CHALMERS
UNIVERSITY OF TECHNOLOGY

Experimental Design for Comparative Metagenomics

Investigating and optimising the trade-off between number of samples and sequencing depth

Master's Thesis in Engineering Mathematics and Computational Science

Sofia Conti
Martina Hermanova Billstein

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2020

Master's Thesis 2020

Experimental Design for Comparative Metagenomics

Investigating and optimising the trade-off between number of samples and sequencing depth

Sofia Conti
Martina Hermanova Billstein

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
Gothenburg, Sweden 2020

Experimental Design for Comparative Metagenomics
Investigating and optimising the trade-off between number of samples and sequencing depth

Sofia Conti
Martina Hermanova Billstein

Supervisors: Erik Kristiansson and Viktor Jonsson
Examiner: Erik Kristiansson, Department of Mathematical Sciences

Master's Thesis 2020
Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
SE-412 96 Gothenburg

Typset in LaTeX
Gothenburg, Sweden 2020

Experimental Design for Comparative Metagenomics
Investigating and optimising the trade-off between number of samples and sequencing depth
Sofia Conti and Martina Hermanova Billstein
Department of Mathematial Sciences
Chalmers University of Technology

Abstract

In comparative metagenomics, samples from different environments are compared with the aim to identify differentially abundant genes. It is important to have a sound experimental design in such studies, including a sufficiently large number of samples from each environment as well as a sufficiently high sequencing depth in each sample.

The aim of this master's thesis was to provide guidance on the required number of samples and sequencing depth for experimental designs in future comparative metagenomic studies. In order to do so, various experimental designs with different number of samples and sequencing depths were evaluated based on their statistical performance. For each design, a large number of artificial datasets were created by resampling real metagenomic data. Three real datasets were used and the analyses were conducted in R.

The performances of all the investigated designs were shown to improve when the effect size of the studied phenomenon was large as well as when the studied genes had high abundance or low variability. It was further found that the performance of the designs increased both with increasing sequencing depth and with increasing number of samples in each group. A sequencing depth of ten thousand reads was generally too low to yield an acceptable performance. Likewise, having only three samples in each group was found to be too few unless the studied genes had high abundance or low variability. The main result was that the performance improved more with increasing number of samples than with increasing sequencing depth. However, when taking the economic aspect into account, a larger amount of samples became less profitable due to the high sequencing cost per sample. A final conclusion was that an experimental design may be less extensive and use fewer samples if the effect size is large or if the studied genes have high abundance or low variability.

Keywords: bioinformatics, performance, statistical power, economic impact, false discovery rate (FDR), effect size, gene abundance, gene variability, differentially abundant genes (DAGs), R.

Acknowledgements

The development of this master's thesis would not have been possible without the help and support that we have received during our work. We would first of all like to thank our supervisor Erik Kristiansson for his steady encouragement and unceasing commitment to our work. His optimism and expertise within the field has been very helpful. We would also like to thank our supervisor Viktor Jonsson for his guidance and valuable inputs and ideas throughout the process. The research previously conducted by Erik, Viktor and their colleagues has also been of great importance when shaping our thesis. Finally, the effort and time that both Erik and Viktor have put into our work has been greatly appreciated and we are thus ever so grateful for their support and involvement.

June 10, 2020
Gothenburg

Sofia Conti

Martina Hermanova Billstein

Table of Contents

Abbreviations

AUC - Area Under the ROC Curve

DAG - Differentially Abundant Gene

FDR - False Discovery Rate

FP - False Positive

FPR - False Positive Rate

GLM - Generalised Linear Model

NGS - Next Generation Sequencing

OGLM - Overdispersed Poisson Generalised Linear Model

ROC curve - Receiver Operating Characteristic curve

TP - True Positive

TPR - True Positive Rate

List of Figures

List of Tables

1. Introduction

Microorganisms such as bacteria, fungi, single-cell eukaryotes and viruses are present in all habitats and are members of almost every ecosystem on earth. There are for example more bacterial cells in our bodies (10^{14}) than human cells (10^{13}) [1]. This shows that the key to understanding human conditions is not only to study the human genome, but also to study the genomes of human microbes [2]. It has for example been found that specific forms of bacteria in the human gut are linked to diseases such as diabetes and cancer [3, 4]. Studying and comparing the present microorganisms in different environments is thus necessary for understanding how genomic differences affect, and are affected by, both the biotic and the abiotic environment [2].

Metagenomics is the study of such microbial communities [5]. It provides a holistic view of the present organisms and genes in an environment by extracting and sequencing random DNA fragments that have been sampled directly from the environment. By quantifying the fragments in the samples, the organisms and molecular functions present in the microbial community can be discovered [2]. The quantification is performed by mapping the sequenced fragments in a sample against a reference database and then counting how many fragments that are mapped to each gene. The total number of gene counts in a sample is in this thesis referred to as the sequencing depth.

In comparative metagenomics, samples from different environments are compared with the aim to identify differentially abundant genes (DAGs). The number of samples collected from each environment are referred to as the group size. The analysis of comparative metagenomic data is challenging due to a large number of genes as well as a large technical and biological variability in the data [6, 7]. These challenges make it hard to distinguish the often few DAGs that are caused by biological effects from genes that only differ due to random fluctuations [8]. It is thus crucial to have a sound experimental design in a comparative metagenomic study. This implies having a sufficiently large number of samples from each environment as well as a sufficiently high sequencing depth in each sample.

1.1 Previous studies

No exact relation between the required group size and sequencing depth could be found in literature. There is, however, an article that mentions how the dependence between these parameters influence the results in metagenomic studies. By comparing results from studies with different amount of samples, the article found indications that it is better to have more samples with a lower sequencing depth compared to having fewer samples with a higher sequencing depth [9]. The studies with larger group sizes had a higher performance compared to studies with approximately the same amount of sequenced fragments distributed over fewer samples with higher sequencing depths [9].

Another article states that the number of samples and sequencing depth required to detect significant differences between sample groups will depend on several factors [10]. Examples of such factors are how consistent different samples from the same environment are, the inherent microbial diversity within each sample and the effect size of the phenomenon being studied. The article further states that decisions regarding the number of samples and sequencing depth can be guided by previous studies in the same type of environment, if such studies exist [10].

1.2 Objectives

This master's thesis aims at providing guidance on experimental designs for future comparative metagenomic studies. In order to do so, various designs with different number of samples and sequencing depths are evaluated.

More specifically, the issues that are investigated are:

1. In what way does the sequencing depth and the group size, i.e. number of samples, affect the performance of an experimental design in comparative metagenomics?
2. How do the requirements of the experimental design change depending on the circumstances of the study?
3. Given these findings, how should the experimental design be constructed when also taking the economic aspect into consideration?

An additional goal of the thesis is to make it as reproducible as possible in order to allow for future validation of the generated results and conclusions.

2. Theory

This chapter presents the basic concepts of metagenomics and the theory behind the statistical methods used in this master's thesis.

2.1 Metagenomics

Only a small percentage of all microorganisms found in nature can be cultivated in laboratories [11]. Hence it is impossible to study most species *in vitro*. Furthermore, microorganisms are often organised in complex microbial communities where interactions occur both between the present species and between the species and their habitat [2]. Such interactions are hard to mimic in laboratories. Thus, besides not being able to cultivate the majority of microbial species, classic genomic studies often fail to report microorganism interactions. Classic genomics also fails to represent the genomic variance and biological functions present in nature [2].

In metagenomics, DNA fragments are randomly sampled directly from natural environments such as soil, the human gut or the ocean [5]. The microorganisms are in this way studied in their natural state and, since every sample contains DNA from many different species, the collective genome of the entire microbial community is studied at once. This collective genome is called the *metagenome* [12]. The advantages of metagenomics is that it is culture-independent and that information about the present microorganisms in an environment, their biological functions and their interactions can be derived [2].

Comparative metagenomics compares groups of samples from different environments with the aim to identify differentially abundant genes (DAGs). DAGs are detected by statistically evaluating the relative abundance, which is obtained by comparing gene abundances in different sample groups [13]. Examples of environments that can be compared are polluted and pristine sites or the human microbiome of sick and healthy individuals [8]. By identifying genes that vary in abundance between the environments, important differences in community structure, diversity and biological functions can be identified [6].

A main approach in metagenomics is shotgun sequencing where the total DNA content in a sample is sequenced. The resulting data consists of DNA fragments, or reads, corresponding to random segments of all the genes present in the metagenome [8]. The reads are matched to known genes in a reference database in a process called *binning*. The abundance of each gene is derived by counting the number of matches for the gene [14].

The field of DNA sequencing, and in particular the field of metagenomics, has grown substantially over the last two decades thanks to the breakthrough in next generation sequencing (NGS) [15, 16]. NGS has enabled a huge progress in terms of speed, read length and throughput along with a reduction in sequencing costs [16]. This has resulted in greater volumes of produced metagenomic data. The drawback of NGS produced data is that it requires short DNA fragments, making the

data fragmented [16]. Apart from being fragmented, metagenomic data contains large technical and biological variability [8]. This, together with the large number of genes and species present in the metagenome, poses a challenge when analysing the data [6, 7].

2.2 Statistical methods

The performance of the different experimental designs are in this master's thesis assessed and compared by measuring the statistical power when identifying DAGs. The main R package used for identifying DAGs is *DESeq2*. The basic concepts of this package are described in Section 2.2.1. Another method for identifying DAGs is the use of overdispersed Poisson generalised linear models (OGLMs) followed by an F-test. The theory of this approach is briefly described in Section 2.2.2.

The statistical power of a study is the likelihood that the study detects an effect when there is an effect to be detected. In other words, the power is represented by all the correctly rejected null hypotheses. Various measures of statistical power are used in this thesis, which are presented in Section 2.2.3. In a study where several tests are performed, errors caused by multiple testing become a problem. In order to control these errors, the false discovery rate (FDR) is estimated. This is further explained in Section 2.2.4.

2.2.1 DESeq2

The R package DESeq2 performs differential analysis of count data [17]. It was originally constructed to analyze RNAseq data but has also become one of the most commonly used packages for analysing metagenomic data [13]. DESeq2 is applied on a count matrix, K, where each row i represents a gene, and each column j represents a sample. The matrix elements, called counts, are denoted by K_{ij} and indicate the number of reads that have been mapped to a gene in a sample. The counts are modelled with a generalised linear model (GLM) of the negative binomial family with a logarithmic link [17]. GLMs are extensions of ordinary linear models, where the dependent variable does not need to be normally distributed [18].

The negative binomial distribution, with mean μ_{ij} and dispersion α_i, that is used to model the read counts, K_{ij}, is given by

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i) \ . \tag{2.1}$$

The mean is calculated according to

$$\mu_{ij} = s_{ij} q_{ij} \ , \tag{2.2}$$

where q_{ij} is a quantity proportional to the concentration of DNA fragments from gene i in sample j and s_{ij} is the normalisation factor. By default, the normalisation factors are considered constant within a sample. Thus,

$$s_{ij} = s_j \ , \tag{2.3}$$

where s_j accounts for differences in sequencing depth between samples.

The normalisation factors are estimated with a median-of-ratios method

$$s_j = \underset{i}{\text{median}} \frac{K_{ij}}{\left(\prod_{j=1}^{m} K_{ij}\right)^{1/m}} \ , \tag{2.4}$$

where m is the number of samples.

The logatithmic link function used in DESeq2 is

$$\log_2(q_{ij}) = \sum_r x_{jr}\beta_{ir} \ , \tag{2.5}$$

where x_{jr} is the element in the so-called design matrix and β_{ir} is the logarithmic fold change for gene i and covariate group r. In the simplest case of comparative metagenomics, where only two environments are compared, the elements in the design matrix indicate from which environment each samples is collected. The GLM fit for such a study returns the $\log_2$ fold change between the two environments for each gene [17]. An analysis with DESeq2 also returns values such as p-values, Benjamini-Hochberg adjusted p-values and estimated gene-wise dispersions.

2.2.2 OGLM and F-test

Similarly to DESeq2, the overdispersed Poisson generalised linear model (OGLM) can also be used to perform differential analysis and is in this case a count-based method [8]. Unlike DESeq2 however, the OGLM assumes an overdispersed Poisson distribution instead of a negative binomial distribution [6].

The gene count of gene i in sample j is noted as Y_{ij} and is modelled with an OGLM according to

$$\log(Exp[Y_{ij}|x_j]) = \alpha_i + \beta_i x_j \ , \tag{2.6}$$

where α_i is the logarithm of the estimated mean count for gene i, β_i is the effect parameter describing the difference in abundance for gene i between the two sample groups and x_j indicates which group sample j belongs to [7].

The model is applied twice on each gene in slightly different forms; once with the effect parameter, β_i, and once without. The two models are then compared with an F-test where the resulting p-value indicates whether the models differ significantly or not. If the models differ, the gene is identified as a DAG [7]. The p-values obtained for each gene can then be adjusted with the Benjamini-Hochberg procedure.

2.2.3 Measures of statistical power

In a comparative metagenomic study of two environments, the identification of DAGs is a binary classification problem. The aim of such a problem is to separate objects in a dataset into two classes of either positives or negatives depending on factors in the data [19]. For a comparative metagenomic study, the aim is thus to identify genes as DAGs (positives) or non-DAGs (negatives) depending on the gene abundances in the samples and the environment from which the samples were collected.

The null hypothesis and the alternative hypothesis are in this case stated as follows for each tested gene;

$$H_0 : \text{The gene abundance is equal in both environments}$$
$$H_A : \text{The gene abundance is not equal in both environments}$$

[20].

Real positives correspond to cases where H_0 is false, while real negatives correspond to cases where H_0 is true [19]. The real classes are usually unknown, and the predicted classes are thus instead based on whether H_0 is rejected or not. Rejecting H_0 results in a positive prediction, while accepting H_0 results in a negative prediction. If H_0 is rejected when it is indeed false, the outcome is a True Positive (TP). False Positives (FPs) correspond to cases where H_0 has been rejected even though H_0 is true [19]. In other words, TPs correspond to genes which differ in abundance between the two environments and that are identified as such, while FPs correspond to genes that are identified to differ even though they do not [20].

The number of TPs and FPs in a study may not be directly comparable if the classes are unbalanced. Thus, their rates can be used instead. The true positive rate (TPR) is obtained by

$$\text{TPR} = \frac{\text{Number of TPs}}{\text{Number of cases where } H_0 \text{ is true}} \tag{2.7}$$

and similarly the false positive rate (FPR) is obtained by

$$\text{FPR} = \frac{\text{Number of FPs}}{\text{Number of cases where } H_0 \text{ is false}} \tag{2.8}$$

[19].

In this thesis, the differential analysis serves as a classifier which, rather than directly identifying genes as positives or negatives, returns a p-value for each gene. The p-values indicate how significantly the genes differs in abundance between the two sample groups in the dataset. This output is continuous, meaning that a classifier boundary, or threshold, has to be selected in order to separate positives from negatives [19]. The threshold corresponds to the significance level, α, at which genes with p-values lower than α are identified to differ between the sample groups and are thus classified as positives.

For a strict significance level, i.e. a low α, the proportion of TPs is likely to be higher than for a less strict significance level. However, at a strict level it is also less likely that all positives are identified, meaning that the proportion of TPs is large but that the amount is small. It is desired to have as many TPs as possible without allowing too many FPs [20].

An alternative to using a specified significance level is to use receiver operating characteristic curves (ROC curves) which visualise the power of a study at different thresholds [21, 22]. This is done by plotting the TPR and FPR values against each other, where the values are obtained at decreasingly strict thresholds of significance [19, 21]. The resulting graph will have two axes raging from 0 to 1, as seen in Figure 2.1 where three examples of ROC curves are presented.

The ROC curves are step functions, where each step corresponds to an increase in either TPR or FPR [19]. A step upwards corresponds to an increase in the TPR, meaning that, for the new threshold, a gene has correctly been identified as a positive. Reversely, a step horizontally implies that the FPR value has increased by classifying an FP. The ROC curve for a perfect classification, visualised by the green line in Figure 2.1, would immediately reach a TPR value of 1, correctly identifying all positives before identifying any negatives as false positives. The purple line in the figure represents a bad classifier which classifies at complete random and thus results in the diagonal line from the bottom left to the upper right corner [19].

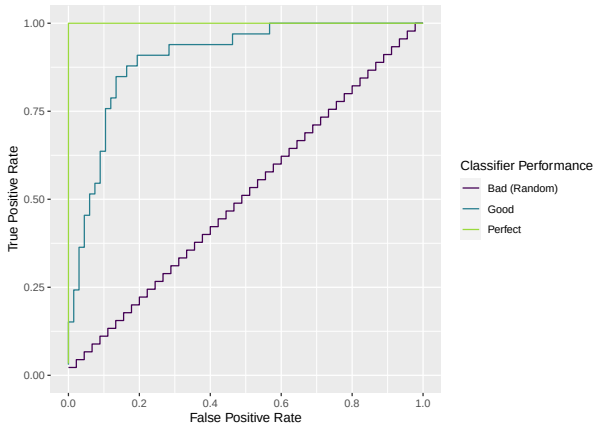


Figure 2.1: *Three ROC curves for classifiers with varying performances.*

A ROC curve can be summarised by computing the area under the ROC curve (AUC) [19, 21]. The AUC is a measure of how well a classifier distinguishes between classes and corresponds to the probability of correctly classifying a random object in a dataset, regardless of which class it belongs to [21]. This is useful when two classes are unequally balanced, which they often are in comparative metagenomics. The AUC is usually computed at an FPR value of 1, resulting in an area under the entire curve, AUC_{tot}. A value of 1 then means that the classification is perfect for the entire dataset, while a value close to 0.5 implies complete randomness [19, 21]. In the case of comparative metagenomics, it may however not always be of importance to look at the total area but rather focus on the beginning of the ROC curve. The reason for this is that metagenomic studies typically only focus on the most significant genes where the false discovery rate (FDR) is sufficiently low. Apart from AUC_{tot}, the area under the ROC curve can thus also be computed at a lower FPR cut-off such as 0.01. The area $\text{AUC}_{0.01}$ is normalised, and a value of 1 thus corresponds to a perfect classifier, while the value corresponding to a completely random classifier is 0.005, i.e lower than for AUC_{tot}. The two areas, AUC_{tot} and $\text{AUC}_{0.01}$, are visualised in Figure 2.2.

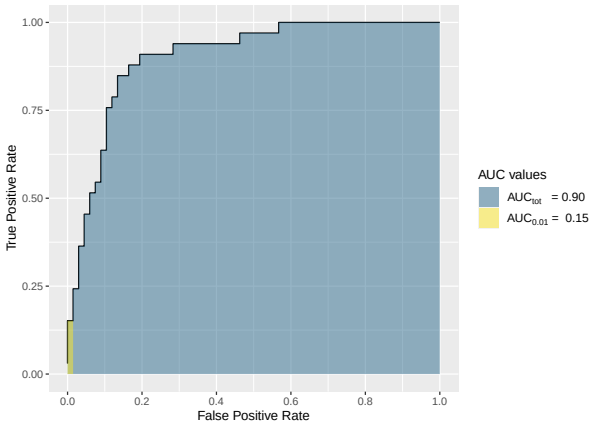


Figure 2.2: *A ROC curve with visualised areas for AUC_{tot} and $AUC_{0.01}$ as well as their computed values.*

2.2.4 False Discovery Rate

When performing simultaneous tests, there will on average be $\alpha * m$ FPs among the predicted positives where α is the significance level and m is the number of performed tests. This problem is present in comparative metagenomic studies since there are large amounts of genes in the data which all are tested individually, leading to multiple testing.

The proportion of FPs among the predicted positives can be estimated with the false discovery rate (FDR), stated as

$$FDR = Exp\left[\frac{\text{Number of FPs}}{\text{Number of cases where } H_0 \text{ is rejected}}\right] \tag{2.9}$$

[23].

The FDR can be controlled by using the Benjamini-Hochberg procedure [23]. The procedure consists of ordering the p-values in increasing order and then, given the ordered list, calculate q-values which are estimations of the FDR-value at each position i of the list. The q-values are calculated by

$$q_{(i)} = \frac{m}{i} p_{(i)} \ , \tag{2.10}$$

where $p_{(i)}$ is the p-value at position i. By choosing a significance level among the q-values, the FDR is controlled below that threshold, and the proportion of FPs is thus controlled.

Given a dataset where the real classes of the genes are known, the true FDR can be calculated and compared with the estimated FDR. By doing so, the reliability of the analysis can be evaluated. The true FDR at position i in an ordered gene list is given by

$$\text{True FDR}(i) = \frac{\text{Number of FPs at position } i}{i} \ . \tag{2.11}$$

3. Methodology

This master's thesis is primarily a simulation study where results and conclusions have been based on observations from a large number of datasets. These datasets were created by resampling real metagenomic data, and were made to include two sample groups which differed due to artificially introduced effects. By varying the group sizes and the sequencing depths when resampling the datasets, different experimental designs were created. The group size corresponds to the number of samples that were drawn for each group when resampling the original datasets, and the sequencing depth corresponds to the total gene count that was drawn for each sample. The experimental designs are described in Section 3.1, while the procedure of creating the datasets is described in Section 3.2. Conclusions regarding the performance of the experimental designs were based on statistical analyses performed on each created dataset. These analyses are described in Section 3.3. An assessment of the economic impact of the designs in relation to their performance was also performed and is described in Section 3.4.

The real metagenomic datasets that, in turn, have been used to analyse the experimental designs are called *Marine, Human Gut II* and *Resistance*. Marine is a dataset included in a large oceanic metagenome study and consists of 243 samples that have been sequenced with the NGS platform *Illumina* [24]. Human Gut II is also an Illumina sequenced dataset that is included in a type 2 diabetes study and consists of 145 human gut samples [3]. Resistance consists of 350 human associated samples and is a sparse dataset retrieved from a study that focuses on the prevalence of antibiotic resistance genes in different environments [25].

In order to create as valid results and conclusions as possible, reproducibility has been an important aspect of this thesis. The results that are presented in Section 4 were created in R (version 3.6.1) within an Anaconda environment (version 4.5.11). The packages that were used in R can be found in Appendix A. The script that was used to generate all results is found on Github at sofiamec/MastersThesis. This script, *Master.R*, runs the entire analysis for each of the three metagenomic datasets depending on which dataset that is selected. The directory structure was inspired by recommendations found in literature [26, 27].

3.1 Set-up of experimental designs

The different experimental designs studied in this thesis were constructed to cover some of the most commonly used designs in comparative metagenomics. The main designs were created by a combination of five different group sizes, m, and five or six different sequencing depths, d (five or six depending on the available sequencing depths in the original dataset). By doing so, a wide variety of both number of samples and sequencing depths could be studied. For each real metagenomic dataset, 100 repeats were performed for each experimental design to ensure reliable results. An overview of the main experimental designs can be found in Table 3.1, where the total amount of reads, $2m * d$, is presented for each design. Some of these designs were used along

with a few additional designs in order to study the dependence between the group size and the sequencing depth. Two cases were studied where the designs had either 6 M or 10 M reads in total. These trade-off designs are presented in Table 3.2.

Table 3.1: *The main experimental designs that have been applied on the Marine and Resistance datasets. Prefixes "k", "M" and "G" denote* 10^3, 10^6 *and* 10^9 *respectively. The total amount of reads are displayed for each design and is derived by multiplying the number of samples in both groups with the sequencing depth. Due to an overall lower sequencing depth in Human Gut II, the designs with a sequencing depth of 10M was not performed for this dataset.*

		Group size (m)				
		3	**5**	**10**	**30**	**50**
Sequencing depth (d)	**10 k**	60 k	100 k	200 k	600 k	1 M
	100 k	600 k	1 M	2 M	6 M	10 M
	500 k	3 M	5 M	10 M	30 M	50 M
	1 M	6 M	10 M	20 M	60 M	100 M
	5 M	30 M	50 M	100 M	300 M	500 M
	10 M	60 M	100 M	200 M	600 M	1 G

Table 3.2: *The trade-off designs used for studying the dependence between group size and sequencing depth. Prefixes "k" and "M" denote* 10^3 *and* 10^6 *respectively. Note that some of the designs are covered by the main designs in table 3.1.*

Group size (m)	**3**	**6**	**15**	**30**	**5**	**10**	**20**	**50**
Sequencing depth (d)	**1 M**	**500 k**	**200 k**	**100 k**	**1 M**	**500 k**	**250 k**	**100 k**
Reads in total (2m*d)	6 M	6 M	6 M	6 M	10 M	10 M	10 M	10 M

3.2 Creation of datasets

The method for creating new datasets from real metagenomic data was based on procedures found in literature [7, 6, 13, 8]. The procedure consists of an initial filtering of the real datasets, resampling of the filtered datasets and finally introducing artificial differences between the sample groups in the new datasets by downsampling gene counts. The filtering is described in Section 3.2.1, while the resampling is described in Section 3.2.2 and the downsampling in Section 3.2.3.

Another approach for creating datasets would have been to simulate the datasets from parametric distributions. The reason why resampling was preferred is that it is a more realistic approach since features such as the underlying read distributions and the gene-gene correlations of the original metagenomic data has been shown to be preserved when resampling [6].

3.2.1 Filtering

An initial filtering of genes was performed on the real metagenomic datasets where genes with zero counts in more than 75% of the samples or an average abundance of less than three reads in all samples were removed. This was done in order to ensure that the step of downsampling genes would be performed correctly since downsampling only can be applied on genes with non-zero counts. Due to the low gene abundances in the Resistance dataset the filtering of genes in this dataset was performed slightly differently, with the single criteria that the genes should have an average abundance of at least 10 counts. Samples with low sequencing depths were also removed from each of the real datasets. The reason behind this was that all experimental designs would

not be able to use these samples for resampling. Consequently, only samples with a sequencing depth higher than what was required for the experimental design with the highest sequencing depth were kept.

After filtering the three real metagenomic datasets, Marine consisted of 202 samples and 4462 genes, Human Gut II consisted of 126 samples and 3591 genes, and Resistance consisted of 348 samples and 18 genes. One of the 18 genes in Resistance was a merged entry of all reads mapped to genes that were not related to antibiotic resistance. The three filtered datasets were then used as original datasets from which new datasets were created by resampling.

3.2.2 Resampling

Given the group size, m, in an experimental design, $2 * m$ samples were randomly chosen from the selected original dataset creating two different groups with m samples each. For each sample, the gene counts were then resampled by randomly drawing d counts without replacement from the original sample.

3.2.3 Downsampling

Finally, artificial effects in the form of DAGs were introduced into the resampled dataset. This was done by randomly selecting approximately 10% of the total amount of genes in the dataset, and then downsampling half of those genes in one of the groups and the other half in the other group. The reason for downsampling gene counts instead of increasing them is that downsampling has been shown to keep the variance structure of the genes [13]. Since genes were downsampled in both sample groups, both positive and negative effects were introduced. The downsampling was performed by drawing new gene counts according to a binomial distribution

$$\tilde{y}_{ij} \sim \mathrm{B}(y_{ij}, \frac{1}{q}) \ , \tag{3.1}$$

where $\tilde{y}_{ij}$ is the new, downsampled, gene count which will be assigned to gene i in sample j. The gene count prior to downsampling is given by y_{ij}, which also determines the number of trials in the binomial distribution. The parameter q determines the degree to which the selected gene is downsampled and $1/q$ corresponds to the probability of success for each trial in the binomial distribution. In this thesis, two fixed values of q have been studied for each dataset. For Marine and Human Gut II, $q = 1.5$ and $q = 3$ were used, while $q = 5$ and $q = 10$ were used for Resistance. These values where chosen in order to study two effect sizes for each experimental design.

3.3 Statistical analysis

The general performance of each experimental design was evaluated based on the statistical power obtained when identifying DAGs in the datasets. The procedure of obtaining the statistical power is described in Section 3.3.1. The results were in turn used to establish the importance of, and trade-off between, group size and sequencing depth. In addition to the general performances of the designs, the impact of gene abundance and gene variability was also taken into account by noting how abundant, respectively varied, the downsampled genes were in the original dataset. This, slightly different, analysis is described in Section 3.3.2.

3.3.1 Evaluating experimental designs

Once an artificial dataset had been created through resampling and downsampling, a statistical analysis was performed to identify DAGs in the dataset. This was mainly done with the R package DESeq2. The reason for choosing this package was that it has been shown to have a higher performance in general compared to other packages, especially when used on Illumina-sequenced data [6]. It was also chosen since it is one of the most commonly used methods in metagenomic studies. The resulting p-values extracted from the analysis with DESeq2 indicated if, and how significantly, genes differed between the two groups in the dataset. Both regular p-values and Benjamini-Hochberg adjusted p-values were extracted. The regular p-values were used for ordering the genes according to significance, while the adjusted p-values were used for identifying the position at which the estimated FDR exceeded 0.05. Apart from DESeq2, an additional approach of using OGLM followed by an F-test was also used to identify DAGs.

In contrast to an analysis of real metagenomic data, a validation of the identified DAGs could in this thesis be performed since the effects had been introduced artificially and were thus known. For increasing significance levels the amount of correctly identified DAGs ,TPs, were noted along with incorrectly identified DAGs, FPs. This was used to compute TPR and FPR values at each possible significance level. The TPRs and FPRs were then used to plot ROC curves, which in turn were used to calculate AUC values at fixed FPR values of 0.01 and 1. The AUC values were calculated with the trapezoidal rule, since this was the simplest method.

Apart from ROC curves and AUC values, true FDR values were also calculated for each artificial dataset based on the number of TPs and FPs at an estimated FDR of 0.05. After repeating the entire process 100 times for each experimental design, median true FDR as well as median AUC values were calculated for each design. The median was chosen as a summary measure since it is less affected by outliers than a mean value. Cases where the median was based on less than 10 observations were noted as "NA". Based on the 100 repeats, mean ROC curves were also obtained by vertical averaging. The curves were plotted with ribbons corresponding to the observed minimum and maximum values.

3.3.2 Evaluating effect of gene abundance and variability

Before any artificial datasets were created or analysed, all genes in the original datasets were assigned different labels depending on their abundance and variability. The abundance and variability parameters were estimated by the mean and dispersion values resulting from analysing the original datasets with DESeq2 without any group comparisons. The genes were then divided into three abundance strata and three variability strata, which were created such that each strata contained approximately the same number of genes. The creation of datasets and identification of DAGs was performed as explained previously, without any influence of the strata. After the identification of DAGs, the correctly, and incorrectly, identified genes were however studied in relation to their different strata. This was done for the trade-off designs used for studying the dependence between group size and sequencing depth, and consisted of plotting different ROC curves as well as computing separate AUC values for each stratum.

3.4 Economic assessment

As a final analysis, a brief economic assessment was performed in order to evaluate the economic impact of the trade-off designs. This was done since the choice of an experimental design in a metagenomic study may be restricted by the economy of the study.

For each trade-off design, the total sequencing cost in US dollars was calculated based on presumed prices of \$100 per sample and \$0.001 per read. These values were roughly estimated from prices for Illumina-sequenced data, and are likely to change in the future. The total sequencing cost of an experimental design was thus calculated by

$$Sequencing\ cost = 100 * 2m + 10^{-4} * d * 2m, \tag{3.2}$$

where m is the number of samples in each group and d is the sequencing depth in each sample. To evaluate the performance of each design in relation to its sequencing cost, a ratio between the $AUC_{0.01}$ value and sequencing cost was also calculated for each design. This ratio, *performance*/\$1000, was compared between the designs in order to find the most cost efficient design with regard to its performance.

4. Results

The results are divided into five parts; the effects of group size and sequencing depth presented in Section 4.1, the dependence between these factors presented in Section 4.2, the effect of gene abundance and variability presented in Section 4.3, the accuracy of the estimated FDR presented in Section 4.4, and finally the economic assessment presented in Section 4.5.

Several aspects have been taken into account in order to study how the performance of a metagenomic study is affected by the experimental design. This has resulted in numerous analyses where three real metagenomic datasets, two effect sizes and two different analysis methods have been used. These analyses have thus yielded a large number of results which in many cases show the same patterns. The results from the different effect sizes show similar trends and the results generated with DESeq2 are generally similar to those generated with OGLM and F-test. Due to the large amount of results, this chapter is thus mainly focused on one dataset and only presents part of the results. Unless otherwise stated, the presented results are obtained from analysing the Marine dataset with DESeq2.

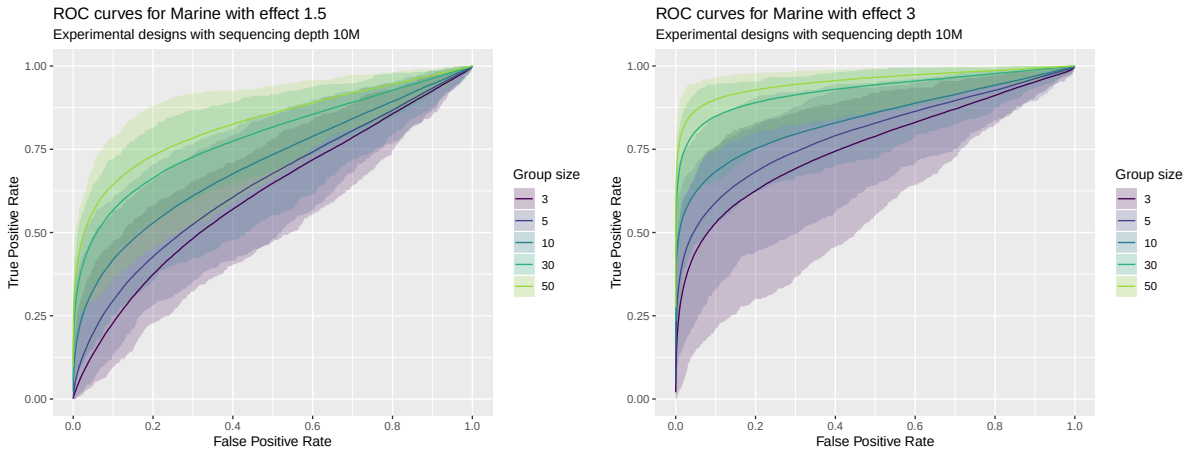
Additional results for Marine are presented in Appendix B where results from the DESeq2 analysis can be found in Appendix B.1, while results from the OGLM and F-test analysis can be found in Appendix B.2. Results obtained from analysing the Human Gut II and Resistance datasets are consistent with the ones obtained for Marine. Apart from the results presented in this chapter, additional results for these two datasets can be found in Appendix C and Appendix D respectively.

4.1 Effects of group size and sequencing depth

The focus of this section is to study how different sequencing depths and group sizes affect the general performance of a study. This was investigated by plotting ROC curves and computing AUC values for each of the main experimental designs. The ROC curves are presented in combined plots for a fixed sequencing depth or group size. The AUC values for all designs are displayed in heatmaps. The results are presented in Section 4.1.1 for Marine, in Section 4.1.2 for Human Gut II and in Section 4.1.3 for Resistance. The results for both effect sizes are presented for Marine, while only the results for the large effect size is presented for Human Gut II and Resistance.

4.1.1 Marine

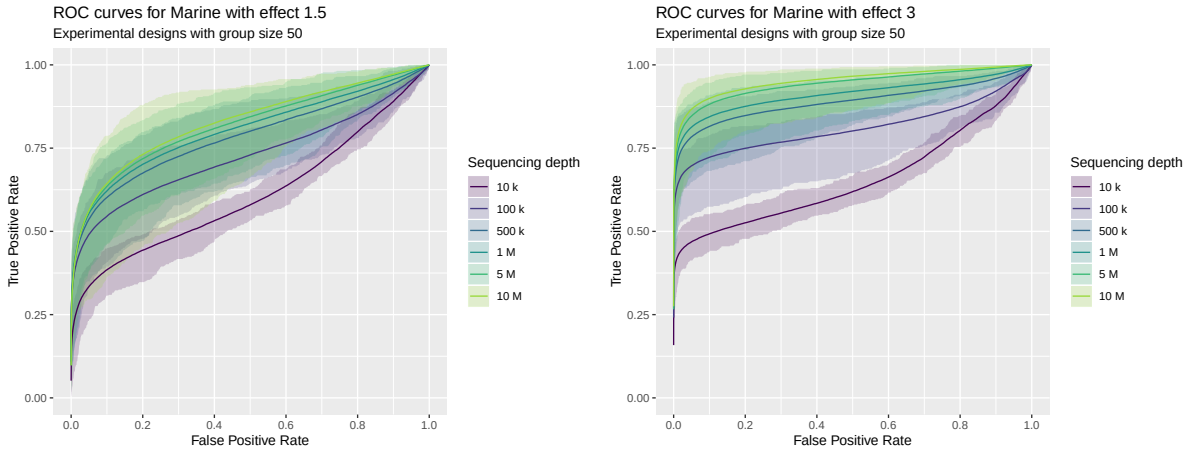
Figure 4.1 presents ROC curves for different experimental designs with a sequencing depth of 10 M. The left plot, Figure 4.1a, corresponds to results obtained when analysing the Marine dataset with the small effect size, while the right plot, Figure 4.1b, corresponds to results obtained with the large effect size. As can be seen when comparing the plots, the curves reach higher TPR values earlier in the right graph where the effect size is large. This means that the general performance is better when the effect size is large. Considering each plot individually, it can also be seen that the performance is increased when the group size is large.



(a) *Small effect size.* **(b)** *Large effect size.*

Figure 4.1: *Mean ROC curves for the main experimental designs with sequencing depth 10 M for Marine analysed with DESeq2. Each design with a different group size is represented by a colored line where the ribbon corresponds to minimum and maximum values. The two plots present curves for different effect sizes.*

Figure 4.2 presents similar ROC curves with the difference that the sequencing depth of the designs varies while the group size is fixed at 50. These plots also show that the general performance is increased when the effect is larger. It can further be seen that the performance increases for increasing sequencing depths. In addition, the design with the lowest sequencing depth, 10 k, performs considerably worse than the other designs in both plots.



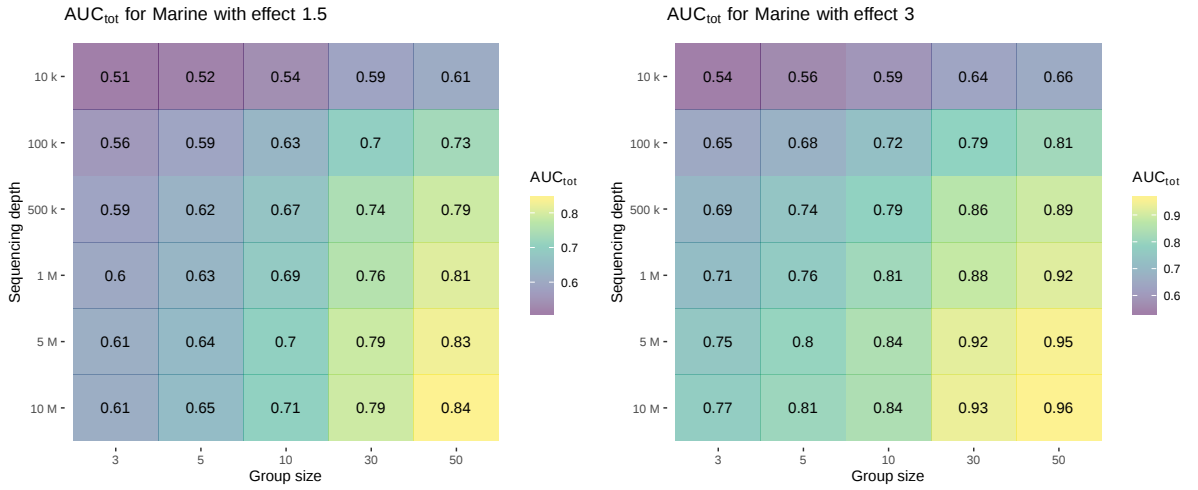
(a) *Small effect size.* **(b)** *Large effect size.*

Figure 4.2: *Mean ROC curves for the main experimental designs with a group size of 50 samples for Marine analysed with DESeq2. Each design with a different sequencing depth is represented by a colored line where the ribbon corresponds to minimum and maximum values. The two plots present curves for different effect sizes.*

Additional plots containing all the generated ROC curves for Marine are presented in Appendix B.1.

The computed AUC_{tot} values for each of the main experimental designs are visualised for both effect sizes in two different heatmaps in Figure 4.3. The performances are seen to be considerably higher in the heatmap corresponding to the large effect size. This is consistent with the previously

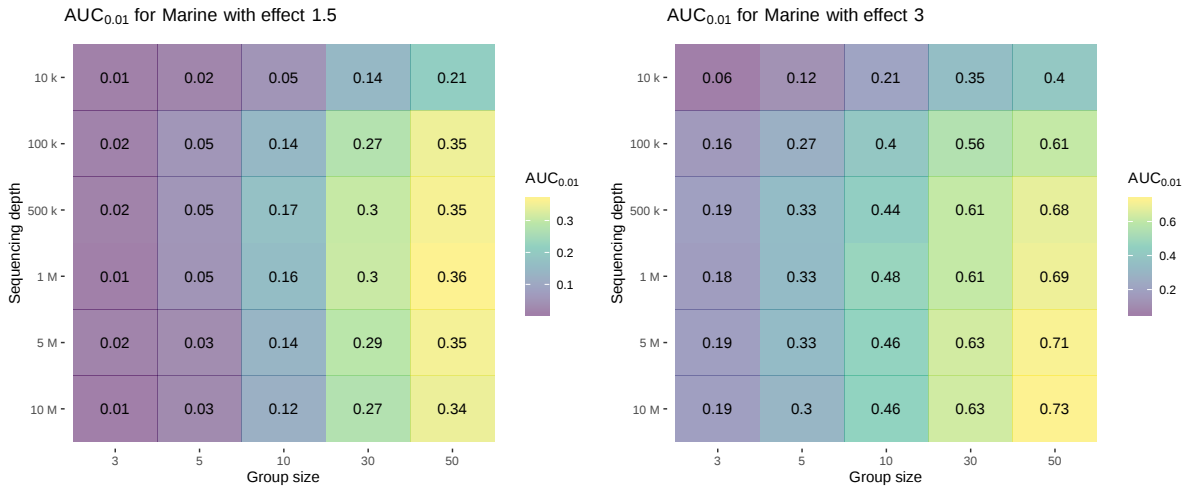
observed results. The heatmaps for both effect sizes also show that the AUC_{tot} values of the experimental designs increase with increasing sequencing depth as well as with increasing group size. Furthermore, designs with the lowest sequencing depth, d= 10 k, are noted to perform remarkably worse compared to designs with higher sequencing depths. Finally, the performance appears to be more affected by the group size than the sequencing depth, since the increase in AUC_{tot} values is more pronounced with increasing group size than it is with increasing sequencing depth.



(a) *Small effect size.* **(b)** *Large effect size.*

Figure 4.3: *Median AUC_{tot} values for the main experimental designs for Marine analysed with DESeq2. The two heatmaps present values for different effect sizes. Note that the color scales are based on the ranges of the values in the heatmaps, and are thus not directly comparable between the two heatmaps.*

The corresponding $AUC_{0.01}$ values for each of the main experimental designs are presented in Figure 4.4 and show similar trends as the heatmaps with AUC_{tot} values. There is however an additional trend in which the performance does not improve for designs with sequencing depths above 1 M, especially for the smaller effect size. Another observation is that the performances of designs with the lowest group sizes, $m = 3$ and to some extent $m = 5$, do not improve regardless of the sequencing depth. This indicates that having a small group size will result in a low performance, seen to $AUC_{0.01}$, even if the sequencing depth is high.

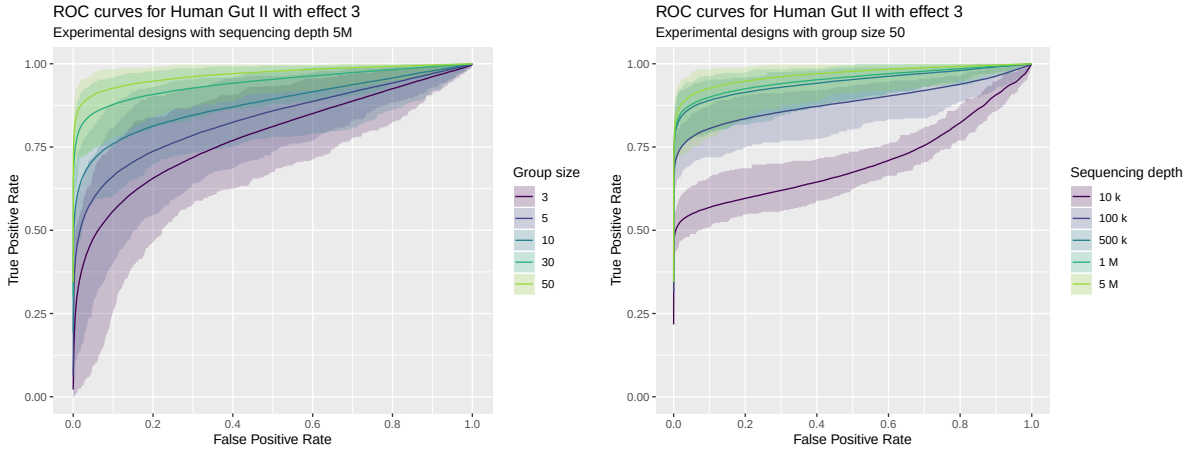


(a) *Small effect size.* **(b)** *Large effect size.*

Figure 4.4: *Median $AUC_{0.01}$ values for the main experimental designs for Marine analysed with DESeq2. The two heatmaps present values for different effect sizes. Note that the color scales are based on the ranges of the values in the heatmaps, and are thus not directly comparable between the two heatmaps.*

4.1.2 Human Gut II

Similarly to previously found results, the ROC curves and AUC values obtained when analysing the Human Gut II dataset show that it is desired to have a high sequencing depth as well as a large group size in order to increase the performance of an experimental design. This is seen in Figure 4.5 where ROC curves for the large effect size are presented either for designs with a sequencing depth of 5 M, or for designs with a group size of 50 samples.



(a) *Designs with sequencing depth 5 M.* **(b)** *Designs with group size 50.*

Figure 4.5: *Mean ROC curves for the main experimental designs for Human Gut II analysed with DESeq2. The designs are represented by colored lines where the ribbon corresponds to minimum and maximum values. The two plots present curves for the large effect size and designs with either a sequencing depth of 5 M or with a group size of 50 samples.*

Additional plots containing all ROC curves generated for Human Gut II with both effect sizes are presented in Appendix C.

Figure 4.6 and Figure 4.7 present heatmaps for the AUC_{tot} and $AUC_{0.01}$ values for the main experimental designs with the large effect size. As seen previously for Marine, these two heatmaps also show that the performance increases more with increasing group size compared to with increasing sequencing depth.

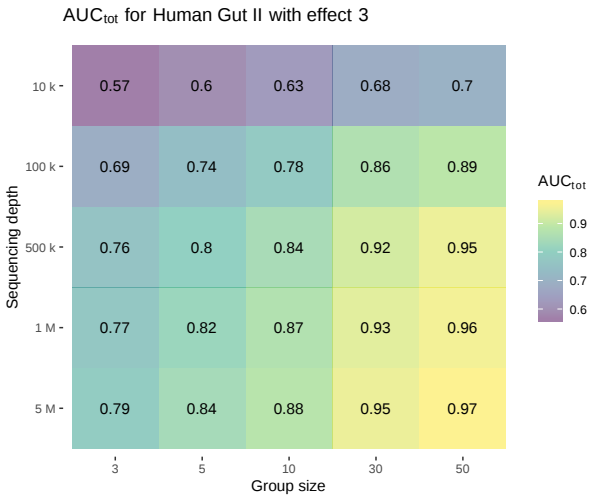


Figure 4.6: *Median AUC_{tot} values for the main experimental designs for Human Gut II analysed with DESeq2. The heatmap presents values for the large effect size.*

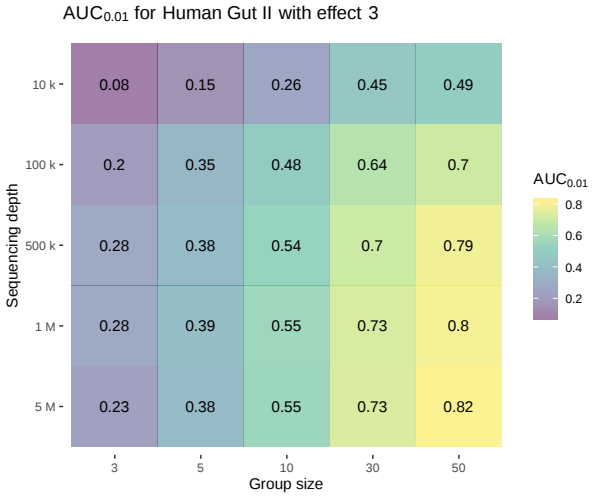
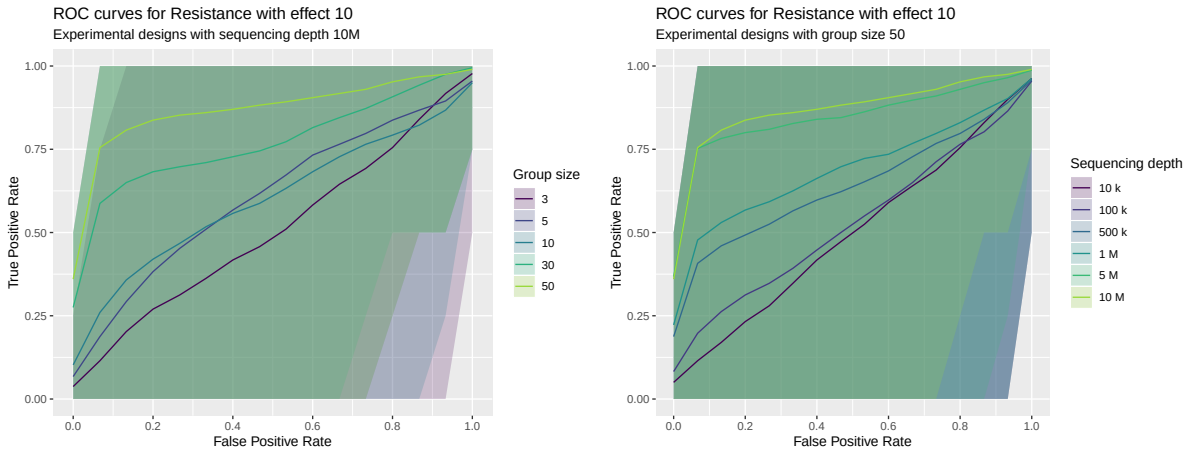


Figure 4.7: *Median $AUC_{0.01}$ values for the main experimental designs for Human Gut II analysed with DESeq2. The heatmap presents values for the large effect size.*

The corresponding figures with the small effect size can be found in Appendix C.

4.1.3 Resistance

Two plots containing ROC curves for the final dataset Resistance are presented in Figure 4.8. As previously seen for Marine and Human Gut II, these plots also show that the performance of the designs increases both with increasing sequencing depth and increasing group size. However, due to the smaller amount of genes in this dataset, the observed minimum and maximum values are wider for these ROC curves compared to the corresponding curves for Marine or Human Gut II.



(a) *Designs with sequencing depth 10 M.* **(b)** *Designs with group size 50.*

Figure 4.8: *Mean ROC curves for the main experimental designs for Resistance analysed with DESeq2. The designs are represented by colored lines where the ribbon corresponds to minimum and maximum values. The two plots present curves for the large effect size and designs with either a sequencing depth of 10 M or with a group size of 50 samples.*

Additional plots containing all the generated ROC curves for Resistance with both effect sizes are presented in Appendix D.

The AUC_{tot} values for the main experimental designs are presented in Figure 4.9 for the large effect size. Compared to the results obtained for Marine and Human Gut II, the AUC_{tot} values are noticeably lower for Resistance. In order to reach high AUC values for this dataset, the experimental design has to have among the highest sequencing depths and largest group sizes available.

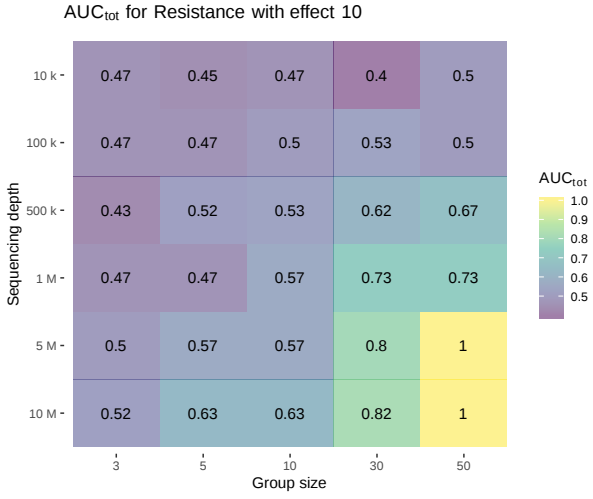


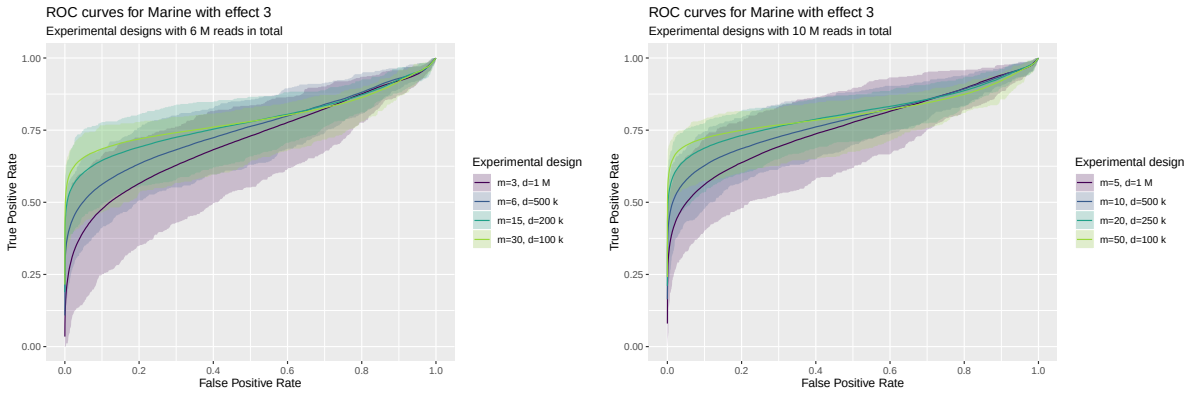
Figure 4.9: *Median AUC_{tot} values for the main experimental designs for Resistance analysed with DESeq2. The heatmap presents values for the large effect size.*

The corresponding heatmap for the smaller effect size is found in Appendix D. $AUC_{0.01}$ values were not calculated for Resistance due to the small amount of genes in the dataset.

4.2 Dependence between group size and sequencing depth

The dependence between group size and sequencing depth is further assessed by studying the trade-off designs that have a fixed amount of reads in total. The designs with small group sizes thus have higher sequencing depths while designs with large group sizes have lower sequencing depths. Two cases are studied for the trade-off designs; one with 6 M reads in total and one with 10 M reads in total.

For the large effect size, two plots with ROC curves for each of the cases are displayed in Figure 4.10. The plots are similar and both cases show that the experimental designs perform better with increasing group size rather than with increasing sequencing depth. Thus, it is preferred to increase the number of samples instead of increasing the sequencing depth.



(a) *6 M reads in total.* **(b)** *10 M reads in total.*

Figure 4.10: *Mean ROC curves for the trade-off designs for Marine analysed with DESeq2. Each design is represented by a colored line where the ribbon corresponds to minimum and maximum values. The two plots present curves for the large effect size for designs with either 6 M or 10 M reads in total.*

The same trend is observed when looking at the $AUC_{0.01}$ values for these ROC curves, presented in Table 4.1.

Table 4.1: *Median $AUC_{0.01}$ values for the trade-off designs for Marine analysed with DESeq2. The table presents values for the large effect size for designs with 6 M and 10 M reads in total.*

6 M reads in total		10 M reads in total	
Experimental design	$AUC_{0.01}$	Experimental design	$AUC_{0.01}$
m=3, d=1 M	0.18	m=5, d=1 M	0.33
m=6, d=500 k	0.35	m=10, d=500 k	0.44
m=15, d=200 k	0.49	m=20, d=250 k	0.54
m=30, d=100 k	0.56	m=50, d=100 k	0.61

The corresponding figure and table for the same designs, but with the smaller effect size, can be found in Appendix B.1.

4.3 Effect of gene abundance and variability

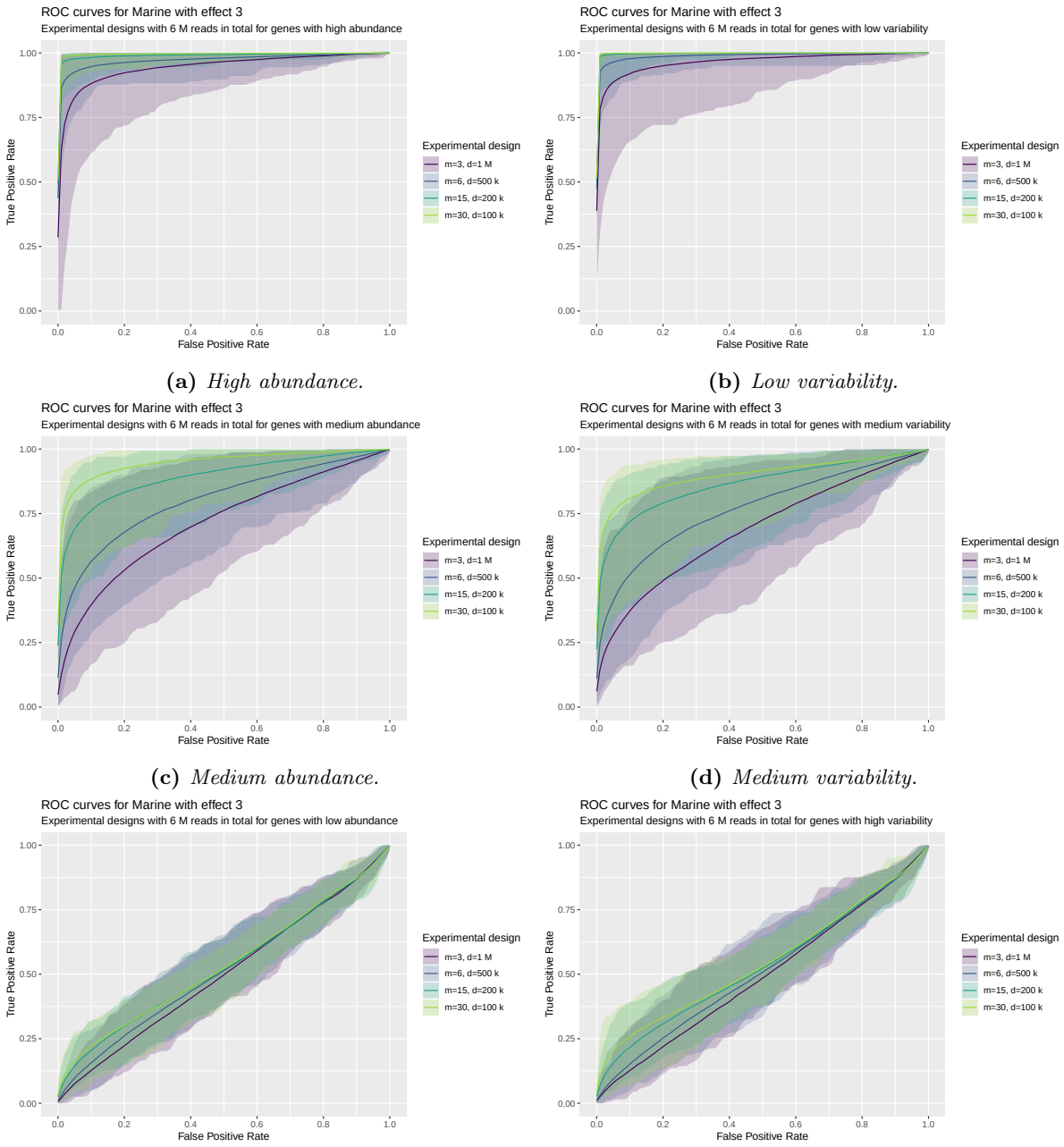
In order to understand how gene abundance and variability influence the performance of an experimental design, the trade-off designs were analysed with regard to these factors. This was done by dividing the genes into three different abundance and variability strata and plotting separate ROC curves and computing separate $AUC_{0.01}$ values for each stratum.

Figure 4.11 presents these ROC curves for the trade-off designs with 6 M reads in total for the large effect size. The plots show that the general performance of the designs increases when the abundance of the studied genes increases. The opposite is seen for the variability, where a low variability gives the best performance. Similarly to previous results, these plots also show a tendency of increased performance for designs with larger group sizes rather than for designs with higher sequencing depths. This is mainly seen for the plots with medium abundance and medium variability. The group size is also seen to be important when studying the plots with low abundance and high variability, where a large group size is required to avoid having a completely random classification genes. However, in the cases where genes with high abundance or low variability are studied, the performance is high for all the designs and it is thus not as important to use a large group size in these cases compared to when genes with lower abundance or higher variability are studied. The plots with high abundance or low variability are further seen to be the only ones that have a high performance for the design with the smallest group size. Thus, it can further be noted that if an experimental design with the smallest group size is used in a study, the identified genes will probably have high abundance or low variability. Genes with lower abundance or higher variability are most likely not detected with such designs. In order to capture such genes, the experimental design is required to have a larger group size.

The same trends can be seen in Table 4.2, where the corresponding $AUC_{0.01}$ values are presented for each of these designs and strata.

Table 4.2: *Median $AUC_{0.01}$ values for different abundance and variability strata for Marine analysed with DESeq2. The table presents values for the large effect size for the trade-off designs with 6 M reads in total.*

	Abundance			Variability		
Experimental Design	Low	Medium	High	Low	Medium	High
m=3, d=1 M	0.010	0.059	0.513	0.759	0.081	0.012
m=6, d=500 k	0.015	0.188	0.848	0.923	0.188	0.015
m=15, d=200 k	0.037	0.429	0.955	0.987	0.417	0.043
m=30, d=100 k	0.053	0.616	0.979	0.995	0.579	0.073



(a) *High abundance.*

(b) *Low variability.*

(c) *Medium abundance.*

(d) *Medium variability.*

(e) *Low abundance.*

(f) *High variability.*

Figure 4.11: *Mean ROC curves for different abundance and variability strata for Marine analysed with DESeq2. The plots present curves for the large effect size for the trade-off designs with 6 M reads in total. Each design is represented by a colored line where the ribbon corresponds to minimum and maximum values.*

The corresponding results for the smaller effect size as well as for the cases with 10 M reads in total can be found in Appendix B.1.

4.4 Accuracy of estimated FDR

The true FDR at an estimated FDR of 0.05 was investigated for the main experimental designs. This makes it possible to further evaluate the performance of the designs with regard to how accurately the FDR is estimated.

The results for the small effect size are presented in a heatmap in Figure 4.12 for the main analysis method DESeq2. The true FDR values are generally shown to be high for designs with small group sizes, while they are closer to the estimated FDR of 0.05 for designs with larger group sizes. It can further be noted that designs with lower sequencing depths estimate the FDR better than designs with higher sequencing depths. This observation is unexpected, and it is thus questioned whether it is an actual consequence of the sequencing depth or if it is a consequence of the chosen analysis method, DESeq2.

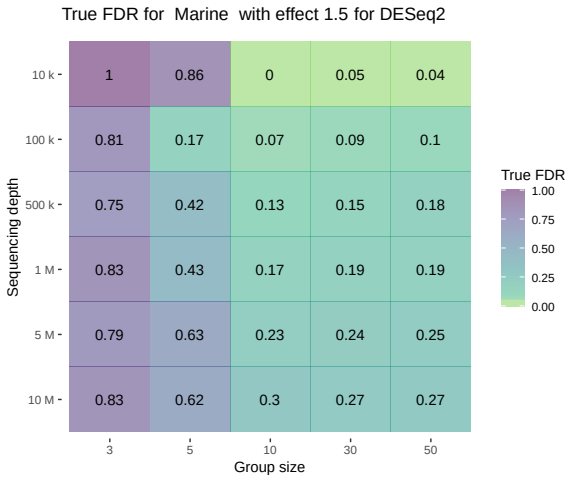


Figure 4.12: *Median true FDR values at an estimated FDR of 0.05 for the main experimental designs for Marine analysed with DESeq2. All values below the estimated value have the same color. The heatmap presents values for the small effect size.*

In order to confirm the results generated with DESeq2, OGLM and F-test were used for the differential analysis. The corresponding FDR results for this method are presented in a heatmap in Figure 4.13. In contrast to the results for DESeq2, designs with lower sequencing depths are in this case found to perform worse than designs with higher sequencing depths. The contradicting results imply that the presented true FDR values depend on what method is used for the differential analysis.

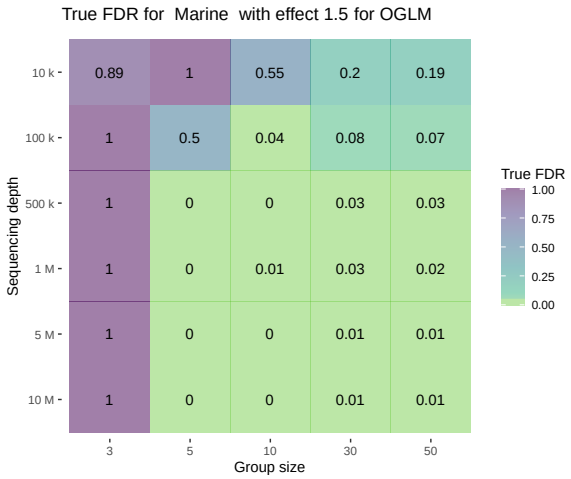


Figure 4.13: *Median true FDR values at an estimated FDR of 0.05 for the main experimental designs for Marine analysed with OGLM and F-test. All values below the estimated value have the same color. The heatmap presents values for the small effect size.*

The heatmaps for both methods are however shown to have considerably higher true FDR values for designs with the smallest group size, $m = 3$, compared to the other group sizes. This observed trend thus appears to depend on the group size rather than on the analysis method, indicating that the designs with the smallest group size estimate FDR values inaccurately.

The corresponding figures with the large effect size are presented in Appendix B.1 for the DESeq2 analysis and in Appendix B.2 for the OGLM and F-test analysis.

4.5 Economic assessment in relation to performance

The choice of the experimental design in a comparative metagenomic study may be restricted by the available budget. It is therefore of interest to investigate the economic impact of the experimental designs and how this relates to their performances.

The sequencing costs for the trade-off designs with 6 M reads in total are displayed in Table 4.3 along with the performance of each design. The performance is displayed for both effect sizes and is measured in $AUC_{0.01}$ while the cost is measured in US dollars. The table also presents ratios of the performances and the sequencing costs, displayed as *performance*/\$1000. The ratio is a measure of "how much performance" that is gained per thousand US dollars. Corresponding results for designs with 10 M reads in total are presented in Table 4.4.

Both tables show that the sequencing cost of a design is increased with increasing group size rather than with increasing sequencing depth. As seen previously, the performance follows the same trend and is shown to be higher for the large effect size.

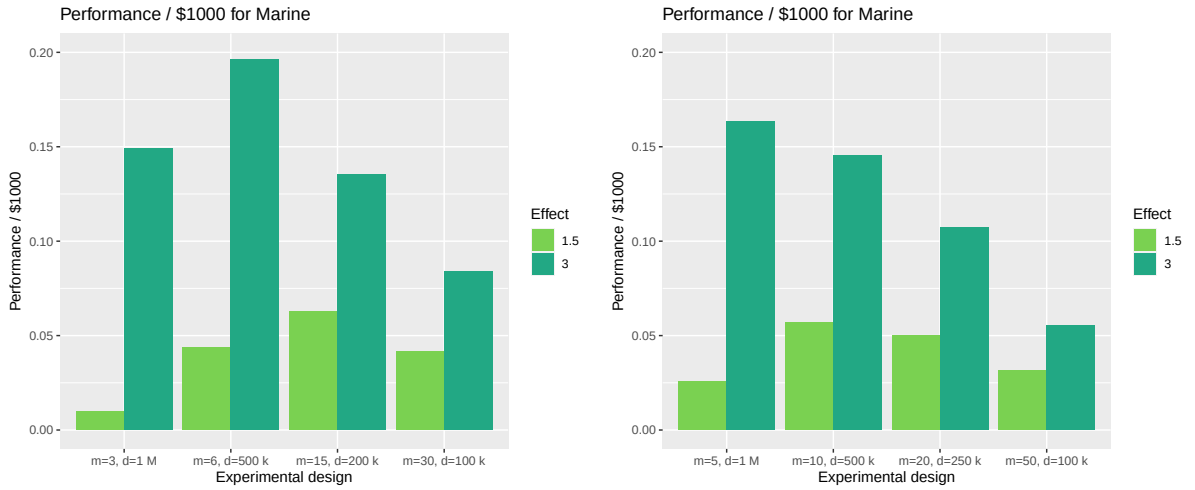
Table 4.3: *Sequencing cost, performance and the ratio performance/\$1000 for the trade-off designs with 6 M reads in total for Marine analysed with DESeq2. The cost is displayed in US dollars and the performance is displayed in $AUC_{0.01}$. The table presents values for both effect sizes.*

		Effect 1.5		Effect 3	
Experimental design	Cost	$AUC_{0.01}$	Performance/\$1000	$AUC_{0.01}$	Performance/\$1000
m=3, d=1 M	1200	0.01	0.01	0.18	0.15
m=6, d=500 k	1800	0.08	0.04	0.35	0.20
m=15, d=200 k	3600	0.23	0.06	0.49	0.14
m=30, d=100 k	6600	0.27	0.04	0.56	0.08

Table 4.4: *Sequencing cost, performance and the ratio performance/\$1000 for the trade-off designs with 10 M reads in total for Marine analysed with DESeq2. The cost is displayed in US dollars and the performance is displayed in $AUC_{0.01}$. The table presents values for both effect sizes.*

		Effect 1.5		Effect 3	
Experimental design	Cost	$AUC_{0.01}$	Performance/\$1000	$AUC_{0.01}$	Performance/\$1000
m=5, d=1 M	2000	0.05	0.03	0.33	0.16
m=10, d=500 k	3000	0.17	0.06	0.44	0.15
m=20, d=250 k	5000	0.25	0.05	0.54	0.11
m=50, d=100 k	11000	0.35	0.03	0.61	0.06

By looking at the tables it can further be seen that the experimental designs with the highest ratio, performance/$1000, are not necessarily the same designs that have the highest $AUC_{0.01}$ values or the lowest sequencing costs. It can also be seen that the designs with the highest ratio varies for the two effect sizes. This is also seen in Figure 4.14 where the ratios are visualised in two barplots for the trade-off designs with either 6 M or 10 M reads in total. In each plot, the design with the smallest group size has the lowest performance/$1000 for the small effect size while it has one of the highest ratios for the large effect size. The difference between the ratios for the two effect sizes is shown to be larger for designs with small group sizes than for designs with larger group sizes. Thus the profitability of designs with small group sizes are more dependent on the effect size. Finally, designs with the largest group size are noted to have among the lowest ratios for both effect sizes. This indicates that designs with large group sizes are less preferable when considering the economic impact.



(a) *6 M reads in total.* **(b)** *10 M reads in total.*

Figure 4.14: *The ratio of performance and sequencing cost, performance/$1000, for the trade-off designs for Marine analysed with DESeq2. The performance is measured in $AUC_{0.01}$ and the cost in US dollars. The two plots present values for either 6 M or 10 M reads in total and each plot contains both effect sizes.*

5. Discussion

Based on the results there are several factors which affect the choice of the experimental design for a comparative metagenomic study. The observed patterns have been consistent for the three different original metagenomic datasets. This indicates that the conclusions are independent of the studied dataset, and can thus be assumed to be general. A trend that has been seen for all the analyses is that the performances of the investigated designs improve when the effect size is large. This means that if a study is performed on data where the effects are known to be large, the requirements on the experimental design are not as high as if the studied effects are small.

By increasing the sequencing depth of the samples in an experimental design, the performance of the design is shown to increase as well. This can be explained by the fact that a high sequencing depth in a sample increases the chance of capturing less common genes in the data, and thus increases the performance when classifying them as DAGs or non-DAGs. However, given that only the most significant genes are of interest, it may not be necessary to have as high sequencing depths as when considering all genes. The reason for this is that the most significant genes already are found with lower sequencing depths. This claim is supported by the observed $\text{AUC}_{0.01}$ values, where all the designs with sequencing depths over one million reads have similar performances. Nevertheless, it is important that the sequencing depth is sufficiently high to provide enough information about the real data. This is seen for the designs with the lowest sequencing depth, $d = 10$ k, which performed considerably worse than the other designs. This indicates that a sequencing depth of ten thousand reads is too low, regardless of group size, and that it should be avoided.

It was also found that the performance of the experimental designs increase with increasing group size. This is an expected result since the variation within the groups is compensated for when more samples are collected for each group. Designs with the smallest group size were always shown to have among the lowest performances out of all the designs. Furthermore, the performances of these designs were not shown to improve with increased sequencing depths, which implies that the smallest group size, $m = 3$, may be too small to yield high performance regardless of the sequencing depth. That a group size of three samples is too small is further supported by the considerably worse FDR values that were obtained for these designs compared to designs with higher group sizes. The poorly estimated FDR values for designs with a group size of three samples imply that these designs would be unreliable if used in a real study. Since the FDR is used to control the number of false positives, a real study that only has a group size of three samples is thus likely to be too optimistic and identify many genes as DAGs even though they do not truly differ between the environments.

Furthermore, if a study is performed with a small group size, the detected DAGs will mainly consist of genes with high abundance or low variability. This is based on the results from the different abundance and variability strata. When studying genes with high abundance or low variability, all designs had a high performance, including the design with the lowest group size.

Thus, the group size $m = 3$, could possibly be used if it is known that the studied genes will have high abundance or low variability. For the medium strata however, a higher group size was required in order to reach an acceptable performance. This indicates that in order to detect significant genes with a lower abundance or higher variability, a group size of three samples is too small. Finally, when studying genes that were known to have low abundance or high variability, the highest group size was required to get the slightest performance. Thus, if such genes are of interest, the experimental design is required to have a large group size.

Given these findings, the best option for studies with unknown circumstances would therefore be to have the highest sequencing depth and largest group size possible, resulting in a large amount of reads in total. This is however rarely possible nor practical. When assessing the dependence between group size and sequencing depth, it was found to be of greater importance to increase the group size rather than the sequencing depth in order to get better results. This could be explained by the fact that the performance of comparative metagenomics is based on the ability to find differences between two sample groups rather than correctly quantifying the gene abundances within each sample. It therefore follows that it is of higher importance to compensate for the biological variability between samples, in order to correctly estimate group means, than it is to correctly estimate the abundances in individual samples.

Taking the economical impact into account, the design with the highest performance may however not be the optimal choice due to a high sequencing cost in relation to its performance. In contrast to previous indications, a large group size is in this regard shown to be considerably worse since the cost of a design is affected more by the number of samples than it is by the sequencing depth. Instead, it is preferable to use a fairly small group size given that it is large enough. What is considered large enough depends on the effect size of the studied phenomenon as well as the gene abundance and variability in the data. Since the consequence of choosing a group size that is too small is a drastically worsened performance, it may be risky to choose a small group size and more prudent to select a slightly larger group size, even though this will be more costly.

To conclude, there are several aspects that affect the performance of an experimental design. If the studied genes are known to have high abundance or low variability, the requirements on the experimental design are not as high as if other genes are studied. This also applies to the effect size of the studied phenomenon, where a high effect size requires a less extensive experimental design. In order to select an appropriate design, it is thus important to estimate such aspects before conducting a full scale study. This could be done through literature research as well as with a small pre-study. A sequencing depth of ten thousand reads is generally too low to yield an acceptable performance. Likewise a group size of three samples is too small, unless the studied genes have high abundance or low variability, and will likely produce unreliable results. Furthermore, the performance is shown to improve more with increasing group size than with increasing sequencing depth. However, when taking the economic impact into account, a larger group size becomes less profitable due to the increased sequencing cost per sample. It should finally be noted that this cost is likely to decrease in the future, which hopefully will make it cheaper to use experimental designs with larger group sizes.

Author Contributions

Both authors, Sofia Conti and Martina Hermanova Billstein, have contributed equally to all parts of this thesis work. This includes reviewing literature, planning experiments, programming in R, analysing results and writing the report.

Bibliography

[1] Berg RD. The indigenous gastrointestinal microflora. Trends in microbiology. 1996;4(11):430–435. Available from: https://doi.org/10.1016/0966-842X(96)10057-3.

[2] Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol. 2010;6(2):e1000667. Available from: https://doi.org/10.1371/journal.pcbi.1000667.

[3] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490(7418):55–60. Available from: https://doi.org/10.1038/nature11450.

[4] Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. The ISME journal. 2012;6(2):320–329.

[5] Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Applied and Environmental Microbiology. 2000;66(6):2541–2547. Available from: https://doi.org/10.1128/AEM.66.6.2541-2547.2000.

[6] Jonsson V, Österlund T, Nerman O, Kristiansson E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC genomics. 2016;17(1):78. Available from: https://doi.org/10.1186/s12864-016-2386-y.

[7] Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. BMC genomics. 2018;19(1):274. Available from: https://doi.org/10.1186/s12864-018-4637-6.

[8] Boulund F, Pereira MB, Jonsson V, Kristiansson E. Computational and statistical considerations in the analysis of metagenomic data. In: Metagenomics. Elsevier; 2018. p. 81–102. Available from: https://doi.org/10.1016/B978-0-08-102268-9.00004-5.

[9] Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, et al. Unlocking the potential of metagenomics through replicated experimental design. Nature biotechnology. 2012;30(6):513. Available from: https://doi.org/10.1038/nbt.2235.

[10] Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nature biotechnology. 2017;35(9):833. Available from: https://doi.org/10.1038/nbt.3935.

[11] Rappé MS, Giovannoni SJ. The uncultured microbial majority. Annual Reviews in Microbiology. 2003;57(1):369–394. Available from: https://doi.org/10.1146/annurev.micro.57.030502.090759.

[12] Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & biology. 1998;5(10):R245–R249. Available from: https://doi.org/10.1016/S1074-5521(98)90108-9.

[13] Jonsson V. Statistical analysis and modelling of gene count data in metagenomics. Sweden:

Göteborg. 2017;.

[14] Österlund T, Jonsson V, Kristiansson E. HirBin: high-resolution identification of differentially abundant functions in metagenomes. BMC genomics. 2017;18(1):316. Available from: https://doi.org/10.1186/s12864-017-3686-6.

[15] Scholz MB, Lo CC, Chain PS. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Current opinion in biotechnology. 2012;23(1):9–15. Available from: https://doi.org/10.1016/j.copbio.2011.11.013.

[16] Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends in genetics. 2014;30(9):418–426. Available from: https://doi.org/10.1016/j.tig.2014.07.001.

[17] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15(12):550. Available from: https://doi.org/10.1186/s13059-014-0550-8.

[18] McCullagh P, Nelder JA. Generalized Linear Models. 2nd ed. Chapman & Hall; 1989.

[19] Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006;27(8):861–874. Available from: https://doi.org/10.1016/j.patrec.2005.10.010.

[20] Odintsova V, Tyakht A, Alexeev D. Guidelines to statistical analysis of microbial composition data inferred from metagenomic sequencing. Curr Issues Mol Biol. 2017;24:17–36. Available from: https://doi.org/10.21775/cimb.024.017.

[21] Centor R, Keightley G. Receiver Operating Characteristics (ROC) Curve Area Analysis Using the ROC ANALYZER. In: Proceedings. Symposium on Computer Applications in Medical Care. American Medical Informatics Association; 1989. p. 222–226.

[22] Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. Archives of pathology & laboratory medicine. 1986;110(1):13–20.

[23] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289–300. Available from: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[24] Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348(6237). Available from: https://doi.org/10.1126/science.1261359.

[25] Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DJ. The structure and diversity of human, animal and environmental resistomes. Microbiome. 2016;4(1):54. Available from: https://doi.org/10.1186/s40168-016-0199-5.

[26] Noble WS. A quick guide to organizing computational biology projects. PLoS Computational Biology. 2009;5(7). Available from: https://doi.org/10.1371/journal.pcbi.1000424.

[27] Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. Good enough practices in scientific computing. PLoS Computational Biology. 2017;13(6). Available from: https://doi.org/10.1371/journal.pcbi.1005510.

A. Packages used in R

The following packages were used in R:

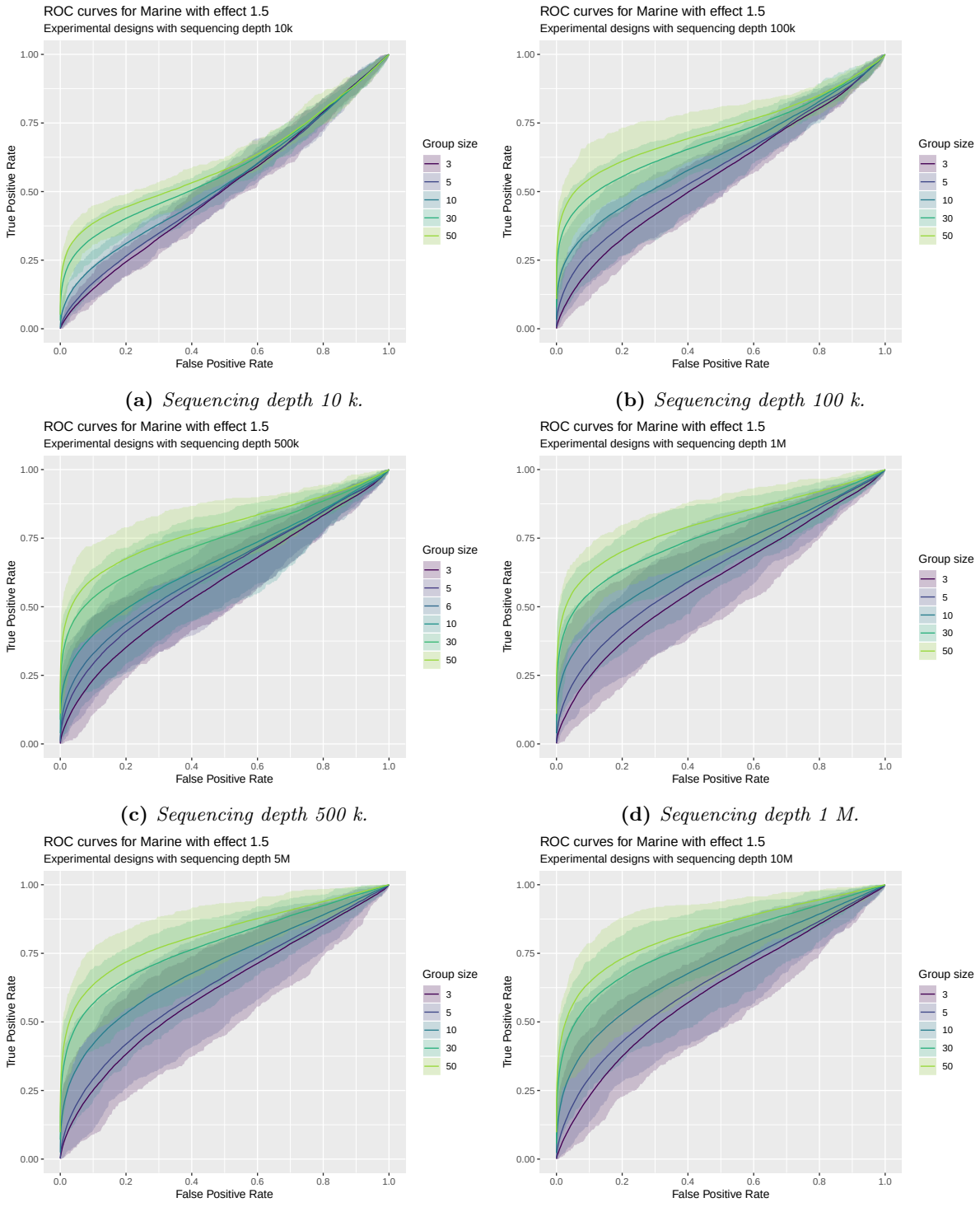
DESeq2	(version 1.26.0)
ggplot2	(version 3.2.1)
plyr	(version 1.8.5)
pracma	(version 2.2.9)
readxl	(version 1.3.1)
viridis	(version 0.5.1)
xtable	(version 1.8-4
RColorBrewer	(version 1.1-2)
DelayedArray	(version 0.12.2)
matrixStats	(version 0.55.0)
GenomicRanges	(version 1.38.0)
IRanges	(version 2.20.2)
BiocGenerics	(version 0.32.0)
viridisLite	(version 0.3.0)
SummarizedExperiment	(version 1.16.1)
BiocParallel	(version 1.20.1)
Biobase	(version 2.46.0)
GenomeInfoDb	(version 1.22.0)
BiocGenerics	(version 0.24.3)
S4Vectors	(version 0.5.1)
reshape2	(version 1.4.3)
tidyverse	(version 1.3.0)

B. Additional results for Marine

The Marine dataset was analysed by identifying artificially introduced effects in the data with DESeq2 as well as with OGLM and F-test. Two cases were studied; one with a smaller effect size, q = 1.5, and one with a larger effect size, q = 3. Results that were not presented for the Marine dataset in Section 4 are presented in this appendix. The results generated with DESeq2 are presented in Section B.1, while the results generated with OGLM and F-test are presented in Section B.2.

B.1 Results generated from analyses with DESeq2

All the generated ROC curves are presented in Figures B.1, B.2, B.3 and B.4.



(a) *Sequencing depth 10 k.*

(b) *Sequencing depth 100 k.*

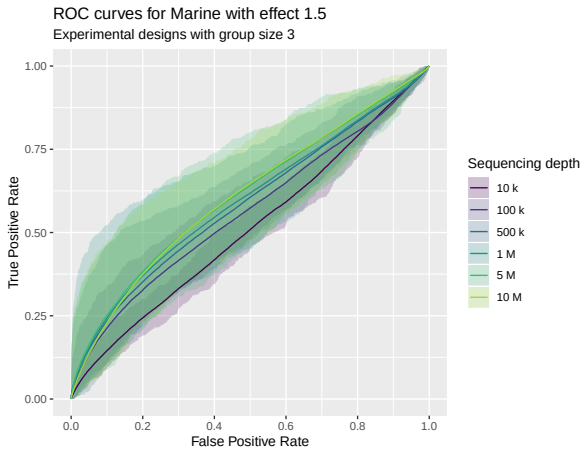
(c) *Sequencing depth 500 k.*

(d) *Sequencing depth 1 M.*

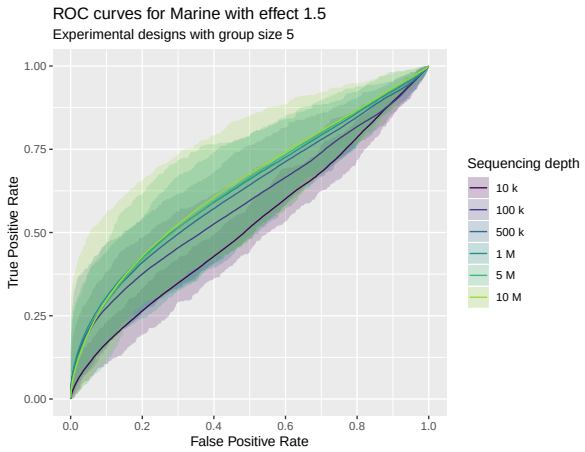
(e) *Sequencing depth 5 M.*

(f) *Sequencing depth 10 M.*

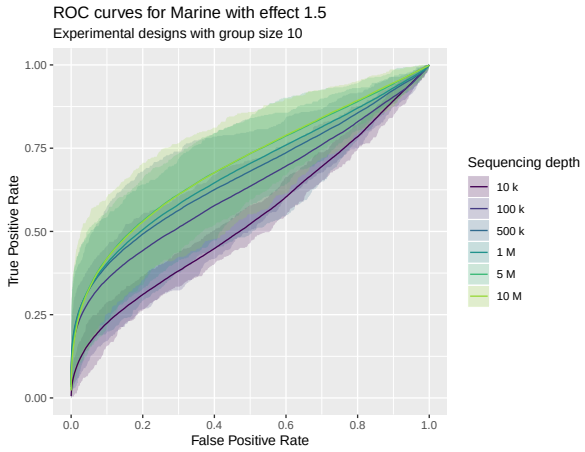
Figure B.1: *Mean ROC curves for the main experimental designs with fixed sequencing depths for Marine analysed with DESeq2. Each design with a different group size is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the small effect size for different sequencing depths.*



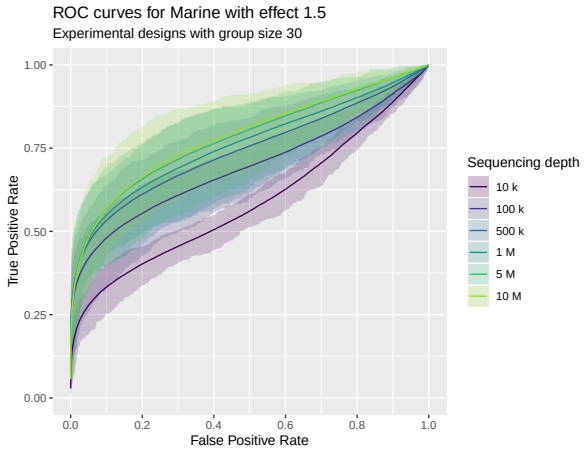
(a) *Group size 3.*



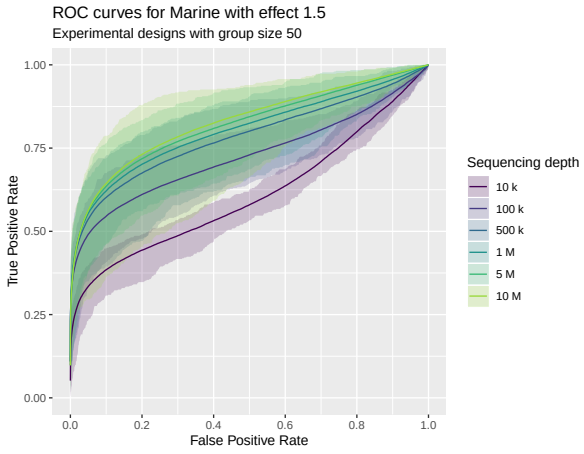
(b) *Group size 5.*



(c) *Group size 10.*

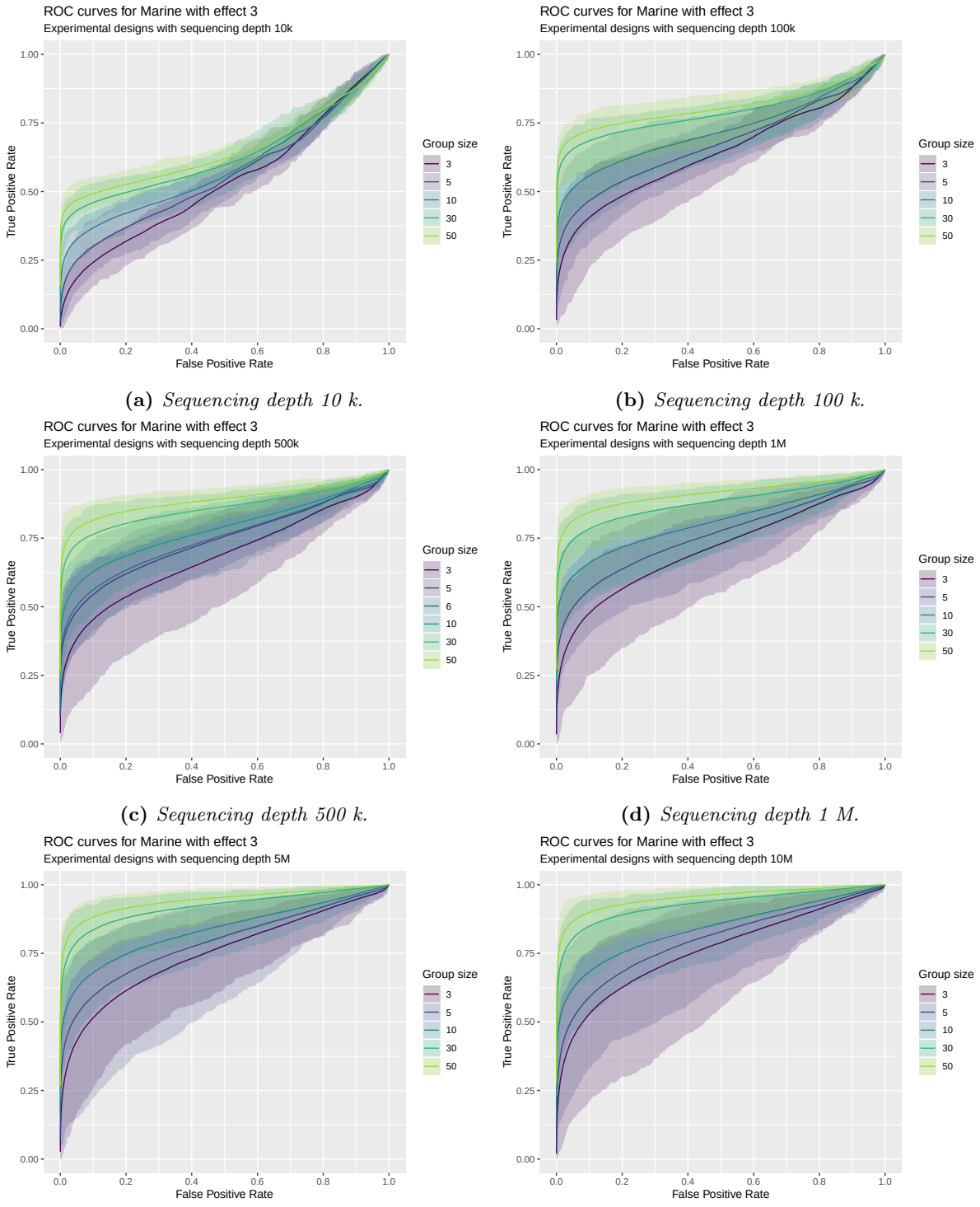


(d) *Group size 30.*



(e) *Group size 50.*

Figure B.2: *Mean ROC curves for the main experimental designs with fixed group sizes for Marine analysed with DESeq2. Each design with a different sequencing depth is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the small effect size for different group sizes.*



(a) *Sequencing depth 10 k.*

(b) *Sequencing depth 100 k.*

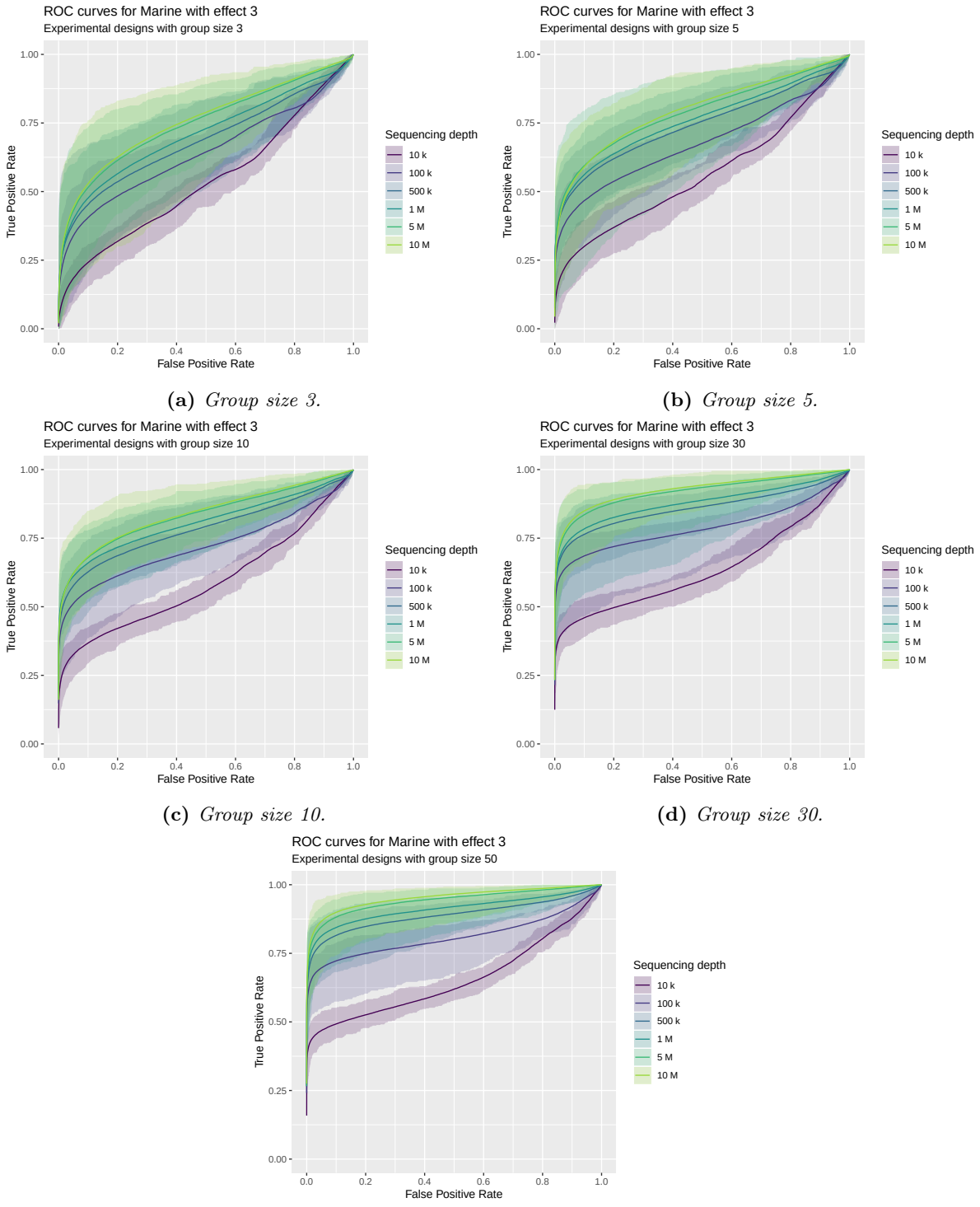
(c) *Sequencing depth 500 k.*

(d) *Sequencing depth 1 M.*

(e) *Sequencing depth 5 M.*

(f) *Sequencing depth 10 M.*

Figure B.3: *Mean ROC curves for the main experimental designs with fixed sequencing depths for Marine analysed with DESeq2. Each design with a different group size is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the large effect size for different sequencing depths.*



(a) *Group size 3.*

(b) *Group size 5.*

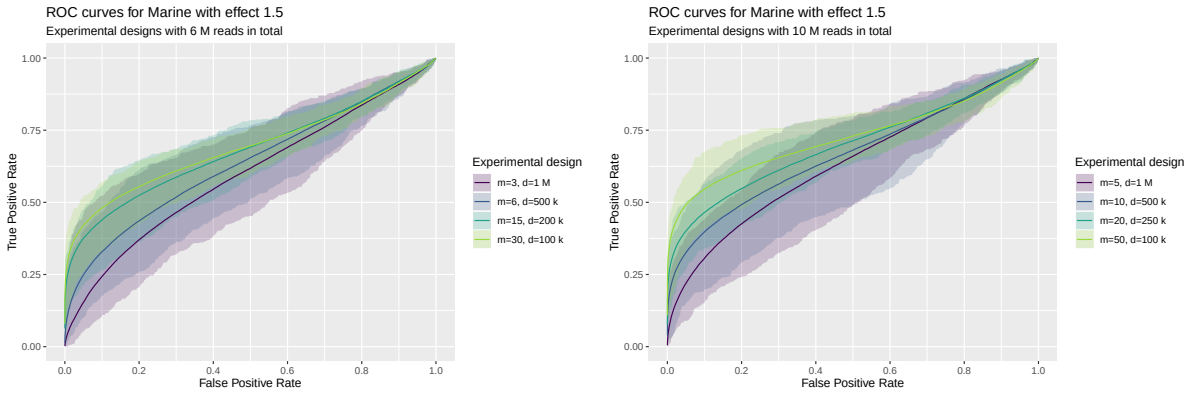
(c) *Group size 10.*

(d) *Group size 30.*

(e) *Group size 50.*

Figure B.4: *Mean ROC curves for the main experimental designs with fixed group sizes for Marine analysed with DESeq2. Each design with a different sequencing depth is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the large effect size for different group sizes.*

Figure B.5 presents ROC curves for the trade-off designs with 6 M and 10 M reads in total for the small effect size. Table B.1 presents $AUC_{0.01}$ values for these designs.



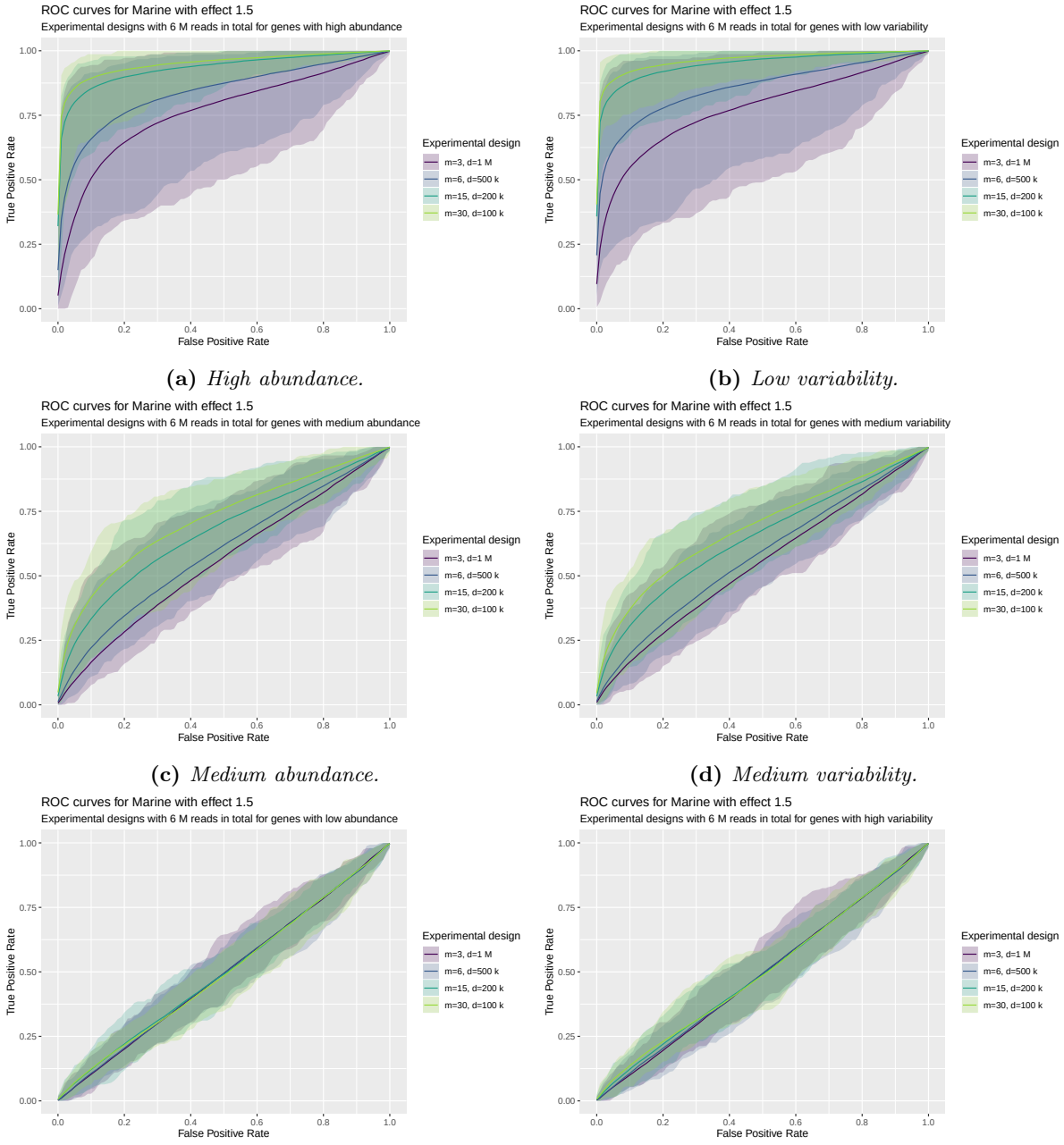
(a) *6 M reads in total.* **(b)** *10 M reads in total.*

Figure B.5: *Mean ROC curves for the trade-off designs for Marine analysed with DESeq2. Each design is represented by a colored line where the ribbon corresponds to its minimum and maximum values. The two plots present curves for the small effect size for designs with either 6 M or 10 M reads in total.*

Table B.1: *Median $AUC_{0.01}$ values for the trade-off designs for Marine analysed with DESeq2. The table presents values for the small effect size for designs with 6 M and 10 M reads in total.*

6 M reads in total		10 M reads in total	
Experimental design	$AUC_{0.01}$	Experimental design	$AUC_{0.01}$
m=3, d=1 M	0.01	m=5, d=1 M	0.05
m=6, d=500 k	0.08	m=10, d=500 k	0.17
m=15, d=200 k	0.23	m=20, d=250 k	0.25
m=30, d=100 k	0.27	m=50, d=100 k	0.35

Figure B.6 and Table B.2 show the results for the different abundance and variability strata for the trade-off designs with 6 M reads in total and the small effect size. The corresponding results for designs with 10 M reads in total are presented in Figure B.7 and Table B.3 for the small effect size and in Figure B.8 and Table B.4 for the large effect size.



(a) *High abundance.*

(b) *Low variability.*

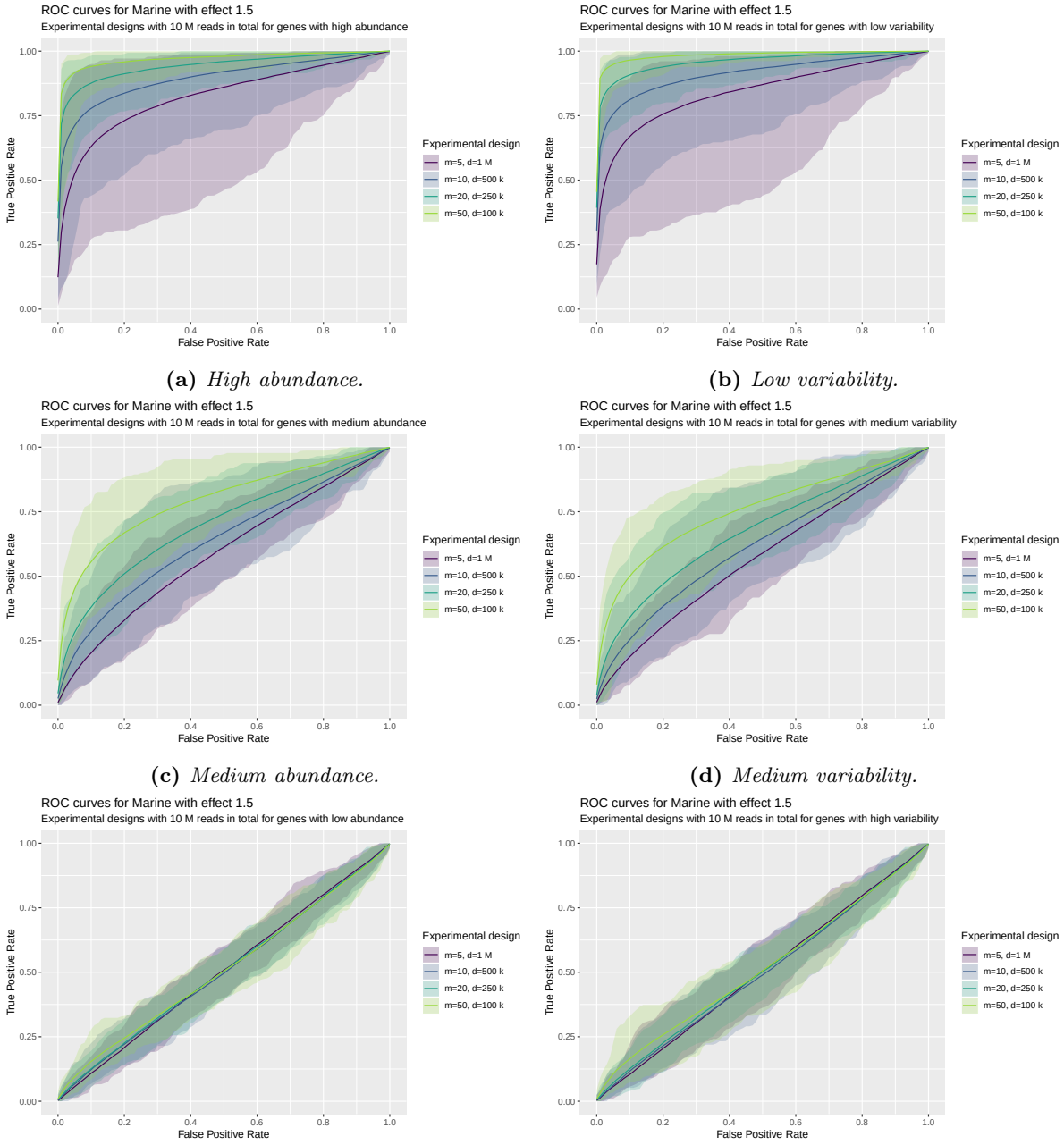
(c) *Medium abundance.*

(d) *Medium variability.*

(e) *Low abundance.*

(f) *High variability.*

Figure B.6: *Mean ROC curves for different abundance and variability strata for Marine analysed with DESeq2. The plots present curves for the small effect size for the trade-off designs with 6 M reads in total. Each design is represented by a colored line where the ribbon corresponds to its minimum and maximum values.*



(a) *High abundance.*

(b) *Low variability.*

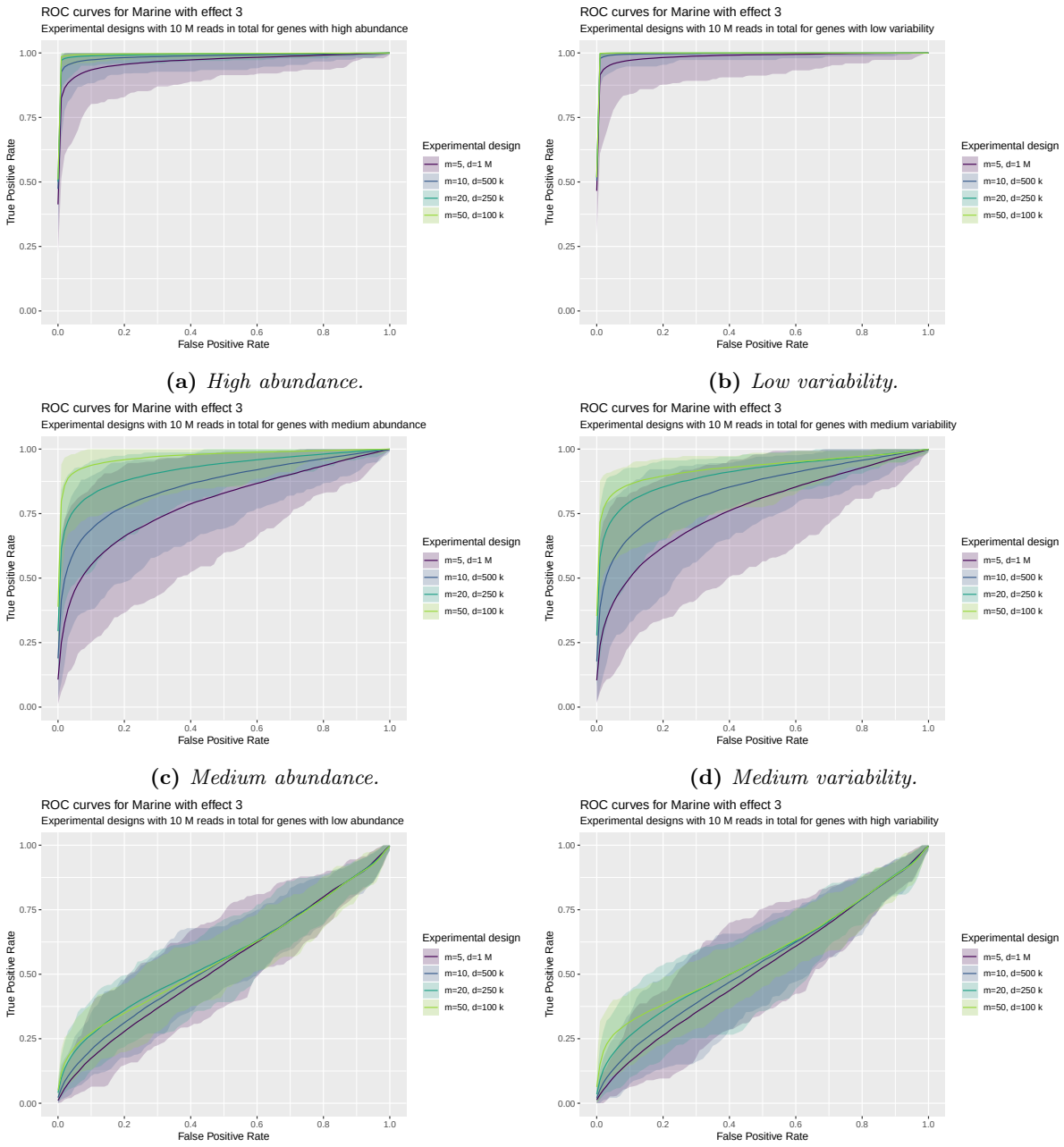
(c) *Medium abundance.*

(d) *Medium variability.*

(e) *Low abundance.*

(f) *High variability.*

Figure B.7: *Mean ROC curves for different abundance and variability strata for Marine analysed with DESeq2. The plots present curves for the small effect size for the trade-off designs with 10 M reads in total. Each design is represented by a colored line where the ribbon corresponds to its minimum and maximum values.*



(a) *High abundance.*

(b) *Low variability.*

(c) *Medium abundance.*

(d) *Medium variability.*

(e) *Low abundance.*

(f) *High variability.*

Figure B.8: *Mean ROC curves for different abundance and variability strata for Marine analysed with DESeq2. The plots present curves for the large effect size for the trade-off designs with 10 M reads in total. Each design is represented by a colored line where the ribbon corresponds to its minimum and maximum values.*

Table B.2: *Median $AUC_{0.01}$ values for different abundance and variability strata for Marine analysed with DESeq2. The table presents values for the small effect size for the trade-off designs with 6 M reads in total.*

	Abundance			Variability		
Experimental Design	Low	Medium	High	Low	Medium	High
m=3, d=1 M	0.005	0.009	0.036	0.130	0.013	0.005
m=6, d=500 k	0.005	0.016	0.237	0.360	0.020	0.006
m=15, d=200 k	0.009	0.057	0.633	0.675	0.057	0.007
m=30, d=100 k	0.011	0.089	0.709	0.778	0.078	0.011

Table B.3: *Median $AUC_{0.01}$ values for different abundance and variability strata for Marine analysed with DESeq2. The table presents values for the small effect size for the trade-off designs with 10 M reads in total.*

	Abundance			Variability		
Experimental Design	Low	Medium	High	Low	Medium	High
m=5, d=1 M	0.005	0.014	0.194	0.305	0.019	0.005
m=10, d=500 k	0.008	0.040	0.468	0.572	0.042	0.007
m=20, d=250 k	0.010	0.080	0.659	0.740	0.063	0.010
m=50, d=100 k	0.016	0.179	0.818	0.876	0.141	0.015

Table B.4: *Median $AUC_{0.01}$ values for different abundance and variability strata for Marine analysed with DESeq2. The table presents values for the large effect size for the trade-off designs with 10 M reads in total.*

	Abundance			Variability		
Experimental Design	Low	Medium	High	Low	Medium	High
m=5, d=1 M	0.016	0.166	0.819	0.913	0.178	0.017
m=10, d=500 k	0.034	0.333	0.922	0.975	0.303	0.034
m=20, d=250 k	0.072	0.550	0.967	0.999	0.516	0.058
m=50, d=100 k	0.086	0.759	0.985	1.000	0.686	0.106

The true FDR values at an estimated FDR of 0.05 for the main experimental designs for the large effect size are presented in Figure B.9.

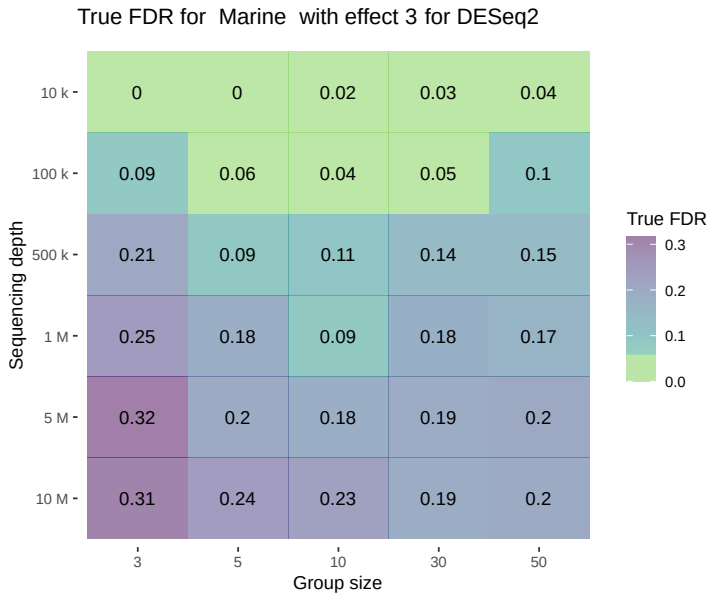
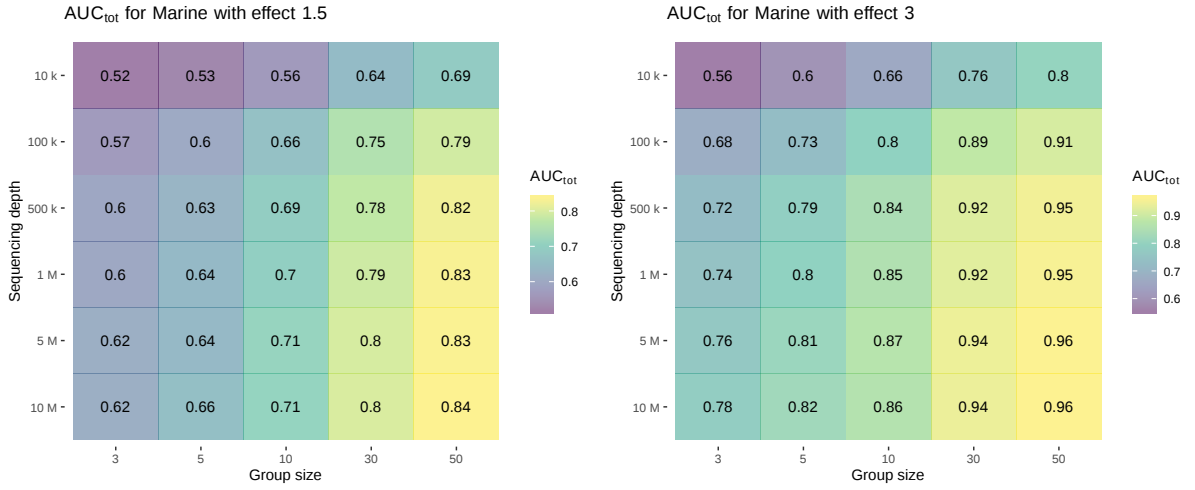


Figure B.9: *Median true FDR values at an estimated FDR of 0.05 for the main experimental designs for Marine analysed with DESeq2. All values below the estimated value have the same color. The heatmap presents values for the large effect size.*

B.2 Results generated from analyses with OGLM and F-test

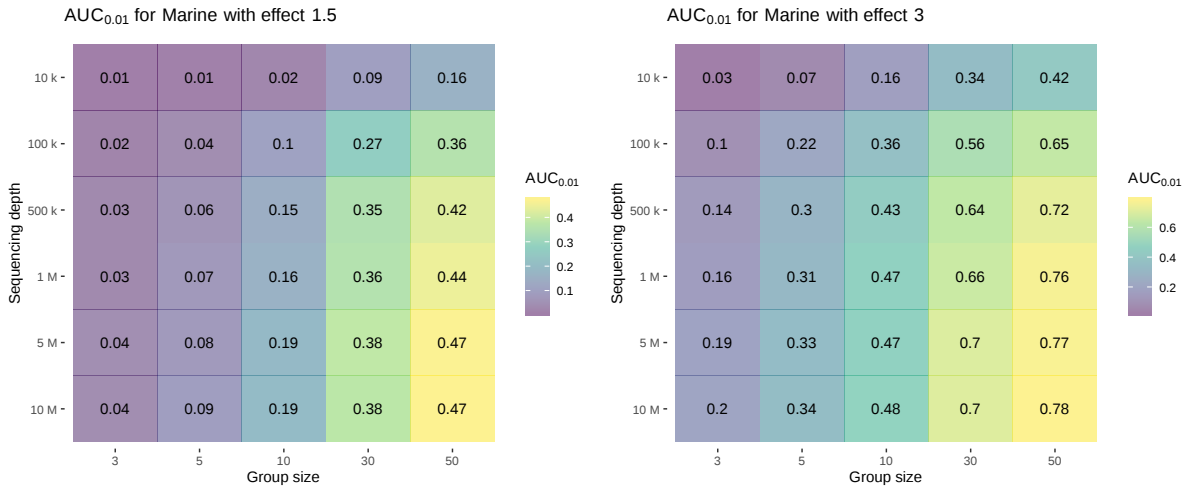
Figure B.10 and Figure B.11 presents heatmaps with AUC_{tot} and $\text{AUC}_{0.01}$ values, respectively, for all the main experimental designs and both effect sizes.



(a) *Small effect size.*

(b) *Large effect size.*

Figure B.10: *Median AUC_{tot} values for the main experimental designs for Marine analysed with OGLM and F-test. The two heatmaps present values for different effect sizes. Note that the color scales are based on the ranges of the values in the heatmaps, and are thus not directly comparable between the two heatmaps.*



(a) *Small effect size.*

(b) *Large effect size.*

Figure B.11: *Median $AUC_{0.01}$ values for the main experimental designs for Marine analysed with OGLM and F-test. The two heatmaps present values for different effect sizes. Note that the color scales are based on the ranges of the values in the heatmaps, and are thus not directly comparable between the two heatmaps.*

The true FDR values at an estimated FDR of 0.05 for the main experimental designs for the large effect size are presented in Figure B.12.

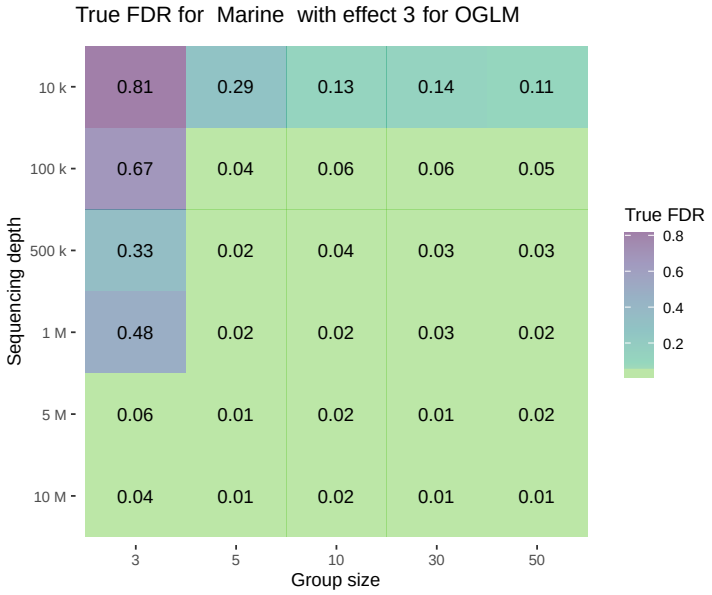
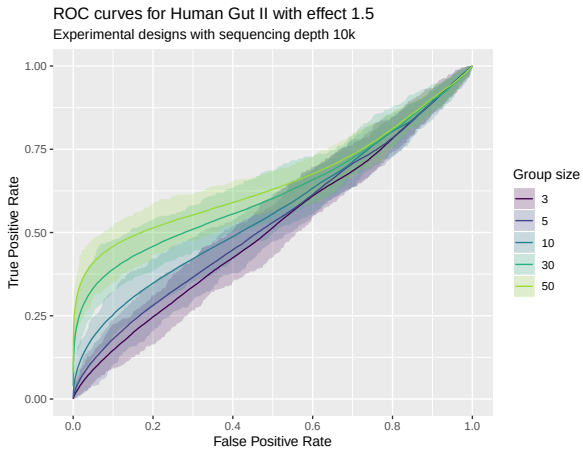


Figure B.12: *Median true FDR values at an estimated FDR of 0.05 for the main experimental designs for Marine analysed with OGLM and F-test. All values below the estimated value have the same color. The heatmap presents values for the large effect size.*

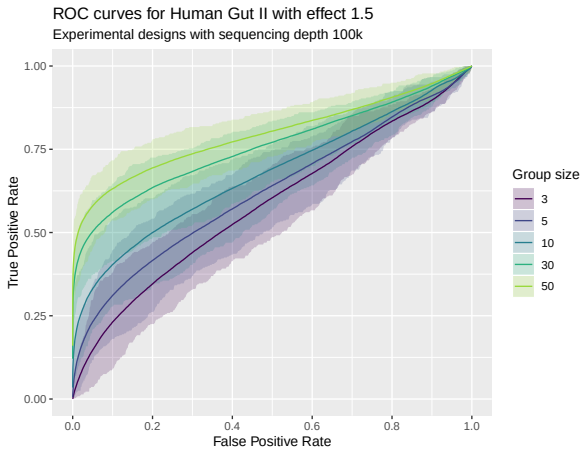
C. Additional results for Human Gut II

The Human Gut II dataset was analysed by identifying artificially introduced effects in the data with DESeq2. Two cases were studied; one with a smaller effect size, q = 1.5, and one with a larger effect size, q = 3. Results that were not presented for the Human Gut II dataset in Section 4 are presented in this appendix.

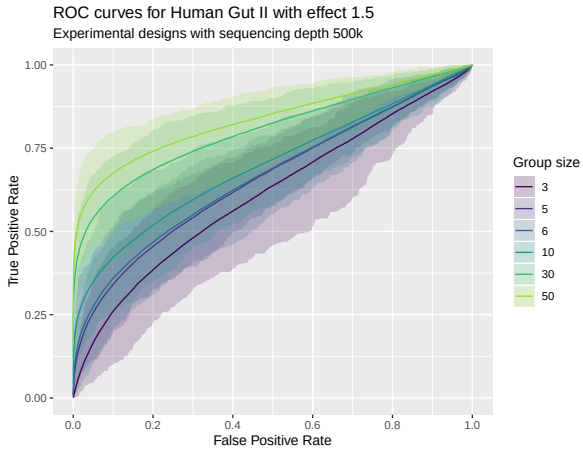
All the generated ROC curves are presented in Figures C.1, C.2, C.3 and C.4.



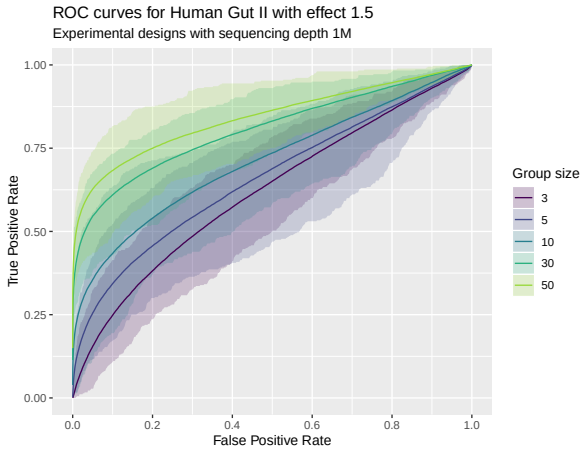
(a) *Sequencing depth 10 k.*



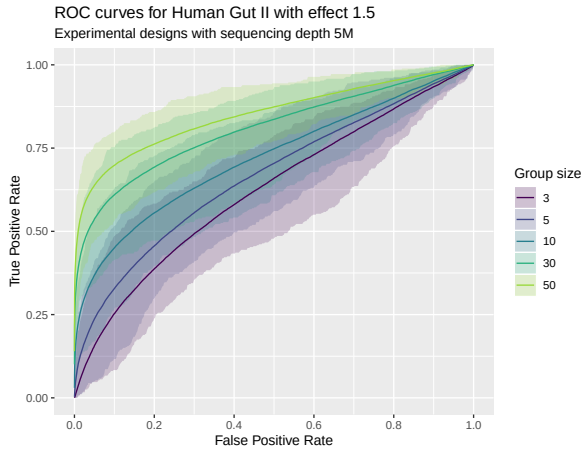
(b) *Sequencing depth 100 k.*



(c) *Sequencing depth 500 k.*

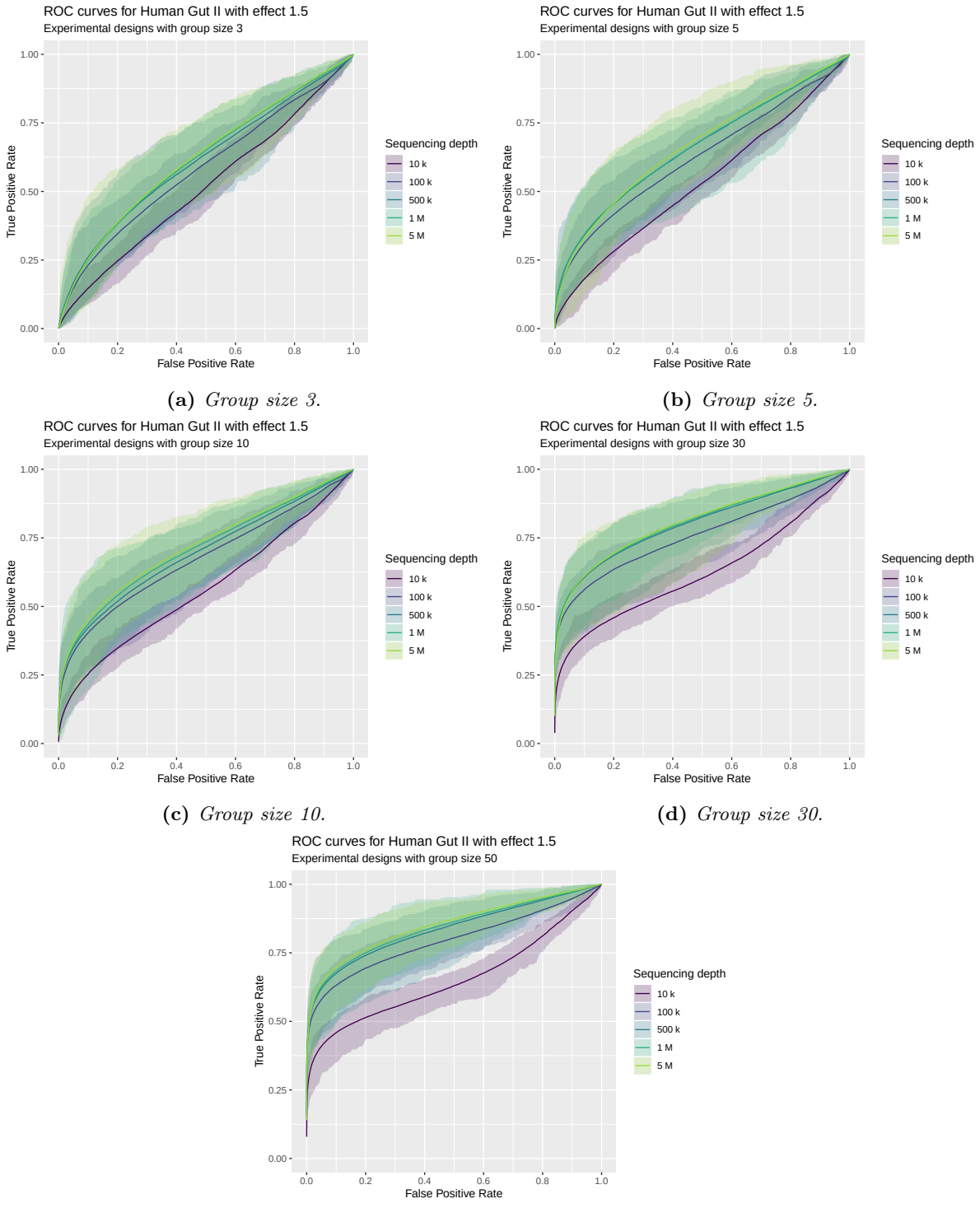


(d) *Sequencing depth 1 M.*



(e) *Sequencing depth 5 M.*

Figure C.1: *Mean ROC curves for the main experimental designs with fixed sequencing depths for Human Gut II analysed with DESeq2. Each design with a different group size is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the small effect size for different sequencing depths.*



(a) *Group size 3.*

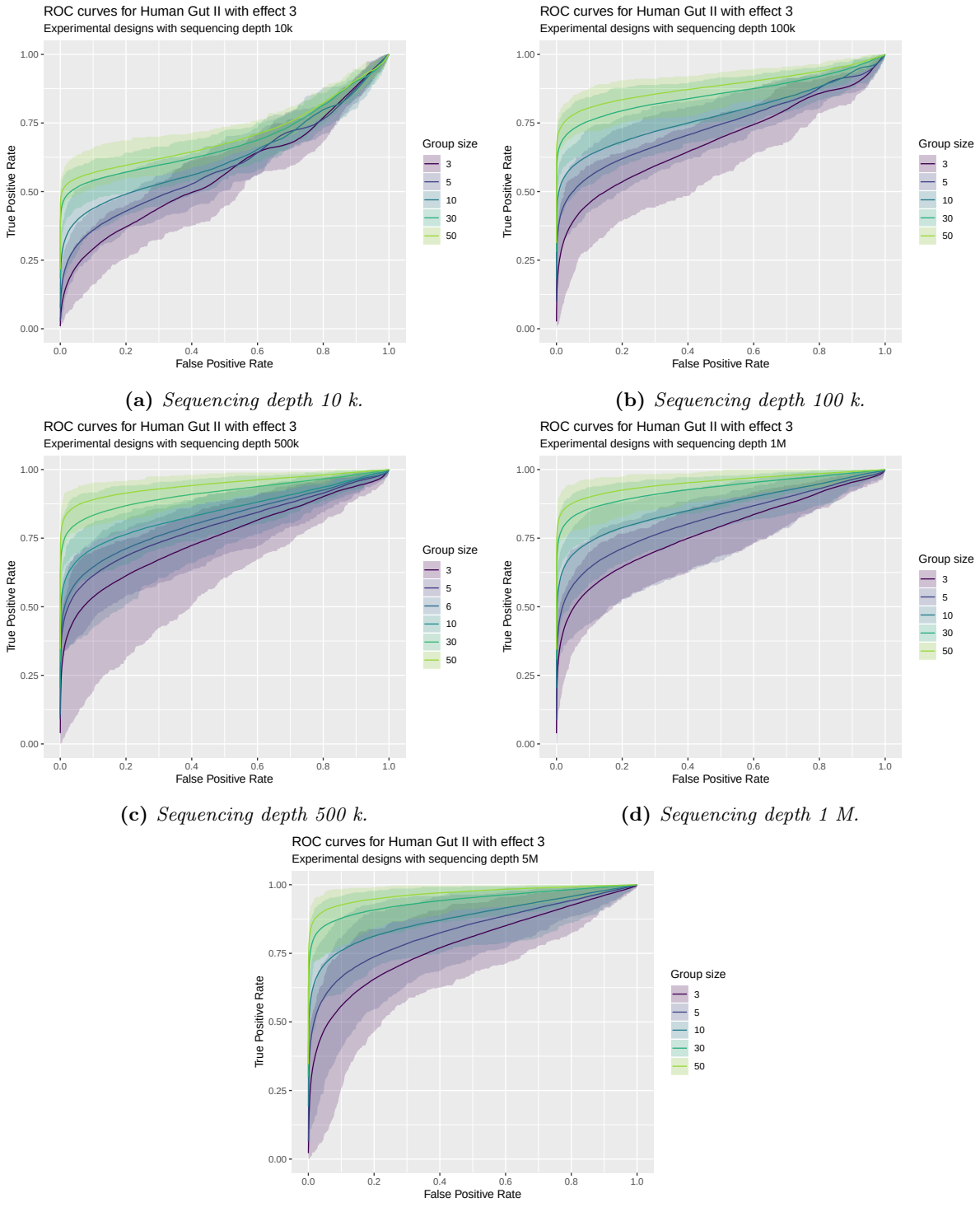
(b) *Group size 5.*

(c) *Group size 10.*

(d) *Group size 30.*

(e) *Group size 50.*

Figure C.2: *Mean ROC curves for the main experimental designs with fixed group sizes for Human Gut II analysed with DESeq2. Each design with a different sequencing depth is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the small effect size for different group sizes.*



(a) *Sequencing depth 10 k.*

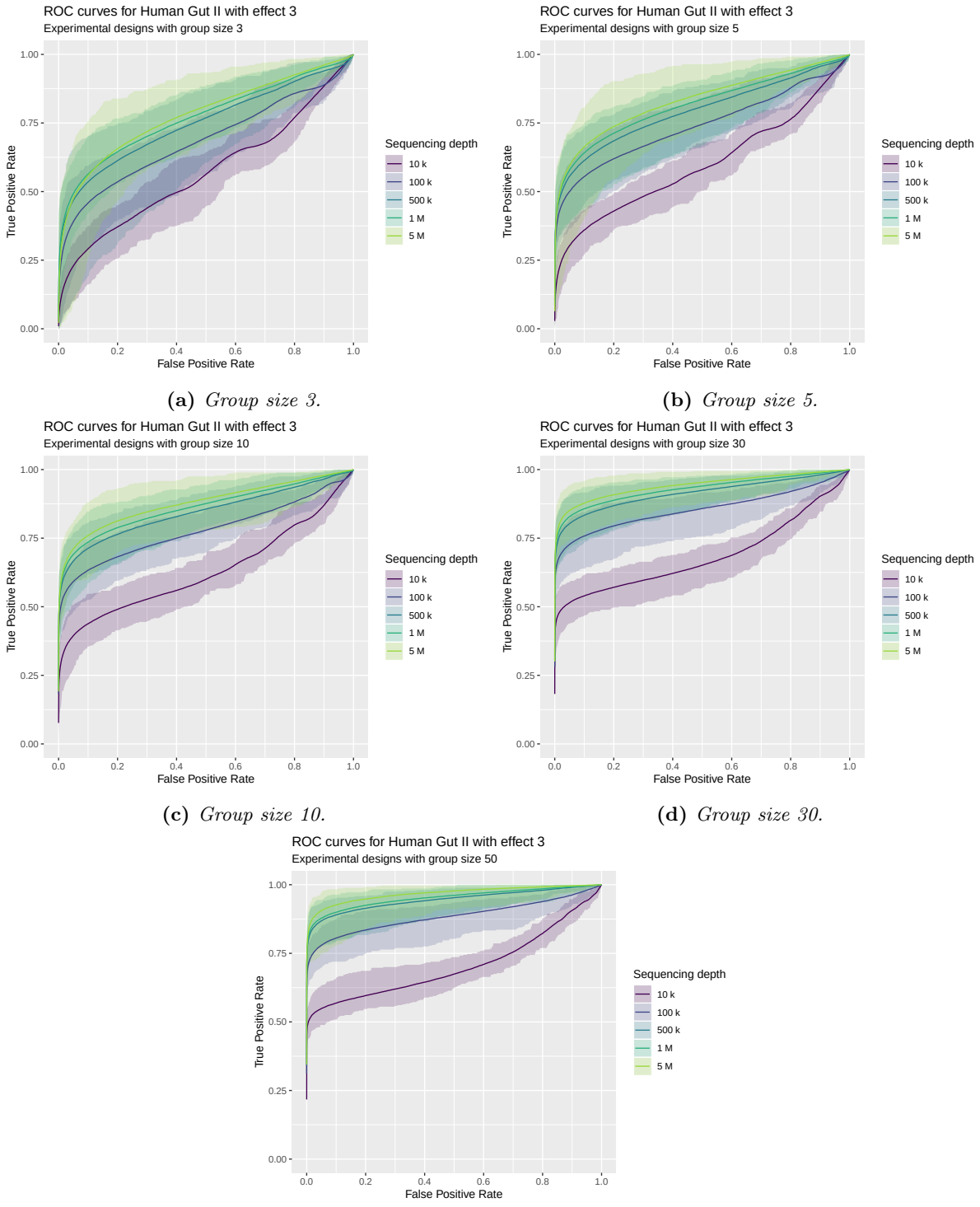
(b) *Sequencing depth 100 k.*

(c) *Sequencing depth 500 k.*

(d) *Sequencing depth 1 M.*

(e) *Sequencing depth 5 M.*

Figure C.3: *Mean ROC curves for the main experimental designs with fixed sequencing depths for Human Gut II analysed with DESeq2. Each design with a different group size is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the large effect size for different sequencing depths.*



(a) *Group size 3.*

(b) *Group size 5.*

(c) *Group size 10.*

(d) *Group size 30.*

(e) *Group size 50.*

Figure C.4: *Mean ROC curves for the main experimental designs with fixed group sizes for Human Gut II analysed with DESeq2. Each design with a different sequencing depth is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the large effect size for different group sizes.*

The ROC curves are summarised in heatmaps in Figure C.5 and Figure C.6 for the small effect size, where the computed AUC_{tot} and $\mathrm{AUC}_{0.01}$ values are presented respectively.

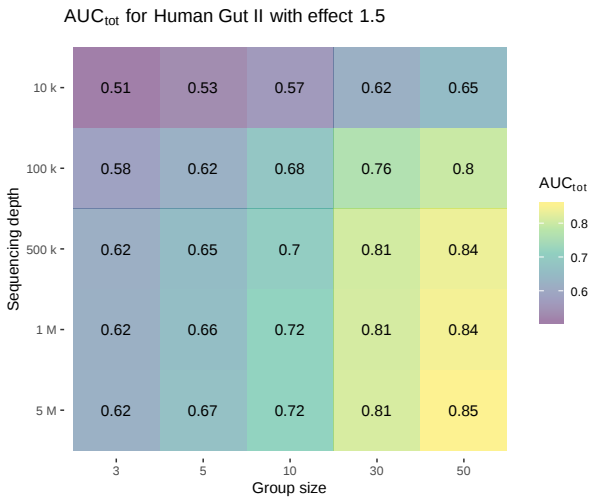


Figure C.5: *Median AUC_{tot} values for the main experimental designs for Human Gut II analysed with DESeq2. The heatmap presents values for the small effect size.*

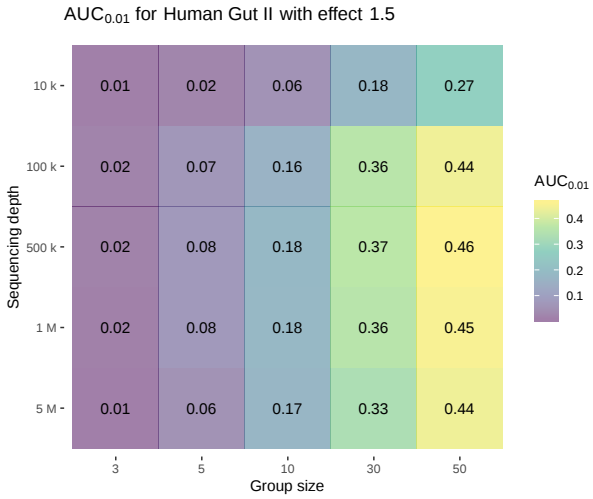
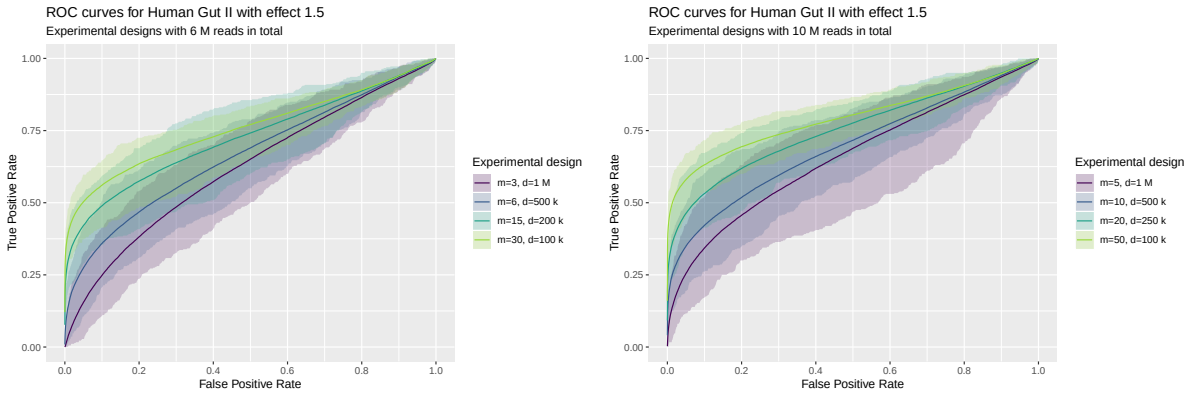


Figure C.6: *Median $AUC_{0.01}$ values for the main experimental designs for Human Gut II analysed with DESeq2. The heatmap presents values for the small effect size.*

The ROC curves for the trade-off designs with 6 M and 10 M reads in total are presented in Figure C.7 for the small effect size and in Figure C.8 for the large effect size. Table C.1 and Table C.2 present $AUC_{0.01}$ values for these designs for the small and large effect size respectively.

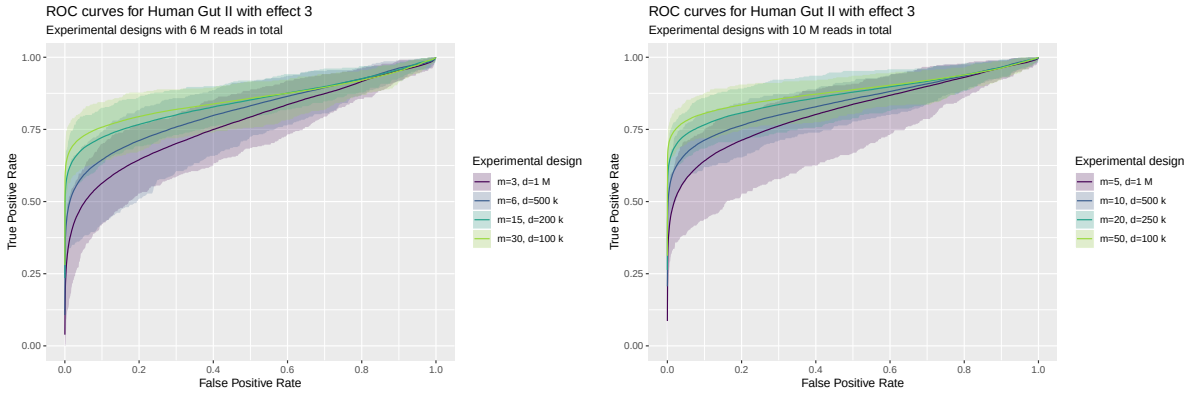


(a) *6 M reads in total.* **(b)** *10 M reads in total.*

Figure C.7: *Mean ROC curves for the trade-off designs for Human Gut II analysed with DESeq2. Each design is represented by a colored line where the ribbon corresponds to minimum and maximum values. The two plots present curves for the small effect size for designs with either 6 M or 10 M reads in total.*

Table C.1: *Median $AUC_{0.01}$ values for the trade-off designs for Human Gut II analysed with DESeq2. The table presents values for the small effect size for designs with 6 M and 10 M reads in total.*

6 M reads in total		10 M reads in total	
Experimental design	$AUC_{0.01}$	Experimental design	$AUC_{0.01}$
m=3, d=1 M	0.02	m=5, d=1 M	0.08
m=6, d=500 k	0.10	m=10, d=500 k	0.18
m=15, d=200 k	0.26	m=20, d=250 k	0.30
m=30, d=100 k	0.36	m=50, d=100 k	0.44



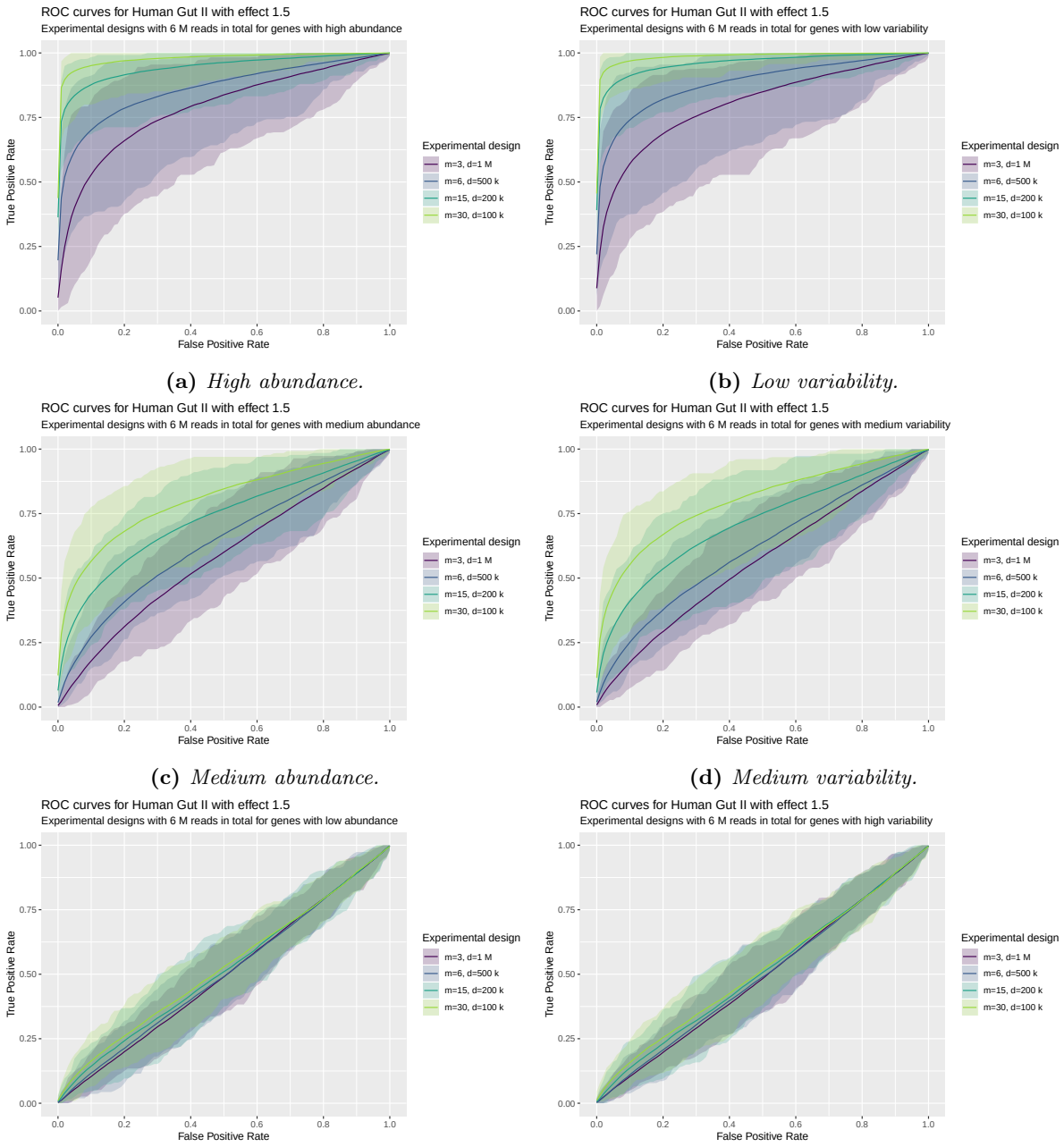
(a) *6 M reads in total.* **(b)** *10 M reads in total.*

Figure C.8: *Mean ROC curves for the trade-off designs for Human Gut II analysed with DESeq2. Each design is represented by a colored line where the ribbon corresponds to minimum and maximum values. The two plots present curves for the large effect size for designs with either 6 M or 10 M reads in total.*

Table C.2: *Median $AUC_{0.01}$ values for the trade-off designs for Human Gut II analysed with DESeq2. The table presents values for the large effect size for designs with 6 M and 10 M reads in total.*

6 M reads in total		10 M reads in total	
Experimental design	$AUC_{0.01}$	Experimental design	$AUC_{0.01}$
m=3, d=1 M	0.28	m=5, d=1 M	0.39
m=6, d=500 k	0.43	m=10, d=500 k	0.54
m=15, d=200 k	0.57	m=20, d=250 k	0.63
m=30, d=100 k	0.64	m=50, d=100 k	0.70

Figure C.9 and Table C.3 show the results for the different abundance and variability strata for the trade-off designs with 6 M reads in total and the small effect size, while Figure C.10 and Table C.4 show the results for the large effect size. The corresponding results for designs with 10 M reads in total are presented in Figure C.11 and Table C.5 for the small effect size and in Figure C.12 and Table C.6 for the large effect size.



(a) *High abundance.*

(b) *Low variability.*

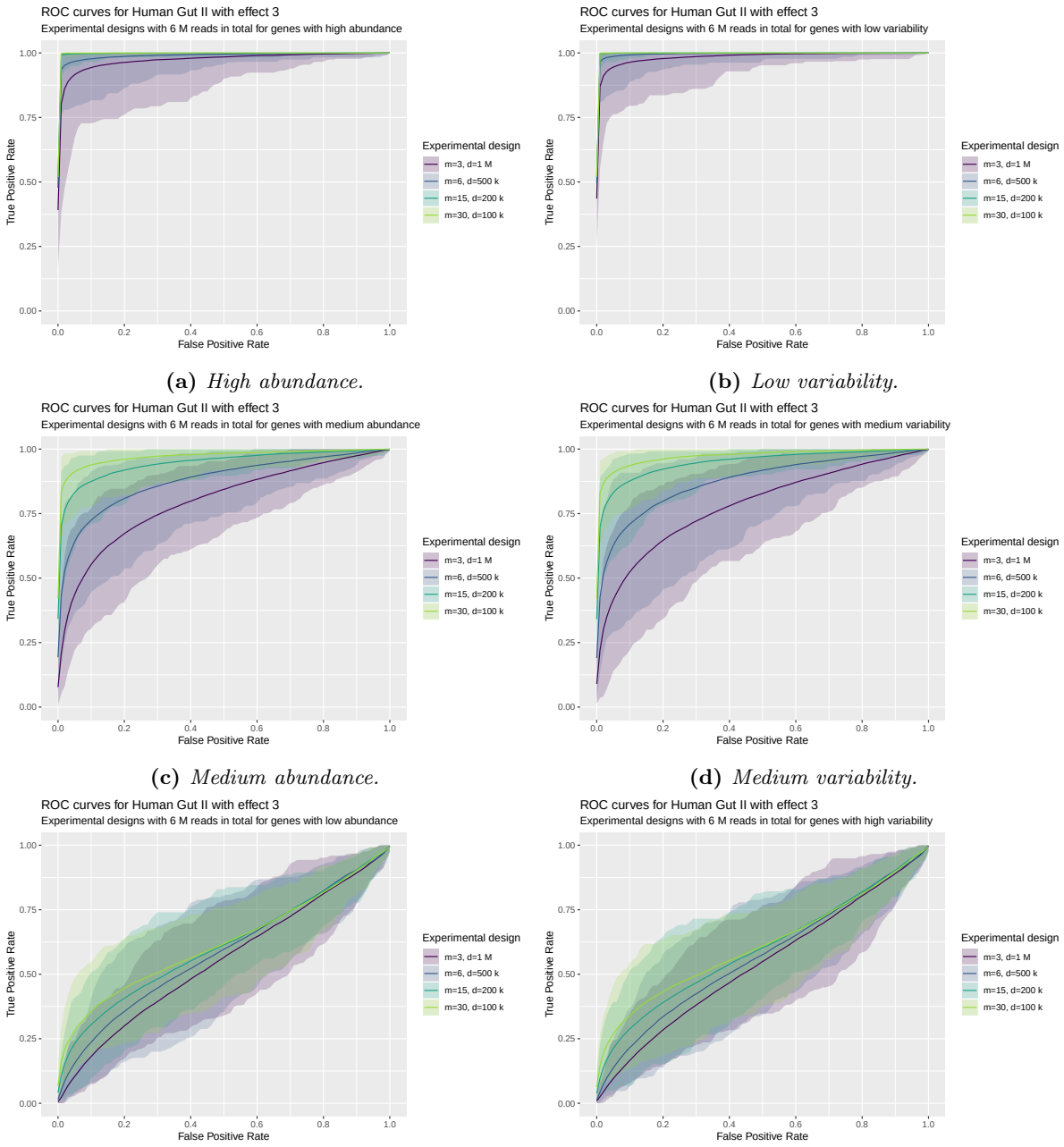
(c) *Medium abundance.*

(d) *Medium variability.*

(e) *Low abundance.*

(f) *High variability.*

Figure C.9: *Mean ROC curves for different abundance and variability strata for Human Gut II analysed with DESeq2. The plots present curves for the small effect size for the trade-off designs with 6 M reads in total. Each design is represented by a colored line where the ribbon corresponds to minimum and maximum values.*



(a) *High abundance.*

(b) *Low variability.*

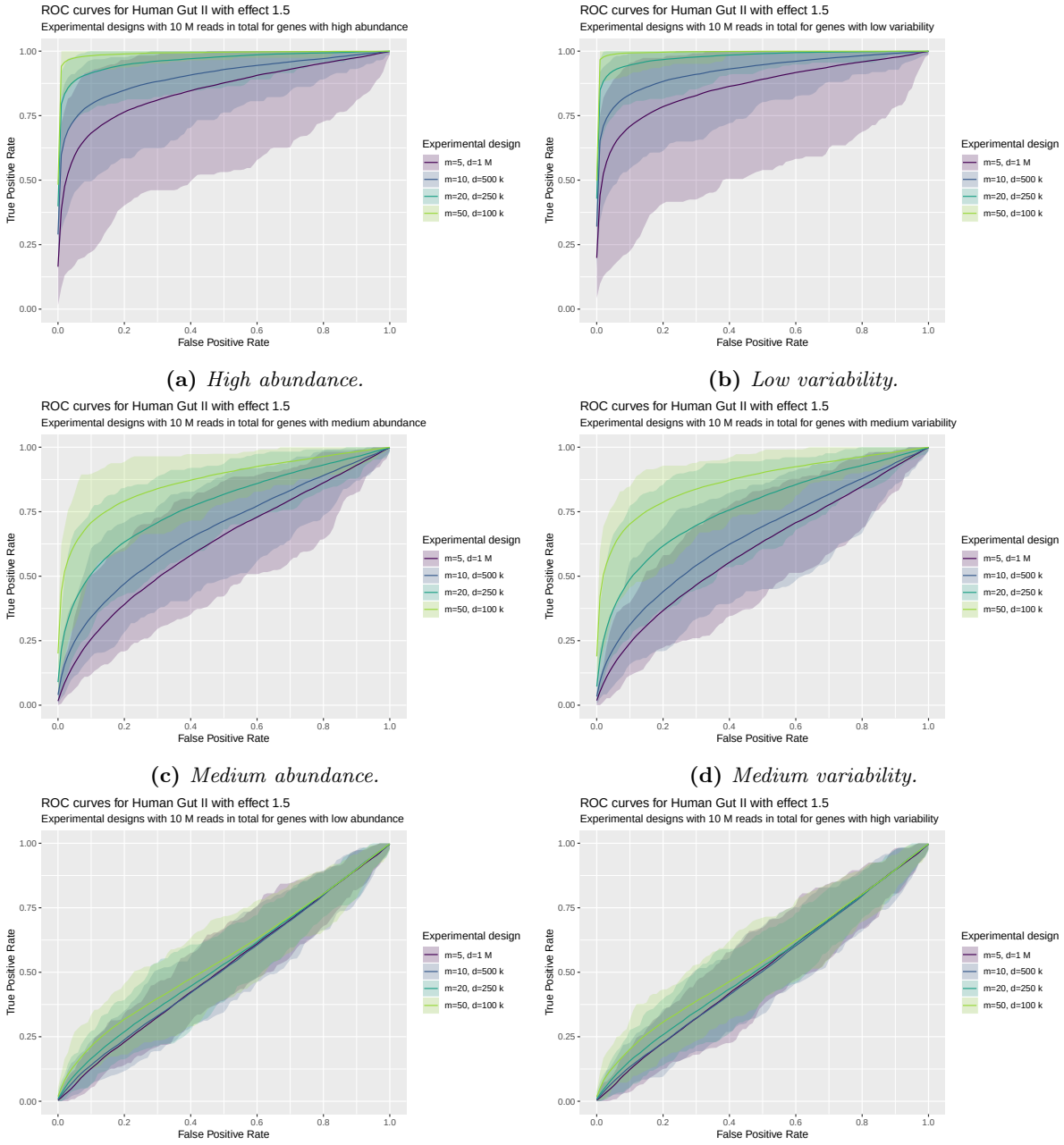
(c) *Medium abundance.*

(d) *Medium variability.*

(e) *Low abundance.*

(f) *High variability.*

Figure C.10: *Mean ROC curves for different abundance and variability strata for Human Gut II analysed with DESeq2. The plots present curves for the large effect size for the trade-off designs with 6 M reads in total. Each design is represented by a colored line where the ribbon corresponds to minimum and maximum values.*



(a) *High abundance.*

(b) *Low variability.*

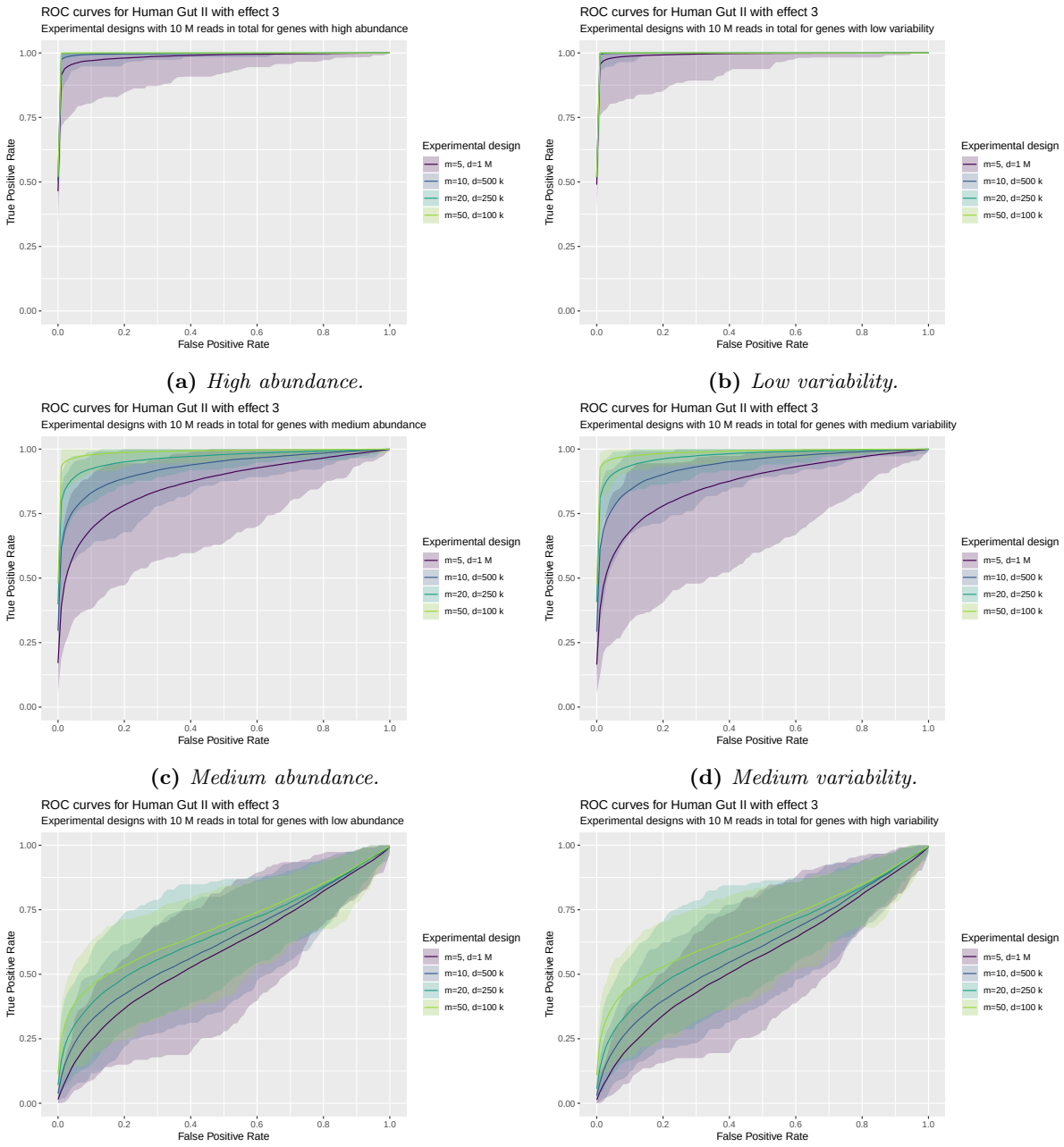
(c) *Medium abundance.*

(d) *Medium variability.*

(e) *Low abundance.*

(f) *High variability.*

Figure C.11: *Mean ROC curves for different abundance and variability strata for Human Gut II analysed with DESeq2. The plots present curves for the small effect size for the trade-off designs with 10 M reads in total. Each design is represented by a colored line where the ribbon corresponds to minimum and maximum values.*



(a) *High abundance.*

(b) *Low variability.*

(c) *Medium abundance.*

(d) *Medium variability.*

(e) *Low abundance.*

(f) *High variability.*

Figure C.12: *Mean ROC curves for different abundance and variability strata for Human Gut II analysed with DESeq2. The plots present curves for the large effect size for the trade-off designs with 10 M reads in total. Each design is represented by a colored line where the ribbon corresponds to minimum and maximum values.*

Table C.3: *Median $AUC_{0.01}$ values for different abundance and variability strata for Human Gut II analysed with DESeq2. The table presents values for the small effect size for the trade-off designs with 6 M reads in total.*

	Abundance			Variability		
Experimental Design	Low	Medium	High	Low	Medium	High
m=3, d=1 M	0.00	0.01	0.07	0.13	0.01	0.01
m=6, d=500 k	0.01	0.03	0.35	0.41	0.03	0.01
m=15, d=200 k	0.01	0.11	0.67	0.73	0.09	0.01
m=30, d=100 k	0.02	0.22	0.83	0.87	0.19	0.02

Table C.4: *Median $AUC_{0.01}$ values for different abundance and variability strata for Human Gut II analysed with DESeq2. The table presents values for the large effect size for the trade-off designs with 6 M reads in total.*

	Abundance			Variability		
Experimental Design	Low	Medium	High	Low	Medium	High
m=3, d=1 M	0.01	0.13	0.74	0.83	0.15	0.01
m=6, d=500 k	0.02	0.33	0.91	0.96	0.33	0.02
m=15, d=200 k	0.07	0.63	0.99	1.00	0.63	0.06
m=30, d=100 k	0.11	0.80	1.00	1.00	0.80	0.10

Table C.5: *Median $AUC_{0.01}$ values for different abundance and variability strata for Human Gut II analysed with DESeq2. The table presents values for the small effect size for the trade-off designs with 10 M reads in total.*

	Abundance			Variability		
Experimental Design	Low	Medium	High	Low	Medium	High
m=5, d=1 M	0.01	0.03	0.28	0.36	0.03	0.01
m=10, d=500 k	0.01	0.06	0.51	0.58	0.05	0.01
m=20, d=250 k	0.01	0.15	0.75	0.80	0.12	0.01
m=50, d=100 k	0.03	0.35	0.92	0.96	0.33	0.03

Table C.6: *Median $AUC_{0.01}$ values for different abundance and variability strata for Human Gut II analysed with DESeq2. The table presents values for the large effect size for the trade-off designs with 10 M reads in total.*

	Abundance			Variability		
Experimental Design	Low	Medium	High	Low	Medium	High
m=5, d=1 M	0.03	0.28	0.89	0.95	0.29	0.02
m=10, d=500 k	0.07	0.54	0.97	1.00	0.52	0.04
m=20, d=250 k	0.11	0.75	1.00	1.00	0.77	0.09
m=50, d=100 k	0.19	0.91	1.00	1.00	0.90	0.18

The true FDR values at an estimated FDR of 0.05 for the main experimental designs are presented in Figure C.13 for both effect sizes.

(a) *Small effect size.* **(b)** *Large effect size.*

Figure C.13: *Median true FDR values at an estimated FDR of 0.05 for the main experimental designs for Human Gut II analysed with DESeq2. All values below the estimated value have the same color. Cases with too few observed values are displayed as "NA". The heatmaps present values for different effect sizes. Note that the color scales are based on the ranges of the values in the heatmaps, and are thus not directly comparable between the two heatmaps.*

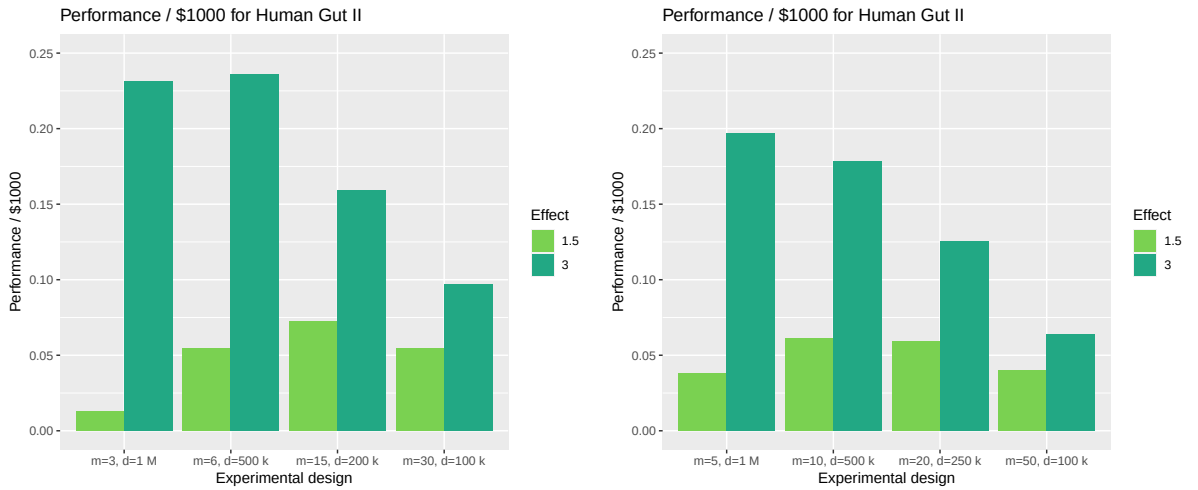
Table C.7 and Table C.8 presents the sequencing cost, the performance and the ratio *performance/$1000* for the trade-off designs with 6 M and 10 M reads in total respectively, for both effect sizes. Figure C.14 visualises the ratios in two barplots for either 6 M or 10 M reads in total.

Table C.7: *Sequencing cost, performance and the ratio performance/$1000 for the trade-off designs with 6 M reads in total for Human Gut II analysed with DESeq2. The cost is displayed in US dollars and the performance is displayed in $AUC_{0.01}$. The table presents values for both effect sizes.*

		Effect 1.5		Effect 3	
Experimental design	Cost	$AUC_{0.01}$	Performance/$1000	$AUC_{0.01}$	Performance/$1000
m=3, d=1 M	1200	0.02	0.01	0.28	0.23
m=6, d=500 k	1800	0.1	0.05	0.43	0.24
m=15, d=200 k	3600	0.26	0.07	0.57	0.16
m=30, d=100 k	6600	0.36	0.05	0.64	0.10

Table C.8: *Sequencing cost, performance and the ratio performance/$1000 for the trade-off designs with 10 M reads in total for Human Gut II analysed with DESeq2. The cost is displayed in US dollars and the performance is displayed in $AUC_{0.01}$. The table presents values for both effect sizes.*

		Effect 1.5		Effect 3	
Experimental design	Cost	$AUC_{0.01}$	Performance/$1000	$AUC_{0.01}$	Performance/$1000
m=5, d=1 M	2000	0.08	0.04	0.39	0.20
m=10, d=500 k	3000	0.18	0.06	0.54	0.18
m=20, d=250 k	5000	0.3	0.06	0.63	0.13
m=50, d=100 k	11000	0.44	0.04	0.7	0.06



(a) *6 M reads in total.*

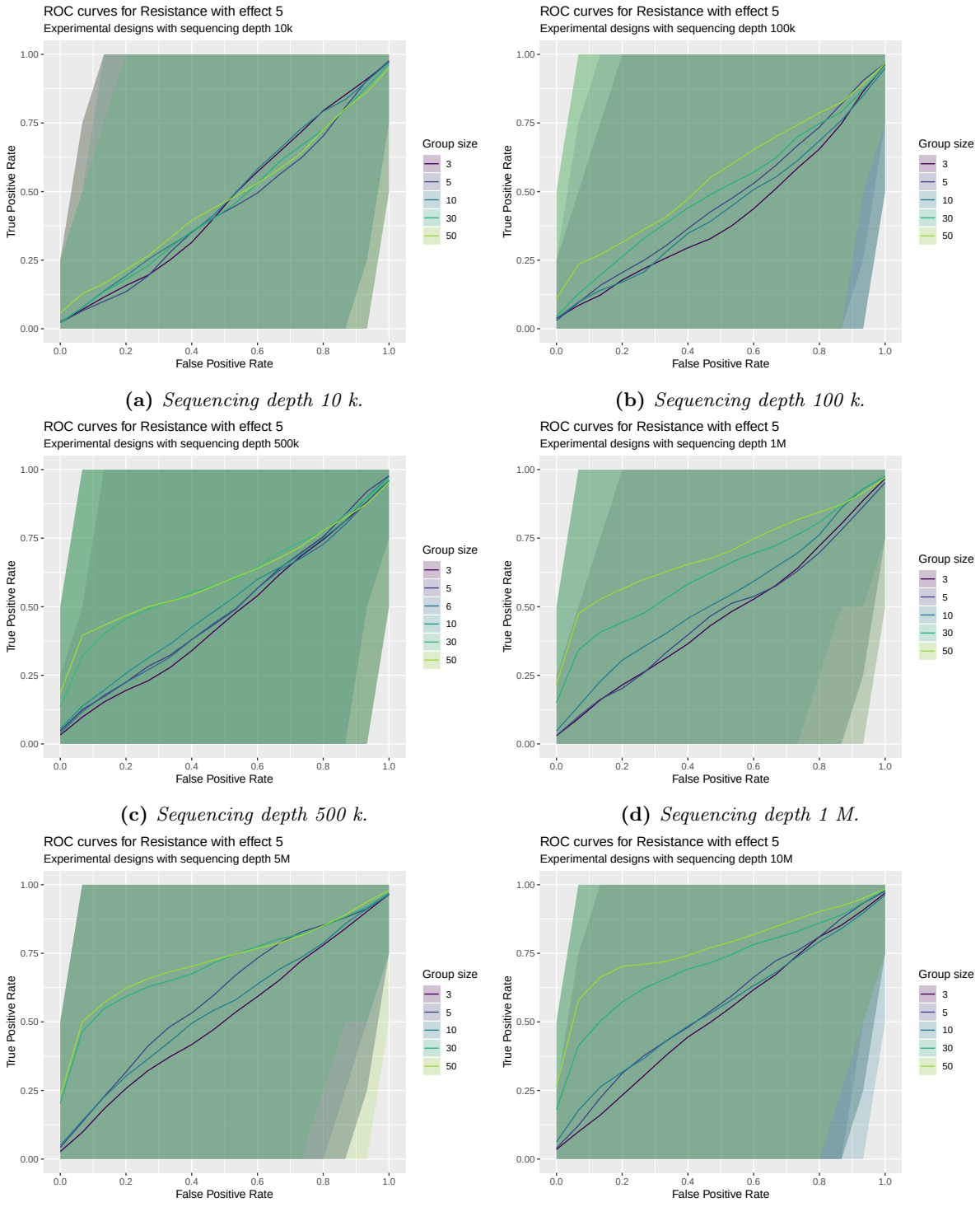
(b) *10 M reads in total.*

Figure C.14: *The ratio of performance and sequencing cost, performance/$1000, for the trade-off designs for Human Gut II analysed with DESeq2. The performance is measured in $AUC_{0.01}$ and the cost in US dollars. The two plots present values for either 6 M or 10 M reads in total and each plot contains both effect sizes.*

D. Additional results for Resistance

The Resistance dataset was analysed by identifying artificially introduced effects in the data with DESeq2. Two cases were studied; one with a smaller effect size, q = 5, and one with a larger effect size, q = 10. Results that were not presented for the Resistance dataset in Section 4 are presented in this appendix.

All the generated ROC curves are presented in Figures D.1, D.2, D.3 and D.4.



(a) *Sequencing depth 10 k.*

(b) *Sequencing depth 100 k.*

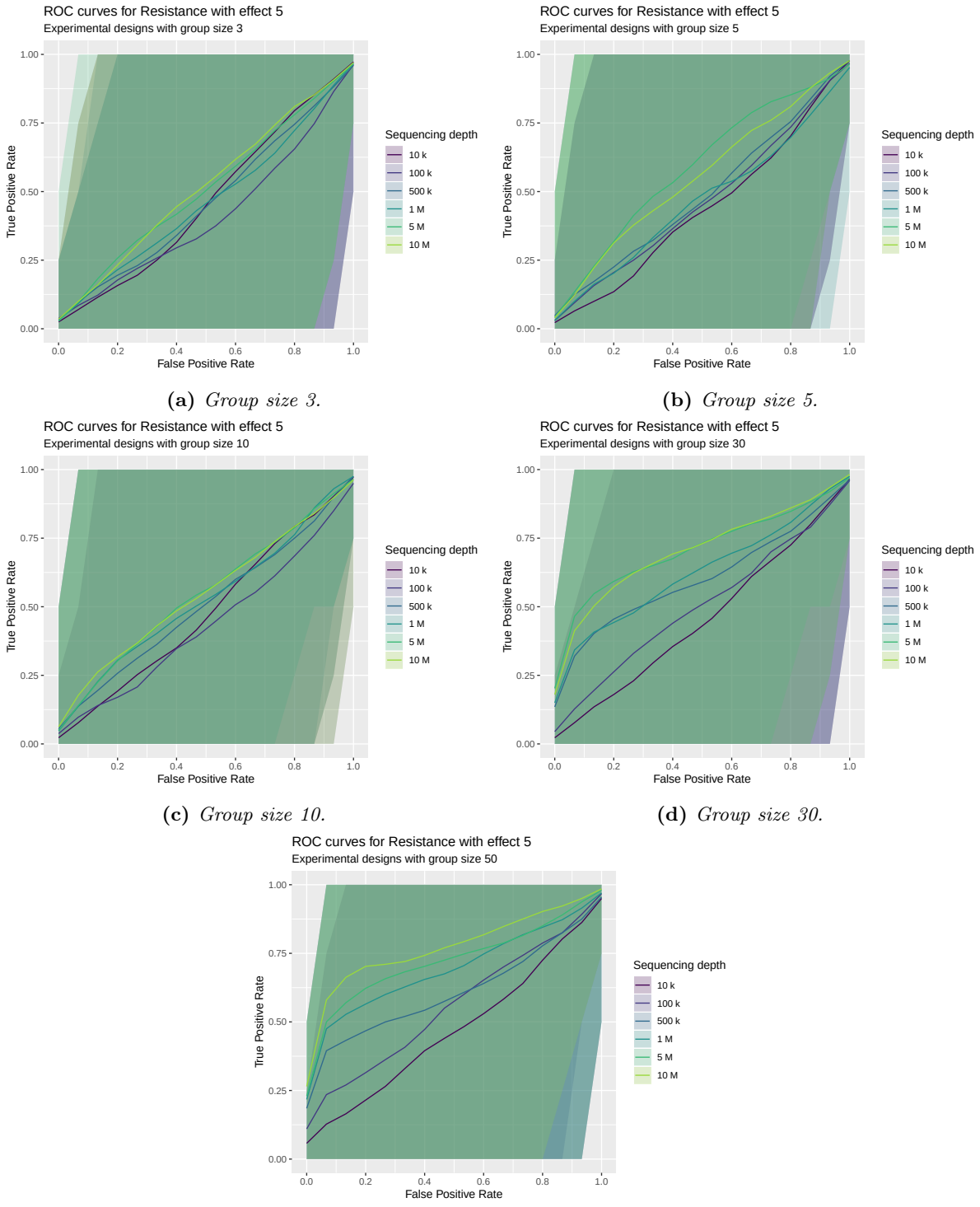
(c) *Sequencing depth 500 k.*

(d) *Sequencing depth 1 M.*

(e) *Sequencing depth 5 M.*

(f) *Sequencing depth 10 M.*

Figure D.1: *Mean ROC curves for the main experimental designs with fixed sequencing depths for Resistance analysed with DESeq2. Each design with a different group size is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the small effect size for different sequencing depths.*



(a) *Group size 3.*

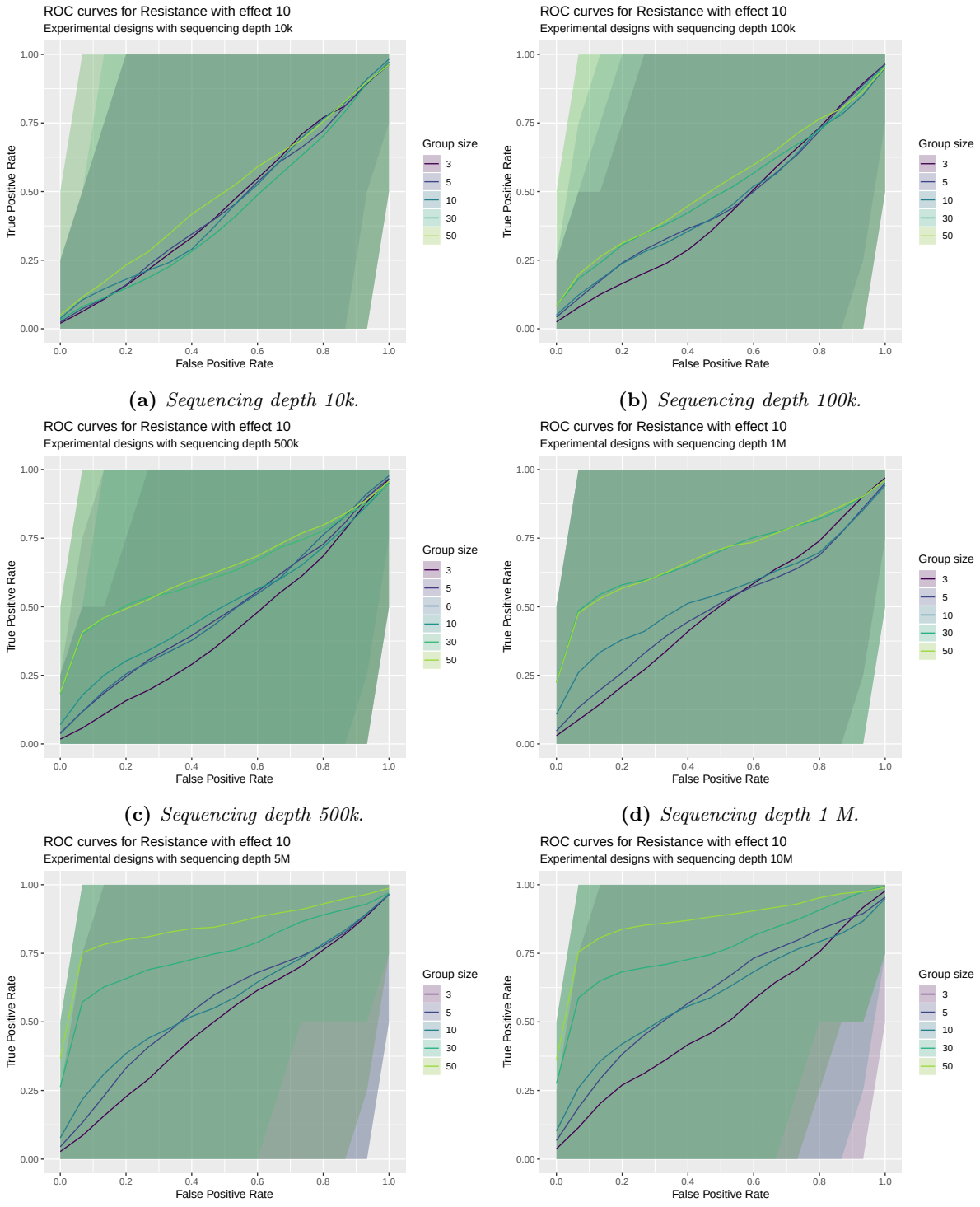
(b) *Group size 5.*

(c) *Group size 10.*

(d) *Group size 30.*

(e) *Group size 50.*

Figure D.2: *Mean ROC curves for the main experimental designs with fixed group sizes for Resistance analysed with DESeq2. Each design with a different sequencing depth is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the small effect size for different group sizes.*



(a) *Sequencing depth 10k.*

(b) *Sequencing depth 100k.*

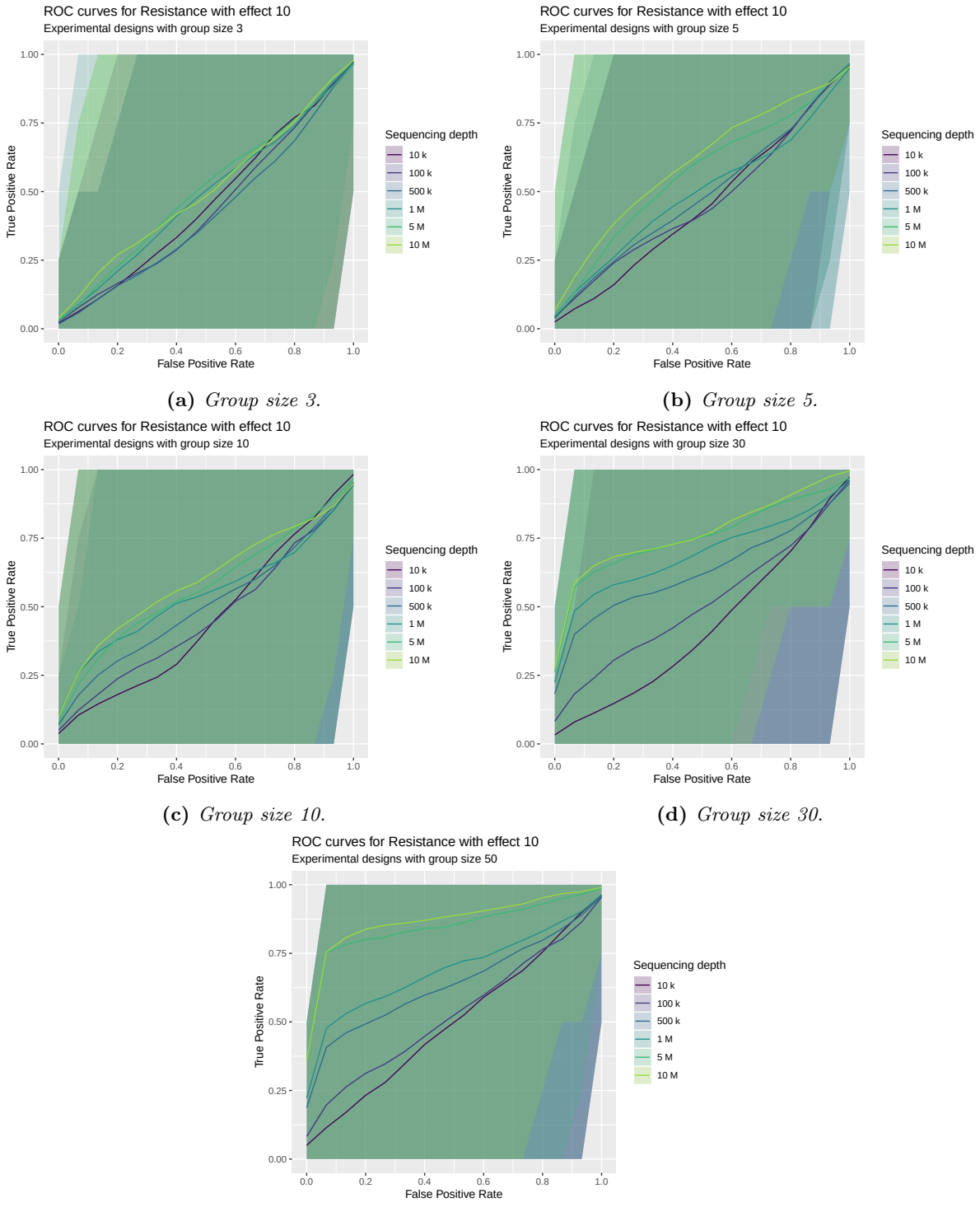
(c) *Sequencing depth 500k.*

(d) *Sequencing depth 1 M.*

(e) *Sequencing depth 5 M.*

(f) *Sequencing depth 10 M.*

Figure D.3: *Mean ROC curves for the main experimental designs with fixed sequencing depths for Resistance analysed with DESeq2. Each design with a different group size is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the large effect size for different sequencing depths.*



(a) *Group size 3.*

(b) *Group size 5.*

(c) *Group size 10.*

(d) *Group size 30.*

(e) *Group size 50.*

Figure D.4: *Mean ROC curves for the main experimental designs with fixed group sizes for Resistance analysed with DESeq2. Each design with a different sequencing depth is represented by a colored line where the ribbon corresponds to minimum and maximum values. The plots present curves for the large effect size for different group sizes.*

The ROC curves are summarised in a heatmap in Figure D.5, where the computed AUC_{tot} values are presented.

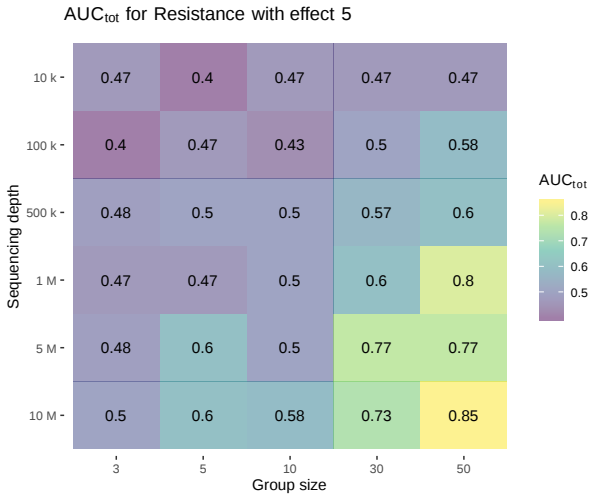


Figure D.5: *Median AUC_{tot} values for the main experimental designs for Resistance analysed with DESeq2. The heatmap presents values for the small effect size.*