UNIVERSITEIT GENT

Faculty of Sciences

Evaluation of normalization and analysis methods for microbiome data

Stijn Hawinkel

Master dissertation submitted to obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Ir. Olivier Thas
Department of Mathematical Modelling,
Statistics and Bioinformatics

Academic year 2014-2015

Faculty of Sciences

Evaluation of normalization and analysis methods for microbiome data

Stijn Hawinkel

Master dissertation submitted to obtain the degree of
Master of Statistical Data Analysis

Promoter: Prof. Dr. Ir. Olivier Thas
Department of Mathematical Modelling,
Statistics and Bioinformatics

Academic year 2014-2015

Foreword

Acknowledgements: Professor Olivier Thas is the first and foremost person I want to thank, for tutoring me through this thesis project with help and advice and for bringing me into touch with his rich network of fellow scientists. I'm looking forward to continue working with him the following years. Also a big thanks to Bie Verbist, Luc Bijnens and all other people from Janssen Pharmaceutics to invite me to the workshops and teleconferences and giving me the opportunity to continue working on microbiomics as a PhD-student. I'm also grateful to Marcus Rauch and Jeroen Raes for sharing their insights in microbiomics during the workshops, and to Karoline Faust, Marcus Rauch, Brindha Lekshmisaran and Nabeetha Nagalingam for kindly sharing their data. And finally also a whole-hearted "thank you" to my father Chris for reading my thesis so carefully and providing me with comments and corrections.

Data sources: The 16S rRNA sequencing count data of the Human Microbiome project (HMP) project were shared the by VIB lab for Bioinformatics and (eco-)systems biology led by Prof. Jeroen Raes. The sequence count data mock community were obtained from Marcus Rauch's team at the Janssen Prevention Center. The code for the simulations was based on the code found in the supplementary materials of the paper 'Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible' by McMurdie and Holmes, 2014.

Own contribution: My own work consisted in exploring the HMP dataset and testing and estimating overdispersion of the Dirichlet multinomial. Further I adapted the code by McMurdie and Holmes to implement Dirichlet multinomial sampling, t-test and Wilcoxon rank sum test, to accomodate the combination of different normalization and analysis methods and I added the upperquartile normalization method.

List of Tables

List of Figures

Contents

1 Summary

The healthy human body is inhabited by billions of bacteria, viruses and fungi on all of its outer and inner surfaces, such as oral cavity, skin and gut. Thereby each body site has its own unique community of micro-organisms adapted to its environmental conditions. The ensemble of these communities of non-human beings living on our bodies is called the human microbiome. Under normal circumstances these organisms are inoffensive and even useful since they contribute to food digestion and maturation of the immune system, among other things. On the other hand, perturbations of the normal composition of the microbiome are often observed together with diseases such as gut inflammation and diabetes. Moreover, transplantation of gut micro-organisms from healthy individuals has been shown to cure cases of accute diarrhea. Together with the fact that each person has a unique microbiome composition, these examples show us the potential of microbiome science for the future of (personalized) medicine.

To investigate the composition of the microbiome one needs to count its micro-organisms in some way. This can be accomplished by counting the number of times a marker molecule, unique for each species, is present in a sample. DNA is a molecule that stores the genetic information of a cell, and its sequence is a unique fingerprint of an organism. Reading off the sequence of a small part of a cell's DNA is sufficient to know from which species it has originated. By reading the same part of all DNA molecules from all micro-organisms in a sample and counting the number of reads, researchers can get an estimate of the abundance of the different species in the microbiome.

Microbiologists often want to monitor changes in the composition of the microbiome as a result of e.g. disease development or drug treatment. To be able to recognize these changes, it is important to know how the microbiome normally varies under undisturbed conditions. At this point microbiome statistics come into play to distinguish systematic changes in the composition of the microbiome from random fluctuations. In this thesis we focus on two main steps in this process: normalization and analysis of abundance. Normalization is a pre-processing step that aims to eliminate technical effects of different samples of the microbiome in order to render them comparable. After all the scientist is interested in biological differences between samples, and not in artefacts due to technical differences. Subsequently, analysis methods are needed to determine which micro-organisms are present in lesser or greater numbers in one sample versus another, e.g. as a result of a drug treatment.

Several normalization and analysis methods have been proposed but it is still under debate which are the best ones. One of the main ways to compare the performance of these methods is through simulation. Simulation means the creation of artificial datasets with some degree of randomness, but still based on known parameters. The structure of these created datasets must resemble as closely as possible that from real microbiome

data. The researcher then changes the abundance of some species of micro-organisms in a subset of the generated samples. The normalization and analysis methods are then applied to these datasets and compared based on how well they manage to detect these changes. The advantage in comparison with real dataset is that for simulated datasets the researcher *knows* which species' counts have been changed, which allows a detailed evaluation of the methods used.

In this thesis a dataset with DNA-counts of the microbiome of 16-19 body sites of 242 healthy individuals is described. Next, simulations are performed based on this dataset to evaluate a number of normalization and analysis methods. Finally the methods that performed best in the simulation were applied to two body sites from the original dataset.

2 Introduction

2.1 The human microbiome as forgotten organ

We, human beings, are not living on our own. Although sterile during gestation, the human body gets colonized by a wide variety of bacteria, archaea, fungi and viruses after birth. They live, grow and reproduce on inner and outer surfaces of the human body such as skin, oral cavity, airways, gut and vagina. This community is called the human microbiome and contains about a ten-fold as many cells as the human body has of its own, eukaryotic cells, and several orders of magnitude more genes [1]. Each colonized body site has its own typical community composition that is preserved between individuals. Any community from one of those body sites is usually dominated by a few signature taxa well adapted to this niche, with a large number of other taxa present at much lower frequencies (we say that the communities are skewed to rare taxa) [2, 3]. Still, the interpersonal diversity in microbiomes is huge, even between healthy individuals, and remains largely unexplained [1]. Under normal circumstances, these communities are inoffensive and even useful for digesting food [2], providing resistance to infection [4] and stimulating the maturation of the immune system [5, 6] and anatomic development [4]. Because of these crucial functions, the microbiome has been coined the "forgotten organ" [7]. On the other hand, (locally) disturbed microbiome composition and structure are known to be associated with a broad range of disease statuses such as gut inflammation [4], vaginosis [8], diabetes [9] and periodontal disease [10]. Even when the causal relationship remains unclear, this reveals a tremendous potential to use the taxonomic composition of the microbiome as prognostic or diagnostic biomarker [1]. Disturbances of the microbiome can occur very quickly, within a few hours, revealing that the microbiome is a plastic and adaptive entity [11]. This observation opens opportunities for active interventions in the microbiome that may cause drastic changes in health status. Inoculation of germ-free mice with gut microflora from obese humans causes a greater increase in total body fat than inoculation with flora from lean humans, suggesting that the gut microbiome may be a key to fighting obesity [12]. Patients suffering from chronic diarrhea caused by *Clostridium difficile* can very often be cured by a fecal transplant from a healthy person which reestablishes the normal gut flora [13]. These examples, together with the great interpersonal variability of microbiomes, illustrate how the science of microbiomics may contribute largely to the advancement of personal medicine [1].

2.1.1 Characterization of the microbiome

Historically, members of microbial communities were identified by physiological characteristics and hence identification depended on culturing of whole colonies of microbes. This severely limited the scope of microbial ecology studies since only a small minority of micro-organisms can be grown under lab conditions. In the 1980s methods based on nucleic acids emerged that allowed assesment of both taxonomic and metabolic diversity of communities without the need for prior culturing. Fluorsecent in situ hybridization (FISH) allowed studying uncultured communities without prior DNA extraction, by hybridizing a fluorescent probe to DNA present in the community [1]. From the 1990s, hybridization of extracted DNA or cDNA (DNA derived from extracted RNA) on microarrays has enjoyed widespread popularity. This technique consists of bringing the DNA or cDNA of a whole community into contact with an array onto which probes of complementary DNA have been spotted on a known location. After some hybridization period, the remaining, unbound DNA is washed away and the hybridization is visualized, usually through fluorescence. This yields a continuous signal for every spot, proportional to the amount of DNA each particular sequence present in the sample [14]. The microarray technology suffers from unspecific hybridization and is limited to interrogating presence of known DNA fragments for which complementary probes are available, excluding the discovery of new elements [15]. Thanks to the emergence of high-throughput DNA-sequencing technology, quantification of DNA composition of a population (or the RNA content of a cell) is nowadays done by direct sequencing. If the research goal is to gain insight in the different metabolic functions a bacterial community as a whole can perform, the entire pool of present DNA is sequenced. This branch of microbiomics is called functional metagenomics. Contrarily, if the research goal is just to quantify the composition of the community in terms of taxa, amplifying and sequencing a highly discriminative single marker gene can be sufficient. The most popular marker is by far the 16S rRNA gene [1], as discussed in the next section.

2.1.2 The 16S rRNA molecule records evolutionary distances as a molecular clock

The 16S rRNA gene is present among almost all bacteria and has a conserved function among them. It consists both of regions conserved among species as well as highly variable regions, making it an excellent marker gene for species determination. The conserved regions assure correct folding and thus functioning of the RNA molecule and are under strong negative selective pressure as a result. These regions can be used as primer annealing sites for PCR-amplification of part of the gene. The 16S rRNA gene also contains 9 variable regions, in which random mutations can occur without being eliminated by selective pressure. These regions serve as a molecular clock, since these random mutations are assumed to appear at a constant pace over time. This way these mutations keep track of evolutionary distances between taxa. Few differences between two 16S rRNA genes indicate that the two bacteria diverged comparatively recently, since little time has passed to acquire different mutations. Based on accordance of the variable regions, bacteria can be grouped into species and other taxonomic levels. In practice, the term operational taxonomic unit (OTU) is used for bacteria instead of species, since no clear-cut definition of a bacterial species is available. By convention, sequencing data of the variable regions that show at least 97% identity are grouped into the same OTU [1,16]. These OTUs can then be mapped to a reference genome database to see from which bacterial taxon the reads originate. Off course, some OTUs will not be found in these databases since most microbiome samples contain previously unknown species. In this case they are assigned to the lowest taxonomic division as possible. Counting the number of reads per taxon yields a measure of abundance of this taxon. These counts are represented in count matrices as discussed in the next section.

Before starting the discussion of microbiome marker gene count data and its analysis, it is important to note that most of these methods have been derived from techniques developed for RNA-Seq or (to a lesser extent) microarray. In RNA-Seq, the aim is to quantify the gene expression levels of a cell by sequencing cDNA, which is DNA derived from present RNA. It is obvious that, despite its different scientific aim, the statistical techniques applied to both types of data are analogous, since both are in essence read count data mapped to a known database. Often even the same sequencing machines are used for both ends. Important differences with RNA-Seq are the sparsity of the data matrix and the fact that all reads have the same length and GC-composition. Marker gene count data matrices are sparse because most OTUs are found only in a minority of the samples, resulting in zero counts for the other samples. On the contrary, most genes have some degree of basic expression so that their read counts are less often zero. For marker gene assays it is always the same gene that is sequenced, with only minor differences between the different OTUs. However, the genes sequenced in RNA-Seq represent the whole genome and may differ in read length and GC-content, which may cause amplification bias [1,17].

2.2 Microbiome marker gene count data and analysis

2.2.1 Data structure

To facilitate overview and inference on read count data, they are usually tabulated as follows. For each sample, a vector of read counts is constructed with length n, equal to the total number of different OTUs found in all samples combined. For sample j, this vector $\boldsymbol{C}_j = (c_{1j}, c_{2j}, ..., c_{nj})$ has as elements c_{ij} representing the raw read count for taxon i in sample j with as total sum (called sequencing depth or library size) $N_j = \sum_{i=1}^{n} c_{ij}$. This is done for all of the m samples and all these vectors are then combined into an n by m contingency table where the rows represent the taxa and the columns the samples. An example of such a microbiome contingency table is shown in Table 1. Since samples often only contain a fraction of the complete diversity of the group of samples, the matrix is sparse, i.e. most cells are zero [18,19]. One of the main goals of microbiome statistics is to decide which taxa show different (relative) abundances between groups of samples. These groups are known to the investigator, e.g. diseased vs. healthy persons or treatment vs. control groups. Many normalization and analysis methods have been devised for this goal,

	Sample						
Taxon	1	2	...	j	...	m	**Total**
1	c_{11}	c_{12}	...	c_{1j}	...	c_{1m}	$c_{1.}$
2	c_{21}	c_{22}	...	c_{2j}	...	c_{2m}	$c_{2.}$
...	...	...	$\ddots$	...	$\ddots$	...	...
i	c_{i1}	c_{i2}	...	c_{ij}	...	c_{im}	$c_{i.}$
...	...	...	$\ddots$	...	$\ddots$	...	...
n	c_{n1}	c_{n2}	...	c_{nj}	...	c_{nm}	$c_{n.}$
Total	N_1	N_2	...	N_j	...	N_m	$N_.$

Table 1: An example of a contingency table for microbiome data of n taxa and m samples. c_{ij} represents the read count for taxon i in sample j, $c_{i.}$ the total count of taxon i over all samples and N_j the library size of sample j.

which will be discussed in the next two sextions.

2.2.2 Normalization renders read counts of different samples comparable

Read counts from different samples in sequencing assays are mostly not directly comparable due to technical noise. A larger sequencing depth (library size) of one sample may result in a higher read count over all taxa. Directly comparing counts from two samples with different sequencing depth may lead to the conclusion that most taxa are differentially expressed, even when the composition of both samples is the same [20, 21]. In addition, larger samples carry more information and parameters estimated from them have thus lower variances [22]. To account for this, normalization is needed to render the read counts comparable and to eliminate this heteroscedasticity [22–24]. Over the years, many normalization methods have been proposed, and some attempts have been made to compare them [17, 20, 22, 25], but no consensus exists about the optimal method [20]. A normalization procedure which has enjoyed wide popularity but which is unadvisable, is *rarefying*. This method consists of random subsampling of taxa counts from a sample without replacement to the size of the smallest sample to equalize the library sizes. Apart from introducing additional variability through the random sampling step, it inflates the uncertainty associated with the read counts by throwing away data of the larger samples. Rarefying eliminates heteroscedasticity, but only by reducing the information of all samples to the level of the lowest one. As a result McMurdie and Holmes consider rarefying as normalization technique prior to differential abundance analysis “statistically inadmissable” [22]. In most other normalization methods, a distinct scaling, size or normalization factor f_j, reflectinthe true sequencing depth of this sample j, is estimated [20, 22]. Using the library sizes N_j as scaling factors f_j appears an intuitive and logical way to normalize count data. This procedure (also called Total Sum Scaling, TSS) does not suffer from the additional uncertainty of the random subsampling used in rarefying. However, scaling by library sizes still has the drawback of reducing all counts to proportions and thus discarding all information on sample size and related variance of the proportions’ estimators [22].

In addition, the library size is very sensitive to changes in the frequencies of abundant taxa. In fact, the estimated proportions will not only depend on the frequency of the taxon itself but on the abundance of the other taxa as well. An increase (true or coincidental) in read counts for a few numerous taxa causes the proportions of the other taxa to shift downward. When comparing these decreased proportions with those of other samples, the difference may seem statistically significant. This leads to a high number of false postives when testing for different abundance of taxa as well as to a loss of power to detect true differences [20, 23–25]. Several more robust normalization methods have been proposed that restrict the influence of those very abundant taxa, founded on different assumptions. A first strategy relies on the assumption that some quantile(s) of the read counts distribution coincide between all the samples, even though the overall distribution of the read counts differs, especially at the highest read counts [20]. A simple choice is to use the upper-quartile ($75th$ percentile) of the read counts of each sample after removing taxa with zero counts in all rows as its scaling factor f_j . This quantile was chosen because it was considered to be high enough to fall outside the range of the zero and low-counts that are uninformative on the sequencing depth. On the other hand it is still low enough not to be affected by the high and extremely high counts, as is the problem with proportion normalization (TSS) [25]. An adapted form of this idea utilizes a data-driven selection of the appropriate quantile and then sums all the read counts smaller than or equal to this quantile for each sample to calculate the scaling factor. This approach thus assumes a common count distribution across samples up to a certain quantile. The quantile used is the point where the distribution of a sample begins to deviate strongly from the reference distribution of all other samples. This method is called cumulative sum scaling (CSS) normalization [17]. A second approach is to assume that most taxa are equally abundant across samples and differentially abundant samples constitute only a minority [20]. The trimmed mean of M-values (TMM) method is an instance of this way of thought. From all samples m, a reference sample r is chosen. Next for each taxon i of the remaining m-1 samples the M_{ij}-value is calculated as follows (taxa for which N_j or N_r are zero are discarded):

$$M_{ij} = \log_2 \frac{c_{ij}/N_j}{c_{ir}/N_r} \tag{1}$$

These M_{ij} values of a sample j are trimmed (authors propose by 30%) and a weighted mean of the remaining M-values is calculated. The weights are inversely proportional to the asymptotic variance of the M_{ij}s. This weighted mean is then used as the scaling factor f_j. Because of the trimming, this scaling factor is robust with respect to taxa that are very abundant or very rare compared to the reference sample [24]. A very comparable procedure is termed the relative log expression (RLE) method. First the ratios of observed non-zero cell counts are calculated for every taxon i as follows:

$$\frac{c_{ij}}{(\prod_{j=1}^{m} c_{ij})^{1/m}} \tag{2}$$

The denominator, a geometric mean across samples, can be seen as a pseudo-reference sample. Subsequently the median of this measure over all taxa n is used as scaling factor f_j of the sample. In this case the median, which is supposed to lie within the range of the non-differentially abundant taxa, renders the method robust with respect to extreme abundances [23]. It should be noted that TMM and RLE normalization were not intended to transform the data, but rather to calculate the size factors to be used in statistical models [23, 24]. Methodologically different approaches to normalization exist, such as entail spiking-in of known quantities of DNA for comparison [26], but this method will not be discussed here since they require information not present in the count data matrix.

2.2.3 Testing for differential abundance

Once the normalization factors are estimated, one can set out to analyse the read count data. There exist two main methods for comparison of taxa vectors from different samples, one looking at the structure of the population , the other at its distribution. The method that focuses on community structure investigates the species abundance distribution (SAD). This kind of analysis drops all taxon labels and taxa with zero counts, and orders the remaining taxa according to descending abundance for each sample. This represents the SAD that is only depending on the structure of the community and not on its composition. The structure of a community may predict its response to perturbation and its flexibility [3]. The SAD theory was initially developped in the field of ecology and will not be further discussed in this thesis, partly since contemporary microbiologists are really interested in community composition rather than just structure.

The second main method keeps the taxon labels and zero counts and compares their composition through the relative abundance of taxa. These methods are said to test differential abundance of taxa and are the methods of interest of in this thesis. A very simple way to compare the composition two microbiomes is to calculate some measure of population diversity and then compare both measures and test for siginificant differences. Several measures of this so-called alpha-diversity exist, some of them even accounting for taxa that were not observed but may also be present in the population based on the sampling distribution. On the other hand, measures of beta-diversity attempt to summarize differences between samples over many taxa into pairwise distances [1], such as Bray-Curtis [27] or Unifrac [28]. Despite being easy to calculate and convenient to compare, these measures entail a large reduction in information, and exclude comparisons on the taxa level. However, a main goal of microbiome research is to determine more precisely which taxa are present to a greater or lesser extent (i.e. differentially abundant) in some sample compared to another [29]. For this end, parametric methods are preferred since their results are easier to interpret. Parametric methods have the advantage that they can quantify the size of the difference in abundance. Moreover parametric tests were found to have higher power and do not rely on the assumptions of equal variablility between groups [19]. To distinguish systematic differences from random noise in the data by parametric tests, assumptions on sampling distributions of the read counts are needed. One approach is to model the variations in read counts taxon by taxon over the different samples, another to model the composition of the whole community at once. We will discuss both approaches in the following paragraphs.

Taxon-by-taxon modelling It has been shown that the variations in read counts of a taxon between technical replicates (i.e. the same sample analyzed repeatedly) follow a Poisson distribution [15]. The Poisson distribution has only one parameter, its mean is equal to its variance with probability mass function

$$f(c;\gamma) = \frac{\exp(-\gamma)\gamma^c}{c!} \tag{3}$$

where $\mathrm{E}(c_{ij}) = \gamma_{ij}$ can be factorized as $f_j q_i$ with f_j the size factor of sample j and q_i the mean proportion for taxon i in this sample [22]. However, for biological replicates (i.e. different samples originating from the same or a comparable source, e.g. same body site or environmental condition, further referred to as experimental condition $\rho(j)$ of sample j) this distribution no longer holds, because the observed variance is larger than the mean, a phenomenon called overdispersion. The Negative Binomial (NB) distribution is an extension of the Poisson that provides a better fit for biological replicates of sequence count data by allowing for overdispersion [30, 31]. It has two free parameters instead of one which allows its variance to deviate from the mean to accomodate for additional, biological variation. The NB has as probability mass function

$$f(c;\gamma,\phi) = \frac{\Gamma(c+\frac{1}{\phi})}{c!\Gamma(\frac{1}{\phi})}\Big(\frac{1}{1+\phi\gamma}\Big)^{\frac{1}{\phi}}\Big(\frac{\gamma}{\frac{1}{\phi}+\gamma}\Big)^{c} \tag{4}$$

with Γ the gamma-function, $\mathrm{E}(c_{ij}) = \gamma_{ij} = f_j q_{i,\rho(j)}$ and $\mathrm{Var}(c_{ij})=\gamma_{ij}+\phi_i\gamma_{ij}^2$ where ϕ_i is called the overdispersion parameter of taxon i. In case $\phi_i=0$ this taxon again follows the Poisson distribution. γ and ϕ uniquely define the NB distribution. The NB also arises as the marginal distribution from hierarchical model whereby the Poisson rate parameters γ are Gamma distributed. This is biologically meaningful since the taxon counts from biological replicates are derived from subcommunities with their own Poisson distributions with different means [22].

A completely different approach to marker gene count modelling was inspired by the fact that count matrices from marker gene surveys contain a high fraction of zeroes [17, 32]. Methods interpreting these zeroes solely as absence of the taxon may suffer from bias since those zeroes may as well be a result of undersampling (i.e. the taxon is present but does not get detected because of the shallow sequencing depth). To tackle this issue, it was proposed to model the continuity corrected $\log_2$ of the raw count data

$$y_{ij} = \log_2(c_{ij} + 1) \tag{5}$$

via a Zero-inflated Gaussian (ZIG) distribution, i.e. a mixture of a point mass at zero and a Gaussian distribution. A ZIG has as probability density function

$$f(y_{ij}; N_j, \beta_0, \beta_1, \mu_i, \sigma_i^2) = \pi_j(N_j) I_0(y_{ij}) + (1 - \pi_j(N_j)) g(y_{ij}; \mu_i, \sigma_i^2) \tag{6}$$

where π_j is the fraction of counts fixed at zero, 1-π_j the fraction following the normal distribution and g the Gaussian density function. This partition is modelled as

$$log\Big(\frac{\pi_j(N_j)}{1 - \pi_j(N_j)}\Big) = \beta_0 + \beta_1 \log(N_j). \tag{7}$$

The mean model given the experimental condition $\rho(j)$ is

$$E\big(y_{ij}|\rho(j)\big) = (1 - \pi_j)\Big(b_{i0} + \log_2(f_j + 1) + b_{i1}k(j)\Big) \tag{8}$$

with $k(j)$ a dummy variable referring to the biological condition $\rho(j)$. Parameters to be estimated for this model are the mean μ_i, the variance σ_i^2, the probability function of the mixture membership per sample $\pi_j(N_j)$ (determined by β_0 and β_1), the offsets of each taxon b_{i0} and the log-fold change per taxon b_{i1}. This ZIG model and matching tests are implemented in the R-package *metagenomeSeq* [17].

Parameter estimation For the NB, the mean counts per taxon and condition $E(c_{i,\rho(j)}) = \mu_{i,\rho(j)} = q_{i,\rho(j)} f_j$ are a product of the scaling factor f_j and a condition dependent factor $q_{i,\rho(j)}$. This condition dependent factor is estimated by the average of the normalized taxon counts c_{ij}/f_j over the experimental condition ρ. The variances can be estimated taxon per taxon [30], but it has been shown to be more efficient to share information on the variance (and hence on the dispersion) over the different taxa [33]. The estimation procedure implemented in the R-package *edgeR* first estimates a common overdispersion parameter ϕ for all taxa, with a mean-variance relationship given by:

$$\sigma_{ij}^2 = \mu_{ij} + \phi \mu_{ij}^2 \tag{9}$$

It is clear that for $\phi = 0$ we are back at the Poisson model. The ϕ is estimated using conditional Maximum Likelihood from a derived set of pseudo-data, obtained by normalizing the raw data by quantile normalization [24, 33]. The taxon-wise overdispersion parameters are estimated through Maximum Likelihood estimation [31]. A weighted mean of these two estimates is calculated for each taxon using Weighted Likelihood. The more dissimilar the taxon-wise dispersion estimates, the heavier the weight of each taxon-wise estimate in the calculation [33]. R-packages *DESeq* and its successor *DESeq2* rely on a more data-driven model using Empirical Bayes, that also allows the overdispersion to vary across experimental conditions [23]. They modelled the variance as follows:

$$\sigma_{ij}^2 = \mu_{ij} + f_j^2 v_{i,\rho(j)} \tag{10}$$

with $v_{i,\rho(j)}$ a smooth function of $q_{i,\rho(j)}$. This smooth function is then estimated using local regression [23]. For small size factors f_j (e.g. due to small library sizes), the Poisson model holds again. This variance estimation method is based on the assumption that equally abundant taxa have equal variance and is implemented in the R-package *DESeq* [23]. This method was refined by the same authors as follows: first the taxon-wise dispersions are estimated by Maximum Likelihood(ML) estimation. Next, a smooth curve is fitted to represent the average dispersion and finally the ML estimates are shrunk towards this fitted line based on an empirical Bayes procedure. This dispersion estimation technique is implemented in the R-package *DESeq2* [34]. Estimation of the parameters of the ZIG happens through an Expectation-Maximization(EM)-algorithm [17].

Statistical testing Formal testing for differential taxa abundance between two groups of samples has been done by simple t-test by relying on asymptotic normality of the sample means of normalized read counts [30, 33], but more advanced tests exist based on the NB and ZIG distributions described above. In *edgeR*, an exact test based on the NB distribution is implemented. For each taxon i, this test considers the total count of taxon i over all samples $c_{i.} = \sum_{j=1}^{m} c_{ij}$ as fixed. Under the null hypothesis of no differential abundance, the sum of NB-distributed counts in either group also follows a NB distribution with expectation half the total count. The P-value of the observed proportions is then calculated using a two-tailed test as the chance to find the observed or a more extreme number of taxon counts in one group [33]. In *DESeq*, the same testing procedure is used [23]. In the successor package, *DESeq2*, a Wald test on the estimated shrunken coefficients of a Negative binomial GLM is implemented to test for differential abundance [34]. Both *DESeq* and *DESeq2* contain functionalities to pre-select potentially differentially abundant taxa using Independent filtering to reduce power loss due to multiple testing adjustment. Only taxa with an average abundance over all samples exceeding a certain threshold pass the filter [23, 34]. For the ZIG, the significance of the log-fold parameter b_{i1} is tested using a moderated t-test to assess whether group membership is predictive on the taxon counts [17]. Since all above methods rely on separate tests for each taxon, comparing two groups of samples implies a lot of statistical tests. This calls for multiple testing correction to control the rate of false positives, which is usually done by controlling the False discovery rate according to the method of Benjamini and Hochberg [35].

Whole-community modelling The previous methods try to model the sampling distribution of a single taxon over different technical or biological replicates. Hence, parameter estimation and testing occur for each taxon separately. This ignores possible dependencies between taxa counts, e.g. in case of symbiotic bacteria where one would expect positive correlations between their counts. Also, since we are interested in *relative* taxon abundances, an increased presence of one taxon is bound to reduce the relative abundances of the others. Moreover, the taxon-per-taxon comparisons involve a lot of separate tests, necessitating multiple testing corrections that reduce power. Multivariate methods are thus indicated to model the sampling distribution of the whole community of taxa over the different samples [36]. A natural attempt would be to model the composition of a sample by a vector of proportions following a multinomial distribution [36], which has the following probability mass function

$$f(c_{1j}, c_{2j}, ..., c_{nj}; \boldsymbol{\phi}_j | N_j) = \binom{N_j}{\mathbf{c}_j} \prod_{i=1}^{n} \phi_{ij}^{c_{ij}} \tag{11}$$

with $\boldsymbol{\phi_j} = (\phi_1, \phi_2, ..., \phi_n)$ the vector of underlying taxon proportions of the sample j and $\mathbf{c}_j = (c_{1j}, c_{2j}, ..., c_{nj})$ the taxon count vector. The library size N_j is considered fixed here and not a random variable, i.e. the distribution of the c_{ij} is conditional on the library size N_j. The expectation of this distribution is $E(c_{ij}) = N_j \phi_{ij}$ and the variance is $\mathrm{Var}(c_{ij}){=}N_j \phi_{ij}(1 - \phi_{ij})$ [37]. Modelling microbiome data using the multinomial assumes that the underlying distributions are fixed and thus only applies to technical replicates. However, separate samples from the same or comparable biological origin(biological replicates) exhibit heterogeneity in true composition because of spatial, temporal and individual-to-individual variation of the microbiome. As a result their taxa counts exhibit much larger variation than predicted by the multinomial distribution [37]. To account for this, also the taxa probability vector is considered as a random variable, being sampled from the slightly different subcommunities present in the same biological condition. These probability vectors are assumed to follow a Dirichlet distribution. The Dirichlet multinomial(DM) is the resulting marginal distribution function of the taxa count vectors for this hierarchical model [18], with as probability mass function

$$f(c_{1j}, c_{2j}, ..., c_{nj}; \boldsymbol{\phi}_j, \theta | N_j) = \binom{N_j}{\mathbf{c}_j} \frac{\prod_{i=1}^{n} \prod_{k=1}^{c_{ij}} \Big(\phi_{ij}(1-\theta) + (k-1)\theta\Big)}{\prod_{k=1} N_j \Big(1 - \theta + (k-1)\theta\Big)} \tag{12}$$

The expectation is the same as for the multinomial ($E(c_{ij}) = N_j \phi_{ij}$), but the variance is $\mathrm{Var}(c_{ij}) = N_j \phi_{ij}(1 - \phi_{ij})(\theta(N_j - 1) + 1)$. The library size N_j is again considered fixed. θ is the overdispersion parameter, for $\theta = 0$ we are back at the multinomial distribution [37]. Overdispersion not only means that the variance is larger, it also has implications on the convergence towards true population values as the number of samples increases. If the sample proportions follow a multinomial distribution, then the mean of the proportions of each taxon is less variable for larger collections of samples. Asymptotically, for an infinitely large number of samples, this variance becomes zero and the mean proportions converge to the true population proportion vector $\boldsymbol{\phi}$. When the counts are overdispersed, the taxa frequencies in different samples do not converge to $\boldsymbol{\phi}$ as more samples are gathered, and the asymptotic variance is larger than zero. This can again be explained by the fact that biological samples are samples from different subcommunities, whose underlying taxa distribution vectors $\boldsymbol{\phi_j}$ differ slightly [19, 38]. A goodness-of-fit test to compare a multinomial with a DM fit has been developped recently, as well as methods to test for equality of taxa composition vectors [19]. However, since the DM-approach to microbiome data is still relatively unexplored, no test for differential abundance of single taxa exists to our knowledge. Both the multinomial and the DM have received criticism for imposing negative correlations among taxa and for their failure to address model positive correlations [39].

2.3 Comparison of statistical methods for microbiome research

In the previous section, several normalization and analysis methods were discussed, where evidently dozens of others exist, such as quantile normalization and baySeq, PoissonSeq, MetaStats and DEGseq for differential expression analysis [40]. The methods discussed and compared in this thesis were chosen because of their frequent use in the field of microbiome research (Total count normalization, rarefying, t-test) [20, 22, 25, 29, 41, 42], positive evaluation in comparative studies (UQ, RLE, TMM, *DESeq*, *edgeR*) [20, 22] or novelty (CSS, ZIG) [17]. Comparison of normalization and analysis methods can occur through either simulation [20, 22, 40] or by application to real calibration data [20].

2.3.1 Simulation as a tool to evaluate microbiome methods

Simulation is a very useful tool to assess the performance of different analysis methods because the true underlying parameters that generated the data are known. A drawback of this approach is that it is based on some distributional assumptions that may not be generally true for all microbiome count data. In addition, generating samples with the same distribution as used in the analysis methodology can of course lead to an overly optimistic evaluation this method [29]. Simulation studies in the past have been done using a taxon-by-taxon approach based on the Poisson [20, 40, 43], the Negative binomial [30, 43], the beta-binomial(an over-dispersed binomial) [29, 30] or the normal distributions [17, 21] and by whole-community modelling through the multinomial distribution [22]. After making a distributional assumption, the necessary parameters are usually estimated based on an existing dataset. Library sizes may be fixed [17, 20], randomly sampled from a distribution [29, 30] or randomly sampled from library sizes from the dataset [22]. After the data generation, the read counts for some taxa are multiplied with some factor to create true positives of differentially abundant taxa [17, 20, 22, 29, 30].

2.3.2 Evaluation of microbiome methods through real data

It may appear impossible to evaluate normalization analysis methods by applying them to real data, since in this case the true underlying abundances are unknown. In practice, another, more precise method (e.g. microarray [44] or qRT-PCR [25]) may be used as gold-standard to determine the 'true' abundances. However, even when true abundances are unknown, normalization methods can be compared. One way is to assess how well the normalization methods equalize within-condition variablility of biological replicates [20]. Another is to apply a multidimensional visualization technique such as multi-dimensional scaling or correspondence analysis to normalized data and evaluate how well samples from different biological sources are separated [17]. In RNA-Seq, housekeeping genes are often used as control since they are assumed to be non-differentially expressed [20, 24], but no equivalent to this is used for microbiome data to our knowledge.

2.3.3 RLE and TMM normalization have been positively evaluated

The use of TSS and rarefying as normalization methods has been dissuaded by comparative studies [20, 22]. On the contrary, RLE and TMM have a very good record, in terms of minimizing the variance of non-differentially expressed genes in RNA-Seq [20] as well as for differential abundance analysis [20, 22]. *edgeR* is known to report more differentially abundant taxa than other methods [44], but also to have an inflated false positive rate [22, 44] and a false discovery rate not controlled at the nominal level [21]. *DESeq* was found to be a rather conservative method with low rate of false positives [44] and a well controlled false discovery rate [21]. It has also been observed by others that neither *DESeq* nor *edgeR* managed to control the false discovery rate at the nominal level [43]. *metagenomeSeq* testing based on the ZIG was found to perform poorly when the number of replicates is low, and has an inflated false positive rate when this number is high. Overall it performs worse than the methods based on the negative binomial (*DESeq2* and *edgeR*) [22]. A large scale comparison of differential expression methods yielded a variance stabilizing transformation combined with a moderated t-test (implemented in the *limma* package) as the top performing method, together with the non-paramteric SAMseq method based on resampling and the Wilcoxon rank sum test [21].

3 Materials and methods

3.1 Materials

3.1.1 HMP Dataset

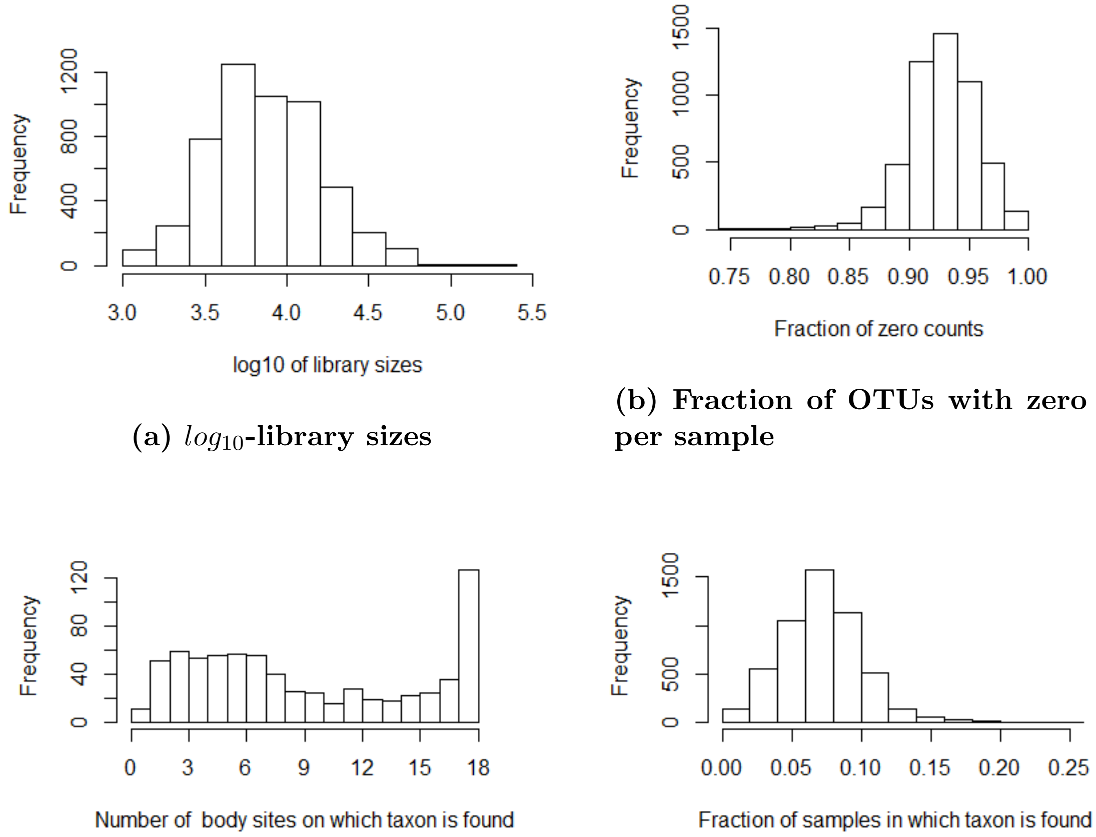
The Human Microbiome project (HMP) was undertaken in 2005 and aimed to exploit high-throughput sequencing technologies to characterize the healthy human microbiome [45]. In total, microbiomes of 16-19 body sites of a population of 242 healthy adults were sampled up to three times. From the oral cavity, samples were taken from the throat, hard palate, keratinized gingiva, saliva, subgingival plaque, supragingival plaque, palatine tonsils, tongue dorsum and buccal mucosa. For the skin, samples were taken from the left and right retroauricular crease (ear) and left and right antecubital fossa (elbow). A stool sample was taken to represent the gut microbiota, a sample from the anterior nares to represent the airways and finally also a blood sample was taken in some cases. For female subjects, additional samples were taken from the vaginal introitus, the mid vagina and posterior fornix as representatives of the vaginal region. These sites were chosen because sampling there minimally disturbed the existing microbiota and sampling held minimal risk for the participants. DNA was extracted from the whole community present at the body site, yielding the metagenome of the bodysite: genetic material of all cells and viruses present. This metagenome was subjected to both shotgun sequencing and 16S rRNA survey [46]. Up to five variable regions of the 16S rRNA gene of the metagenome were sequenced and the reads assigned to OTUs from kingdom up to genus level using the Greengenes database. Because of the choice of marker gene, only the kingdoms of bacteria and archaea were detected, excluding eukaryots and viruses. The resulting dataset is a contingency table that contains 5232 samples from 19 different body sites in the columns and a total of 725 taxa identified up to some taxonomic level. The cells of the table contain integers denoting the number of times a read was assigned to a certain taxon in each sample [46]. The numbers of samples per body site are shown in Table 2a. Because of the scarcity of the blood samples (6) we drop them from the analysis altogether. We use the general notation introduced in Section 2.2.1 to describe the dataset. We have then n=725 taxa and m=5226 remaining samples. A histogram of the library sizes N_j on the log_{10}-scale is shown in Figure 1a. Even though most library sizes are around 10^4 reads, some libraries are up to 10 times smaller or larger. The dataset is very sparse: per sample the majority of read counts are zero (see Figure 1b). This could indicate either absence of the taxon in the sample or undetected presence, e.g. due to insufficient sequencing depth. The high number of zeroes is referred to as the sparsity of the count matrix, and is due to the low prevalence of taxa over the samples. Even though a majority of the taxa is found at several sites (see Figure 1c), they're found in very little samples overall (see Figure 1d). The fractions of missclassification on the different taxonomic levels are shown in Table 2b. It is clear that most unclassified OTUs are only unclassified on the genus level.

3.2 Methods

All analyses and simulations were performed with the statistical software R, version 3.2.0 [47], using the packages phyloseq [48], *edgeR* [49], *DESeq* [23], DESeq2 [34], *metagenomeSeq* [17], foreach [50], plyr [51], ROCR [52], reshape2 [53] and HMP [54]. Plotting was done with the ggplot2 [55] and VennDiagram [56] packages.

3.2.1 Correspondence analysis

Correspondence analysis (CA) is a visualization technique for contingency tables, designed to explore relationships between categorical variables. It can be seen as the analog of



(a) log_{10}-library sizes

(b) Fraction of OTUs with zero counts per sample

(c) Prevalence of taxa over body sites, i.e. number of body sites in which the individual taxa were found

(d) Prevalence of taxa over samples, i.e. fraction of samples of the whole dataset in which individual taxa were found

Figure 1: Exploratory graphs of the HMP dataset

principal component analysis for discrete data. In a contingency table, the differences or distances over rows or columns can be quantified using some distance measure. The aim of CA is to summarize these distances in a number of independent dimensions. The rows and columns can then be represented as points in a two dimensional space spanned by the dimensions in which the samples exhibit the strongest deviation from independence between rows and columns. The result is called an ordination plot and allows a graphical summary of a large dataset [57]. Here we summarized the distances between the samples in terms of taxa read counts through CA, without prior normalization and after application of rarefying, TSS, RLE, TMM, UQ and CSS normalization. As distance measure we use the Bray-Curtis distance [27].

3.2.2 Testing and estimating overdispersion

All sites of the HMP dataset were tested for overdispersion using the method by Kim and Margison [58], which determines if the Dirichlet multinomial (DM) provides a better fit to the data than a regular multinomial. For significant sites, the parameters of the DM and

Body site	Samples
Mid vagina	150
Posterior fornix	149
Vaginal introitus	151
Hard palate	342
Keratinized gingiva	345
Saliva	327
Subgingival plaque	346
Supragingival plaque	353
Palatine Tonsils	351
Tongue dorsum	357
Throat	339
Buccal mucosa	348
L Retroauricular crease	318
R Retroauricular crease	310
L Antecubital fossa	190
R Antecubital fossa	194
Anterior nares	302
Blood	6
Stool	354

(a)

Taxonomic level	Fraction of unclassified entries
Kingdom	0
Phylum	0.0014
Class	0.0069
Order	0.026
Family	0.041
Genus	0.16
Total	0.232

(b)

Table 2: Number of samples per body site. R=right, L=left (a) and fraction of OTUs that could not be matched with the Greengenes database on this taxonomic level (b) of the HMP dataset

their standard deviations were then estimated using the method of moments implemented in the *dirmult* package [38, 59].

3.2.3 Simulation

The R-code for running the simulations and analysing the results was largely based on the simulations for differential abundance detection by McMurdie and Holmes, 2014 [22]. In order to generate fake communities that resemble true data as much as possible, the DM was used with the parameters estimated from the different bodysites and library sizes sampled randomly from the same bodysite. Resulting communities with less than 2 taxa were discarded and simulation was repeated until the community contained at least 2 taxa. Data from the Posterior fornix were not used since it was too difficult to generate the desired communities from them. Each generated community was divided into 2 classes and a number of taxon counts was multiplied by an effect size in one of the classes. Parameters that were varied across simulations to generate communitites were number of communities per class or number of replicates (3,5,10 and 20) and effect size (3,5 and 10). The fraction of differentially abundant taxa was kept fixed at 10% of the library size. This number of differentially abundant taxa was rounded upwards to have at least one differentially abundant taxon per community. Simulations were repeated 60 times for each body site and every combination of simulation parameters (number of replicates and effect size) to increase reliability of the estimate. So in total 1020 datasets were generated per combination of effect size and number of replicates.

Analysis parameters that were varied were type of normalization used (none, rarefying, TSS or proportion normalization, CSS, RLE, TMM and UQ) and the subsequent analysis method (t-test, Wilcoxon rank sum test, *DESeq*, *DESeq2*, *edgeR* and *metagenomeSeq*), with which differential abundance was tested between the two predefined classes. All these combinations of normalization and analysis methods were applied to each of the 12240 generated datasets (3 different numbers of replicates x 4 effect sizes x 17 body sites x 60 repeats=12240). The EM-algorithm for fitting the zero-inflated Gaussian of *metagenomeSeq* does work when all normalization factors are 1, so the normalization methods for which this is the case, rarefying and no normalization, could not be combined with the *metagenomeSeq* analysis method. Both *metagenomeSeq* and *DESeq2* sometimes fail to fit their model onto the simulated datasets. In this case the analysis was stopped and the result discarded. This happened in around 7% of the cases for *metagenomeSeq* for all normalization methods and in 27% of the cases for *DESeq2* with TSS normalization. The results for these methods are a summary of the cases were the model could be fit. Correction for multiple testing was done by the method by Benjamini and Hochberg to control the False Discovery Rate at a the nominal level of 5% [35]. Independent filtering was applied only for *DESeq2* since it is part of its default algorithm. Estimates of interest were averaged over the body sites and the 60 replicates.

The performance of a classification method can be characterized by its sensitivity (fraction of differentially abundant taxa detected), specificity(fraction of non-differentially abundant taxa reported as such) and false discovery rate. The specificity is equal to 1-FPR with FPR the false positive rate (fraction of non-differentially abundant taxa reported as differentially abundant). Usually, the settings of a method such as significance level can be changed to improve on of these criteria (sensitivity or specificity), but this comes at the cost of reducing the other one. To visualize this trade-off, the sensitivity is often plotted versus the FPR in a receiver operating characteristic (ROC) curve, constructed by varying some cut-off parameter such as the nominal false discovery rate and plotting the resulting sensitivity-FPR combinations. An example of such a ROC curve is shown in Figure 2. To summarize an ROC in one number, one can use the area under the ROC curve (area-under-the-curve or AUC), which will be 1 for a perfect classifier and 0.5 for a random classifier. The AUC value thus summarizes the performance of the method over all sensitivity-FPR combinations and is independent of the cut-off value used to construct the ROC curve [52, 61]. Another measure for method performance is the false discovery rate (FDR). It is calculated as the fraction of false positives among the taxa reported as differentially abundant (the discoveries) and reflects the reliability of the reported discoveries [35].

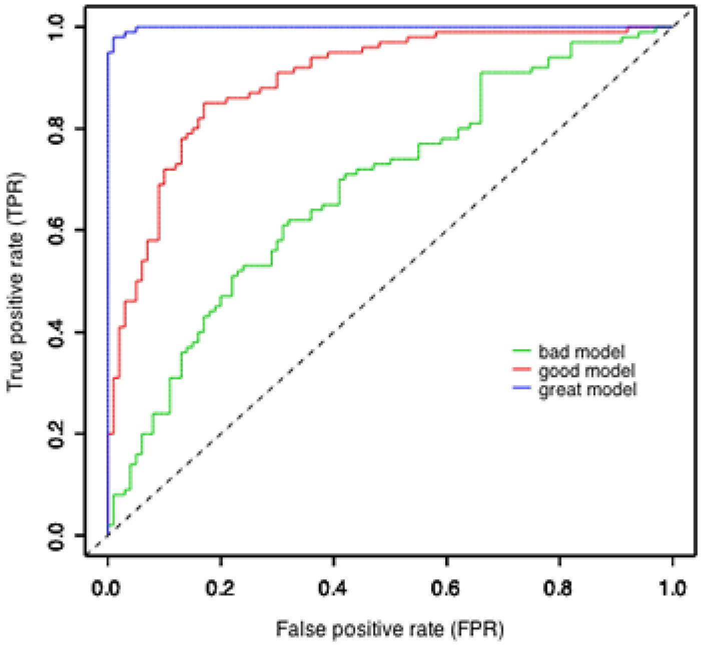


Figure 2: Example of an ROC curve, showing the trade-off between sensitivity and specificity. Diagonal dashed line shows performance of a random classifier. It is clear that the area below the curves (AUC) comes closer to 1 for better methods [60].

3.2.4 Differential abundance detection

The methods that performed best in the simulation study were applied to two sites from the HMP dataset, the tongue dorsum and the palatine tonsils, to test for differential abundance. These sites were chosen because we assume they differ little in bacterial community composition because of their proximity.

4 Results

4.1 Correspondence analysis

A plot of the correspondence analysis prior to normalization for the first two dimensions is shown in Figure 3. We see that the vaginal samples are very different from all other samples, whereas skin and airways on the one hand and oral cavity and gut on the other hand resemble each other more. The third dimension separates the gut samples from the others (plot not shown). From this exploratory analysis it is clear that the vaginal microbiome is most unique, and that there is a large degree of homogeneity between communities found at different sites from the same body region(gut, skin, oral cavity, airways and vagina). Most normalization methods do not have a big impact on the outcome of the correspondence analysis, only CSS normalization slightly improves the separation of the different regions (results not shown).

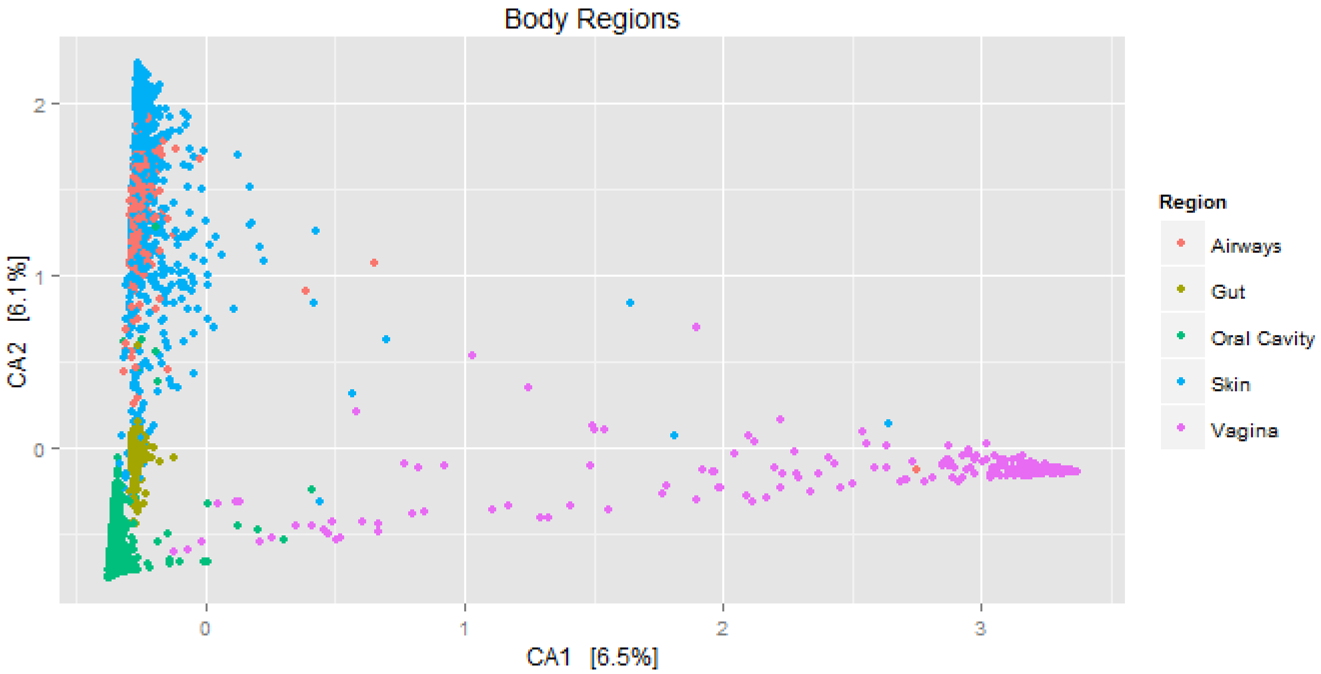


Figure 3: Ordination plot of the first two independent dimensions of the correspondence analysis of unnormalized data. Percentages indicate the fraction of the Pearson Chi-squared statistic explained by this dimension. The Bray-Curtis method was used for distance calculation

4.2 All body sites of the HMP exhibit overdispersion of thier read counts

All 18 communities of the HMP dataset were significantly better fit by the Dirichlet multinomial than by the multinomial on the 5% significance level (all $P < 0.0001$). The estimated overdispersion parameters are shown in Figure 4. We see that oral cavity and stool communities have a lower overdispersion than samples from the vagina and skin. This can be caused either by the differences in sampling procedure or by true differences in heterogeneity of the communities.

4.3 Simulation study results were summarized through AUC curves and FDR

The AUC-values are shown for all combinations of effect size, number of samples per group (replicates), normalization method and analysis method in Figure 5. Analogous curves were constructed for sensitivity (Figure 6) specificity (Figure 7) and FDR (Figure 8). In what follows we refer with *DESeq*, *DESeq2*, *edgeR* and *metagenomeSeq* to the model fitting and differential abundance testing methods implemented in those R-packages and discussed in Section 2.2.3, but NOT to the normalization methods implemented in these packages.

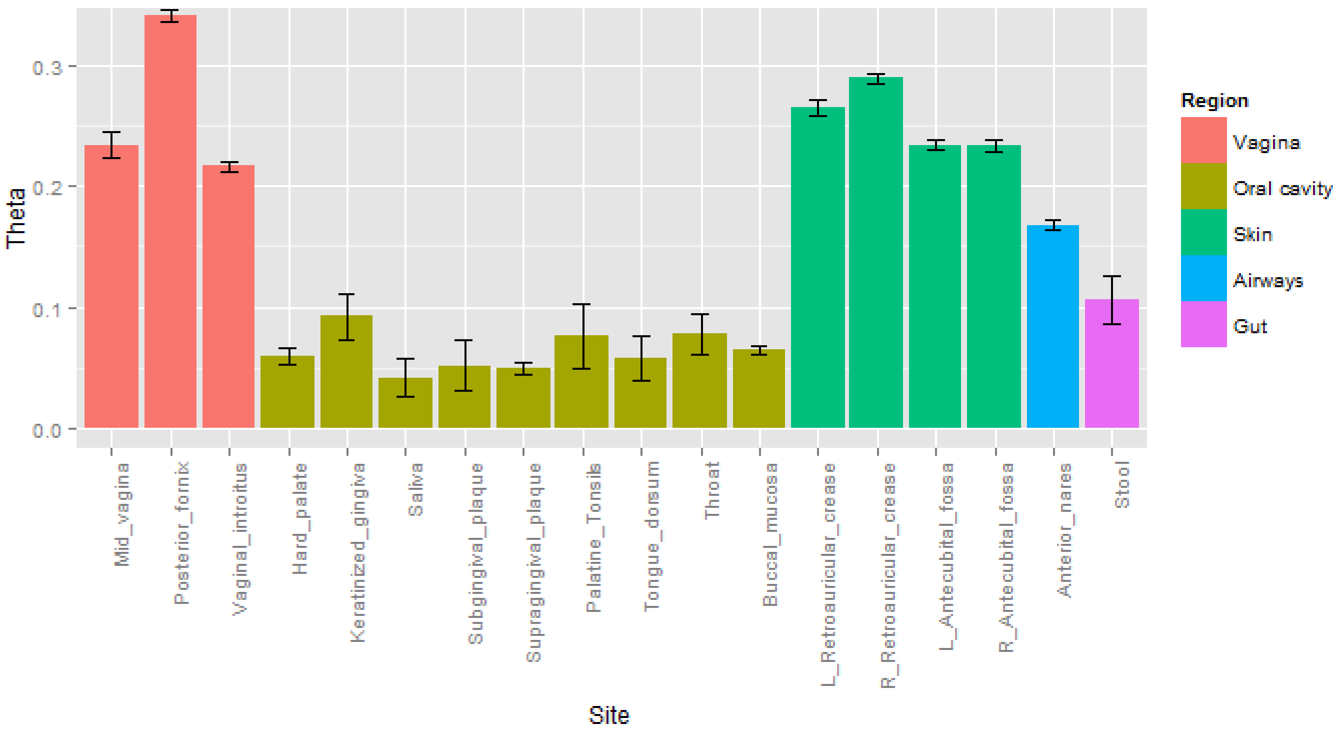


Figure 4: Overdispersion parameters θ of the Dirichlet multinomial distribution for every sampled body site of the HMP dataset, estimated by the method of moments. Error bars represent ± one standard deviation

4.3.1 Optimal normalization method depends on subsequent analysis method

Rarefying and proportion(TSS) normalization are widely used but have been criticised for increasing false positive and false negative rates [22]. Our simulation results confirm that TSS performs worse in terms of AUC values than the other investigated normalization methods for *edgeR*, *DESeq2* and *metagenomeSeq* over all tested effect sizes and number of samples per group. For Welch t-test and Wilcoxon rank sum test it performs well for effect size 3 and very badly for effect size 20. It also has an increased FDR for Welch t-test, Wilcoxon rank sum test and *edgeR*. Notably, it also performs much worse than when no normalization is applied at all in most cases. The differences with other methods grow more pronounced as the effect sizes increases and are mainly due to lower specificity when TSS is used. This corresponds well with the hypothesis that larger differences in abundance of other taxa (due to the larger effect size) affect the library size and thus the proportions of other, non-differentially abundant taxa so much that they are also marked to be differentially abundant [20, 23–25].

Rarefying works rather well on the simulated datasets for *DESeq*, but in combination with *edgeR* and *DESeq2* it shows a mediocre performance. For Welch t-test and Wilcoxon rank sum tests rarefying has a slightly increased false positive rate and a strongly increased FDR. Skipping the normalization step leads to surprisingly good results for *DESeq*, *DESeq2*, Welch t-test and Wilcoxon rank sum test, espcially for larger effect sizes. In combination with *edgeR* it performs intermediately, but worse than rarefying. UQ-normalization is

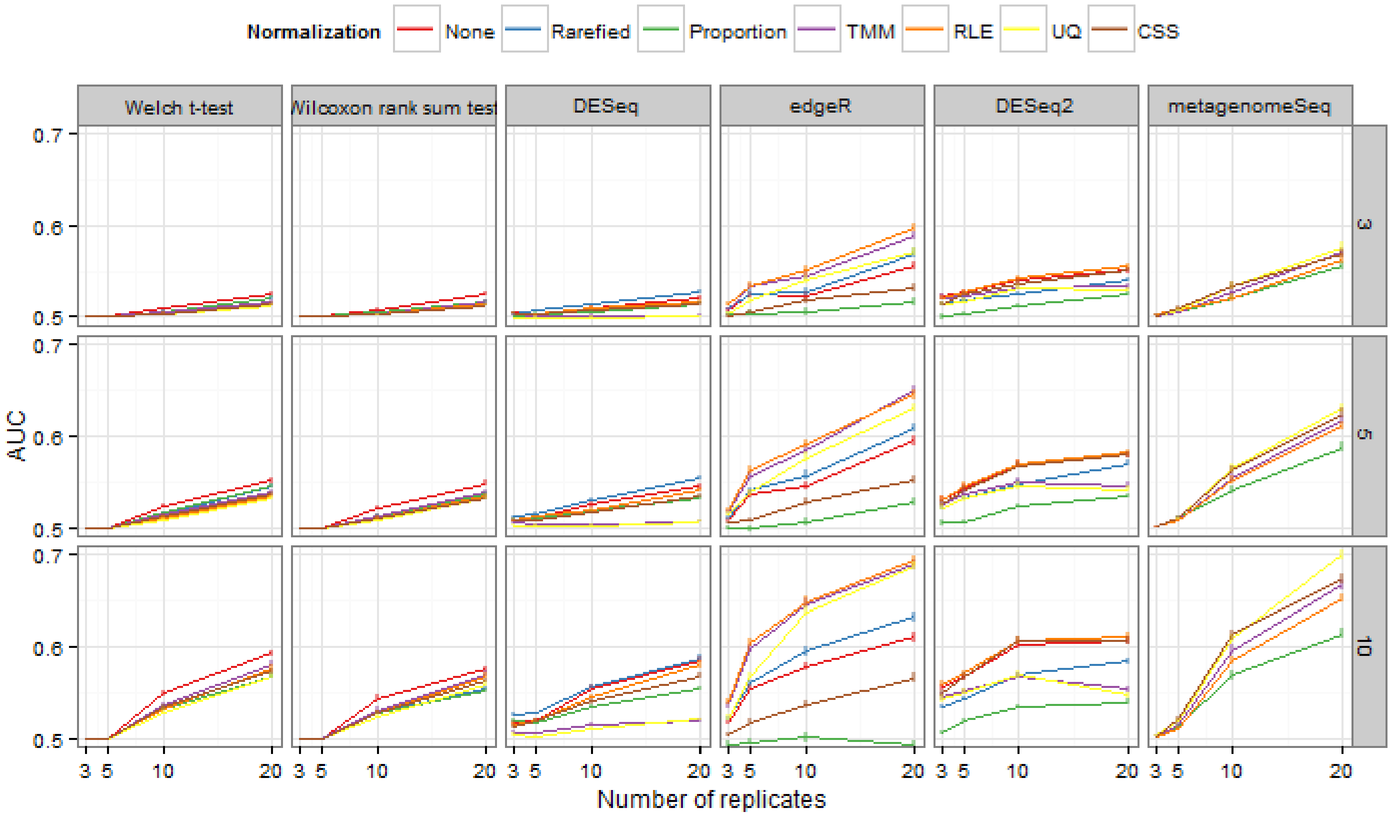


Figure 5: Results of simulation study for AUC values for all combinations of number of samples per group (x axis), normalization method(plotting colour), analysis method (top facets) and effect size (facets on the right). Error bars represent ± 1 standard error of the mean. *metagenomeSeq* results could not be obtained with no normalization or rarefying.

performant in combination with *edgeR* and *metagenomeSeq*, but not at all with the other methods. In combination with *DESeq*, UQ-normalization leads to a much higher FDR, with *metagenomeSeq* it renders this method more specific. CSS normalization works well for *metagenomeSeq*, *DESeq* and *DESeq2*, but not for Welch t-test and Wilcoxon rank sum test and for *edgeR* it is outright bad. The CSS normalization method renders *DESeq2* and *edgeR* very specific, but it reduces *edgeR*'s power drastically. On the other hand it renders *metagenomeSeq* more powerful. RLE and TMM normalization are clearly the best normalization methods for *edgeR* in terms of AUC values, but only because they make a very sensitive method. With these normalization methods *edgeR* is also least specific. TMM also combines well with *metagenomeSeq*, but not with *DESeq* and *DESeq2*. On the other hand, RLE goes well together with *DESeq2*. Summarizing, for a true positive fraction of 10%, *edgeR* performs best with RLE or TMM normalization, *DESeq2* with CSS, RLE or no normalization and *metagenomeSeq* with CSS, UQ and TMM normalization. These simulation results suggest that optimality of normalization methods for differential abundance analysis depends on the subsequent analysis method, and that combining normalization and analysis methods from the same authors does not necessarily lead to the best result.

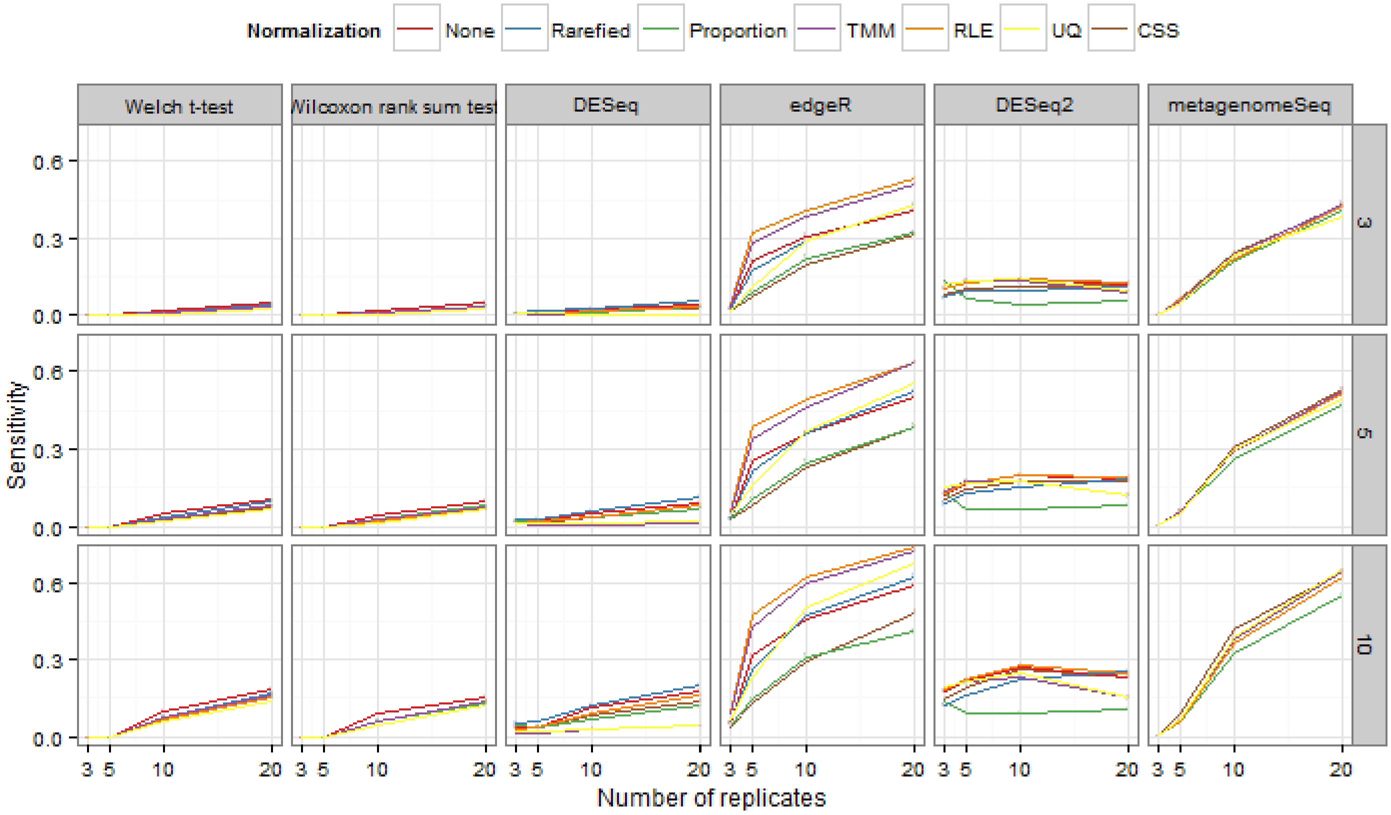


Figure 6: Results of simulation study for sensitivity for all combinations of number of samples per group (x axis), normalization method (plotting colour), analysis method (top facets) and effect size (facets on the right). Error bars represent ± standard error of the mean. *metagenomeSeq* results could not be obtained with no normalization or rarefying.

4.3.2 *edgeR* performs best in terms of AUC values but suffers from low specificity and elevated FDR

DESeq, the Welch t-test and Wilcoxon rank sum test are no match for the other investigated methods in terms of AUC values, especially when the number of replicates is low. Their performance varies very little with normalization methods, but they are almost powerless to detect differential abundance, especially for small effect sizes and number of replicates. A possible partial explanation for this is that the Welch t-test and Wilcoxon rank sum test do not share information on the variance over the different taxa as *DESeq*, *DESeq2* and *edgeR* do, which may make them less efficient. When there are only 3 samples per group, only *DESeq2* has any power at all to detect differential abundance. However, once there are 5 replicates per group, *edgeR* in combination with RLE or TMM normalization (and UQ to a lesser extent) gains enormously in sensitivity and continues to do so for more replicates. Perhaps the variance estimation algorithm of *edgeR* overestimates the variance when the number of replicates is small because the taxon-wise correction of the variance estimates through empirical Bayes does not work well, as has been suggested previously [23]. On the contrary, *DESeq2*'s power hardly depends on the number of replicates and is hence lower than *edgeR*'s power for 5 replicates or more. *metagenomeSeq*'s power is less than 0.1 for 3 or 5 replicates, but surpasses 0.3 for 10 or 20 replicates. *DESeq2* becomes more specific when there are more replicates,

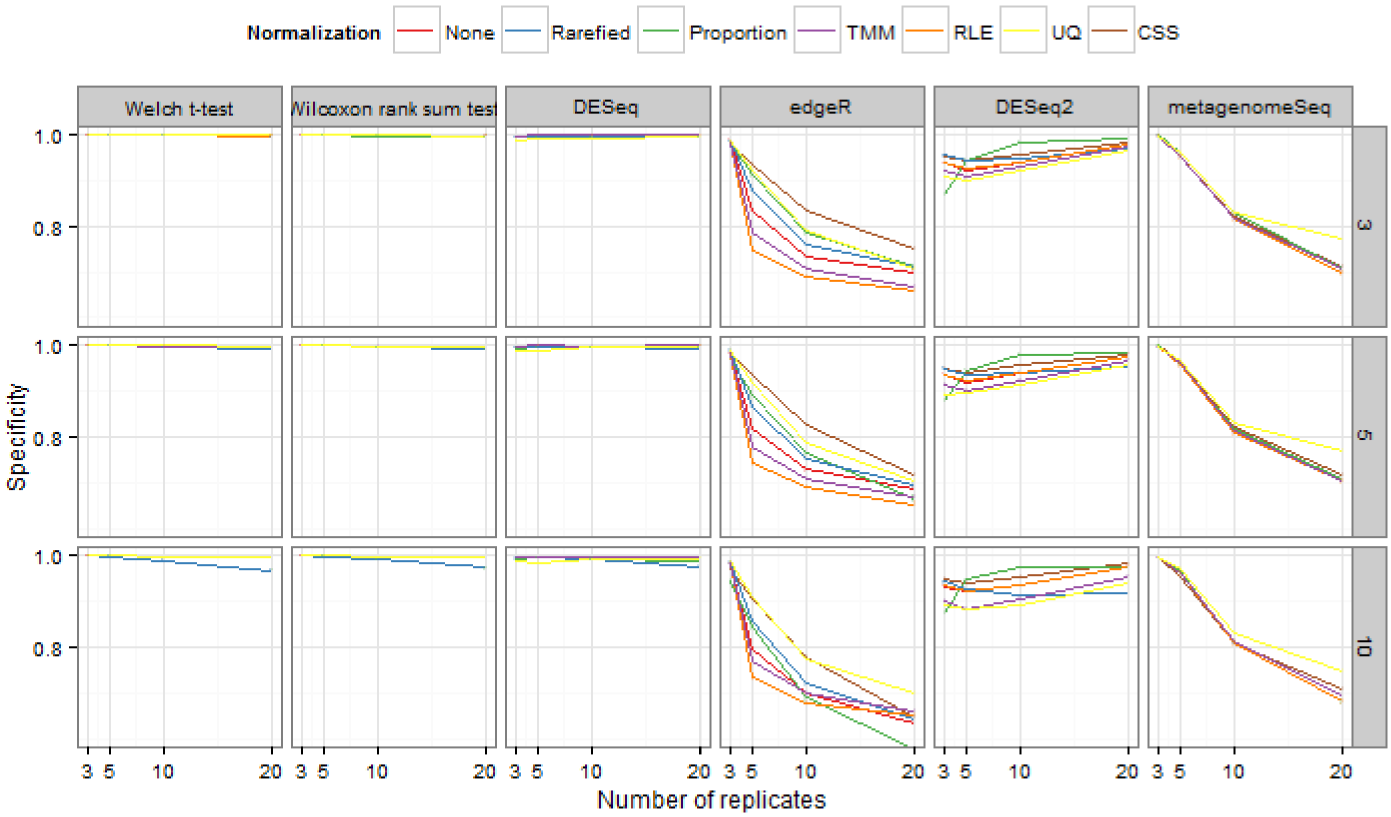


Figure 7: Results of simulation study for specificity of all combinations of number of samples per group (x axis), normalization method(plotting colour), analysis method (top facets) and effect size (facets on the right). Error bars represent ± standard error of the mean. *metagenomeSeq* results could not be obtained with no normalization or rarefying.

and in particular with CSS, RLE and no normalization its FDR decreases quickly with the number of replicates. For these normalization methods the FDR also decreases with effect size, but for an effect size of 3 it is TSS that has the lowest FDR. On the contrary, starting from 5 and 10 replicates respectively, *edgeR* and *metagenomeSeq* become much less specific, and for 5 replicates and more their FDR hovers around 0.75, independent of effect size. For *DESeq2* and *metagenomeSeq* the sensitivity and specificity depend little on the normalization method, but for *edgeR* the RLE and TMM methods make it less specific but more sensitive, whereas the UQ and CSS methods render it more specific and less sensitive. In terms of AUC values the *edgeR* analysis method slightly outperforms *metagenomeSeq* when there are many replicates and by a great margin when the numer of replicates is low. *DESeq2* only has higher AUC values when there are only 3 samples per group.

The number of replicates is thus a crucial for the evaluation of analysis methods, especially when this number is low. Larger effect sizes improve the sensitivity of all methods, but result in a slightly lower specificity for the *edgeR* method.

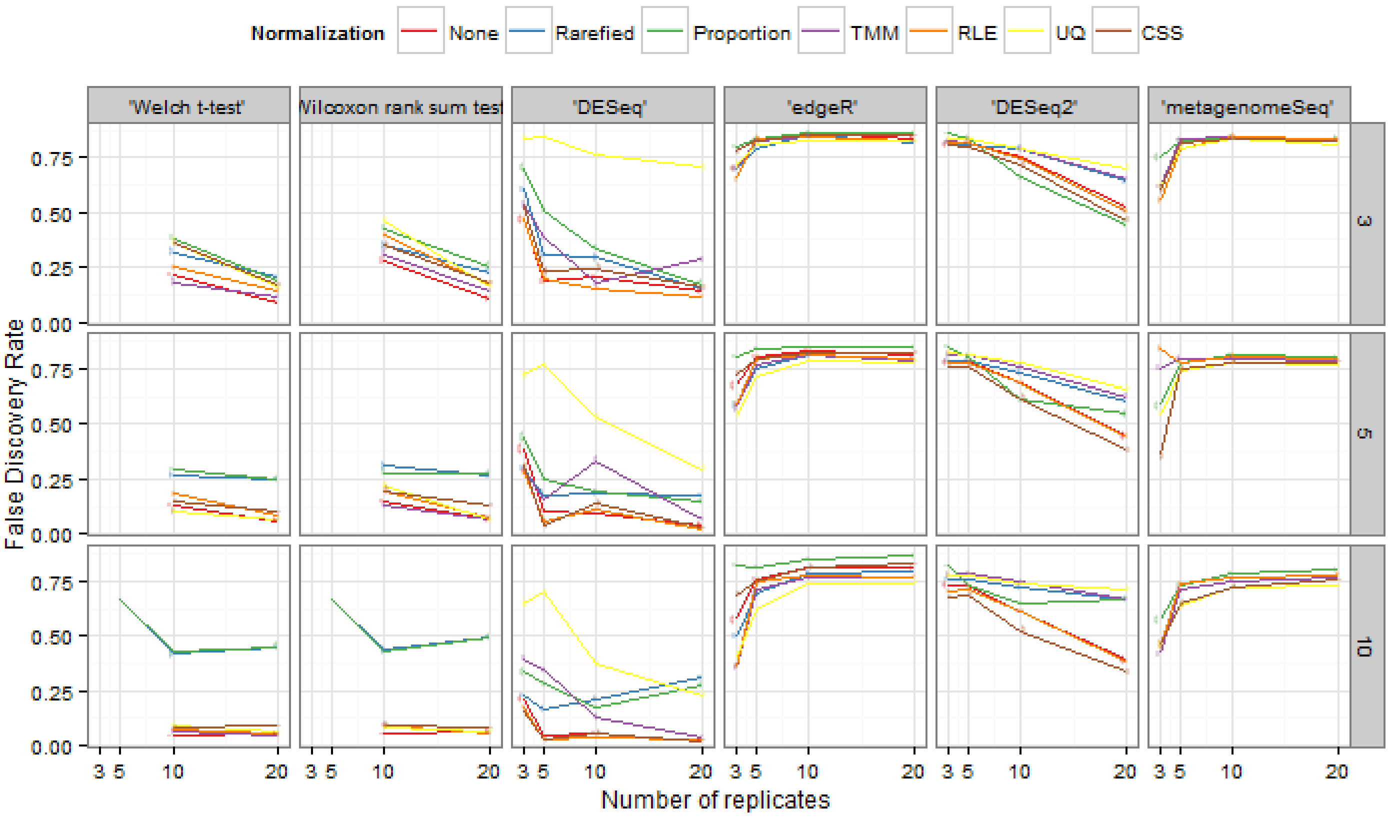


Figure 8: Results of simulation study for False discovery rate (FDR) of all combinations of number of samples per group (x axis), normalization method(plotting colour), analysis method (top facets) and effect size (facets on the right). Nominal FDR was ≤0.05. Error bars represent ± standard error of the mean. *metagenomeSeq* results could not be obtained with no normalization or rarefying. Empty spaces indicated no taxa were reported differentially abundant, so FDR could not be calculated

4.4 *edgeR* reports more differentially abundant taxa than *DESeq2* on real data

The communities at the tongue dorsum and palatine tonsils were analysed with*edgeR* combined with RLE and TMM normalization and *DESeq2* combined with CSS, RLE and no normalization. In the tongue dorsum 250 different taxa were found, in the palatine tonsils 307. The number of taxa reported as differentially abundant is shown in Figure 9. *edgeR* reported exactly the same 89 taxa to be differentially abundant with TMM and RLE normalization, with 64 taxa more abundant in the palatine tonsils and 25 more abundant on the tongue dorsum. 21 differentially abundant taxa were detected by all 5 analysis combinations. An additional 24 taxa were found siginificant by all combinations except for CSS normalization with *DESeq2*. 2 taxa were only marked to be differentially abundant by *DESeq2* with all normalization methods and another 2 only by *DESeq2* and

RLE and no normalization. The biological significance of these differential abundances fall beyond the scope of this thesis, but these results confirm that *edgeR* reports more differentially abundant taxa than *DESeq2*. The apparent conservativeness of *DESeq2* in combination with CSS normalization was however not seen in the simulation study.

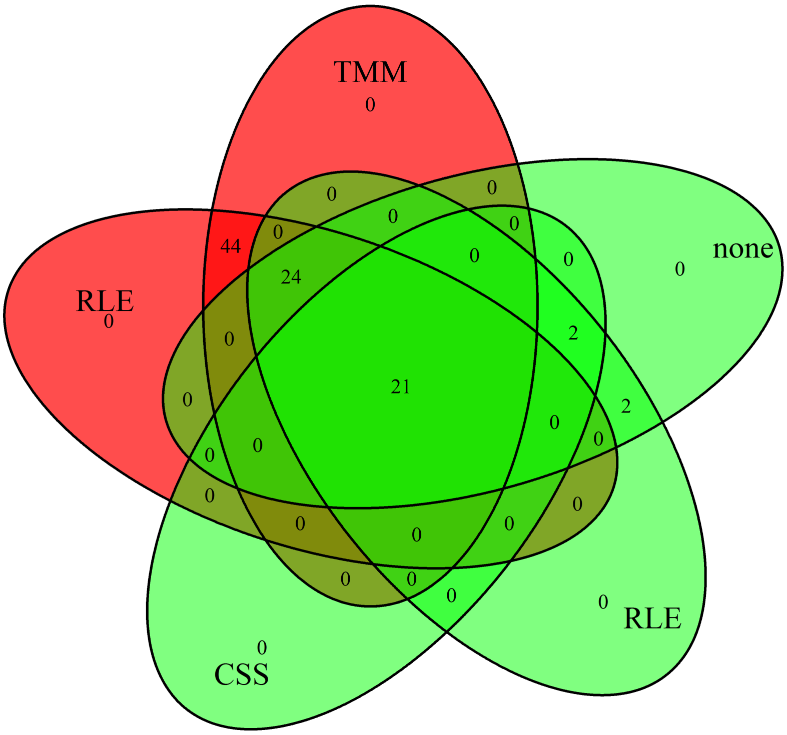


Figure 9: Venn diagram of number of taxa found to be differentially abundant between tongue dorsum and palatine tonsils according to different analysis methods. Red circles indicate analysis by *edgeR*, green circles by *DESeq2*. Normalization methods are indicated on the circles

5 Discussion

Simulation studies to evaluate performance of normalization or analysis methods for sequence read count data most often generate data based on some distribution that real sequence data are supposed to follow. Past simulation studies have used a broad range of distributions to model the counts taxon-by-taxon between the different samples [17, 20, 21, 29, 30, 40, 43]. However, whole-community modelling (e.g. through the multinomial distribution [22]) requires less parameters to be estimated and might represent a more efficient approach to simulation. We found that all communities sampled in the HMP project exhibit overdispersion compared to the multinomial, although this phenomenon is less pronounced in the samples from the oral cavity. Hence we simulated microbiome datasets based on the Dirichlet multinomial distribution, whereby taxon proportions and overdispersion were estimated per environmental condition.

Over the years, dozens of normalization and analysis methods for differential abundance have been devised, and some attempts have been made to compare both the normalization [20,22] and analysis methods [21,22]. In this thesis a simulation study was performed to assess the performance of combinations of some of the most popular normalization and analysis methods over a number of effect sizes and number of replicates per sample, whereby the fraction of differentially abundant taxa was kept fixed at 10%. Of course the results of such a study should be interpreted carefully, since they rely on assumptions about count distribution, number of differentially abundant taxa and scope of effect sizes that may not be representative of all microbiome data.

The testing algorithm implemented in the *edgeR* package in combination with the TMM and RLE normalization methods performed best in our simulation study in terms of AUC values. It is a very sensitive method (except when there are only 3 replicates), but especially for high number of replicates its specificty drops to 80%. Also its false discovery rate (FDR) is very high (75%, where the nominal level was 5%), which means that 3 out of 4 taxa reported as differentially abundant are not. The analysis method of *DESeq2* combines well with the RLE and CSS normalization methods and with no normalization at all. It is more specific than *edgeR* and its FDR is lower than for *edgeR* when there are 10 or more replicates, but also has a much lower power. Even though devised especially for microbiome read count data, *metagenomeSeq* is slightly outperformed by *edgeR*, developped for RNA-Seq. It lacks power when the number of replicates is low, is little specific when this number is large and also has a FDR of around 75%. The CSS normalization method from the *metagenomeSeq* package works well in combination with its analysis method, but our results show that also the TMM and UQ normalization methods collaborate well with it. The Welch t-test, Wilcoxon rank sum test and *DESeq* lack power to detect differential abundance of taxa, even at high effect sizes and number of replicates. On the other hand these methods do have a much lower FDR of around 10%l. These results correspond relatively well with findings from previous research [20–22,44].

For differential abundance analysis, our simulation results suggest to use *edgeR* with TMM or RLE normalization when there are 5 replicates or more per group and the researcher definitly wants to avoid missing differentially abundant taxa, even when this comes at the cost of a high false positive rate and a high FDR. Normalizing with the UQ method instead renders the method a bit more specific,less sensitive and lowers its FDR, but still preserves its overal good performance. *DESeq2* together with CSS, RLE or no normalization seems indicated when the number of replicates is less than 5 or when one wants to restrict the false positive rate and FDR, even at the cost of a lower power.

Since most authors of analysis methods for RNA-seq or microbiome abundances also compose a new normalization method and implement them in the same package, they are sometimes seen as an undivisible unit. We have tried to combine different normalization and analysis methods as far as possible to assess their compatibility. Strikingly, very often combinations of normalization methods and analysis methods devised by different people performed as well or better than in the original combinations in our simulation study. On the other hand we did not establish an "optimal" normalization method for all analysis methods, which suggests that one should not regard normalization and subsequent analysis separately. It might be more useful to look for an optimal *combination* of normalization and analysis method for differential abundance. Therefor it might be a good idea to implement more than one normalization method in released R-packages for microbiome count data analysis, or to facilitate their implementation by the user. These results also confirm the merits of the Negative Binomial distribution in testing for differential taxon abundance, although the Zero-inflated Gaussian that was proposed recently also performs almost as well [17]. In our simulation study, another interesting parameter to vary is the fraction of differentially abundant taxa. Also one might think of a way to include correlations between taxa counts in the simulated datasets.

Control of false discovery rate fails completely for the most performant methods in our simulation study (*edgeR*, *DESeq2* and *metagenomeSeq*). Welch t-test, Wilcoxon rank sum test and *DESeq* only manage to control FDR at the nominal level for the largest effect sizes and number of replicates with the appropriate normalization method. This problem has been noted before [21, 40] and calls for a revision of the current FDR control methodology. Further research into normalization and analysis methods is definitely indicated, preferably also by other methods than simulation only.

6 References

[1] Xochitl C. Morgan and Curtis Huttenhower. Chapter 12: Human microbiome analysis. *PLoS Comput Biol*, 8(12):e1002808, Dec 2012. PCOMPBIOL-D-12-01454[PII].

[2] Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, Jun 2012.

[3] Brian J. McGill, Rampal S. Etienne, John S. Gray, David Alonso, Marti J. Anderson, Habtamu Kassa Benecha, Maria Dornelas, Brian J. Enquist, Jessica L. Green, Fangliang He, Allen H. Hurlbert, Anne E. Magurran, Pablo A. Marquet, Brian A. Maurer, Annette Ostling, Candan U. Soykan, Karl I. Ugland, and Ethan P. White. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10(10):995–1015, 2007.

[4] Inna Sekirov and B. Brett Finlay. The role of the intestinal microbiota in enteric infection. *J Physiol*, 587(Pt 17):4159–4167, Sep 2009. 19491248[pmid].

[5] Ivaylo I. Ivanov, Koji Atarashi, Nicolas Manel, Eoin L. Brodie, Tatsuichiro Shima, Ulas Karaoz, Dongguang Wei, Katherine C. Goldfarb, Clark A. Santee, Susan V. Lynch, Takeshi Tanoue, Akemi Imaoka, Kikuji Itoh, Kiyoshi Takeda, Yoshinori Umesaki, Kenya Honda, and Dan R. Littman. Induction of intestinal th17 cells by segmented filamentous bacteria. *Cell*, 139(3):485–498, Oct 2009. 19836068[pmid].

[6] Ivaylo I. Ivanov and Dan R. Littman. Segmented filamentous bacteria take the stage. *Mucosal Immunol*, 3(3):209–212, May 2010. 20147894[pmid].

[7] Ann M. O'Hara and Fergus Shanahan. The gut flora as a forgotten organ. *EMBO Rep*, 7(7):688–693, Jul 2006. 16819463[pmid].

[8] Jacques Ravel, Pawel Gajer, Zaid Abdo, G. Maria Schneider, Sara S. K. Koenig, Stacey L. McCulle, Shara Karlebach, Reshma Gorle, Jennifer Russell, Carol O. Tacket, Rebecca M. Brotman, Catherine C. Davis, Kevin Ault, Ligia Peralta, and Larry J. Forney. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A*, 108(Suppl 1):4680–4687, Mar 2011. 20534435[pmid].

[9] Christina Tobin Kahrstrom. Microbiome: Gut microbiome as a marker for diabetes. *Nat Rev Micro*, 10(11):733–733, Nov 2012.

[10] Jose U. Scher, Carles Ubeda, Michele Equinda, Raya Khanin, Yvonne Buischi, Agnes Viale, Lauren Lipuma, Mukundan Attur, Michael H. Pillinger, Gerald Weissmann, Dan R. Littman, Eric G. Pamer, Walter A. Bretz, and Steven B. Abramson. Periodontal disease and the oral microbiota in new-onset rheumatoid arthritis. *Arthritis Rheum*, 64(10):3083–3094, Oct 2012. 22576262[pmid].

[11] Les Dethlefsen and David A. Relman. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc Natl Acad Sci U S A*, 108(Suppl 1):4554–4561, Mar 2011. 20847294[pmid].

[12] Peter J. Turnbaugh, Ruth E. Ley, Michael A. Mahowald, Vincent Magrini, Elaine R. Mardis, and Jeffrey I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–131, Dec 2006.

[13] A. Khoruts. Changes in the composition of the human fecal microbiome following bacteriotherapy for recurrent clostridium difficile-associated diarrhea, 2010.

[14] Melissa B. Miller and Yi-Wei Tang. Basic concepts of microarrays and potential applications in clinical microbiology. *Clin Microbiol Rev*, 22(4):611–633, Oct 2009. 0019-09[PII].

[15] John C. Marioni, Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9):1509–1517, Sep 2008. 18550803[pmid].

[16] J. Michael Janda and Sharon L. Abbott. 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J Clin Microbiol*, 45(9):2761–2764, Sep 2007. 1228-07[PII].

[17] Joseph N. Paulson, O. Colin Stine, Hector Corrada Bravo, and Mihai Pop. Robust methods for differential abundance analysis in marker gene surveys. *Nat Methods*, 10(12):1200–1202, Dec 2013. 24076764[pmid].

[18] Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE*, 7(2):e30126, 02 2012.

[19] Patricio S. La Rosa, J. Paul Brooks, Elena Deych, Edward L. Boone, David J. Edwards, Qin Wang, Erica Sodergren, George Weinstock, and William D. Shannon. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE*, 7(12):e52078, 12 2012.

[20] Marie-Agns Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Cline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Lalo, Caroline Le Gall, Brigitte Schaffer, Stphane Le Crom, Mickal Guedj, and Florence Jaffrzic. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[21] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 14:91–91, Mar 2013. 1471-2105-14-91[PII].

[22] Paul J. McMurdie and Susan Holmes. Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4):e1003531, Apr 2014. PCOMPBIOL-D-13-01815[PII].

[23] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106–R106, Oct 2010. gb-2010-11-10-r106[PII].

[24] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25–R25, Mar 2010. gb-2010-11-3-r25[PII].

[25] James H. Bullard, Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11:94–94, Feb 2010. 1471-2105-11-94[PII].

[26] Davide Risso, John Ngai, Terence P. Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nat Biotech*, 32(9):896–902, Sep 2014. Computational Biology.

[27] J. Roger Bray and J. T. Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349, Oct 1957.

[28] Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, 71(12):8228–8235, Dec 2005. 1021-05[PII].

[29] James Robert White, Niranjan Nagarajan, and Mihai Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol*, 5(4):e1000352, Apr 2009. 08-PLCB-RA-0894R3[PII].

[30] Jun Lu, John K. Tomfohr, and Thomas B. Kepler. Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, 6:165–165, Jun 2005. 1471-2105-6-165[PII].

[31] Mark D. Robinson and Gordon K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[32] Alyssa C. Frazee, Ben Langmead, and Jeffrey T. Leek. Recount: A multi-experiment resource of analysis-ready rna-seq gene count datasets. *BMC Bioinformatics*, 12:449–449, Nov 2011. 1471-2105-12-449[PII].

[33] Mark D. Robinson and Gordon K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

[34] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550, Dec 2014. 25516281[pmid].

[35] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, Jan 1995. ArticleType: research-article / Full publication date: 1995 / Copyright © 1995 Royal Statistical Society.

[36] Patricio S. La Rosa, Yanjiao Zhou, Erica Sodergren, George Weinstock, and William D. Shannon. *Metagenomics for Microbiology.* Academic Press, 1 edition, Nov 2014.

[37] Jun Chen and Hongzhe Li. Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis(). *Ann Appl Stat*, 7(1):10.1214/12–AOAS592, Mar 2013. 24312162[pmid].

[38] Torben Tvedebrink. Overdispersion in allelic counts and -correction in forensic genetics. *Theoretical Population Biology*, 78(3):200 – 210, 2010.

[39] Siddhartha Mandal, Will Van Treuren, Richard White, Merete Eggesb, Rob Knight, and Shyamal Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26(0), 2015.

[40] Jun Li, Daniela M. Witten, Iain M. Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13(3):523–538, Jul 2012. 22003245[pmid].

[41] Joshua S. Bloom, Zia Khan, Leonid Kruglyak, Mona Singh, and Amy A. Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, 10:221–221, May 2009. 1471-2164-10-221[PII].

[42] Peter A. C. 't Hoen, Yavuz Ariyurek, Helene H. Thygesen, Erno Vreugdenhil, Rolf H. A. M. Vossen, Renée X. de Menezes, Judith M. Boer, Gert-Jan B. van Ommen, and Johan T. den Dunnen. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res*, 36(21):e141–e141, Dec 2008. 18927111[pmid].

[43] Vanessa M. Kvam, Peng Liu, and Yaqing Si. A comparison of statistical methods for detecting differentially expressed genes from rna-seq data. *American Journal of Botany*, 99(2):248–256, 2012.

[44] Intawat Nookaew, Marta Papini, Natapol Pornputtapong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen. A comprehensive comparison of rna-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in saccharomyces cerevisiae. *Nucleic Acids Res*, 40(20):10084–10097, Nov 2012. gks804[PII].

[45] The NIH HMP Working Group, Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A. Schloss, Vivien Bonazzi, Jean E. McEwen, Kris A. Wetterstrand, Carolyn Deal, Carl C. Baker, Valentina Di Francesco, T. Kevin Howcroft, Robert W. Karp, R. Dwayne Lunsford, Christopher R. Wellington, Tsegahiwot Belachew, Michael Wright, Christina Giblin, Hagit David, Melody Mills, Rachelle Salomon, Christopher Mullins, Beena Akolkar, Lisa Begg, Cindy Davis, Lindsey Grandison, Michael Humble, Jag Khalsa, A. Roger Little, Hannah Peavy, Carol Pontzer, Matthew Portnoy, Michael H. Sayre, Pamela Starke-Reed, Samir Zakhari, Jennifer Read, Bracie Watson, and Mark Guyer. The nih human microbiome project. *Genome Res*, 19(12):2317–2323, Dec 2009. 19819907[pmid].

[46] A framework for human microbiome research. *Nature*, 486(7402):215–221, Jun 2012.

[47] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2015.

[48] Paul J. McMurdie and Susan Holmes. phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, 8(4):e61217, 2013.

[49] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010. btp616[PII].

[50] Revolution Analytics and Steve Weston. *foreach: Foreach looping construct for R*, 2014. R package version 1.4.2.

[51] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011.

[52] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881, 2005.

[53] Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007.

[54] Patricio S. La Rosa, Elena Deych, Berkley Shands, and William D. Shannon. *HMP: Hypothesis Testing and Power Calculations for Comparing Metagenomic Samples from HMP*, 2013. R package version 1.3.1.

[55] Hadley Wickham. *ggplot2: elegant graphics for data analysis.* Springer New York, 2009.

[56] Hanbo Chen. *VennDiagram: Generate high-resolution Venn and Euler plots*, 2014. R package version 1.6.9.

[57] Nadia Sourial, Christina Wolfson, Bin Zhu, Jacqueline Quail, John Fletcher, Sathya Karunananthan, Karen Bandeen-Roche, François Baland, and Howard Bergman. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *J Clin Epidemiol*, 63(6):638–646, Jun 2010. 19896800[pmid].

[58] Byung Soo Kim and Barry H. Margolin. Testing goodness of fit of a multinomial model against overdispersed alternatives. *Biometrics*, 48(3):711–719, Sep 1992. ArticleType: research-article / Full publication date: Sep., 1992 / Copyright © 1992 International Biometric Society.

[59] B. S. Weir and W. G. Hill. Estimating f-statistics. *Annual Review of Genetics*, 36(1):721–750, 2002. PMID: 12359738.

[60] Jack Weiss. Background on roc curves, 2010.

[61] Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. Technical report, 2004.