Internship report

In-silico benchmark of methods for detecting differentially abundant features between metagenomics samples

Léonard Dubois

Master 2 GENIOMHE 2018-2019 - Université Paris Sud

Institut National de la Recherche Agronomique - INRA

Domaine de Vilvert, Jouy-en-Josas

Supervisors:

Magali Berland - Unité MetaGenoPolis

Mahendra Mariadassou - Unité MaIAGE

Acknowledgements

I would like to thank first my supervisors, Dr. Magali Berland and Dr. Mahendra Mariadassou for their guidance and all their relevant advice. For the freedom they gave me and the trust they put in me while conducting this project.

Many thanks also to Dr. Stanislav Dusko Ehrlich for having created the most interesting lab in France for studying microbiome. For the open-mindedness of mixing wet and dry lab, sequencing and storing inside the very same place. A very special thanks also for the wonderful talk a few years ago that lit up in me the spark of insatiable curiosity about the wonders of bacterial world.

Many special thanks to Drs. Nicolas Pons, Mathieu Almeida and Florian Plaza Oñate for the leads they gave me concerning what to do after the Master.

Then, to the people of the IBS team at Metagenopolis for the warm welcome, the nice atmosphere, for teaching me how to play tarot, for the cakes and all the laugh: Susie Guilly, Samar Berreira Ibraim, Dr. Victoria Meslier, Sébastien Fromentin, Marie Jeammet, Laurie Alla, Florence Thirion, Houssem Gharbi, Franck Gauthier, Fatoumata-Adama Traoré, Ariane Bassignani, Sana Zaghouani, Nicolas Maziers MD, Karine Valeille, Kévin Weiszer, Florence Haimet, Dr. Emmanuelle Le Chatelier and Dr. Joël Doré.

Finally, many thanks to Océane Kacimi for the cigarette breaks.

Contents

1 Presentation of the laboratory **1**

1.1 Institut National de la Recherche Agronomique 1

1.2 MetaGenoPolis 1

1.3 MaIAGE 1

2 Introduction **2**

2.1 Metagenomics data 2

2.1.1 Data generation 2

2.1.2 Particularities of metagenomic data 3

2.2 Differential analysis 4

2.3 Objective of the work 5

3 Material and methods **6**

3.1 Methods for differential analysis 6

3.1.1 Methods not included in the benchmark 6

3.1.2 Detailed methods 7

3.2 Creation of the package `metaDAF` 10

3.2.1 Package workflow 10

3.2.2 Dataset generation 11

3.3 Experimental design 12

3.4 Performances metrics 12

4 Results **13**

4.1 Area Under Curve - True Positive and False Positive Rates 13

4.2 About False positives 16

4.3 Example on real dataset 18

5 Discussion and possible improvements **20**

6 Conclusion **21**

7 References **22**

8 Supplementary Materials **26**

R session 26

Intersection figures 27

AUC boxplot - Specific case 28

1 Presentation of the laboratory

1.1 Institut National de la Recherche Agronomique

The National Institute of Agricultural Research (Institut National de la Recherche Agronomique - INRA) is a French public research institute founded in 1946 under the joint authority of the Ministries of Research and Agriculture. The INRA researches focuses on the quality of agriculture and environment, covering wide topics such as climate change, human nutrition, sustainable agriculture, preserved environment. Its goals range from discovering, gathering and spreading knowledge and innovation to scientific training and providing expertise for companies or institutions. It employs more than 8000 permanent staff combined with more than 500 PhD students and 2500 trainees. In terms of knowledge production, the INRA is among the top 1% most cited institutions in the world.

1.2 MetaGenoPolis

MetaGenoPolis (MGP) is a "Pre-industrial Demonstrator" project funded by the French initiative Future investments and created in 2013. It investigates the impact of the human gut microbiota on health and disease. MGP developed a wide field of skills spread across several teams. SAMBO is in charge of sample collection, management and biobanking, MetaQuant runs the sequencing platform, MetaFun focuses on functional metagenomics through wet-lab approaches and InfoBioStat deals with bioinformatics and biostatistic analysis of metagenomics big data as well as methodological developments.

1.3 MaIAGE

The research Unit "Applied Mathematics and Computer Science, from Genomes to the Environment" (MaIAGE) gathers mathematicians, computer scientists, bioinformaticians and biologists. By using mathematics, statistics and computer science, it develops methods for solving various problems ranging from the molecular level to the ecosystem level. It is split across 4 research teams : Bibliome, working on text-mining and knowledge extraction. Dynenvie for dynamic and statistical modelling of ecosystems, epidemiology and agronomy. BioSys for systems biology and finally StatInfOmics for bioinformatics and statistic of omics data. In addition to the research teams, the unit provide bioinformatic expertise and service though the Migale platform.

The supervisors of this internship belong respectively to the InfoBioStat and StatInfOmics teams.

2 Introduction

The popularity of high-throughput sequencing (HTS) techniques is soaring while its prices are always decreasing. More and more omics data are generated and rapid adoption of whole process mastery leads to many new applications. Metagenomics focuses on sequencing genetic material from culture-independent microorganisms communities. This is particularly interesting in the case of the human gut microbiota, of the utmost importance because of all its links with health and diseases [Ehrlich, 2016]. The microbiota is indeed linked, among others, to obesity [Ley et al., 2006], type 2 diabetes [Qin et al., 2012], inflammatory bowel disease [Morgan et al., 2012]. Several large scale studies describe the diversity of the gut flora and its link with nutrition [Wu et al., 2011, The Human Microbiome Project Consortium, 2012].

Even though the word "metagenomics" appeared 20 years ago [Handelsman et al., 1998], the field is still changing while seeking golden standard in all its aspect, from DNA extraction, conservation, sequencing, data processing and analysis. One particular point of discussion is the difference between 16S metagenomics, also called metabarcoding, which targets a marker-gene and Whole Genome Shotgun (WGS) metagenomics, also called shotgun metagenomics, which targets all the genes present in the microbiota and the pros and cons of each approach. Such open questions led to the creation of consortia like International Human Microbiome Standards (IHMS) whose endeavours is to coordinate the development of standard operating procedures [Santiago et al., 2014, Cardona et al., 2012]
The aim of this internship project is to study the very last part of the whole microbiome data analysis pipeline: comparing groups based on the count table of metagenomics features (genes, species...) in order to formulate hypothesis about the link between specific microorganisms abundances and traits in the sample (e.g. a disease in samples coming from gut microbiota). In order to understand the ins and outs of such analysis, the whole data generation process must be clearly explained.

2.1 Metagenomics data

2.1.1 Data generation

The overall process of data generation for shotgun metagenomics data follows the workflow presented in Figure 1. Though every step can be made by following a lot of different protocols, tools or method, the pipeline is always roughly the same. The biggest difference between shotgun metagenomics and metabarcoding occurs during the sequencing part: metabarcoding focuses on a marker gene, usually the 16S rRNA for bacteria which acts as a taxonomic marker whereas shotgun metagenomics does not focus on a specific gene. Both approaches are able to create the taxonomic profiles of a sample but they have specific needs in terms of computational method and hardware resources [Jovel et al., 2016].

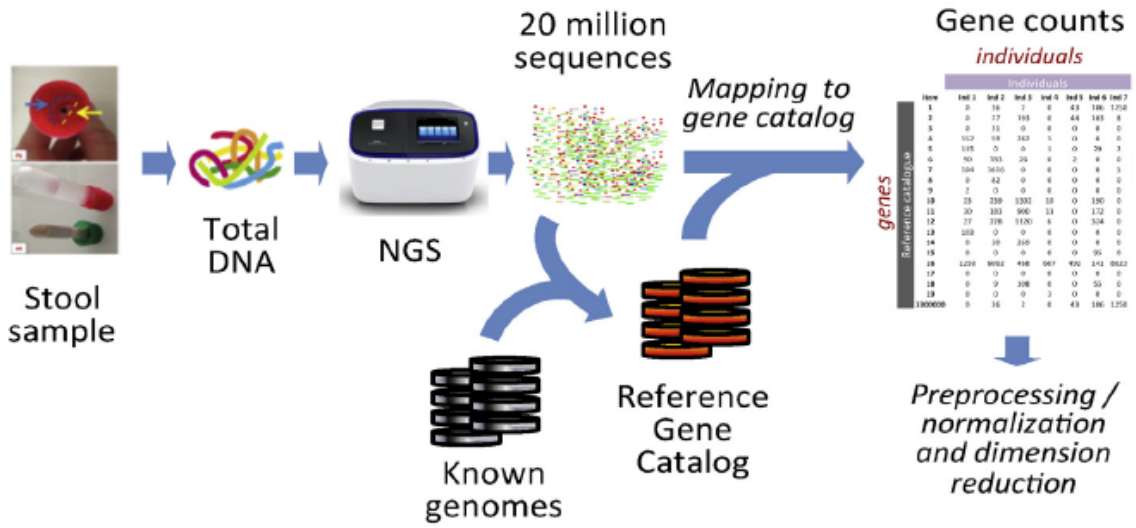

Figure 1: **Overview of shotgun metagenomic dataset generation**
Raw DNA is extracted from stool samples and all the genetic material is sequenced from it. Sequences are then processed to get rid of contaminant (human DNA) and mapped to a reference catalog. Thus, gene or species count tables are generated. Source [Ehrlich, 2016]

Based on their similarity, sequences can be clustered according to similarity with previously annotated sequences from references databases or clustered *de novo* into Operational Taxonomic Units (OTUs). In the case of shotgun metagenomics, co-abundant genes can be clustered into gene repertoire of microbial species. A new method allow to identify species core part as well as accessory genes sets and thus reconstitute Metagenomic Species Pan-genomes (MSPs) [Oñate et al., 2018]. In both cases, the taxonomic profiling end up generating count tables with samples as individuals and genes/species/OTUs/MSPs as features.

In some way, such data does not differ a lot from RNA-seq results. It consists in discrete count values. For both, counts are the number of reads mapped to a specific biological entity. Such count data often present similar distribution, sparsity.. as RNA-seq data. Thus, a lot of methods for differential analysis developed for RNA-seq have been used to analyses metagenomics count tables.

2.1.2 Particularities of metagenomic data

Several difficulties arise when working with metagenomics count tables. These can be divided into two main types : those explained by the biological background and those due to statistical properties of the data.

During the sequencing process, all samples do not generate the same amount of reads. These "differences in library sizes" are usually solved by rarefying, which is throwing away sequences from the largest libraries so they all have the same size. A threshold higher than the smallest

library size will also results in getting rid of some of the samples with the smallest library sizes. Although widely used for certain kind of analyses, this rarefying process does not make a consensus in the community [McMurdie and Holmes, 2014]. It is seen as sacrificing useful information. This relationship between the information available as reads from a sequencing run and the "real" information present in the environment is at the crux of problems occurring in metagenomics analysis.
Another solution would be to convert read counts to proportions within each sample (*i.e.* dividing it by the library size) [Gloor and Reid, 2016]. However this is also not unanimously recognized as the best method as it raises problems inherent to the peculiarities of so-called *compositional* data [Aitchison, 1982]. A dataset is compositional when the sum of the values for each sample is predefined. Hence the data points do not map to Euclidean space, but instead to a simplex. Both the total sum of counts and the absolute differences between observation are not really informative. The ALDEx2 method studied in this project [Fernandes et al., 2014] focuses on this compositional aspect.
Moreover, problems also arise in the transformation of the genomic information from sequences to genes and genes to species. The average gene and genome size can induce a strong bias in the final count table [Beszteri et al., 2010]. This is particularly true if MSPs or OTUs are generated by aggregating read count from a set of genes. Indeed, more reads map to longer genes thus increasing their counts if length is not properly accounted for.

Among the other property of metagenomics count table is the fact that it does not fit well with normal distribution. On top of having only non-negative discrete values, the sparsity of the count table is often high [Odintsova et al., 2017]. In fact, many zero values are due to the absence of a species from a sample or to a low sequencing depth making the detection of under-represented species impossible. Some distributions were used to model such data, overdispersed Poisson distribution, Negative binomial distribution, zero-inflated models... each trying to take tackle some or all of these issues.

2.2 Differential analysis

The analysis process extensively studied during this project is the differential analysis of features which is comparing, for each feature, the distribution of count values across samples. The data used are metagenomics count tables, matrices of numeric values, with individuals as one dimension and taxa/metagenomics species as the second dimension. Metadata are usually attached, containing qualitative and quantitative variables providing details about individuals. These variables can be used to generate groups inside the count matrix. For the sake of simplicity, the number of groups is limited to two in this study.
The basic statistical question is to test, at a given confidence threshold, whether or not some location value is different between the two groups. The focus is put on the mean for parametric

methods and median for non-parametric ones.
The results of the analysis, whatever method is used, is a p-value for each feature conveying the information on the probability of having a more extreme difference simply due to random fluctuations. The lower the p-value, the higher the confidence that the two groups have different abundances (count values) for a given feature.

2.3 Objective of the work

As said before, there is no golden standard in metagenomics analysis, including the differential analysis of features. Several studies already tried to compare the performances of existing methods [Jonsson et al., 2016, Jonsson et al., 2018, Lee et al., 2017]. Across these previous benchmarks, the methods compared are not the same nor the data type. The Differentially Abundant Features (DAF) could be genes, taxons, MGS... Moreover, they were several issues with these comparisons. First, the number of datasets tested is small (from 1 to 3), the reproducibility and availability of the code is limited and it is not easy to update the results by adding a new method to the benchmark.

A first part of the work proposed here aims to offer a reproducible support by providing a framework for efficient comparison of several popular methods in the form of an R package wrapping those methods and the analyses. Further methods could easily be added to the package to keep it up to date. On the other hand, the methods' performances were tested on a wide variety of artificial datasets ranging to very easy to very difficult. This allowed us to assess which methods are best depending on the characteristics of the dataset at hand.

Having a precise idea of the capacities of the method in terms of detection of differentially abundant features is of the utmost importance when such methods are used to detect abundance of specific microorganisms as biomarkers for pathology. All these links with sensible data and outcomes urge to identify powerful methods whose pitfalls and limits are well-known.

3 Material and methods

3.1 Methods for differential analysis

Over the past decade, a lot of methods for assessing differentially abundant features have been proposed in the literature, mostly as packages for the R Project for Statistical Computing or for the Python programming language. The project conducted during this internship focuses on the method implemented in R, as it is more commonly used than Python for statistical analyses.

We can distinguish three generations of solutions implemented. The first one containing simple solutions such as Wilcoxon test and some adaptations.
The second one, is composed of methods firstly designed for RNA-seq dataset analysis and subsequently adapted to metagenomics data. The count tables display in both cases a high percentage of rather low values and some uncommon high values. This results in a non-normal overall distribution that must be taken into account. The solution chosen often involves distributions with long tails. These methods include DESeq and DESeq2 [Anders and Huber, 2010, Love et al., 2014], edgeR [Robinson et al., 2010], voom [Law et al., 2014] and ALDEx2 [Fernandes et al., 2014]. However, metagenomics data tends to have an higher sparsity than RNA-seq data.
These methods were progressively put in competition with newest ones, emphasizing this abundance in low values by using zero-inflated distributions. This includes metagenomeSeq and its updates [Paulson et al., 2013], the mbzinb package [Chen et al., 2018a], ZIBseq [Peng et al., 2016], RAIDA [Sohn et al., 2015].

3.1.1 Methods not included in the benchmark

There are other methods worth mentioning even if they are not part of the benchmark conducted in this project. Several reasons explain the absence of such methods in the analysis.

First, the Metastats package [White et al., 2009]. It normalizes the data by replacing the raw count data by the relative abundance of each feature in each library then applies a non-parametric t-test. Then, in order to apply multiple hypothesis testing correction, it computes a q-value, an individual measure of the False Discovery rate (FDR) for each test. Even if this package is not be included in the benchmark because no longer maintained, it is important to keep in mind that it the oldest package for Differentially Abundant Features analysis in metagenomics samples.
The ZIBseq method [Peng et al., 2016] based on Zero-inflated beta regression is wrapped in the package developed in this project but absent from the final benchmark. The method is quite time-demanding and has been put aside. The ANCOM method [Mandal et al., 2015] focuses on the relative abundance of taxa in a sample. It compares two or more samples whithout making any distributional assumptions. Since it is not available on an official repository and the different

version of the code available here and there can not be easily included in the benchmark.
The QIIME software (Quantitative Insights Into Microbial Ecology) suite is implemented in Python [Caporaso et al., 2010] and also put aside because it can't be easily wrapped in an R package.
A couple of other methods came out recently and it still is difficult to assess whether they will take off : NBZIMM [Zhang et al., 2018], AMDA [Banerjee et al., 2019], metamicrobiomeR [Ho et al., 2019]. On top of that, lots of them are incremental variations on the negative binomial distribution and/or zero-inflated model, already tested here with mbzinb.
The package presented in section 3.2 was build from the ground up and makes it easy to add new methods.

3.1.2 Detailed methods

DESeq2 [Love et al., 2014] The DESeq method is not specifically designed for metagenomics data analysis but rather for high-throughput data provided as "quantitative readout in the count data". It can thus also be RNA-Seq, SAGE, ChIP-Seq, barcode data... The most common way to transform data prior to the analysis is using logarithmic transformation which leads to variance stabilization. DESeq is based on a negative binomial distribution, better suited for lower count values while allowing overdispersion (which is not the case of the Poisson distribution expected in the case of low count). The package demands the raw count values as basic input. Then, the differential analysis is performed by calling the function nbinomTest. DESeq provides a list of p-values among other information such as Fold-change (FC) between groups or FC standard error.
The DESeq2 package works in a similar way but adds a few subtleties in the overall organization of the function's R code. On top of that, it compute a Generalized Linear Model (GLM) for each gene. The hypothesis testing offers two alternatives. First one is a Wald test to estimate whether a log2 FC is equal to zero. The other is a Likelihood Ratio test (LRT) testing a full model and a reduced model where some terms are removed. It assesses if using the full model increases likelihood. Thus, a p-value can be obtained for each gene conveying whether its count value is significantly different across groups/conditions.
DESEq2 can test extended model with multiple factors influencing the count. It could test the effect of confounder variables such as library size for example.

edgeR [Robinson et al., 2010] and voom [Law et al., 2014] These two methods were first developed for RNA-Seq analysis but can be easily transposed to metagenomic data for they are both working with table of count values.
edgeR (empirical analysis of Differential Gene Expression in R) fit GLM based on negative Binomial (NB) model. It tests the hypothesis using a quasi-likelihood F-test after having normalize the data to correct the sequencing depth (library size) bias. The real specificity of the method is the use of the Trimmed Mean of M-values (TMM) as a new normalization method

developed by the authors [Robinson and Oshlack, 2010]. It aims to correct the composition effect when a particular feature with high values makes the other features seem they are under-represented. TMM is used to find scaling factors minimizing the log-fold change between samples for most features.
voom is a function now part of the `limma` R package. It consists in a transformation of the data prior to the GLM fitting and analysis. It transforms count data to log2-counts per million (logCPM), estimates the mean-variance relationship and uses this to compute observation-level weights.
Both of those method allow multiple factor comparison.

ALDEx2 [Fernandes et al., 2014] The ALDEx2 method (ANOVA-Like Differential Expression tool for high throughput sequencing data) is first based on the fact that metagenomic abundance matrix are not count data *per se* but rather a proxy for compositional data. This means that raw counts are only used insofar as they provide information on relative abundances. It estimates per-feature technical variation using Monte-Carlo instances drawn from the Dirichlet distribution which maintains the proportional nature of the data. Because of the compositional aspect of the data, they need to be transformed using Centered log-ratio (CLR) transformation : the read counts within a sample are first log-transformed (usually after addition of a pseudo count) and then centered.

metagenomeSeq [Paulson et al., 2013] This method has been specifically designed for analysis of metagenomic data, described by its creator as "sparse high-throughput microbial marker-gene survey". It aims to deal with the sparsity of data and the uneven sequencing depth (corrected through normalization). The normalization to correct the bias of varying depths of sequencing across samples is made using a cumulative-sum scaling (CSS). It uses a scaling factor for each sample based on the sum of the count of this sample up to a given quantile. Then, the differential abundance testing uses a zero-inflated Gaussian (ZIG) distribution mixture model allowing frequent zero values. So it seeks to estimate if a zero comes from the undersampling or the count distribution (absence in the microbial community).
After an update, a new method sometimes called metagenomeSeq2, is based on a zero-inflated log-normal (ZIL) mixture model instead of the ZIG one.

RAIDA [Sohn et al., 2015] RAIDA stands for Ratio Approach for Identifying Differential Abundance and uses the ratio between features in a zero-inflated log-normal model. The particularity of such approach is that is it not affected by unbalanced conditions, i.e. a large difference between the total abundance of DAF (Differentially Abundant Features) across conditions. However, the RAIDA package suffers from the absence of vignette or official repository. It must be downloaded as an archive on the author's website.

mbzinb [Chen et al., 2018a] This package implements an omnibus test for microbiome analysis testing simultaneously the abundance, prevalence and dispersion of features. It also uses on a Zero-Inflated Negative Binomial regression model (ZINB) for differential analysis and a likelihood-ratio test. Normalization for correcting variable library size is done using a method developed by the authors called GMPR for Geometric Mean of Pairwise Ratio [Chen et al., 2018b]. It consists in the same steps as the DESeq normalization but in the opposite order. First, pairwise ratio across sample are computed and then the geometric mean of all gives the scaling factor. By doing so, the information used is maximized. The method also deals with outliers using Cook's distance or Winsorization (replacing extreme values by the 3^{rd} or 97^{th} percentile of the distribution). Getting rid of outliers improves the ZINB fitted model. Otherwise, the risk is an increase of false positive.

Wilcoxon test The last method, the Wilcoxon–Mann–Whitney test, is fundamental for the benchmark and serves as a baseline. It is a nonparametric test for assessing whether the median value is different between two groups. The pvalue resulting of this test is then corrected using the Benjamini-Hochberg correction (BH) procedure which controls the False Discovery Rate at a given level. Here, no prior normalization is performed. High variability in the library size could leads to poor performances.

3.2 Creation of the package `metaDAF`

In order to easily compare the performances of all these methods, a specific R package was developed called `metaDAF` for metagenomic Differentially Abundant Features. It wraps datasets (count values and metadata) in R objects and methods in functions having standardized input and output.

3.2.1 Package workflow

The Figure 2 summarizes what the `metaDAF` package is capable of. First, count table data and metadata defining groups can be loaded and formatted or generated as described in section 3.2.2. Multiple datasets can be generated at once for repeated analysis.

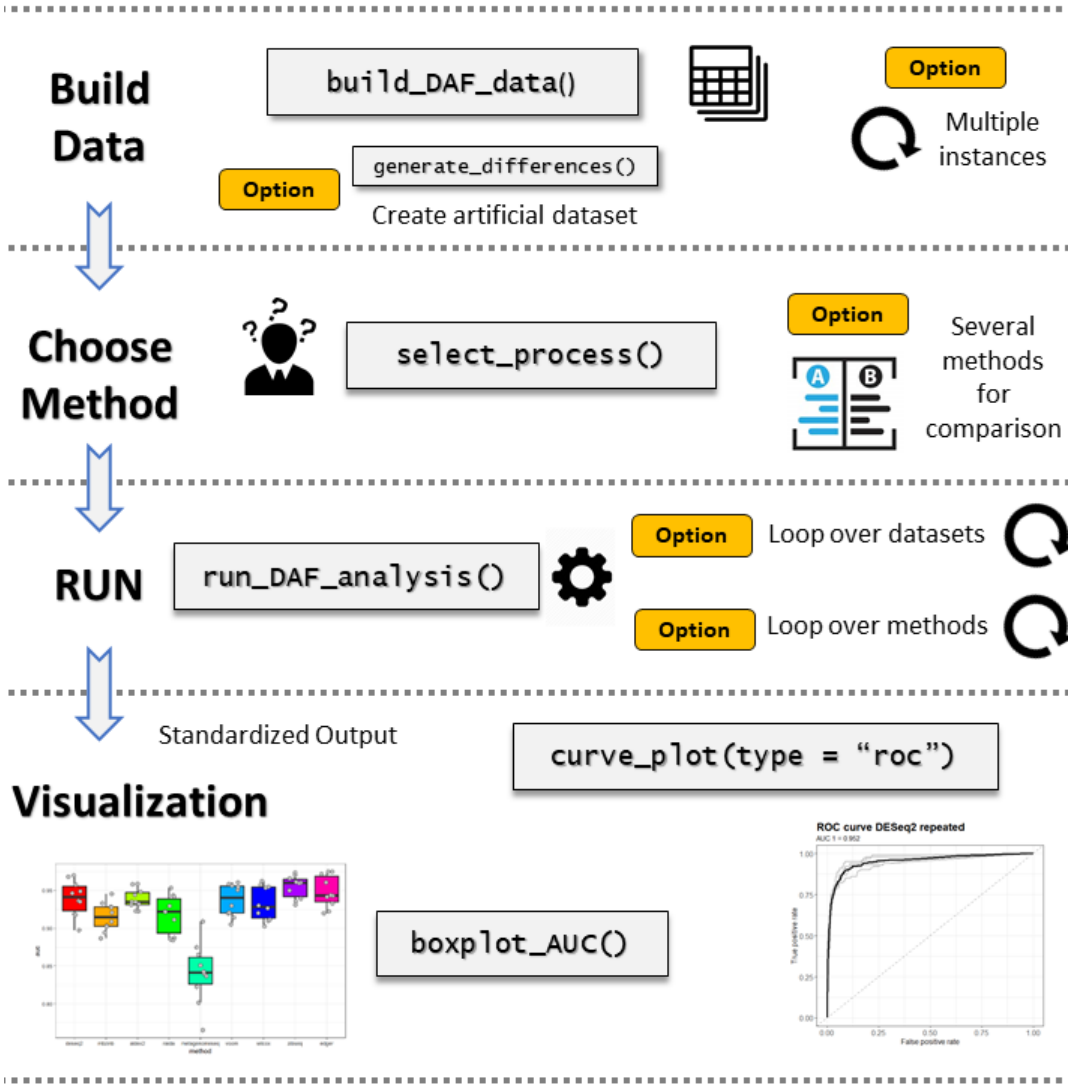

Figure 2: **metaDAF package - Worflow of the analysis**
Example of analysis workflow and option when using the metaDAF package for differential analysis, data generation, comparison and benchmark.

Then, one or several method is selected. All methods selected will analyze all datasets provided/generated. Thanks to the wrapping process, all the outputs have the same format and can then be used for visualization. The functions provided in the package draw Receiver Operating Characteristic curve (ROC curve), Precision-Recall Curves for any kind of analysis (repeated datasets, comparison of method, mix of both...). When comparing method over several datasets, the most informative plot remains the boxplot of a relevant metric such as Area Under Curve (AUC).

3.2.2 Dataset generation

Another package, firstly designed for RNA-seq analysis, offers a great tool for count table generation [Juntao et al., 2014]. The package, called `EDDA` for Experimental Design in Differential Analysis includes a `generateData()` function allowing a lot of flexibility in the generation process. It generates synthetic counts based on Negative Binomial distribution and/or multinomial distribution. The function can also take as input real experimental data and learn parameters from it. Hence, the artificial data will have a profile similar to the real one.
EDDA generates an abundance profile for a condition, then based on the fold-change distribution provided, generates the abundance profile for the other condition. According to the number of replicates and the variability specified, the entire dataset (count matrix) is built.

`generateData()` most interesting arguments are :

Argument	Role
ControlRep	Number of samples in the control group
CaseRep	Number of samples in the case group
EntityCount	Number of features
FC	Fold change type. It can be "Norm(mu,sigma)", "logNorm(mu,sigma)", "log2Norm(mu,sigma)" or "Unif(a,b)"
perDiffAbund	Percentage of features beeing differentially abundant
numDataPoints	Number of data point (similar to sequencing depth)
SampleVar	Fix the value of the shape parameter of a gamma distribution used to compute count dispersion of a feature across samples

Table 1: Parameters for the `generateData()` function

With these parameters, a lot of different datasets are generated in order to compare the performances of all the methods.

3.3 Experimental design

The datasets for the benchmark are generated using sets of parameters. The experimental design choice was a factorial design which results in testing all possible combinations of these parameters. The values are presented in Table 2. The number of features and data points are fixed to 1,000 and 20,000 and correspond to a library size (total count per sample) of 2,000,000. Such number has been estimated from good quality shotgun metagenomic dataset privately used in MGP. On top of that, each parameter set is used to generate 50 replicated datasets, allowing to average results. This leads to $6 \times 5 \times 4 \times 5 \times 3 = 1800$ parameters sets for a total of 90,000 datasets.

Parameter	Values
Sample size	10, 20, 50, 100, 200, 500
Disequilibrium between groups	10, 20, 30, 40, 50% of samples in group 1
Effect size (FC between groups)	Normal distribution with μ = 1, 2, 5, 10
Percentage of DAF	1, 2, 5, 10, 20%
Dispersion	`generateData()` values: "low", "medium" or "high"
Number of features	1000
Number of data points	20000

Table 2: Experimental grid design

The dataset generation and all the analysis were run on the ProActive cluster of MetaGenoPolis (High Performance Computing - HPC). The execution time is variable between methods and dataset. DESeq2, edgeR, voom, Wilcoxon and mbzinb are really fast. ALDEx2 and metagenomeSeq take a bit more time. ZIBseq and RAIDA are particularly time-consuming. Besides, time consumption also depends on the size of the dataset.

3.4 Performances metrics

Methods performances will be assessed using several metrics. Using prediction results, a Receiver Operating Characteristic curve (ROC curve) curve can b built. It describes the prediction ability with variations of a p-value threshold. It involves two metrics :

True Positive Rate also called Recall or Sensitivity. $TPR = \frac{TP}{TP+FN}$

False Positive Rate also called Fall-out. $FPR = \frac{FP}{FP+TN}$

The information conveyed by such curve can be summarized by the Area Under Curve or AUC ranging for 0 (worse) to 1. An AUC of 0.5 corresponds to a pure random prediction.
Precision-Recall curves are also used and act in a similar way, they involve Recall described above and *Positive Predictive Power* also called Precision. $PPV = \frac{TP}{TP+FP}$

4 Results

Because of the experimental design chosen, results quickly become overwhelming. One of the reasons why is that they suffer from the curse of dimensionality, the experimental. Since the outcome depends on various factors (all the parameters used for dataset generation) and we have access to 90,000 ROC and PR curves. We decided to summarize the results of each replicate by a single numeric value: the Area Under the ROC curve (AUC).

4.1 Area Under Curve - True Positive and False Positive Rates

As seen in figure 3, the overall distribution of the AUC values across all datasets does not bring any relevant information. It can however be seen that all methods, except for metagenomeSeq, are quite good at detecting DAF for the parameters values we evaluated. We then focus and use small multiple plots to split the results and assess the effect of each parameter on accuracy.

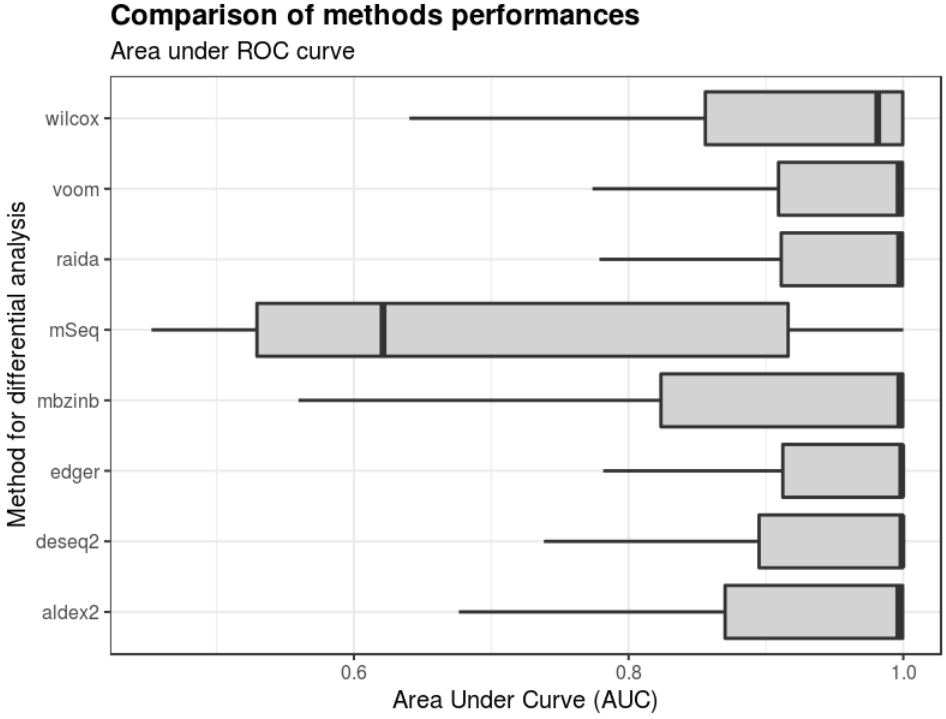

Figure 3: **Overview of the results - AUC distribution**
Distribution of the area under the ROC curve across all the 90,000 datasets generated.

For example, by separating the effect of dispersion, effect size and sample size as done on figure 4 patterns appear even if some information is still summarized. It appears that with high effect size (mean fold-change between groups of 5 or 10), the dispersion or the sample size do not really matter anymore. Since the differences between groups are high, they are easy to detect. The AUC are high (near 1). Similarly and as expected, the overall performance increases with

sample size. Interestingly, it increases sharply when from 10 to 50 samples but less so from 50 to 500 samples.

The granularity of the results can be again increased as in figure S3. In this figure, metagenomeSeq results are removed. At medium or high dispersion methods tend simply to become more efficient (higher AUC) as the sample size increases. The exception is the mbzinb method whose particularity is the use an omnibus test including dispersion as one of the variable studied (as well as abundance and prevalence), it is expected that its sensitivity to dispersion shifts differs from other methods.

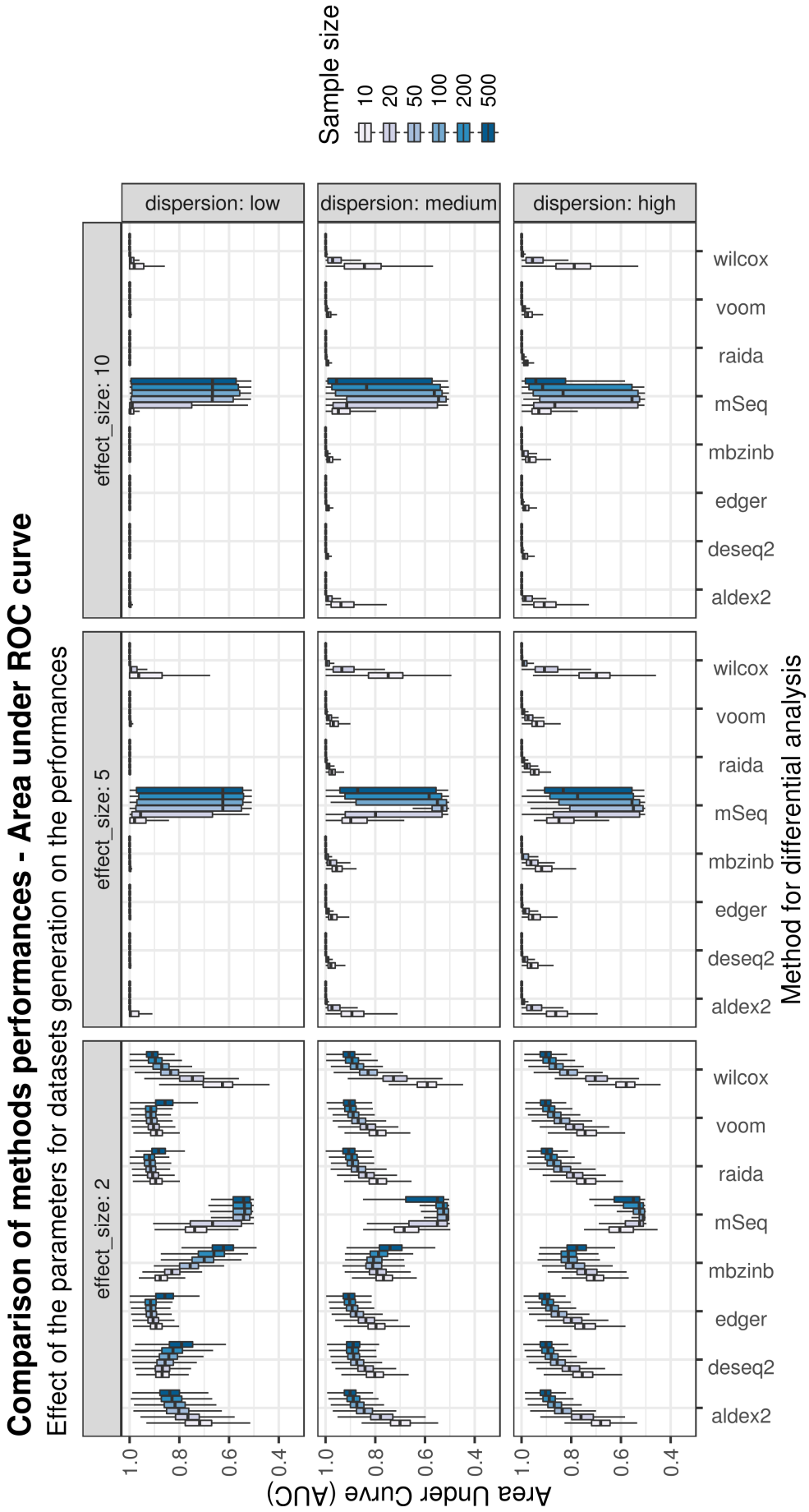

Figure 4: **Scattered view of the results - AUC distribution**

Influence of effect size, sample size and dispersion on the area under the ROC curve.

4.2 About False positives

Apart from the AUC metric, focusing on the evolution of TPR with regards to the FPR, it is also relevant to also focus on a raw FPR value at a given p-value threshold. This represents the amount of false positives in the features considered significant. For this, a p-value threshold is necessary (here the consensus of 0.05 is chosen). To the same extent of AUC, the overall distribution in Figure 5 do not discriminate methods a lot: all methods have a nominal FDR rate well below the target of 0.05.

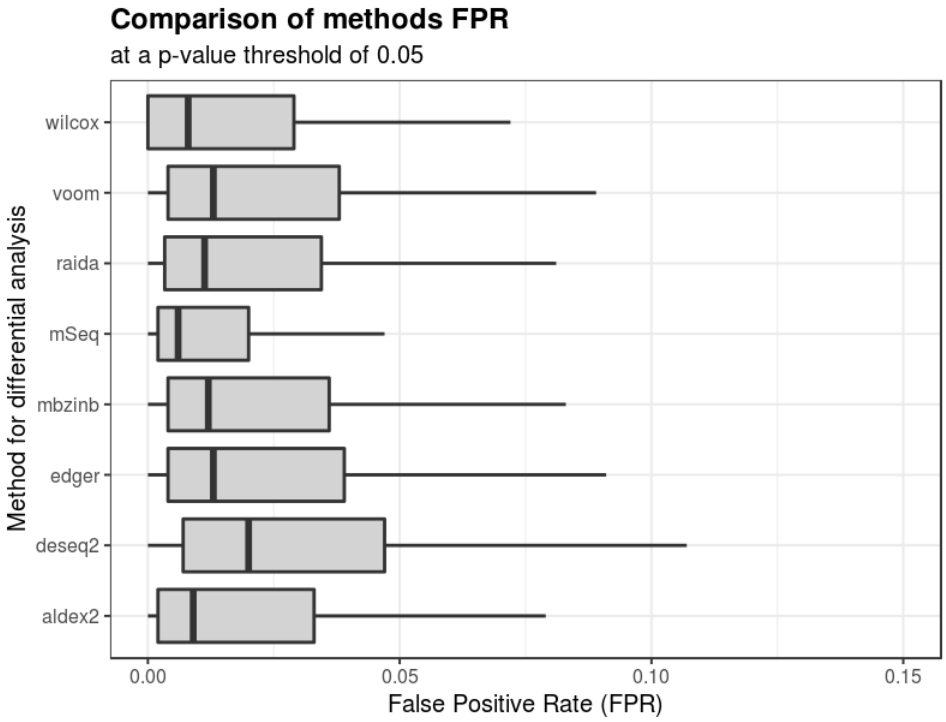

Figure 5: **Overview of the results - FPR distribution**
Distribution of the False Positive Rate at a p-value threshold 0.05 across all the 90,000 datasets generated.

In this case again, results must be scattered. Figure 6 shows the influence of effect size, dispersion and the percentage of truly differentially abundant features on the FPR. The sample size does have an impact. The bigger the dataset, the greater the FPR. A higher percentage of truly DAF also increases it.

Strikingly, the metagenomeSeq method seemed to perform badly when one focuses on the AUX metric. it appears that the method is particularly good at controlling the FPR which remains most of the time lower than for the other methods. However. the dispersion parameter seem to have a slight impact on the overall FPR of metagenomeSeq results. Increasing the dispersion makes the FPR slightly decrease.

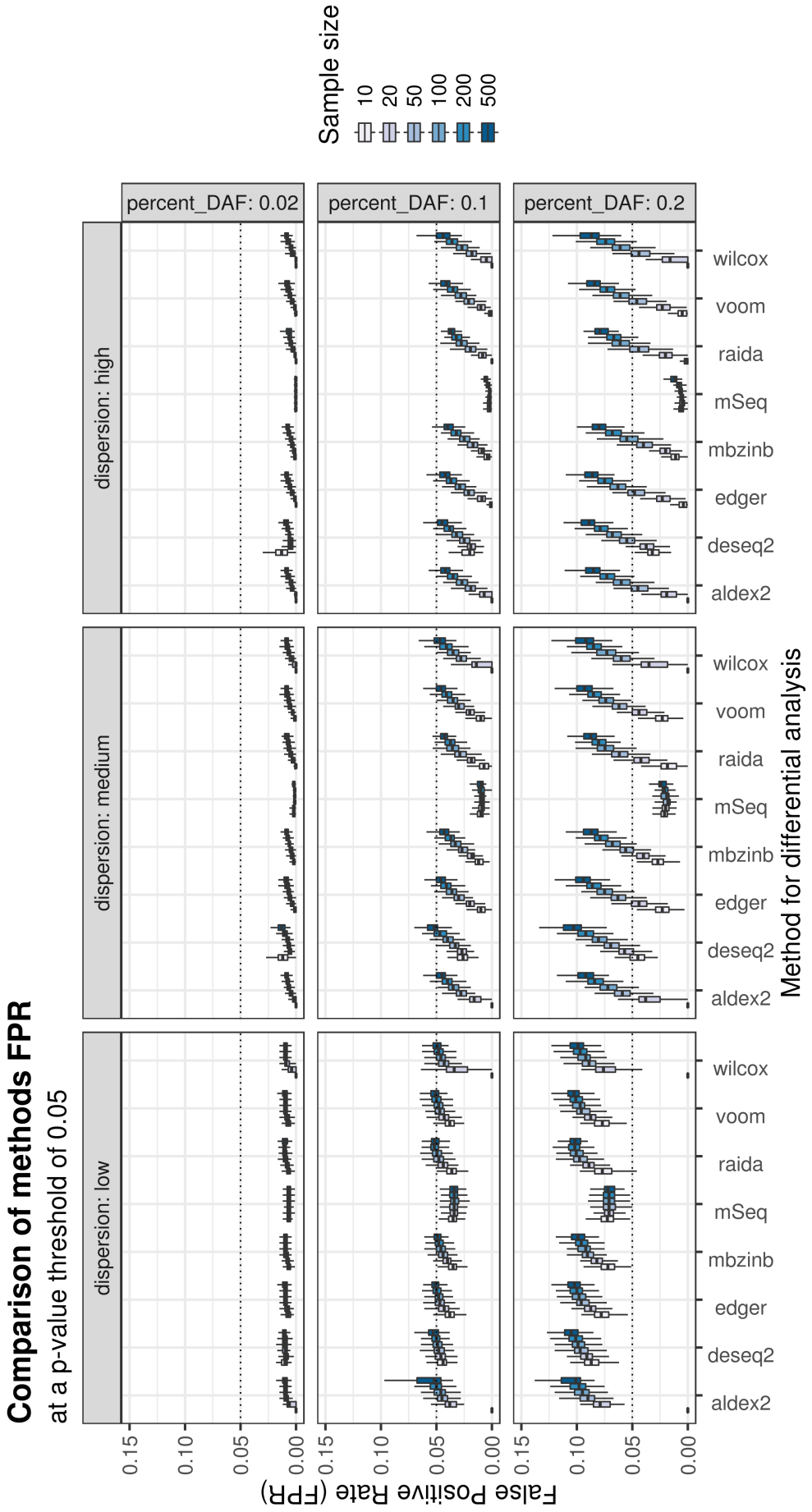

Figure 6: **Scattered view of the results - FPR distribution**

Influence of effect size, sample size and dispersion on the FPR. The dotted line is the expected FPR at a p-value threshold 0.05.

4.3 Example on real dataset

The comparison of methods is extended to results with "real" (non-simulated) datasets. Metagenomics datasets are not always easy to find. Moreover, datasets can have various formats and are not always available as a count table.

The `curatedMetagenomicData` for BioConductor [Pasolli et al., 2017] offers a lot of datasets as count tables and provides curated metadata allowing for fast analysis. In the next part, focus is put on a dataset of 199 individuals (66 healthy, 133 CRC) and 604 features (bacterial species) first published in [Zeller et al., 2014] .

Using non-simulated dataset prevents us from computing metrics such as AUC or FPR because true positives (features that are truly differentially abundant between groups) are unknown. Only the raw results are comparable. A common point between methods is that they all provide a p-value for each feature. However, it would be irrelevant to compare the results using a p-value threshold to assess what is significant or not [Amrhein et al., 2019]. Instead, the features with the top 10 % lowest p-values are kept and compared across methods.

Because of the high number of methods, we don't compare results across methods using Venn diagrams but resort to the `UpSetR` package, which provides alternative way to visualize intersection among many groups, as shown in Figure 7.

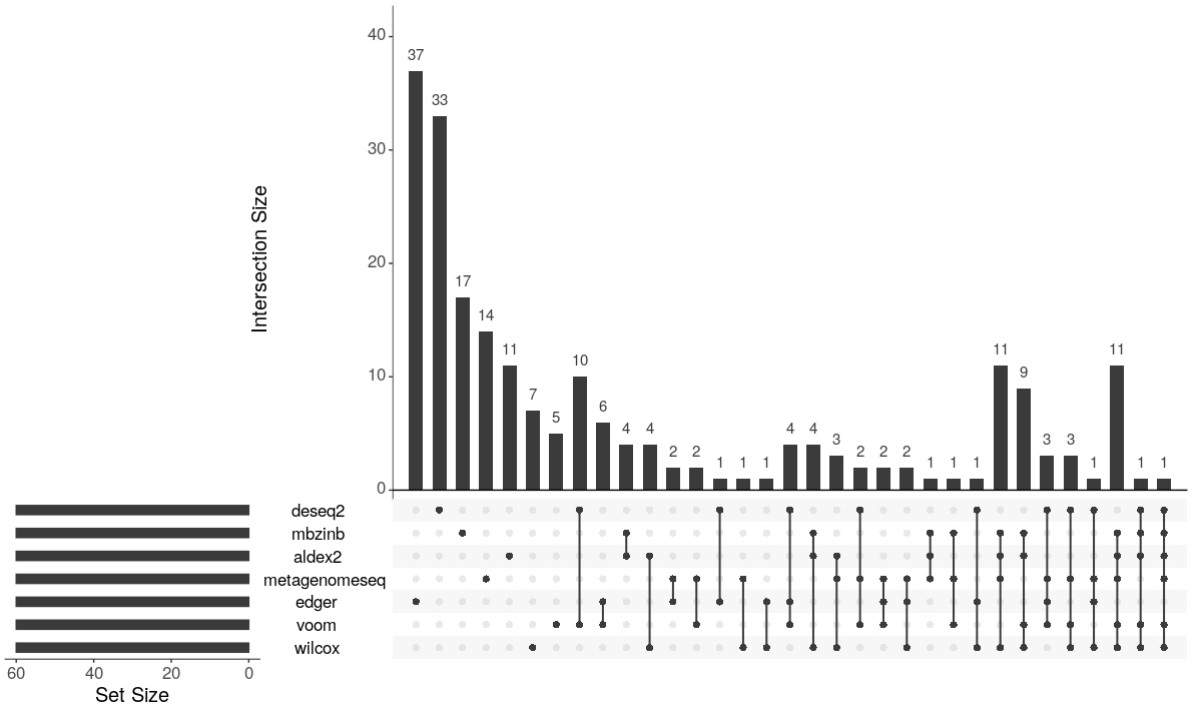

Figure 7: **Intersection of methods results**

Distribution of the size of the intersection between top features across methods. Only the 60 lowest p-values are kept for each method. The RAIDA method failed to complete on this example due to technical issues.

Some methods display high dissimilarities. More than half of the top features from DESeq2 or edgeR are specific to them. However, the figure conveys the idea that, albeit for those features

specific to DESeq2 and edgeR, most of the features are detected by more than one method. In regards to the same figure for sets of features defined using a 0.05 p-value threshold, it appears relevant to not use raw p-values. Intersection results with sets of features defined by a p-value threshold of 0.05 are available as Figure S1.

An easier way of comparing significant features among method is to use the Jaccard similarity measure as a dissimilarity (1 - Jaccard similarity) [Jaccard, 1901] and group the using hierarchical clustering (Figure 8). Strikingly two clear groups appear: on the one hand edgeR, DESeq2, metagenomeSeq and voom are all based on Generalized Linear Model using specific distributions (Negative Binomial, ZIL) and focus on the "mean value" of the distribution, on the other hand mbzinb performs an omnibus test (mean, dispersion, prevalence), Wilcoxon is a non-parametric test and ALDEx2 focuses on the compositional aspect of the data.
The same dendrogram with sets of features defined by a p-value threshold of 0.05 are available as Figure S2. Here, edgeR, DESeq2 and voom, using GLM are separated in one cluster. metagenomeSeq and nbzinb, specifically designed for metagenomics and using zero-inflated distribution are another cluster.

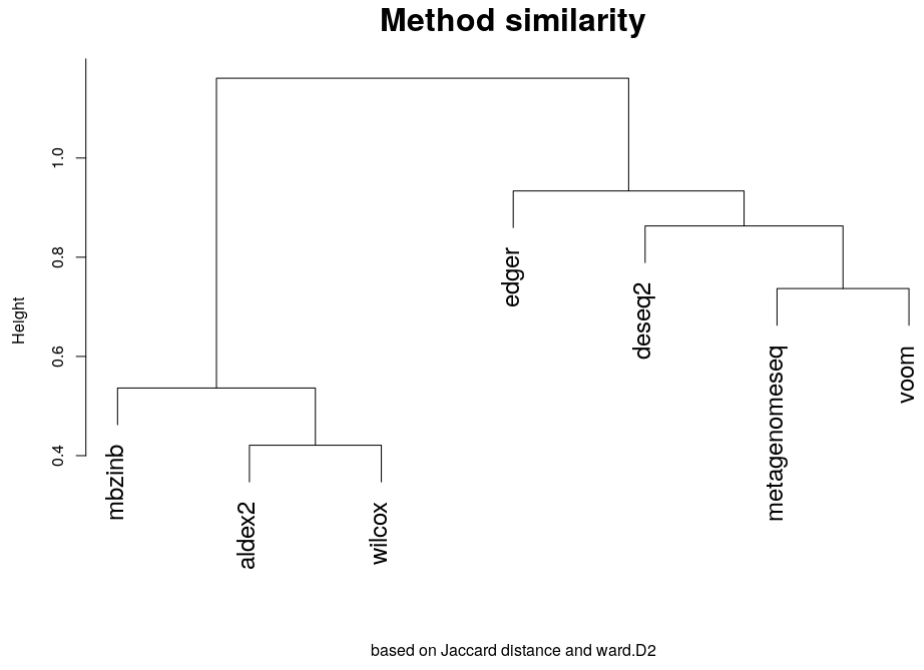

Figure 8: **Similarity of methods results**
Hierarchical cluster analysis drawn using Jaccard distance between the sets of the features with the top 60 lowest p-values for each method. Aggregation criterion is Ward's method. `ward.D2` in `hclust()` function.

5 Discussion and possible improvements

As shown previously, the results of the simulation study constitute a treasure trove that we only barely explored. One can still extract a lot from the raw results to answer more specific questions. For example, another metrics apart from AUC and FPR could be computed and compared across methods and datasets. Or the comparison across artificial datasets could be more qualitative instead of quantitative and focus on which features end up significant using which methods, which overlaps and differences are found between methods.
Then, thanks to the `metaDAF` package developed for this project, other methods could be added to the comparison. In that case, results can be easily concatenated. Further improvements could also be added to the `metaDAF` package for methods able to compare more than 2 groups or to take into account confounding variables.

Moreover, the experimental design chosen, even if already computationally demanding, can be still extended. Parameters such as library size (sequencing depth/number of data point) or the number of outliers are considered equals across all datasets. Such parameters could also impact the final results and their effect is not represented in this study. However, adding such parameters will make the number of datasets soar. Indeed, for each value of library size tested, 90,000 datasets should be generated or the experimental grid has to be redesigned. A more practical approach consists in (i) identifying a reasonable parameter set (neither too hard nor too easy) and (ii) varying sequencing depths across samples for that parameter only.

Finally, two more points were eluded in this report but are of great interest. The first one is the running time of each method. Even if we dealt with it during the numerical experiments, we did not keep precise records as different paramaters sets were performed on different nodes of the MGP HPC facility. Besides, time was a limiting factor regarding the amount of runs needed for this benchmark. In simple analysis (as the one the methods are designed to be used for), a couple of runs is enough.
The second point eluded is the combinations of parameters in dataset generation that makes a method unable to run. For example, a sample size of 10 with a group disequilibrium of 10% leads to a 1-individual-vs-9 analysis. Some of the methods are not able to run under these conditions.

As presented in section 3.1.1, the project has been conducted so that other methods could be easily added. It would however be interesting to pursue this benchmark in order to cross-validate previous results stating an improvement due to the use of zero-inflated models [Jonsson et al., 2018].

6 Conclusion

Metagenomics data analysis is still a new field of research and new tools are constantly created or adapted from previously existing ones. The problem of identifying the differentially abundant features in a count table matrix is only one of the many computational and statistical issues researchers have to deal with.

This project endeavours to provide a summary of the performances of the most commonly used methods. It appears that such performances are greatly influenced by the datasets characteristics (sample size, effect size, dispersion of counts...) It appears that in a lot of cases, a Wilcoxon test is not on par with more complex methods involving data transformation. However, in the case of small datasets, such a non-parametric method should be avoided. RAIDA, voom, edgeR or DESeq2 are preferred as they have better statistical power.
Once again, there is no golden standard. It appears that on real dataset, methods results could differ and it would be an interesting decision to use an ensemblist strategy and run not only one but several well-performing method and compare the consensus and disagreement within their outputs. In such case, involving a method controlling its FPR such as metagenomeSeq for corroboration would be wise.

This benchmark presents a first step and the data generated can still be stirred and visualized from a lot of different angles. One advantage here is that the package developed offers a good starting material for adding up new the methods that are yet to be created or for extending the benchmark to datasets pre-processing, library sizes, features number, effect of confounding variables...

7 References

[Aitchison, 1982] Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

[Amrhein et al., 2019] Amrhein, V., Greenland, S., and McShane, B. (2019). Scientists rise up against statistical significance.

[Anders and Huber, 2010] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

[Banerjee et al., 2019] Banerjee, K., Zhao, N., Srinivasan, A., Xue, L., Hicks, S. D., Middleton, F. A., Wu, R., and Zhan, X. (2019). An Adaptive Multivariate Two-Sample Test With Application to Microbiome Differential Abundance Analysis. *Frontiers in Genetics*, 10.

[Beszteri et al., 2010] Beszteri, B., Temperton, B., Frickenhaus, S., and Giovannoni, S. J. (2010). Average genome size: a potential source of bias in comparative metagenomics. *The ISME Journal*, 4(8):1075–1077.

[Caporaso et al., 2010] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336.

[Cardona et al., 2012] Cardona, S., Eck, A., Cassellas, M., Gallart, M., Alastrue, C., Dore, J., Azpiroz, F., Roca, J., Guarner, F., and Manichanh, C. (2012). Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC microbiology*, 12(1):158.

[Chen et al., 2018a] Chen, J., King, E., Deek, R., Wei, Z., Yu, Y., Grill, D., and Ballman, K. (2018a). An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics*, 34(4):643–651.

[Chen et al., 2018b] Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018b). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*, 6:e4600.

[Ehrlich, 2016] Ehrlich, S. D. (2016). The human gut microbiome impacts health and disease. *Comptes Rendus Biologies*, 339(7-8):319–323.

[Fernandes et al., 2014] 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(1):15.

[Gloor and Reid, 2016] Gloor, G. B. and Reid, G. (2016). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62(8):692–703.

[Handelsman et al., 1998] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, 5(10):R245–R249.

[Ho et al., 2019] Ho, N. T., Li, F., Wang, S., and Kuhn, L. (2019). metamicrobiomeR: an R package for analysis of microbiome relative abundance data using zero-inflated beta GAMLSS and meta-analysis across studies using random effects models. *BMC Bioinformatics*, 20(1):188.

[Jaccard, 1901] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

[Jonsson et al., 2016] 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(1).

[Jonsson et al., 2018] Jonsson, V., Österlund, T., Nerman, O., and Kristiansson, E. (2018). Modelling of zero-inflation improves inference of metagenomic gene count data. *Statistical Methods in Medical Research*.

[Jovel et al., 2016] Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., Perry, T., Kao, D., Mason, A. L., Madsen, K. L., et al. (2016). Characterization of the gut microbiome using 16s or shotgun metagenomics. *Frontiers in microbiology*, 7:459.

[Juntao et al., 2014] Juntao, L., Huaien, L., Burton, C., and Nagarajan, N. (2014). Edda: experimental design in differential abundance analysis. *Genome Biol*, 15:527.

[Law et al., 2014] Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29.

[Lee et al., 2017] Lee, C., Lee, S., and Park, T. (2017). A comparison study of statistical methods for the analysis metagenome data. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1777–1781.

[Ley et al., 2006] Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: Human gut microbes associated with obesity. *Nature*, 444(7122):1022–1023.

[Love et al., 2014] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12).

[Mandal et al., 2015] Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health & Disease*, 26(0).

[McMurdie and Holmes, 2014] McMurdie, P. J. and Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology*, 10(4):e1003531.

[Morgan et al., 2012] Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., Bousvaros, A., Korzenik, J., Sands, B. E., Xavier, R. J., and Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):R79.

[Odintsova et al., 2017] Odintsova, V., Tyakht, A., and Alexeev, D. (2017). Guidelines to Statistical Analysis of Microbial Composition Data Inferred from Metagenomic Sequencing. *Current Issues in Molecular Biology*, pages 17–36.

[Oñate et al., 2018] Oñate, F. P., Le Chatelier, E., Almeida, M., Cervino, A. C., Gauthier, F., Magoulès, F., Ehrlich, S. D., and Pichaud, M. (2018). Mspminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *bioRxiv*, page 173203.

[Pasolli et al., 2017] Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., Huttenhower, C., Morgan, M., Segata, N., and Waldron, L. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, 14(11):1023–1024.

[Paulson et al., 2013] Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202.

[Peng et al., 2016] Peng, X., Li, G., and Liu, Z. (2016). Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data. *Journal of Computational Biology*, 23(2):102–110.

[Qin et al., 2012] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K., and Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60.

[Robinson et al., 2010] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[Robinson and Oshlack, 2010] Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

[Santiago et al., 2014] Santiago, A., Panda, S., Mengels, G., Martinez, X., Azpiroz, F., Dore, J., Guarner, F., and Manichanh, C. (2014). Processing faecal samples: a step forward for standards in microbial community analysis. *BMC microbiology*, 14(1):112.

[Sohn et al., 2015] Sohn, M. B., Du, R., and An, L. (2015). A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics*, 31(14):2269–2275.

[The Human Microbiome Project Consortium, 2012] The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.

[White et al., 2009] White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Computational Biology*, 5(4):e1000352.

[Wu et al., 2011] Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y.-Y., Keilbaugh, S. A., Bewtra, M., Knights, D., Walters, W. A., Knight, R., Sinha, R., Gilroy, E., Gupta, K., Baldassano, R., Nessel, L., Li, H., Bushman, F. D., and Lewis, J. D. (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science*, 334(6052):105–108.

[Zeller et al., 2014] Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology*, 10(11):766.

[Zhang et al., 2018] Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., and Yi, N. (2018). Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data. *Frontiers in Microbiology*, 9.

8 Supplementary Materials

R session

- R version 3.5.2 (2018-12-20), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=French_France.1252`, `LC_CTYPE=French_France.1252`, `LC_MONETARY=French_France.1252`, `LC_NUMERIC=C`, `LC_TIME=French_France.1252`
- Running under: `Windows >= 8 x64 (build 9200)`
- Matrix products: default
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: abind 1.4-5, baySeq 2.16.0, Biobase 2.42.0, BiocGenerics 0.28.0, BiocParallel 1.16.6, DelayedArray 0.8.0, DESeq 1.34.1, DESeq2 1.22.2, devtools 2.0.2, dplyr 0.8.0.1, EDDA 1.20.1, edgeR 3.24.3, forcats 0.4.0, foreach 1.4.4, GenomeInfoDb 1.18.2, GenomicRanges 1.34.0, ggplot2 3.1.1, glmnet 2.0-16, gplots 3.0.1.1, IRanges 2.16.0, jaccard 0.1.0, lattice 0.20-38, limma 3.38.3, locfit 1.5-9.1, MASS 7.3-51.3, Matrix 1.2-17, matrixStats 0.54.0, mbzinb 0.2, metaDAF 0.7.6, metagenomeSeq 1.24.1, pkgdown 1.3.0, protoclust 1.6.3, pscl 1.5.2, purrr 0.3.2, qvalue 2.14.1, RAIDA 1.0, RColorBrewer 1.1-2, Rcpp 1.0.1, readr 1.3.1, ROC 1.58.0, ROCR 1.0-7, S4Vectors 0.20.1, snow 0.4-3, stringr 1.4.0, SummarizedExperiment 1.12.0, tibble 2.1.1, tidyr 0.8.3, tidyverse 1.2.1, UpSetR 1.4.0, usethis 1.5.0
- Loaded via a namespace (and not attached): acepack 1.4.1, annotate 1.60.1, AnnotationDbi 1.44.0, assertthat 0.2.1, backports 1.1.4, base64enc 0.1-3, bit 1.1-14, bit64 0.9-7, bitops 1.0-6, blob 1.1.1, broom 0.5.2, callr 3.2.0, caTools 1.17.1.2, cellranger 1.1.0, checkmate 1.9.1, cli 1.1.0, cluster 2.0.8, codetools 0.2-16, colorspace 1.4-1, commonmark 1.7, compiler 3.5.2, crayon 1.3.4, data.table 1.12.2, DBI 1.0.0, desc 1.2.0, digest 0.6.18, foreign 0.8-71, Formula 1.2-3, fs 1.2.7, gdata 2.18.0, genefilter 1.64.0, geneplotter 1.60.0, generics 0.0.2, GenomeInfoDbData 1.2.0, glue 1.3.1, grid 3.5.2, gridExtra 2.3, gtable 0.3.0, gtools 3.8.1, haven 2.1.0, Hmisc 4.2-0, hms 0.4.2, htmlTable 1.13.1, htmltools 0.3.6, htmlwidgets 1.3, httr 1.4.0, iterators 1.0.10, jsonlite 1.6, KernSmooth 2.23-15, knitr 1.22, latticeExtra 0.6-28, lazyeval 0.2.2, lubridate 1.7.4, magrittr 1.5, memoise 1.1.0, modelr 0.1.4, munsell 0.5.0, nlme 3.1-137, nnet 7.3-12, pillar 1.3.1, pkgbuild 1.0.3, pkgconfig 2.0.2, pkgload 1.0.2, plyr 1.8.4, prettyunits 1.0.2, processx 3.3.0, ps 1.3.0, R6 2.4.0, RCurl 1.95-4.12, readxl 1.3.1, remotes 2.0.4, reshape2 1.4.3, rlang 0.3.4, roxygen2 6.1.1, rpart 4.1-13, rprojroot 1.3-2, RSQLite 2.1.1, rstudioapi 0.10, rvest 0.3.3, scales 1.0.0, sessioninfo 1.1.1, splines 3.5.2, stringi 1.4.3, survival 2.44-1.1, testthat 2.0.1, tidyselect 0.2.5, tools 3.5.2, withr 2.1.2, xfun 0.6, XML 3.98-1.19, xml2 1.2.0, xtable 1.8-3, XVector 0.22.0, yaml 2.2.0, zlibbioc 1.28.0

Intersection figures

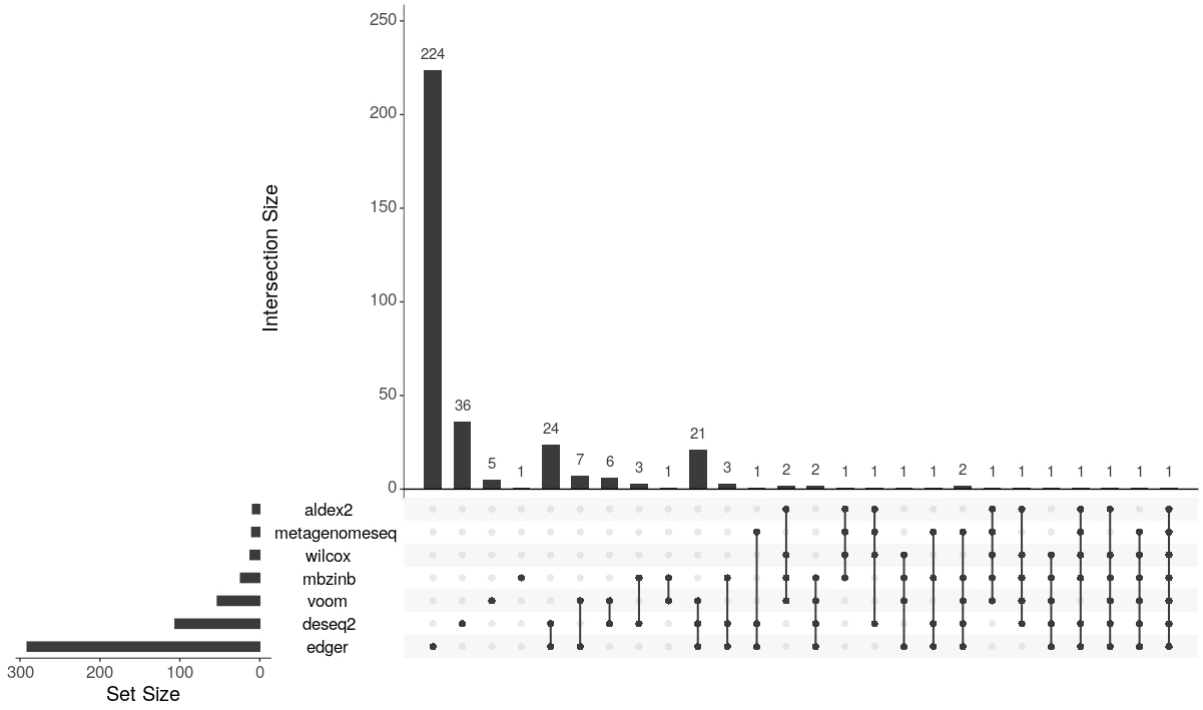

Figure S1: **Intersection of methods results**

Distribution of the size of the intersection between top features across methods. Only the features with p-values <= 0.05 are kept. In this example, the RAIDA method did not run because of technical issues.

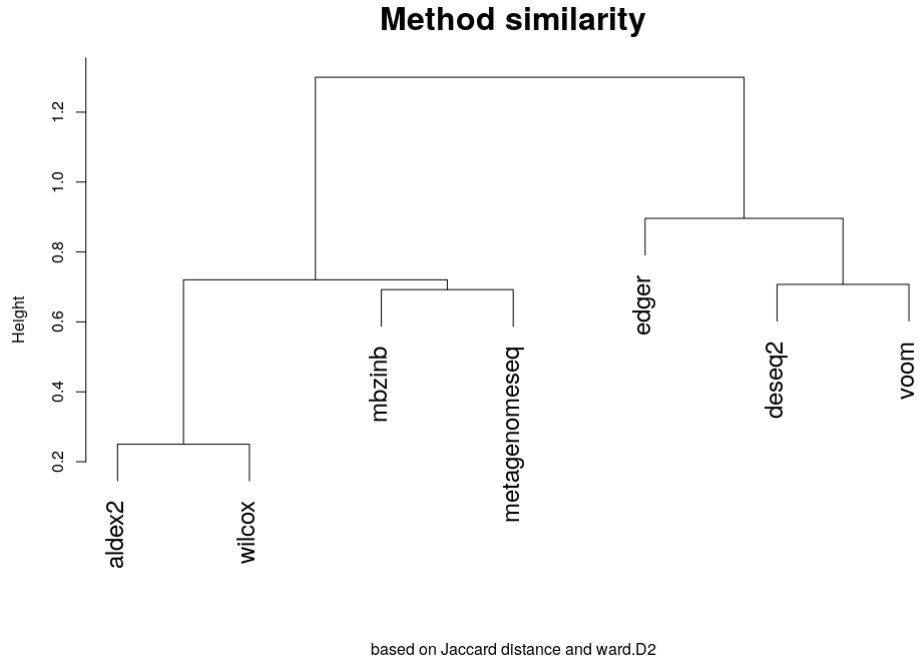

Figure S2: **Similarity of methods results**

Hierarchical cluster analysis drawn using Jaccard distance between the sets of the features with the p-values <= 0.05. Aggregation criterion is Ward's method. `ward.D2` in `hclust()` function.

AUC boxplot - Specific case

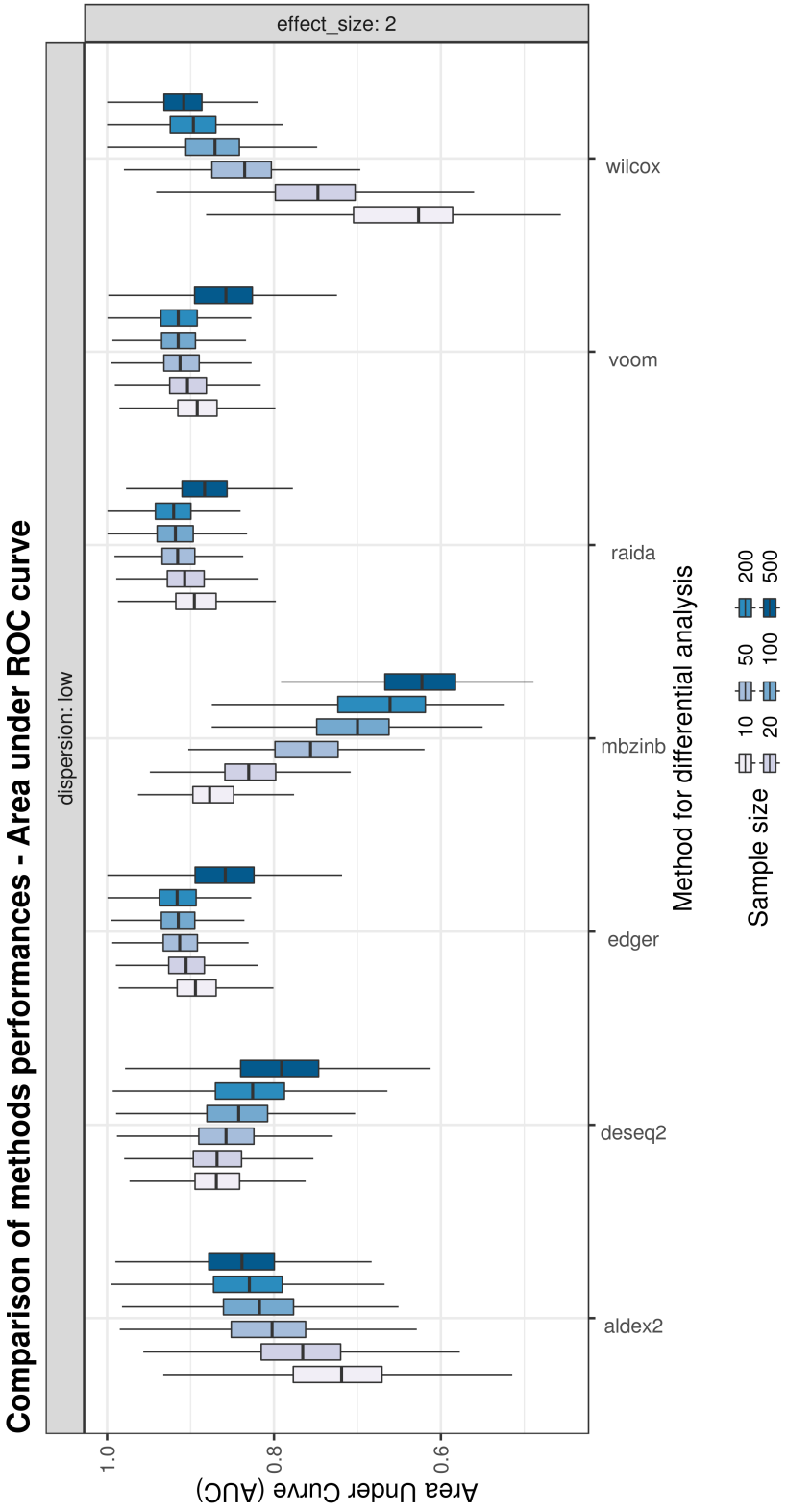

Figure S3: **Case specific AUC distribution**

Distribution of the Area Under Curve at a blocked effect size = 2 and low dispersion. It corresponds to a parameter set where the differences are not easy to detect.

Abstract

Metagenomics studies microbial communities by sequencing their genetic material. This is done by targeting either a marker-gene (barcoding) or all the genes present in the samples (shotgun sequencing). It has been extensively used to characterize taxonomic and functional profiles of many ecosystems including the human gut microbiota. In the later, the shifts in the abundancy of specific species are used as biomarkers for many biological or clinical conditions such as cancer, diabetes or inflammatory bowel disease. However a lot of technical difficulties arise when working with such data: high-dimension, noise, high sparsity, low number of replicates... Thus, detecting these shifts with satisfying precision and recall is challenging. Many statistical methods were introduced in the past decade, each trying to overcome specific constraint of the data. In this study we present a benchmark of the most commonly used methods for detecting differentially abundant features between samples. By using simulated data, statistical performances such as true/false positive or negative rates are assessed. In addition, results on real microbiome datasets are qualitatively discussed.

Keywords : comparative metagenomics, differential abundance, benchmark