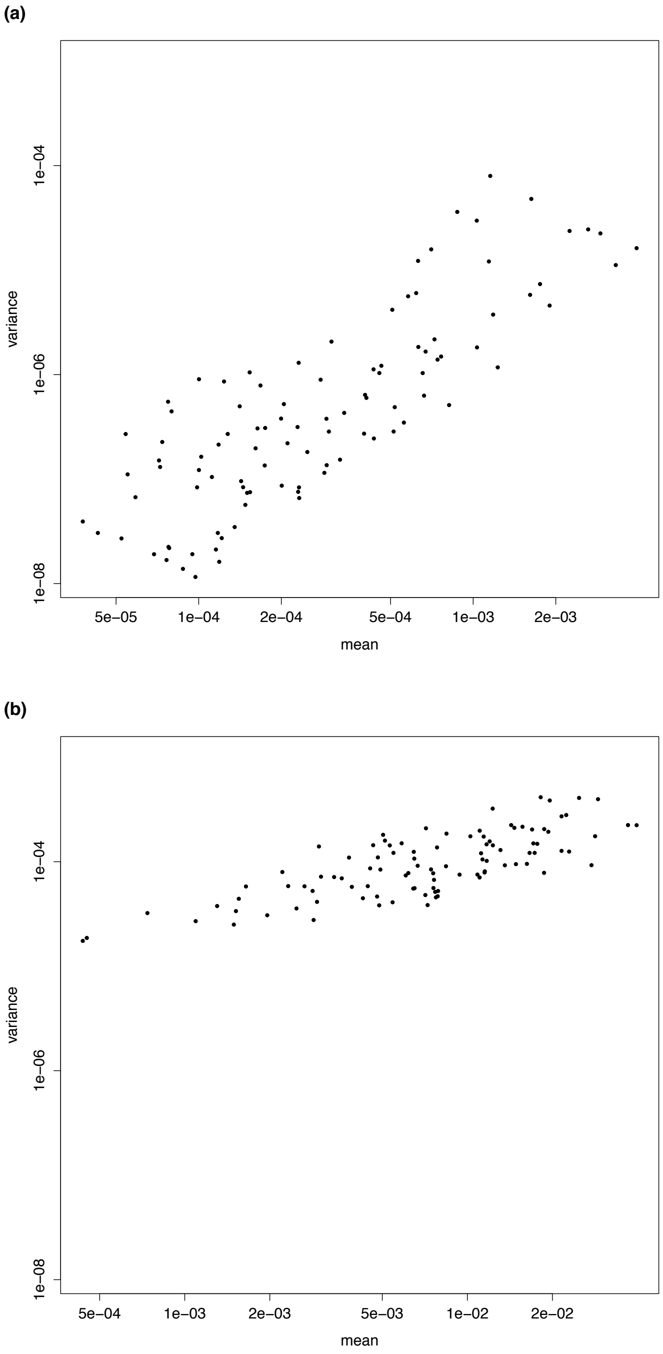

Figure 2.2 Relation between variance and mean of species relative abundance in logarithmic scale. (a) Normalized by the total sum of reads per sample, (b) normalized by the total sample sum log-transformed data (with adding one 'pseudocount'). The scatter plots are constructed for the published data on gut microbiota of healthy people and patients with diabetes (Egshatyan *et al*., 2016); the points correspond to 100 randomly selected OTUs.

Wang *et al.*, 2016). The dependence of variance on the mean value after log-transformation is illustrated in Figure 2.2b. It was previously shown (Wang *et al.*, 2016) that a simple means comparison of log-transformed data is a powerful method compared with some other parametric and non-parametric methods. Regression analysis paired with arcsine-square-root transformation was also more powerful compared to the methods that used untransformed data. In method comparison (Jonsson *et al.*, 2016), square-root transformation followed by the Student's test showed results comparable to the results of methods designed for metagenomic data.

Normal distribution does not fit well to the metagenomic data

The other specific property of metagenomic data is that the normal distribution implied in many statistical tests does not describe this type of data well. First of all, normal distribution is continuous while the metagenomic sequencing data have discrete nature. Second, components of a feature vector cannot be negative. At the same time the mean value of taxon abundance is often comparable to its variance and the abundance matrix is sparse (has many zero values). For example, in the case of gut microbiota, due to the wide diversity of microbial community structures, a typical bacterial species has zero or low abundance in microbiota of the majority of people while only a small fraction of the population carry high abundance (tens of per cent) of this species. A typical histogram of taxon abundance across samples is presented in Fig. 2.3a.

One way of overcoming the problem is to apply non-parametric methods that are not based on any underlying distribution. However, their flaw is relatively low sensitivity, which leads to the inflated rate of type II error probability (La Rosa *et al.*, 2012).

Another method is transforming the data so that they fit a normal distribution better. Partially this problem may be solved by applying the above-mentioned variance-stabilizing transformation (Fig. 2.3b), but the scarcity of metagenomic data remains a problem (Paulson *et al.*, 2013; Wang *et al.*, 2016) (Fig. 2.3c and d). The quality of the fit can be subsequently assessed using the Shapiro–Wilk test of normality.

The third way is to choose a distribution other than normal and apply appropriate parametric methods. Often this distribution is based on the modification of a multinomial distribution that describes the number of successes in a series of n independent experiments with k possible outcomes, each with probability p_k,

$$\sum_{i=1}^{k} p_k = 1.$$

In the case of metagenomic data, each experiment corresponds to the classification of a single read, possible outcomes – to the taxa and the total number of experiments – to the sequencing depth. In contrast to the normal distribution, the multinomial one is discrete.

As the number of species in a complex microbial community is high, interactions between them are often neglected during statistical analysis and a binomial distribution is used for the component-wise description of relative abundance vector. The binomial distribution is a restriction of the multinomial for the case of two possible outcomes, meaning that a read can be classified as originating from a species or not. At the limit – in the case of high sequencing depth – the binomial distribution becomes a continuous Poisson distribution. Unlike the normal distribution, the mentioned parameterizations do not allow negative values of the random variable. On the other hand, the methods

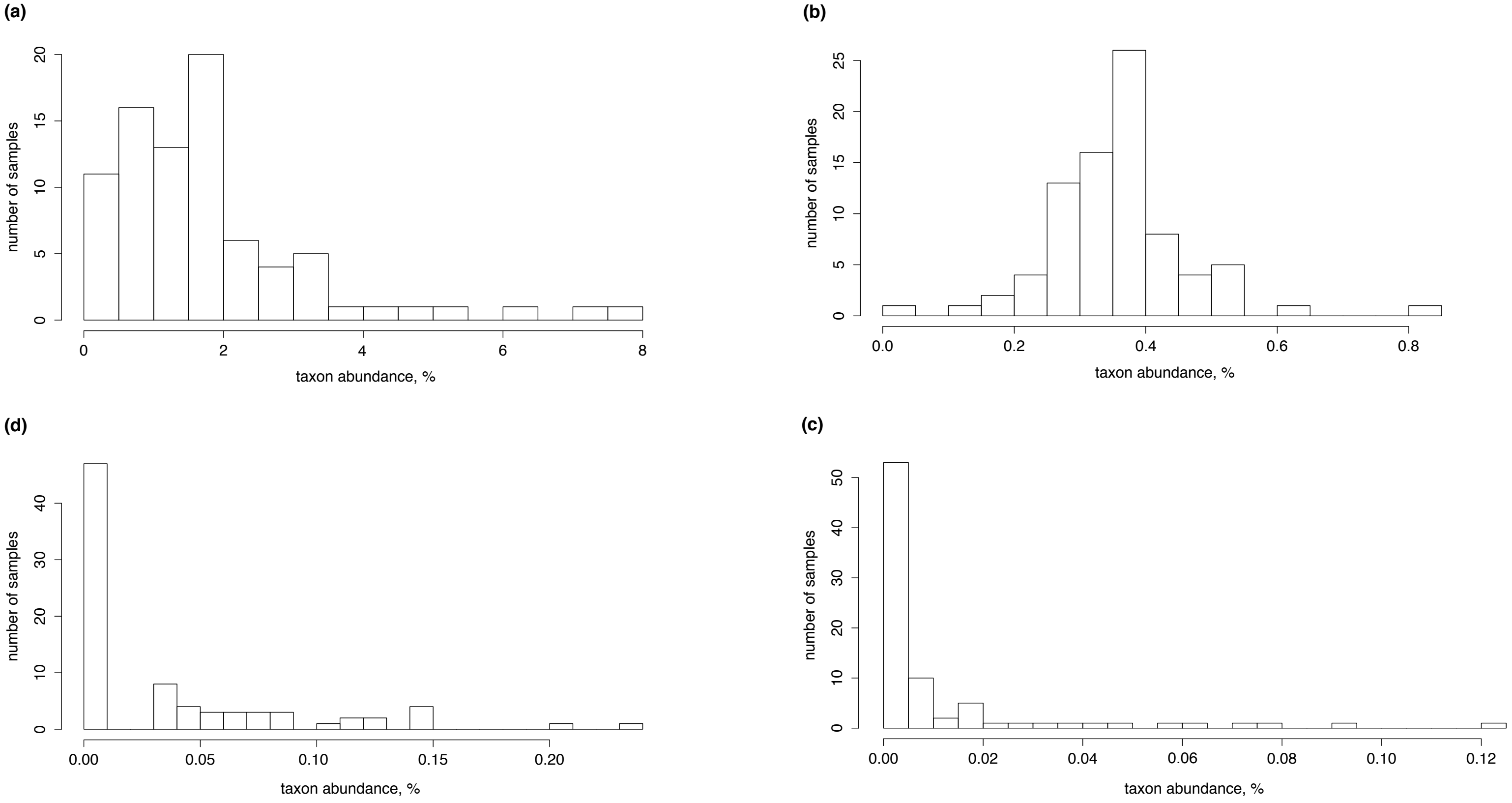

Figure 2.3 Typical form of a histogram for the distribution of a single OTU abundance across samples. (a) For the data normalized to 100% for one of the major OTUs, (b) for the data log-transformed and then normalized to 100% for one of the major OTUs, (c) for the data normalized to 100% for one of the minor OTUs, and (d) for the data log-transformed and then normalized to 100% for one of the minor OTUs. The plots are constructed using 82 gut metagenomic datasets of healthy individuals (Egshatyan *et al*., 2016).

based on binomial, multinomial and Poisson distributions have recently been shown to produce a greater number of false discoveries than normal-based and others (Jonsson *et al.*, 2016). Apparently, the effect is caused by a low capability to fit the variance to the data independently of the mathematical expectation, as these distributions contain only one parameter to fit the data. An alternative is to use mixture distributions, such as univariate negative binomial or multivariate Dirichlet-multinomial distributions, that account for the overdispersion (McMurdie *et al.*, 2014).

Metagenomic sequencing is a costly experimental approach so commonly there are not so many samples to estimate the parameters of a model with high precision. For this reason, Bayesian approach is widely applied in metagenomic statistics: additionally to the available data, certain a priori knowledge about the distribution is used. If the parameters of the a priori distribution are estimated according to the data, the approach is called an empirical Bayesian one.

The above-described models are one-dimensional: they are suitable for the description of each component (species) of a feature vector describing the composition of a microbial community. However, certain models describe the whole feature vector as one random variable. Metagenomic analysis based on such models does not allow us to test if the relative abundance of a specific taxon is associated with certain factors but rather if the whole community structure is associated with them. This approach may be useful to analyse the beta diversity between microbial communities.

R packages for statistical analysis of metagenomic compositional data

R programming language is widely used for omics-data analysis due to the large number of free packages. Here we will describe R packages specifically intended for metagenomic analysis: basic methods commonly used for a comparison of two or more groups [on the example of their implementation in *ALDEx2* package (Fernandes *et al.*, 2014)] and advanced approaches based on generalized linear models allowing both continuous and discrete factors [*metagenomeSeq* (Paulson *et al.*, 2013), *edgeR* (McCarthy *et al.*, 2012), *DESeq2* (Love *et al.*, 2014), *MaAsLin*, *shotgunFunctionalizeR* (Kristiansson *et al.*, 2009)]. Finally, the methods for vector-wise rather than component-wise comparison will be introduced [*HMP* (La Rosa *et al.*, 2012), *vegan* (Oksanen *et al.*, 2012), *micropower* (Kelly *et al.*, 2015)]. The summary of all mentioned packages is presented in Table 2.2.

Component-wise analysis

Comparison of two groups is one of the most common tasks in microbiota analysis. The comparison can be paired or unpaired. The paired comparison is applied when each sample from one group corresponds to a single sample from the other group of equal size (e.g. samples from two body parts of the same individual or from the same person before and after medication, microbiota of mothers and their children). The unpaired comparison is applied for two independent groups, such as healthy subjects versus patients with certain disease or vegetarians versus omnivores; the groups may have different size.

Metagenomic profiling provides an extensive portrait of the community displaying global species-level composition. For simplicity, interactions between the individual species is often ignored and the abundance is compared individually for each of the hundreds of

Table 2.2 Characteristics of the methods for statistical analysis of metagenomic data discussed in the article

Package	Method	Type of comparison	Correction for varying sequencing coverage	Approach to differences in dispersions	Parameterization	Notes	Power and required sample size estimation	Two-group comparison (one discrete factor with two allowed values)	Multiple-group comparison (one discrete factor with >2 allowed values)	Many continuous or discrete factors
stats	Wilcoxon, Kruskal–Wallis, Welch and Student's tests	Component-wise	Should be controlled manually (rarefaction or normalization)	Variance stabilizing transformation is recommended for all tests except for the Welch test designed for unequal dispersions	Wilcoxon and Kruskal–Wallis test are non-parametric, Welch and Student's test suggest normal distribution	Variance stabilizing data transformation is recommended for non-parametric methods and Student's test	n.ttest() from *samplesize* package for Welch and Student's tests; https://fedematt.shinyapps.io/shinyMB/ and [22,23] for unpaired Wilcoxon test	+	–	–
ALDEx2	Wilcoxon, Welch and Kruskal–Wallis tests, ANOVA	Component-wise	Each sample is substituted with Monte Carlo samples from Dirichlet distribution	Monte Carlo sampling provides more accurate estimation of the dispersion; variance stabilizing clr-transformation, Welch test is valid for data with non-uniform variance	Wilcoxon and Kruskal–Wallis test are non-parametric, Welch test and ANOVA suggest normal distribution	Allows accurate estimate of the significance for small sample sizes and correct comparison of abundance vectors after neglecting low-abundance taxa	*samplesize* and *pwr* packages for Welch and Student's tests and ANOVA; https://fedematt.shinyapps.io/shinyMB/ and [22,23] for unpaired Wilcoxon test	+	+	+
metagenomeSeq	Generalized linear model	Component-wise	Percentile normalization to avoid biases caused by the preferable amplification of certain nucleotide sequences	Variance stabilizing logarithmic transformation, used statistics is valid for data with non-uniform variance	Zero-inflated Gaussian distribution	Avoids biases caused by the preferable amplification of certain nucleotide sequences; designed for sparse data inherent for 16S rRNA sequencing; needs sufficient sample size and sequencing depth	No	+	+	+

edgeR	Generalized linear model	Component-wise	Normalization: normalizing factors are selected to minimize the log-fold changes for the majority of taxa	Empirical Bayesian approach to dispersion estimation, used statistics is valid for data with non-uniform variance	Negative binomial distribution	Some results suggest the method yields too many false positives caused by too low estimate of the dispersion	No	+	+	+
DESeq2	Generalized linear model	Component-wise	Normalization: normalization factors take into account the sequencing depth	Empirical Bayesian approach to dispersion estimation, used statistics is valid for data with non-uniform variance	Negative binomial distribution	Provides an empirical Bayesian approach to effect size estimation	No	+	+	+
HMP	Generalized Wald-type statistics	Community-level	Should be rarefied manually	Dirichlet-multinomial distribution, used statistics is valid for data with non-uniform variance	Dirichlet-multinomial distribution	Takes into account the compositional nature of metagenomic data	Is implemented in the package; the calculations of the required sample size may be performed by the wrapper https://fedematt.shinyapps.io/shinyMB/	+	+	−
vegan	PERMANOVA, ANOSIM	Community-level	Should be rarefied manually in the case of weighted metrics	The method is not designed for unequal variances. PERMANOVA is more robust than ANOSIM and some other non-parametric methods	Non-parametric	Takes into account the compositional nature of metagenomic data	*micropower* package (for Jaccard and UniFrac metrics)	+	+	−

taxa resulting in hundreds of statistical tests. Therefore, multiple testing correction is necessary for adjusting the resulting *P*-values. However, if the study initially aims to examine only a single species of interest from the totality of community members and the other species are not analysed further, the procedure is not required. The power of the comparison should be calculated according to the adjusted significance level.

The common solution for the comparison of the two groups is the Student's *t*-test for samples from normal distributions with equal variances. As it was mentioned above, this method cannot be directly applied to metagenomic data, because of the weak conformance to normal distribution and heterogeneous variance. The latter problem may be overcome with the use of variance-stabilizing transformation of the feature vector or Welch's test – a generalization of the Student's test for the case of unequal variances. These approaches provide acceptable results for the high-abundance taxa (Jonsson *et al.*, 2016). The Student's and Welch's tests for paired and unpaired comparisons are implemented in a standard R package *stats*. The power or needed sample size for such comparison without multiple testing correction can be calculated using *samplesize* package.

An approach that overcomes the non-normality of the data distribution is the nonparametric Wilcoxon test. It also has modifications for paired and unpaired comparisons. No assumptions about the type of distribution are used for them. This method is also implemented in *stats* package. The power calculations for Wilcoxon test in case of single-feature abundance comparison is described in detail elsewhere (Mattiello *et al.*, 2016; Rahardja *et al.*, 2009). An online resource for calculating the required sample size for multiple features is available at https://fedematt.shinyapps.io/shinyMB/(Mattiello *et al.*, 2016).

The package *ALDEx2* combines a variance-stabilizing transformation, Welch's and Wilcoxon tests and an instrument for obtaining a more correct *P*-value estimate for small sample size. In order to compare metagenomes with different sequencing depth, each feature vector is associated with a Dirichlet distribution with mean value equal to the normalized feature vector. A random vector from this distribution has nonnegative components summing to 1. The probability density of Dirichlet distribution is:

$$f(x_1,\ldots,x_n;a_1,\ldots,a_n)=\frac{1}{B(a)}\prod_{i=1}^{n}x_i^{a_i-1},$$

where $x=(x_1,\ldots,x_n)$ is a random vector, $a=(a_1,\ldots,a_n)$ is a feature vector (with an additional pseudocount of 0.5 added to each component) and $B(a)$ is the multivariate beta function. The greater the taxon abundance and the less the whole number of reads for the sample, the greater the variance (Fernandes *et al.*, 2014). Substituting the original feature vector with several random vectors generated from the corresponding Dirichlet distribution leads to a more correct estimation of variance and thus of significance of differences. At the next step, centred log-ratio transformation is applied to the obtained random vectors. Besides stabilizing the variance, such transformation ensures the proper comparison of two components of the same vector (e.g. which of two species is higher in abundance within a single community) even if low-abundance taxa are excluded from the study (as it is often done) (Fernandes *et al.*, 2014). The transformed abundances may be compared using either Wilcoxon or Welch's tests. For small groups, the authors of the package recommend to use Welch's test, as its power is less sensitive to the sample size, while for the large groups the results for the two tests are similar.

ALDEx2 contains two methods for multiple group comparisons. They allow us to test the hypothesis stating that a taxon abundance is equal among all groups against the alternative hypothesis stating that for at least one of the groups the taxon abundance is different. The first method is the non-parametric Kruskal–Wallis test which is a generalization of the Wilcoxon test for the case of more than two groups. The second method is the one-way ANOVA (analysis of variance) approach that is based on the comparison of inter-group and intra-group differences. It is a parametric method that suggests that the transformed data are normally distributed. Both ANOVA and Kruskal–Wallis methods require the equality of variances in all groups.

The mentioned methods are appropriate only for the studies with discrete factors (e.g. disease severity index or country of residence). But it is often necessary to identify the associations between the microbial community structure and the factors that take continuous values – for example, body mass index, age, drug dosage. One of the approaches is a nominal division into groups according to some intervals of the factor's values to make them discrete. For example, age groups can be used instead of the age in years. Then the methods used for discrete methods can be implemented.

Another approach that allows both discrete and continuous factor analysis and, moreover, multifactor analysis is based on generalized linear models (GLMs). Essentially, the method suggests that the mathematical expectation of the dependent value (microbial composition components) is a function of linear combination of covariates (factors): $E(y) = g^{-1}(X\beta)$, where $g(y)$ is a so-called link function which should be defined for the model, -1 denotes the inversion of a function, X is a predictor vector, β is a coefficient vector that is estimated from the input data, y is a dependent variable and E is a mean. The dependent value should correspond to an exponential class of distributions (examples include normal, log-normal, Dirichlet, Poisson, binomial or negative binomial). In the case of the identity link function and normally distributed random variable, the GLM degenerates into a standard linear model. Paired comparison can be performed by including N additional binary factors, where N is the number of pairs, with each of the factors reflecting if a metagenome belongs to a certain pair. In the case of a GLM, each of the null hypotheses states that the coefficient preceding the factor equals zero. It is worth noting that such a model is useful when the underlying association between the factor and the feature abundance (transformed by link function) is linear indeed. Otherwise, this approach is inappropriate and may lead to biased conclusions.

Common function for fitting generalized linear model in R is *glm2* (Marschner, 2014). Adaptation of such a model to metagenomic data is implemented in *metagenomeSeq*, *MaAsLin* and *shotgunFunctionalizeR* packages. The packages *edgeR* and *DESeq2* for differential gene expression analysis based on RNA-seq data also contain implementations of GLM. These are widely used in microbiome studies due to the similarity of data format and statistical properties between RNA-Seq and metagenomics: hundreds to thousands of discrete features with distributions varying within several orders of magnitude.

In the *edgeR* package, the dependent variable – relative abundance of a taxon – is described by a negative binomial distribution. Between-group comparison is performed using an exact test based on this distribution model (in a way similar to a standard t-test). The variance is estimated using empirical Bayesian approach: the estimate obtained from the data for individual taxon is shrunk to the value assessed across all taxa. The degree of shrinkage depends on the mean value. The *edgeR* package includes the correction for sequencing depth that minimizes the log-fold changes for the majority of the taxa. Several methods for determining the GLM coefficients β and their significance are available.

The approach implemented in *DESeq2* package is similar with a few differences. Besides alternative normalization and variance estimation methods, *DESeq2* uses an empirical Bayesian approach to obtain the effect size. The variance estimate is shrunk towards zero – the less abundant the taxon, the stronger the shrinking. The shrinkage also depends on the feature variance. This approach helps to avoid a situation when the majority of the features differentially abundant in the two groups have low abundance. The methods for estimating the significance also differ from the ones in *edgeR*. *DESeq2* tends to make less false positives then *edgeR*. However, unlike the below-described *metagenomeSeq* package, *DESeq2* works slowly on large datasets containing 100 or more samples per group – a typical scenario for a metagenomic study (Weiss *et al.*, 2015).

The *metagenomeSeq* package based on GLMs was developed specifically for 16S rRNA metagenomic data that were found to be more sparse than the gene expression data. To provide the correct comparison of the metagenomes with different library sizes, the taxa abundances are normalized by certain percentile determined from the given data. This method allows us to resolve the problem of varying OTU-specific PCR amplification efficiency, a known technical artefact. The variance-stabilizing logarithmic transformation is applied to the normalized data. The transformed feature abundances are supposed to follow a zero-inflated Gaussian distribution, which takes into account the dependence of the set of the taxa detected in a sample on the sequencing depth and adjusts data for its sparsity. After this correction, the data distribution is closer to the normal type according to Shapiro–Wilk test (Paulson *et al.*, 2013). The empirical Bayesian approach implemented in the *limma* (Ritchie *et al.*, 2015) package is used to test the null hypothesis that the linear model coefficient equals zero and to estimate the significance. The test is based on moderated t-statistics test and involves shrunk variance estimate in a way similar to edgeR (Paulson *et al.*, 2013; Ritchie *et al.*, 2015).

Benchmarking of various methods on the example of two-groups comparison for both simulated and real-world metagenomic datasets showed that *metagenomeSeq* performed better in the terms of AUC (area under curve), especially for the middle- and high-abundance taxa (Button *et al.*, 2013; Jonsson *et al.*, 2016). However, the package underestimates FDR value, especially for the datasets with low sequencing coverage and sample size, and tends to make more false discoveries than other packages (Jonsson *et al.*, 2016; McMurdie *et al.*, 2014; Weiss *et al.*, 2015). Thus, taking into account its high speed as compared to other methods (Weiss *et al.*, 2015), *metagenomeSeq* is recommended for the analysis of 16S rRNA sequencing data providing both high group size and sequencing depth (Jonsson *et al.*, 2016; McMurdie *et al.*, 2014; Paulson *et al.*, 2013; Weiss *et al.*, 2015).

Generalized linear models for metagenomic data are also implemented in the packages *MaAsLin* and *shotgunFunctionalizeR*. While only limited information is published on the details, the overdispersed Poisson generalized linear model realized in *shotgunFunctionalizeR* showed good results on 'shotgun' metagenomes (Jonsson *et al.*, 2016). As for *MaAsLin*, its advantage is a boosting process that rates the factors in the order of contribution to the observed differences in microbiota composition – which is useful in the cases when the number of the factors is high and it is difficult for a researcher to infer the importance of each factor.

Community-level comparison of microbial communities

The comparison of microbial communities can be conducted not only in the component-wise manner, but also viewing the community as a whole, taking into account the possible interdependence of the relative abundance levels between various bacterial species in microbiota. It is reasonable, in the view of many studies pointing out elaborate ecological relations between the species within microbiota (Levy and Borenstein, 2013; Stein *et al.*, 2013). One of such common methods in biostatistics is MANOVA (multivariate analysis of variances), a generalization of the ANOVA method for the multivariate case – but its application to metagenomic data is limited because it requires the normal distribution of taxa abundance levels. A researcher can resort to non-parametric modifications of this approach such as ANOSIM [analysis of similarity, function *anosim* in package *vegan* (Oksanen *et al.*, 2012)] and PERMANOVA (permutational multivariate analysis of variances, function *adonis* in the same package). They are also based on the comparison of within-group and between-group variances. In the case of metagenomics, the researcher may choose specific dissimilarity measures like weighted and unweighted Jaccard distance or UniFrac. The unweighted metrics are less sensitive to differences in sequencing depth between the samples, since they are based only on the presence/absence of a taxon in a sample rather than its abundance levels. The methods differ by their approach to the variance comparison: ANOSIM compares ranks of variances similar to Kruskal–Wallis test, while PERMANOVA compares variance values by estimating the significance via a permutation method. A disadvantage of these methods is that they require equal within-group variances – however, usually this is not the case for metagenomic data (Warton *et al.*, 2012). PERMANOVA was shown to be more robust to the failure of this restriction than ANOSIM (Anderson and Walsh, 2013). R package *micropower* provides functions for statistical power calculations for PERMANOVA based on weighted and unweighted Jaccard distance as well as UniFrac.

A parametric approach to the problem of multivariate group comparison is implemented in package *HMP*. It employs a generalized Wald-type statistics to compare the estimates of statistical model parameters. The Dirichlet–multinomial distribution is used to model the vector abundance across each group. It is a combination of multinomial and Dirichlet distributions. Similarly to the binomial and Poisson distributions, the multinomial one does not provide an instrument for independent estimation of both mean and variance. The Dirichlet-multinomial distribution is able to avoid this flaw, as it provides an overdispersion parameter that can be derived independently of the data; in the case of the zero overdispersion it coincides with the multinomial distribution. It is shown in a paper by La Rosa *et al.* (2012) that this approach is indeed better to describe the metagenomic data. The parameters of the model are estimated with the method of moments or the maximum likelihood method. There is no instrument for correct comparison of samples with different sequencing depth in *HMP*, so one should be careful with the data preprocessing. Package *HMP* allows power calculations. The wrapper for the sample size definition is available at https://fede.shinyapps.io/shinyMB/ (Mattiello *et al.*, 2016).

Conclusions

It is important to emphasize that statistical analysis should be thought of at the very beginning of a metagenomic study, before the sample collection and sequencing procedures.

Proper balance between the number of the samples and sequencing depth will lead to high statistical power and subsequently the results of the study will be more valuable to the scientific community. Preliminary *in silico* experiments with the published metagenomes in similar format and microbiota type as well as on simulated data contribute to the success of the analysis.

The choice of a package for a specific problem in a metagenomic survey depends on several conditions: paired or unpaired design, continuity of factors values, component-wise or vector comparison, the need for power control, suggested sample size and sequencing depth. First of all, it is needed to formulate the aim of comparison – is the researcher interested in the component-wise analysis or in beta-diversity among groups? The packages *ALDEx2, metagenomeSeq, edgeR, DESeq2, MaAsLin* and overdispersed model in *shotgunFunctionalizeR* are designed for the former task, while *HMP* and PERMANOVA coupled with *micropower* package may be used for the latter. For the case of continuous factors or multifactor analysis, the models based on generalized linear model are recommended (*metagenomeSeq, edgeR, DESeq2, MaAsLin* and overdispersed model in *shotgunFunctionalizeR*). In the case of low number of samples, the *DESeq2* package is recommended, while for larger sample sizes and deeper sequencing *metagenomeSeq* is preferable due to higher performance (Weiss *et al.*, 2015). The *ALDEx2* package is suitable for multiple group comparison for the small sized groups, as it is more accurate in significance estimates. The existing evidence suggests that methods based on binomial, multinomial and Poisson distributions are not appropriate for metagenomic statistical evaluation due to a great number of false discoveries. Overall, a researcher should perform an exploratory analysis to check if the distributions for the particularly analysed dataset conform to the parameterization used by the package of choice, as it greatly influences the accuracy of the results. If the quality of model fit is low, the non-parametric methods, such as Wilcoxon test, Kruskal–Wallis test and PERMANOVA should be used.

Acknowledgements

We thank Pavel Mazin for discussions of the text. This work was financially supported by the Ministry of Education and Science of Russian Federation (agreement # RFMEFI60414X0119).

References

Anderson, M.J., and Walsh, D.C.I. (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? Ecol. Monogr. *83*, 557–574. http://dx.doi.org/10.1890/12-2010.1

Baker, M. (2016). Statisticians issue warning over misuse of P values. Nature *531*, 151. http://dx.doi.org/10.1038/nature.2016.19503

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., and Munafò, M.R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. *14*, 365–376. http://dx.doi.org/10.1038/nrn3475

Egshatyan, L., Kashtanova, D., Popenko, A., Tkacheva, O., Tyakht, A., Alexeev, D., Karamnova, N., Kostryukova, E., Babenko, V., Vakhitova, M., *et al.* (2016). Gut microbiota and diet in patients with different glucose tolerance. Endocr. Connect. *5*, 1–9. http://dx.doi.org/10.1530/EC-15-0094

Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., and Gloor, G.B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome *2*, 15. http://dx.doi.org/10.1186/2049-2618-2-15

Goodrich, J.K., Di Rienzi, S.C., Poole, A.C., Koren, O., Walters, W.A., Caporaso, J.G., Knight, R., and Ley, R.E. (2014). Conducting a microbiome study. Cell *158*, 250–262. http://dx.doi.org/10.1016/j.cell.2014.06.037

Hair, J.F., Black, W.C., Babin, B.J., and Anderson, R.E. (2010). Multivariate Data Analysis (Pearson Prentice-Hall, Inc).

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214. http://dx.doi.org/10.1038/nature11234

Jonsson, V., Österlund, T., Nerman, O., and Kristiansson, E. (2016). Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics *17*, 78. http://dx.doi.org/10.1186/s12864-016-2386-y

Kelly, B.J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J.D., Collman, R.G., Bushman, F.D., and Li, H. (2015). Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics *31*, 2461–2468. http://dx.doi.org/10.1093/bioinformatics/btv183

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics *25*, 2737–2738. http://dx.doi.org/10.1093/bioinformatics/btp508

Levy, R., and Borenstein, E. (2013). Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proc. Natl. Acad. Sci. U.S.A. *110*, 12804–12809. http://dx.doi.org/10.1073/pnas.1300926110

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. http://dx.doi.org/10.1186/s13059-014-0550-8

Marschner, I. (2014). glm2: Fitting Generalized Linear Models. R package version 1.1.2. http://CRAN.R-project.org/package=glm2

Mattiello, F., Verbist, B., Faust, K., Raes, J., Shannon, W.D., Bijnens, L., and Thas, O. (2016). A web application for sample size and power calculation in case-control microbiome studies. Bioinformatics *32*, 2038–2040. http://dx.doi.org/10.1093/bioinformatics/btw099

McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. *40*, 4288–4297. http://dx.doi.org/10.1093/nar/gks042

McMurdie, P.J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. PLOS Comput. Biol. *10*, e1003531. http://dx.doi.org/10.1371/journal.pcbi.1003531

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara R.B. *et al.* (2012), vegan: Community Ecology Package. R package version 2.0-5. http://CRAN.R-project.org/package = vegan

Paulson, J.N., Stine, O.C., Bravo, H.C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. Nat. Methods *10*, 1200–1202. http://dx.doi.org/10.1038/nmeth.2658

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rahardja, D., Zhao, Y.D., and Qu, Y. (2009). Sample Size Determinations for the Wilcoxon–Mann–Whitney Test: A Comprehensive Review. Stat. Biopharm. Res. *1*, 317–322. https://dx.doi.org/10.1198/sbr.2009.0016

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. http://dx.doi.org/10.1093/nar/gkv007

La Rosa, P.S., Brooks, J.P., Deych, E., Boone, E.L., Edwards, D.J., Wang, Q., Sodergren, E., Weinstock, G., and Shannon, W.D. (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLOS ONE *7*, e52078. http://dx.doi.org/10.1371/journal.pone.0052078

La Rosa, P.S., Zhou, Y., Sodergren, E., Weinstock, G., and Shannon, W.D. (2015). Hypothesis Testing of Metagenomic Data. In Metagenomics for Microbiology, J. Izard and M.C. Rivera, ed. (Academic Press), pp. 81-96.

Sham, P.C., and Purcell, S.M. (2014). Statistical power and significance testing in large-scale genetic studies. Nat. Rev. Genet. *15*, 335–346. http://dx.doi.org/10.1038/nrg3706

Stein, R.R., Bucci, V., Toussaint, N.C., Buffie, C.G., Rätsch, G., Pamer, E.G., Sander, C., and Xavier, J.B. (2013). Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. PLOS Comput. Biol. *9*, e1003388. http://dx.doi.org/10.1371/journal.pcbi.1003388

Tyakht, A.V., Kostryukova, E.S., Popenko, A.S., Belenikin, M.S., Pavlenko, A.V., Larin, A.K., Karpova, I.Y., Selezneva, O.V., Semashko, T.A., Ospanova, E.A., *et al.* (2013). Human gut microbiota community structures in urban and rural populations in Russia. Nat. Commun. *4*, 2469. http://dx.doi.org/10.1038/ncomms3469

Wang, F., Kaplan, J.L., Gold, B.D., Bhasin, M.K., Ward, N.L., Kellermayer, R., Kirschner, B.S., Heyman, M.B., Dowd, S.E., Cox, S.B., *et al.* (2016). Detecting Microbial Dysbiosis Associated with Pediatric Crohn Disease Despite the High Variability of the Gut Microbiota. Cell. Rep. *14*, 945–955. http://dx.doi.org/10.1016/j.celrep.2015.12.088

Warton, D.I., Wright, S.T., and Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. Methods Ecol. Evol. *3*, 89–101. http://dx.doi.org/10.1111/j.2041-210X.2011.00127.x

Wasserstein, R.L., and Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. Am. Stat. 70, *2*, 129-133. http://dx.doi.org/10.1080/00031305.2016.1154108

Weiss, S.J., Xu, Z., Amir, A., Peddada, S., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vazquez-Baeza, Y., Birmingham, A., *et al.* (2015). Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. PeerJ *230313.* http://dx.doi.org/10.7287/peerj.preprints.1157v1