Master's Thesis

Over-Optimism in Gene Set Analyses: how does the choice of methods and tools influence the detection of enriched gene sets?

Ludwig-Maximilians-Universität München
Department of Statistics

Milena Wünsch

Munich, April 14th, 2022

Submitted in partial fulfillment of the requirements for the degree of M. Sc.
Supervised by Prof. Dr. Anne-Laure Boulesteix

Abstract

Gene set analysis, which is a popular approach to gaining insight into high-throughput data of gene expression, aims at identifying sets of related genes that show significantly enriched or depleted expression patterns between two phenotypes of interest. In addition to the multitude of methods and tools developed to conduct gene set analysis, each of the tools in turn offers a number of changes that can be made in the workflow to adapt to the research question. This variety of options might tempt users of such gene set analysis tools to optimize the parameter setting with the goal to generate maximally interesting and meaningful findings, not being aware that this practice leads to over-optimistic results that are not valid. In order to raise awareness of the potential for over-optimistic results and the corresponding practices, a gene set analysis workflow based on a real gene expression data set is carried out for a number of gene set analysis tools in this thesis. In particular, parameter optimization is performed to illustrate the effect of certain parameter adaptions on the number of gene sets detected as differentially enriched. The findings show that over-optimistic results can be induced in the broad majority of the tools with a variety of parameter adaptions. Moreover, they indicate that the choice of the gene set analysis tool, if made based on which tool provides the highest number of differentially enriched gene sets, can in itself lead to over-optimistic results.

Contents

List of Figures

List of Tables

1. Introduction

Differential expression analysis and gene set analysis, which is abbreviated by "GSA", are both popular approaches to gaining insight into high-throughput gene expression data (Maleki et al., 2020). While differential expression analysis provides a list of differentially expressed genes, i.e. individual genes that show significantly different expression behaviour between two opposing phenotypes of interest (Khatri et al., 2012), in GSA, the list of genes measured in the experiment is categorized into sets of related genes which are usually provided by knowledge bases. This leads to a reduction of the dimensionality of the underlying statistical problem as statistical tests are performed for hundreds of gene sets instead of thousands of individual genes (Ackermann and Strimmer, 2009). Accordingly, the goal of GSA is to detect gene sets that are associated with the phenotypes of interest in the sense that they either show enriched or depleted expression levels across the phenotypes (Maleki et al., 2020). In this context, a higher statistical power as well as an improved interpretability of the results is achieved compared to an analysis on the single gene-level. In general, GSA can be classified into three major approaches, namely Over-Representation Analysis (ORA), Functional Class Scoring (FCS) and Pathway Topology (PT) which utilize the available information from the high-throughput experiment to various extents and therefore differ with regard to the complexity of the underlying methodology (Khatri et al., 2012).

In the last years, a multitude of different methods and tools have been developed to conduct GSA. However, there is a discrepancy between the frequency and reliability of validation strategies utilized in the corresponding benchmark studies, resulting in a lack of guidance concerning the right choice of a GSA tool (Xie et al., 2021). Furthermore, each individual tool additionally offers a number of changes in the workflow that can be made to adapt to the research question. Particularly, the entirety of these researcher degrees of freedom (Simmons et al., 2011) offers flexibility with respect to the detection of differentially enriched gene sets and especially regarding the number of gene sets detected as differentially enriched. This could entice a user of such GSA tools with little statistical experience to utilize the set of possible adaptions, may it be the choice of GSA tool or the parameters in preparation for or within the tool, to produce maximally promising results for his or her research question, not being aware that this approach can lead to over-optimistic findings. As a consequence, the findings contain an optimistic bias and are

reduced in validity (Boulesteix and Strobl, 2009). Whereas studies have been conducted to evaluate the consistency between the results of a number of GSA tools (Maleki et al., 2019), there are no studies to investigate the potential for over-optimistic results, i.e. research findings that result from the optimization of the choice of the GSA approach itself and the corresponding statistical parameter setting.
Consequently, the objective of this thesis is to assess and illustrate the potential for over-optimism across a selection of GSA methods and tools that are classified as ORA or FCS using a real gene expression data set. In this context, this work aims at helping users of GSA tools to develop consciousness of over-optimism in general and practices that lead to over-optimistic results. Furthermore, readers of research publications that contain results generated with GSA shall be encouraged to gain awareness that seemingly interesting and meaningful results could be overly optimistic due to an optimization process of certain parameters.
This thesis is structured in the following way: Chapter 2 provides the general framework of GSA, including necessary mathematical notation and the steps performed prior to and within two major approaches of GSA, namely ORA and FCS. Furthermore, a selection of popular or high-performing GSA methods that either belong to ORA or FCS and the corresponding tools are described in Chapter 3. Then, in Chapter 4, the setup to evaluate the methods' and tools' potential for over-optimistic results is elaborated utilizing a real gene expression data set and a stepwise optimization process. With the objective to maximize the number of differentially enriched gene sets, the optimization process considers all parameters prior to and within the GSA tools which are presumed to provide flexibility regarding this quantity, while still being appropriate in the given statistical and biological context. Afterwards, the results of the optimization process are presented in Chapter 5 and the potential for over-optimistic results is evaluated for each individual tool and across all tools. Finally, the main findings from these results are summarized in Chapter 6 together with a discussion of the strengths and weaknesses of the setup. Moreover, an outlook is given.

The results presented in this thesis were generated using version 4.1.0 of the software `R` (R Core Team, 2021). Besides the `R` packages cited in the text, the package `org.Hs.eg.db` (Carlson, 2021) was required to run a number of the tools under investigation. Furthermore, the packages `ggplot2` (Wickham, 2016) and `dplyr` (Wickham et al., 2021) were utilized in order to generate result illustrations. The results generated are in whole or part based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga).

2. General Framework

This chapter provides the general framework behind GSA. Firstly, necessary mathematical terminology is introduced in Section 2.1, which is followed by a short overview in Section 2.2 over the two major GSA approaches investigated in this thesis as well as the required steps to be carried out before running GSA. In addition to this overview, the components of GSA are described in further detail in the Sections 2.3-2.8.

2.1 Notation

The following mathematical notation will be used repeatedly throughout this thesis to define overall relevant statistical and biological quantities. Since some methods require further notation, additional quantities are introduced locally.

g_i: gene i, $i = 1, \ldots, N$

N: total number of genes in the experiment

U: background data set, i.e. set of all genes in the experiment, such that $|U| = N$

G_k: gene set k, $k = 1, \ldots, n$

n: total number of gene sets

$|G_k|$: Gene set size, i.e. number of genes contained by G_k

S_j: library size of sample j, $j = 1, \ldots, m$

s_j: normalization factor of sample j, $j = 1, \ldots, m$

m: total number of samples in the experiment

q_{ij}: relative abundance, i.e. fraction of all cDNA fragments in sample j that are mapped to gene i

Q_{ij}: quantity that is proportional to relative abundance q_{ij}

L_{u}: unranked list containing the set of genes detected as differentially expressed

$|L_u|$: size of L_u, i.e. number of genes detected as differentially expressed

n_k^0: number of genes in gene set G_k that are present in L_{u}

L_r: list of the entirety of genes in the experiment ranked by their correlation with the phenotype, i.e. magnitude of differential expression

K_{ij}: number of RNA reads in sample j that are unambiguously mapped to gene i

2.2 Overview

The entirety of gene set analysis methods that are put under investigation in this thesis apply the general framework illustrated in Figure 2.1. This figure contains the theoretical workflow for methods classified as ORA or FCS. A short overview of the workflow is given below and the exact details of each step are presented in the following Sections 2.3-2.8. In Chapter 3, the specific gene set analysis methods and tools which are assessed with respect to the potential for over-optimistic results are described.

The goal of GSA is to detect gene sets which are differentially enriched between the phenotypes of interest. In accordance with Maleki et al. (2020), the term "differential enrichment" is used in this thesis to refer to gene sets that are either enriched, i.e. show an increased gene expression activity, or are depleted which means that their gene expression activity decreases. All gene set analysis methods require two input components, namely a gene expression data set and a gene set database. For the gene expression data set, this thesis considers a data set generated from the state-of-the-art RNA Sequencing Technology (Wang et al., 2009). In this data type, gene expression is encoded as the number of read counts that are mapped to each gene in each sample. Additionally, each sample is binary labeled according to its phenotype or condition. A gene set database provides clusters of related genes which are applied to the genes in the gene expression data set to form gene sets. Before starting the analysis, it can be useful to filter out lowly expressed genes, i.e. genes with only few read counts across all samples in the data set, as these genes are unlikely to produce interesting results from the start. This is particularly relevant for differential expression analysis performed prior to the conduct of ORA, but also indicated for FCS for the purpose of consistency.
Methods belonging to ORA require the conduct of differential expression analysis beforehand to generate the input, namely an unranked list of differentially expressed genes. The goal of differential expression analysis is to detect all genes in the experiment that show significantly different gene expression patterns between the considered phenotypes. In contrast to that, FCS methods do not only consider those genes from the experiment detected as differentially expressed, but instead the expression profile of all genes in the experiment. Hence, the RNA-Seq data set is utilized as a whole instead of only a selection of genes. This approach requires, prior to performing the actual gene set analysis, a normalization technique in order to remove technical biases from the data that are a result

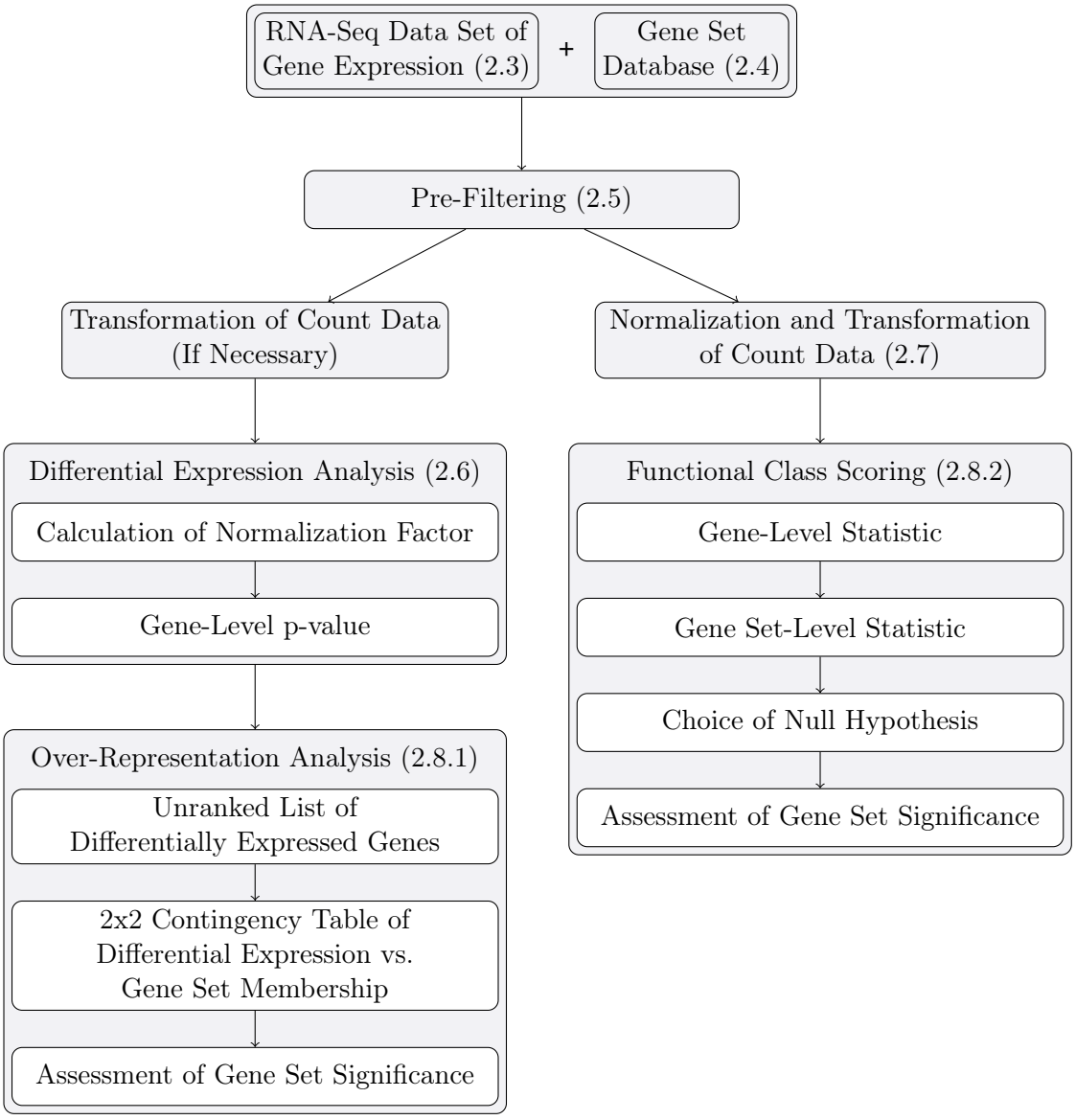

Figure 2.1: General Overview of Gene Set Analysis Approaches

General Overview over gene set analysis approaches, their required input and preparatory steps. Depending on the form of the expression data set and the respective tool, additional preparatory steps might be necessary.

of the sequencing process itself and would hinder comparability between samples unless addressed. Furthermore, oftentimes a transformation of the expression data is necessary in order to fit the distributional assumptions of the gene set analysis methods classified as FCS. It is noted that for gene set analysis methods assigned to ORA, normalization is performed internally in the differential expression analysis process and therefore does not require manual conduct by the user. A transformation of the expression data, on the other hand, is required for ORA if the differential expression technique was not specifically developed for RNA-Seq data. However, due to the choice of `DESeq2` and `edgeR` as differential expression techniques in this thesis, a transformation of the expression data is

not necessary.
The general framework of ORA consists of generating a contingency table for each gene set G_k, $k = 1;:::;n$, with the categories differential expression (differentially expressed, not differential expressed) vs. gene set membership (member of gene set G_k, not a member of gene set G_k). From this contingency table, the p-value of over-representation is computed and a gene set G_k is considered as differentially enriched if the number of differentially expressed genes within G_k is unlikely to be caused by chance.
For FCS, on the other hand, a gene-level statistic for each gene is usually calculated internally which captures the gene's magnitude of differential expression across the phenotypes. Based on this gene-level statistic, a ranking of the genes is generated, whereby up-regulated genes appear at the top and down-regulated genes at the bottom of the ranked list. For some FCS tools, however, this ranked list of genes must be provided as input. Based on this ranking, a gene set-level statistic is computed for each gene set. Finally, depending on the choice of null hypothesis, the significance of the gene set-level statistic of each gene set is assessed. As for all gene set analysis methods, independent of the assignment to ORA or FCS, the procedure is performed for multiple gene sets, multiple test adjustment is necessary in order to reduce the probability of false positive enrichment (Maleki et al., 2020), namely gene sets with a significant p-value of enrichment that results solely from chance.

2.3 RNA-Seq Data of Gene Expression

This thesis utilizes gene expression data generated from the state-of-the-art RNA Sequencing ("RNA-Seq") Technology which has replaced Microarray Analysis Technique as gene expression profiling in the years since its release. In the process of an RNA-Seq experiment, an mRNA strand, which is a gene expression product, is essentially fragmented into shorter snippets that are then, after further processing, sequenced by the high-throughput sequencing technology, resulting in short sequence reads (Wang et al., 2009). By performing this process with an entire sample of mRNA strands, gene expression is assessed as the number of sequence reads that can be assigned to a gene in this particular sample. Overall, an RNA-Seq data set consists of the measurement of N genes in m samples and its $(i; j)$-th entry corresponds to the number of reads mapped to gene i in sample j, also called read counts. Additionally, each sample is assigned a binary label that corresponds to the phenotype of interest.
Utilizing RNA-Seq data as gene expression data entails several challenges that need to be addressed in the context of differential expression analysis as well as gene set analysis. Firstly, different samples in their raw format are in fact not comparable unless normalization is performed. As mentioned above, in the context of ORA normalization is performed internally during the differential expression analysis procedures considered in this thesis

(see Section 2.6). For gene set analysis methods belonging to FCS, however, normalization must be performed beforehand manually. Furthermore, many gene set analysis methods were developed specifically for microarray data which differs from RNA-Seq data in its distributional assumptions. As the input for ORA is an unranked list of differentially expressed genes, this difference becomes immaterial since the unranked list can be obtained using differential expression analysis techniques specifically developed for RNA-Seq data. In FCS, this aspect is substantial in that gene-level statistics were specifically developed for microarray data and therefore assume a normal distribution. Therefore, in addition to normalization, a transformation of the count data is necessary in order to align its distribution to a normal distribution. These two aspects that need to be taken into account when applying an FCS method are further described in Section 2.7.

2.4 Gene Set Databases

As the term itself as well the preceding methodology suggest, gene set analysis is performed by aggregating the information of gene expression of a number of related genes into gene sets. A relationship between genes can be based on various aspects, such as a common chromosomal location or biological function (Subramanian et al., 2005). A meaningful analysis requires a coherent formation of gene sets, called gene set database, especially with respect to the fact that the choice of gene sets naturally influences the final results significantly. Two of the most common gene set databases are called "GO" and "KEGG", however, several GSA tools enable the user to provide user-defined gene sets. This thesis exclusively considers commonly used gene set databases. It is noted that, even though the GSA tools under investigation essentially provide the use of the same gene set databases, they do not necessarily refer to the same versions of the databases as the tools are not maintained simultaneously and in identical time intervals. As a result, the number of gene sets to be provided by a gene set database can differ across tools.

2.4.1 Gene Ontology (GO)

Gene Ontology, abbreviated with "GO", summarizes up-to-date scientific knowledge about the functions of gene products such as RNA molecules or proteins resulting from the gene's expression (Ashburner et al., 2000; The Gene Ontology Consortium, 2021). The overall goal is to understand in a species-independent way how individual genes contribute to the biology of an organism at the molecular, cellular and organism level. Gene Ontology is organized in the form of a directed acyclic graph, where each node corresponds to a GO "term", which in turn refers to a specific gene set. In this context, a directed edge reflects the hierarchical relationship between two terms in the sense that the "child" term is more "specialized", meaning that the contained genes are related by a more specific relationship, and the parent term is more general. Terms are categorized into three different

subontologies, namely Molecular Function, Cellular Component and Biological Process. This thesis only takes the subontology Molecular Function into consideration.
The subontology Molecular Function entails terms in form of activities of a gene product performed at the molecular level. In particular, a term does not describe context, time and location of the activity and instead only focuses on the activity itself. The molecular functions reported in this ontology mostly refer to activities performed by an individual gene product, yet some activities are performed by molecular processes composed of multiple gene products. An example of a broad functional term is "catalytic activity" or "transporter activity", whereas more specialized terms are "adenylate cyclase activity" or "toll-like receptor binding".
Cellular Component, on the other hand, refers to locations in which a gene product is active relative to the cellular structures whereas Biological Process describes pathways and larger processes to which a gene product's activity contributes. As already mentioned, these two subontologies are excluded from the analysis performed in this thesis.

2.4.2 Kyoto Encyclopedia of Genes and Genomes (KEGG)

Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000; Kanehisa et al., 2021) is a resource of 16 databases which was initiated in 1995. These databases are classified into the four categories systems information, genomic information, chemical information and health information. The purpose of KEGG is to impart an understanding of functions and utilities of the biological system on a higher level. Of interest in this thesis is the database KEGG Pathways which is a collection of manually drawn pathway maps that represent the current knowledge of molecular interaction, reaction and relation networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development. Each pathway map is a molecular reaction/interaction network diagram which ensures that experimental evidence in specific organisms can be generalized to other organisms through genomic information. In the following, KEGG Pathways will solely be referred to by the term "KEGG". Furthermore, a second, more specialized database will be utilized for a number of tools, namely KEGG Modules, where modules are gene sets with a more straightforward interpretation.

2.5 Pre-Filtering

As part of differential expression workflows, genes with low counts across all samples are usually removed from the analysis as these genes are unlikely to be detected as differentially expressed from the start and only increase the number of statistical tests to be performed. Consequently, removing lowly expressed genes from the analysis can lead to an increased overall statistical power. While some differential expression techniques require the manual

specification of a threshold value below which genes are omitted from the analysis, others provide functions to filter out genes based on more sophisticated criteria. In this thesis, all genes with a total number of counts across samples below a manually specified threshold value are omitted from the analysis, independent of the choice of the GSA approach and the differential expression technique.

2.6 Differential Expression Analysis

Prior to the conduct of GSA, differential expression analysis on the single-gene level has to be performed to obtain either an unranked list of differentially expressed genes for ORA or a list of all genes in the experiment ranked by their magnitude of differential expression for some FCS tools. The overall aim of differential expression analysis is to assess whether the relative abundance q_{ij} of a given gene i differs between conditions which are in this context defined as "0" and "1". Two of the most popular techniques for differential expression analysis, both implemented in `R` as part of the `Bioconductor` project (Huber et al., 2015), are `DESeq2` and `edgeR`. In the following sections, the default procedures for differential expression analysis of both methods are described. Changes that can be made in the corresponding `R` workflows which are relevant for this work are presented in Section 4.4.

2.6.1 DESeq2

`DESeq2` is a method that was introduced by Love et al. (2014) and implemented in the `R` package `DESeq2`. The core utility of the package is differential expression analysis of genes between phenotypes or conditions. It was specifically developed for RNA-Seq data and therefore addresses challenges arising from count data, such as high variability for low counts that indicate low expression profiles.

The starting point of `DESeq2` is a matrix K, whose entries K_{ij}, $i = 1;:::;N; j = 1;:::m$, correspond to the number of reads from the sequencing experiment that can be unambiguously mapped to gene i from sample j. The counts K_{ij} are modeled using the Negative Binomial distribution such that

$$K_{ij} \quad NB(\text{mean}= \ _{ij}; \text{dispersion}= \ _i): \tag{2.1}$$

Here, mean $\ _{ij}$ is split into $\ _{ij} = s_j \ Q_{ij}$, namely into normalization factor s_j, which is uniformly estimated for all genes within a sample, and quantity Q_{ij}. Quantity Q_{ij} is proportional to the proportion of cDNA fragments of gene i in sample j. The normalization factor s_j is computed via the Median-of-Ratios method to account for differences in sequencing depth between samples which would otherwise impair sample-wise comparisons. The overall reasons for normalization of RNA-Seq data can be found in Appendix A.1 and details on the Median-of-Ratios method are presented in Appendix A.1.1. The dispersion

parameter α_i in Equation (2.1) is assumed to vary across genes but to be constant across all samples.
In general, the expression strength of a given gene i is estimated using a generalized linear model

$$\log_2(Q_{ij}) = \sum_r x_{jr} \beta_{ir}. \tag{2.2}$$

Since differential expression analysis is applied in the framework of GSA, the only comparison is made between two conditions, resulting in $r = 1$. Consequently, the design matrix elements are set as follows:

$$x_{jr} = x_j = \begin{cases} 1 & \text{if sample j belongs to condition 1;} \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

This fit returns the $\log_2$ fold change estimates $\hat{\beta}_{ir} = \hat{\beta}_i$ between both conditions and therefore the overall expression strength of gene i.
In order to perform inference of differential expression, an accurate estimation of α_i is required, especially for small sample sizes. This parameter models the within group-variability via $\mathrm{Var}(K_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$. As for small sample sizes, noisy estimates would compromise accuracy of differential expression testing, the idea is to share information across genes. In this regard, the assumption is made that genes of similar average expression strength have similar dispersion. In the first step, the gene-wise dispersion is estimated via Maximum Likelihood for each gene separately. Then, the location parameter of the distribution of the dispersion estimates of all genes is determined and a smooth curve is fit to allow for a dependence of the dispersion on the average expression strength. To obtain the final dispersion estimate of an individual gene, its gene-wise dispersion estimate is shrunk towards the smooth curve using Empirical Bayes method. The strength of shrinkage depends on firstly how close an estimate of the true dispersion values tends to be to the fit. Secondly, the strength of shrinkage decreases with increasing degrees of freedom in the sense that it decreases as sample size increases.

Another challenge `DESeq2` addresses is a high variance of genes with low read counts across all samples, leading to exaggerated log fold change estimates. This heteroscedasticity would, unless accounted for, lead to complications in the downstream analysis and eventually in the interpretation of the results. `DESeq2` offers three shrinkage estimators to deal with this challenge, the default option being proposed by Zhu et al. (2019). The method is implemented in the `R` package `apeglm` and moreover integrated into the differential expression analysis workflow conducted in `DESeq2`. A heavy-tailed Cauchy distribution

is set as prior distribution to account for large effect sizes:

$$_i \quad \text{Cauchy}(0;\ \text{scale}_i); \qquad i = 1;:::;N: \tag{2.4}$$

The scale parameter scale_i of the prior distribution is then estimated using the Maximum Likelihood Estimates $\hat{}_i$ and the corresponding standard errors resulting from the generalized linear model in Equation (2.2). The estimates $\hat{}_i$ are assumed to follow a normal distribution around their true values $_i$ and, despite the choice of a Cauchy prior for the true values, these are assumed to be normally distributed around 0 for the computation of the scale parameter scale_i. The posterior distribution for the $_i$ is calculated via Laplace Approximation and eventually, the shrinkage estimator of coefficient $_i$ is obtained as the posterior mode.

After fitting a generalized linear model and performing shrinkage to obtain a shrunken log fold change estimate for a given gene, a hypothesis test is performed to assess whether the estimate significantly differs from 0. The default hypothesis test in `DESeq2` is a Wald test, in which the test statistic is the shrunken log fold change divided by the estimated standard error, resulting in a z-score which is then compared to the standard normal distribution. As this testing procedure is conducted for every single gene in the experiment, multiple test correction is performed and the resulting adjusted p-value is eventually used to detect a gene as differentially expressed or not.
Since in `DESeq2`, parametric methods are used for the detection of differentially expressed genes, the log fold change estimates and resulting p-values of the genes can be strongly influenced by few outliers that do not fit the distributional assumptions made for the model. To address this issue, `DESeq2` uses Cook's distance as an outlier diagnostic. For each sample j and gene i, it is defined as the scaled distance that coefficient $\hat{}_i$ would change if the given sample j was removed and the linear model from Equation (2.2) refit. A sample is then declared an outlier if the corresponding Cook's distance is larger than the 99th percentile of the $F(p; m \quad p)$-distribution, where p is equal to the number of model parameters including intercept and m is the number of samples in the experiment. In the case of this thesis, where the expression data set contains a sufficiently high number of samples, an outlier value is replaced with the trimmed mean over all samples which is additionally scaled with the sample's normalization factor. Since this means that outlier values are replaced with values predicted by the null hypothesis, the approach is conservative and leads to a decreased number of differentially expressed genes compared to a scenario in which the outlier detection is turned off.

2.6.2 edgeR

edgeR, which is short for "empirical analysis of digital gene expression data in R", is an R package developed by Robinson et al. (2010). It is specifically developed for data arising from RNA sequencing technologies and focuses on differential expression analysis.

Analogous to DESeq2, edgeR starts from count data K_{ij}, $i = 1,\dots,N$, $j = 1,\dots,m$. Normalization is performed via the Trimmed Mean of M-values method to account for differences in sequencing depth and compositionality effects (see Appendix A.1). This method calculates the normalization factor of a given sample j as a trimmed and weighted mean of log expression ratios between the sample and a reference sample. A more detailed description of the normalization process can be found in Appendix A.1.2.
For count data K_{ij}, the same distributional assumption as in the previous Section 2.6.1 is made, namely

$$\begin{aligned} K_{ij} &\sim NB(\mu_{ij}, \alpha_i) \\ \mathrm{E}[K_{ij}] &= \mu_{ij} = q_{ij} s_j \\ \mathrm{Var}(K_{ij}) &= \mu_{ij} + \alpha_i \mu_{ij}^2. \end{aligned} \tag{2.5}$$

In this method, the intuition behind the Negative Binomial distribution is to see it as a Poisson distribution with over-dispersion: As the dispersion parameter α_i converges to 0, variance $\mathrm{Var}(K_{ij})$ approaches μ_{ij} and the distribution becomes increasingly Poisson-like with parameter μ_{ij}. However, with α_i being non-zero, the variance of counts increases relative to the mean μ_{ij}. It can thus be split into two parts, namely the Poisson-like technical variability μ_{ij} and the over-dispersion that arises from biological and other sources. The expected value of counts $\mathrm{E}[K_{ij}]$ is split into the library size s_j measured in sample j and the relative abundance q_{ij} of gene i in sample j.
By dividing the variance $\mathrm{Var}(K_{ij})$ from Equation (2.5) with μ_{ij}^2, it can be transformed into the form of the squared coefficient of variation

$$CV^2(K_{ij}) := \frac{1}{\mu_{ij}} + \alpha_i. \tag{2.6}$$

The first addend then corresponds to the squared coefficient of variation of the Poisson distribution, referring to the uncertainty with which the relative abundance of gene i in sample j is measured with the sequencing technology. The second addend, on the other hand, is called biological squared coefficient of variation or BCV^2, i.e. the variance of the relative abundance of gene i across all samples. Since technical variability is expected to marginalize as sample size increases to infinity, BCV, which is the square root of α_i, is likely to be the dominant source of variation and therefore needs to be estimated accurately in order to assess differential expression realistically. It is noted that BCV is also

equal to the square root of the dispersion parameter of the Negative Binomial distribution as visible in Equation (2.5). The biological coefficient of variation BCV is estimated by estimating a common dispersion across all genes and samples in the first step and then estimating a gene-wise dispersion afterwards.
The common dispersion is estimated using the quantile-adjusted Conditional Maximum Likelihood (qCML) method. It consists of maximizing the common conditional likelihood in the first step which assumes identical library sizes across all samples. It is calculated by first conditioning on the total counts for each gene and a single condition. The common dispersion estimator eventually maximizes the sum of the corresponding log-likelihoods over the entirety of conditions and genes. As this so-called "common likelihood" is based on the unrealistic assumption of equal library sizes mentioned above, quantile adjustment is applied to generate pseudodata. This pseudodata represents the count data that would occur if the total read counts were equal to the geometric mean of all samples. The pseudodata is then applied to the CML estimate for common dispersion, yielding the quantile-adjusted Conditional Maximum Likelihood. To account for the possibility of different dispersions for individual genes, a weighted likelihood is maximized for each gene which consists of the sum of the individual log-likelihood of the gene and the weighted quantile-adjusted common likelihood. As the weight of the common likelihood increases, the gene-wise dispersion approaches the common dispersion. Eventually, with the estimated parameters of the Negative Binomial distribution, a p-value for differential expression testing can be derived for each gene.
In the context of `edgeR`, the null hypothesis for each gene i is stated as follows:

$$\mathrm{H}_0 : q_{i;0} = q_{i;1} \text{ for } i = 1;:::;N; \tag{2.7}$$

namely equal relative abundance of each gene across both conditions. To evaluate this null hypothesis, the following quantities are defined as

$$\begin{aligned} Z_{i;0} &:= \sum_{j;0} K_{ij} \\ Z_{i;1} &:= \sum_{j;1} K_{ij} \\ Z_i &:= Z_{i;0} + Z_{i;1}: \end{aligned} \tag{2.8}$$

The first quantity of Equation (2.8) is the sum of all counts of gene i ascribed to condition 0, whereas the second quantity refers to the the entirety of counts of gene i that belong to condition 1. Z_i is simply the sum of read counts over all samples that are mapped to gene i. Define $p(a_i;b_i)$ as the joint probability of a pair of total number of read counts

mapped to gene i in two conditions, then the p-value for gene i is defined as

$$p_i = \sum_{\substack{a_i + b_i = Z_i \\ p(a_i;b_i) \quad p(Z_{i;0};Z_{i;1})}} p(a_i;b_i) \tag{2.9}$$

$$= \sum_{\substack{a_i + b_i = Z_i \\ p(a_i;b_i) \quad p(Z_{i;0};Z_{i;1})}} p(a_i) \quad p(b_i): \tag{2.10}$$

The p-value corresponds to the sum of probabilities of combinations of counts mapped to gene i that are more in favor of differential expression than $(Z_{i;0}; Z_{i;1})$, conditioned on the sum of both components being equal to the observed total counts mapped to gene i. The second line of Equation (2.9) is derived using the assumption of independence across samples. Probabilities $p(a_i)$ and $p(b_i)$ can be again estimated with the Negative Binomial distribution.
As usual, multiple test adjustment is performed to obtain adjusted p-values for the entirety of genes in the experiment.

2.7 Normalization and Transformation of Count Data

As mentioned in Section 2.3, for FCS methods the RNA-Seq expression data must be normalized and transformed before carrying out the actual GSA. Normalization itself is performed to remove sample-specific biases that arise from the sequencing process and therefore hinder comparability between samples. This includes differing library sizes between samples, i.e. differences in the number of total read counts mapped to a specific sample. The intuition behind normalization is described in further detail in Section A.1.
The second aspect that needs to be accounted for is that most FCS methods were developed specifically for microarray data which is assumed to be normally distributed. As RNA-Seq expression data contains count values, microarray methods are not applicable without further adjustments. This challenge becomes apparent in the choice of the gene-level statistic to capture a gene's magnitude of differential expression, such as the moderated t-statistic in PADOG (see Section 3.4) or the Signal2Noise Ratio (see Equation (3.4)), since these statistics usually assume a normal distribution. Another factor that hinders the use of microarray methods for RNA-Seq data is the heteroscedasticity of the latter, namely the increase of the variance for high count values. The methods utilized in this thesis to perform RNA-Sequencing transformation are described in Section 4.3.
It is noted that, depending on the choice of the differential expression technique, this issue needs to be addressed in the application of ORA methods, too. While in this thesis, differential expression analysis is performed using `DESeq2` or `edgeR` (see Section 2.6), which were both specifically developed for RNA-Seq data, other techniques such as provided by `limma` (Smyth, 2004) were developed for microarray data which, on the other hand,

require a transformation of the RNA-Seq data.

2.8 Classification of Gene Set Analysis Methods

Gene Set Analysis methods are classified into the three categories Over-representation Analysis (ORA), Functional Class Scoring (FCS), and Topology-based methods (PT). This work only includes methods and tools ascribed to the former two.

2.8.1 Over-Representation Analysis

In ORA, the input consists of an unranked list L_u which contains $|L_u|$ genes detected as differentially expressed by the method used for differential expression analysis on the single-gene level. Furthermore, a gene set database is required, containing n gene sets G_k of the size $|G_k|$. For each gene set G_k, the number of genes that are part of input list L_u is denoted with n_k^0. A gene set G_k is detected as differentially enriched if the number n_k^0 of differentially expressed genes is unlikely to be caused by chance.
Assuming a total number of N genes in the experiment to be contained by the background data set U, such that $|U| = N$, relevant quantities for ORA can be summarized in the following Contingency Table 2.1:

Table 2.1: Contingency Table in Over-Representation Analysis

	Genes in L_u	Genes not in L_u	total
Genes in G_k	n_k^0	$\|G_k\| \quad n_k^0$	$\|G_k\|$
Genes not in G_k	$\|L_u\| \quad n_k^0$	$N \quad \|G_k\| \quad (\|L_u\| \quad n_k^0)$	$N \quad \|G_k\|$
total	$\|L_u\|$	$N \quad \|L_u\|$	N

The null hypothesis states that there is no association between membership of gene set G_k and differential expression. This implies that G_k is the result of random sampling of $|G_k|$ genes from the entirety of genes in the experiment, namely U. Consequently, the probability distribution of n_k^0 differentially expressed genes within G_k can be modeled using the hypergeometric distribution (Drăghici et al., 2003):

$$f(n_k^0; N; |G_k|; |L_u|) = \frac{\binom{|G_k|}{n_k^0} \quad \binom{N \quad |G_k|}{|L_u| \quad n_k^0}}{\binom{N}{|L_u|}}: \tag{2.11}$$

This corresponds to Fisher's exact test and the p-value of over-representation is consequently calculated as

$$P_{G_k} = \sum_{j=n_k^0}^{jG_kj} f(j; N; jG_kj; jL_uj) = 1 \quad \sum_{j=0}^{n_k^0 \; 1} f(j; N; jG_kj; jL_uj): \tag{2.12}$$

Since in the case of a large number of genes in the expression data set, which is typical for high-throughput experiments, the calculation of the hypergeometric distribution tends to be computationally extensive. Consequently, an approximation with the Binomial distribution is used to calculate p-values of over-representation (Drăghici et al., 2003). Eventually, inference is based on adjusted p-values resulting from multiple test correction since a p-value of over-representation is calculated for each gene set in the experiment.

2.8.2 Functional Class Scoring

FCS methods follow the idea that the enrichment of a gene set is not only significantly affected by large changes in individual genes but possibly also by weaker but coordinated changes in the entire gene set (Ackermann and Strimmer, 2009). FCS methods usually follow variations of the following framework:

In contrast to ORA, the entire gene expression data is used as input for methods to belong to FCS. In this context, a gene-level statistic is calculated for each gene from the data set, representing the extent to which it is differentially expressed with respect to the phenotypes of interest. Common statistics to perform this step are the Signal2Noise Ratio as well as the t-Statistic (Ackermann and Strimmer, 2009), which can be found in Section 4.3.3 among other statistics. It is noted that in some FCS tools, such a ranking is not generated internally, but instead, must be provided as input to the tool.

Afterwards, for each gene set, the gene-level statistics of all contained genes are aggregated into a single gene set-level statistic. Usual gene set-level statistics are the Kolmogorov-Smirnov statistic, the sum, mean, or median of the respective gene-level statistics as well as the Wilcoxon Rank sum (Ackermann and Strimmer, 2009).

Before assessing the significance of the gene set-level statistic, the null hypothesis must be chosen, which affects the further conduct of the analysis. The two most common categories are the competitive null hypothesis as well as the self-contained null hypothesis. The choice of the competitive null hypothesis leads to comparing the association of a gene set with the phenotypes to the association of all remaining genes with the phenotypes. On the other hand, the self-contained null hypothesis implies a focus on the given gene set regardless of the remaining genes. In this case, the association of the gene set and the phenotypes is compared to the association that results from randomly assigned phenotypes.

In the next step, the significance assessment of each gene set-level statistic is performed. The competitive null hypothesis implies a fixed association between the samples and the

phenotype, whereas membership of a gene set corresponds to the sampling units. This means that for each gene set G_k separately, a high number of gene sets of the same size jG_kj are randomly drawn from the entirety of genes. Then, the gene set-level statistic for each randomly drawn gene set is computed and all resulting statistics are aggregated into a null distribution of gene set-level statistics. In the case of a positive value of the gene set-level statistic of the observed gene set, its p-value is computed as the fraction of resampled gene set-level statistics that exceeds the observed value. On the other hand, the p-value is computed as the fraction of resampled gene set-level statistics that falls below the observed statistic in case of a negative value of the gene set-level statistic.
In contrast to that, the self-contained null hypothesis implies fixed gene set membership, preserving the correlation structure among the genes in the observed gene sets. In this case, the sample labels are chosen as the sampling units and are therefore permuted a large number of times and the gene set-level statistics are computed for each permutation. In the end, the p-value of the observed gene set-level statistic is calculated as the fraction of gene set-level statistics resulting from permutation that exceeds the observed value in case it is positive or falls below the observed value if it is negative. Eventually, differential enrichment of a given gene set is assessed based on the respective adjusted p-value which is obtained by performing multiple test adjustment.

3. Gene Set Analysis Methods and Tools

The following methods and tools were chosen for their popularity or performance as summarized in Xie et al. (2021). Popularity is quantified as the number of citations, whereas performance studies mostly focus on sensitivity, false positive rate (FPR), prioritization, computational time, and reproducibility. The choice of the methods, as discussed before, is restricted to those belonging to ORA or FCS, whereas topology-based methods are excluded from the analysis. Furthermore, the focus is put on methods implemented in `R` (RStudio Team, 2021), however, some methods implemented in web applications are put under investigation as well. An overview of the methods and tools investigated in this thesis is given in Table 3.1. In this table, the classification into methods and tools is made based on the intuition that each tool implements a particular methodology. For instance, PADOG is considered a method that is implemented in the eponymous `R` package `PADOG`.

Table 3.1: Overview of Gene Set Analysis Methods and Tools

	Method/Tool	Approach	Tool Implementation	Section
DAVID	Tool	ORA	Web	3.1
GOSeq	Method/Tool	ORA	`R`	3.2
Gene Set Enrichment Analysis	Method	FCS	/	3.3
GSEA Web Application	Tool	FCS	Web	3.3.1
GSEAPreranked	Tool	FCS	Web	3.3.2
PADOG	Method/Tool	FCS	`R`	3.4
clusterProfiler	Tool	ORA/FCS	`R`	3.5

3.1 DAVID

Database for Annotation, Visualization, and Integrated Discovery, short DAVID, is a collection of web tools with the goal to provide an understanding of the biological meaning behind lists of genes (Huang et al., 2009a,b). Of interest in this thesis is the functional annotation tool for the identification of enriched biological terms, i.e. gene sets. This tool implements a method that is classified as ORA but slightly modifies Fisher's exact test introduced in Section 2.8.1. According to the website, DAVID was cited 6438 times in 2021.

DAVID assesses the enrichment of a gene set by calculating the EASE score (Hosack et al., 2003), which is a conservative adaptation of Fisher's exact test in the sense that it favors bigger gene sets with respect to the detection of differential enrichment. To be more precise, it modifies the Contingency Table 2.1 by removing one gene within G_k from the list L_u of differentially expressed genes and adding it to the set of genes that are not in L_u. This adaptation is illustrated in the contingency table presented in Table 3.2.

Table 3.2: Contingency Table in DAVID

	Genes in L_u	Genes not in L_u	total
Genes in G_k	n_k^0 1	G_k $n_k^0 + 1$	jG_kj
Genes not in G_k	jL_uj n_k^0	N jG_kj $(jL_uj$ $n_k^0)$	N jG_kj
total	jL_uj 1	N $jL_uj + 1$	N

Afterwards, the p-value of over-representation is recalculated for the new contingency table, analogous to Equation (2.12), resulting in higher p-values for smaller gene sets, whereas p-values of bigger gene sets only increase slightly. In the special case of a gene set consisting of only one gene which is moreover part of list L_u, the corresponding p-value results in 1, i.e. differential enrichment is precluded completely.

3.2 GOSeq

GOSeq is an application for performing ORA on RNA-Seq data that gives special attention to the properties of the corresponding count data that cannot be removed by normalization or rescaling. The method was introduced by Young et al. (2010) and is implemented in `R` as part of the `Bioconductor` project. Concerning popularity measured by the number of downloads, GOSeq lies between PADOG and `clusterProfiler` with 22606 total downloads in 2021 and 12743 downloads from distinct IPs.

The general idea behind GOSeq is based on the previously discussed fact that expression data gathered in RNA-Seq experiments has the form of count data, where the expected read count for a transcript is proportional to the length of the transcript multiplied by the gene's expression level. As for count data, statistical power increases with the magnitude of counts, longer or more highly expressed genes are more likely to be detected as differentially expressed compared to genes whose transcript length is shorter and/or expression profile is lower. This phenomenon of an increased probability for longer genes of being detected as differentially expressed is called length bias and is incorporated into the GOSeq-methodology.

In standard GO analyses, on the other hand, the underlying assumption behind testing for over-representation is that each gene is equally likely to be detected as differentially expressed under the null hypothesis. This assumption implies that the number of genes

from a given gene set that are present in the input list of differentially expressed genes follows a hypergeometric distribution as in Section 2.8.1. Due to length bias, however, the assumption is violated so that standard methods should not be used for GO analysis with RNA-Seq data, according to the authors. Results from such analyses are affected by length bias in that gene sets derived from Gene Ontology (see Section 2.4.1) usually differ greatly in transcript length, resulting in gene sets with generally longer genes having a higher probability of being detected as differentially enriched.

To incorporate length bias into GO analysis, the GOSeq framework consists of the following steps:

Prior to the GOSeq workflow, the set of differentially expressed genes from the entire set of genes in the experiment is identified. This can be carried out with any differential expression technique such as the methods implemented in `DESeq2` and `edgeR` in Section 2.6 or by including genes with $\log 2$-fold change higher than a fixed threshold value between the conditions. As usual, multiple test correction is performed afterwards to detect the set of differentially expressed genes.

Then, a probability weight function, called PWF, is estimated from the data. It quantifies the probability of a gene being detected as differentially expressed as a function of transcript length. The estimation process is conducted as follows: Firstly, each gene is either assigned value 1 if it is detected as differentially expressed or, else, value 0. Then, a monotonic cubic spline with 6 knots is fitted to the binary data series with each gene's length as a predictor. A monotonicity constraint is imposed on this fit since, as already discussed, the power of detecting a gene as differentially expressed increases with gene length or read count. The probability weight function serves as the null hypothesis for the enrichment test.

Finally, enrichment in form of a p-value is assessed for each gene set using the probability weight function as well as resampling. To be precise, a random set of differentially expressed genes of the same size as the actual set of differentially expressed genes is generated. In order to incorporate length bias into this procedure, the probability of each gene being included in the set is weighted by the corresponding value of the probability weight function. This process is repeated N_{ite} times and for each randomly resampled set of differentially expressed genes, the number of genes associated with the given gene set is counted. Eventually, the p-value of over-representation for a given gene set is calculated as

$$P_{\text{GOSeq}}(G_k) = \frac{X_{G_k} + 1}{N_{\text{ite}} + 1}; \tag{3.1}$$

where X_{G_k} is the number of resampled samples with at least as many genes associated with gene set G_k as in the observed set of differentially expressed genes.

Despite accurately accounting for the effect of length bias in RNA-Seq experiments, this resampling technique is computationally expensive and an alternative, approximate method

is proposed: Wallenius approximation is based on the Wallenius non-central hypergeometric distribution. This distribution is an extension of the standard hypergeometric distribution in the sense that all genes within the same gene set have the same probability of being selected, however, this probability differs from the probability of selecting genes from outside the gene set. The mean of the probability weightings for each gene, whether a member of a given gene set or not, is defined as the common probability of choosing a gene within or outside the gene set. Wallenius approximation is significantly closer to the true distribution in comparison with the standard hypergeometric distribution, despite being approximate.

3.3 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) is an FCS method that was proposed by Subramanian et al. (2005). According to Xie et al. (2021), GSEA was cited 4779 times between September 2019 and September 2020 which makes it the most popular method among the methods and tools investigated in this thesis. An illustration of the steps performed in GSEA can be found in Figure 3.1.

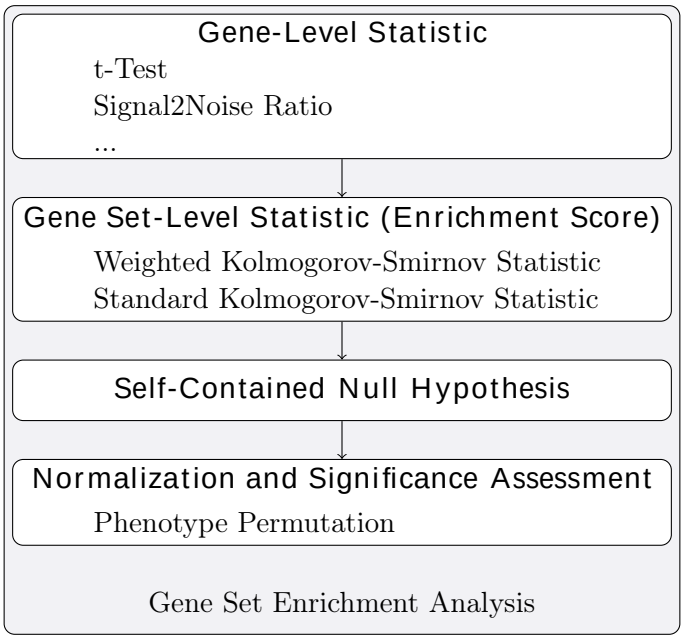

Figure 3.1: Steps Performed in Gene Set Enrichment Analysis

The starting point of GSEA is a ranked list L_r, containing each of the N genes whose expression was measured in the experiment. The ranking is generated based on the correlation between the gene's expression and the phenotype of interest, captured by a suitable ranking metric. Consequently, the top of the ranked list entails genes whose expression shows the strongest positive correlation with the phenotype, whereas the bottom of the

ranked list contains those genes with an expression negatively correlated with the phenotype. A standard ranking metric is the t-statistic or the Signal2Noise Ratio which can be found in Sections 4.3.3 and 3.3.1, respectively.

The general idea of GSEA is to investigate whether members of a given gene set G_k are distributed more towards the top or bottom of L_r, which indicates differential enrichment, or else spread across it randomly. In this context, the overall framework consists of the following three steps:

In the first step, an enrichment score is calculated for each gene set. The enrichment score, abbreviated with "ES", represents the extent to which gene set G_k is represented at the top or bottom of L_r. To describe the process in mathematical terms, the correlation of a gene g_i with the phenotype is denoted by $r_i = r(g_i)$. The genes are then ranked with respect to r_i in a decreasing manner such that the list of ranked genes has the form $L_r = (g_1, \ldots, g_N)$. In particular, gene g_1 has the strongest positive correlation with the phenotype of interest whereas g_N has the strongest negative correlation with it among all genes. For a gene set with $|G_k|$ genes, the enrichment score is then calculated by running down the ranked list L_r and evaluating the following two sums with every step l, $l = 1, \ldots, N$, taken:

$$\begin{aligned} P_{\text{hit}}(G_k, l) &= \sum_{\substack{g_i \in G_k \\ i \le l}} \frac{|r_i|^p}{N_R}, \qquad \text{where } N_R = \sum_{g_i \in G_k} |r_i|^p \\ P_{\text{miss}}(G_k, l) &= \sum_{\substack{g_i \notin G_k \\ i \le l}} \frac{1}{N - |G_k|}. \end{aligned} \tag{3.2}$$

Parameter p determines the weight of the correlation of gene g_i with the phenotype in the calculation of the enrichment score. This means that for $p = 0$, each gene's contribution to the enrichment score is equal and the enrichment score corresponds to a standard Kolmogorov-Smirnov statistic. In contrast to that, the choice of $p = 1$ results in those genes with a higher positive or negative correlation with the phenotype affecting the enrichment score more strongly. The enrichment score of gene set G_k is computed as the maximum deviation of $P_{\text{hit}}(G_k, l) - P_{\text{miss}}(G_k, l)$, namely

$$\text{ES}_{G_k} = \max_{1 \le l \le N} P_{\text{hit}}(G_k, l) - P_{\text{miss}}(G_k, l). \tag{3.3}$$

The intuition behind this statistic is that a gene set whose genes are concentrated towards the top or bottom of L_r has a high absolute enrichment score, whereas gene sets with randomly spread genes result in a lower absolute enrichment score. In this context, a positive enrichment score indicates an increased expression activity of the gene set while a negative enrichment score denotes a depleted expression activity.

The second step of GSEA focuses on assessing the significance of each gene set's enrichment

score. As the default null hypothesis in GSEA is the self-contained null hypothesis, significance of a gene set G_k is assessed by comparing its enrichment score to a null distribution of scores obtained by generating 1000 random permutations of the original phenotypes across all samples. This step is accompanied by generating a new gene-level ranking and recomputing the enrichment score for each random permutation. The p-value of gene set G_k is eventually computed relative to this null distribution.
In the third step, a gene set's enrichment score is normalized for the gene set size, and its p-value is adjusted for multiple comparison. Normalization for gene set size means that the enrichment score is recalculated such that gene sets that differ in size do not have systematically different enrichment scores. In this context, a gene set's enrichment score is divided by the mean of the enrichment scores obtained from the 1000 random permutations generated in the manner described above. The resulting normalized enrichment score, abbreviated with "NES", then builds the base for the interpretation of the results. In the last step, significance of a given gene set is evaluated based on the p-value which is adjusted for multiple testing, as the overall procedure is performed for a multitude of gene sets.

3.3.1 GSEA Web Application

The web application "GSEA" (Subramanian et al., 2005; Mootha et al., 2003) implements the methodology presented in the previous Section 3.3. The default gene-level ranking metric to measure a gene's correlation with the phenotype of interest is the Signal2Noise Ratio

$$\text{Signal2Noise} = \frac{K_{i;0}^{\text{norm}} \quad K_{i;1}^{\text{norm}}}{{}_0 + {}_1}; \tag{3.4}$$

where $K_{i;0}^{\text{norm}}$ and $K_{i;1}^{\text{norm}}$ are the mean transformed and normalized counts of gene i in the samples ascribed to phenotypes 0 and 1, respectively. Similarly, ${}_0$ and ${}_1$ refer to the standard deviations of the normalized counts in both phenotypes. With a default exponent value $p = 1$, the enrichment score, as calculated in the Equations (3.2) and (3.3), corresponds to a weighted Kolmogorov-Smirnov Statistic.

3.3.2 GSEAPreranked

GSEAPreranked (Subramanian et al., 2005; Mootha et al., 2003) is part of the GSEA web application but instead of utilizing the entire gene expression set, a ranked gene list is required as input. Therefore, the user manually creates a ranking of the genes which captures their correlation with the phenotype of interest. Afterwards, the enrichment score for a given gene set G_k is computed as described in Equations (3.2) and (3.3) with a default exponent of $p = 1$. As in the context of a ranked list, the information of the sample labels is lost, the significance of the enrichment score of a given gene set cannot

be computed using phenotype permutation but instead must be calculated using gene set permutation. As described in Section 2.8.2, gene set membership then corresponds to the sampling units.

3.4 PADOG

PADOG, which is short for "Pathway Analysis with Down-weighting of Overlapping Genes" is a GSA method introduced by Tarca et al. (2012). It is categorized as FCS with a self-contained null hypothesis and is implemented in `R` as part of the `Bioconductor` project. In this work, the method has been chosen for its high performance in several benchmark studies of GSA methods summarized by Xie et al. (2021). For example, PADOG overall scores best under the conditions tested by Zyla et al. (2019), such as sensitivity, false positive rate (FPR), prioritization, computational time, and reproducibility. In another benchmark study conducted by Tarca et al. (2013), who is the author of PADOG, the method is reported as one of the top three methods to balance sensitivity and prioritization ability. Despite PADOG's high performance in these studies, the method is less popular than other methods and tools under investigation in this thesis. Accordingly, `Bioconductor`'s download statistics indicate 3356 downloads in 2021 with 1869 downloads from distinct IPs.

The general idea behind PADOG is to down-weight genes present in copious gene sets in the computation of the gene set-level statistic and to accordingly up-weight those genes that can be found in only few gene sets. This idea is based on the intuition that highly gene set-specific genes that are differentially expressed potentially indicate true relevance of the respective gene set with regard to the phenotype of interest.

The methodology that lies behind PADOG is as follows and the performed steps are additionally illustrated in Figure 3.2.

In an experiment with n gene sets G_k of size jG_kj under investigation, G is defined as set of genes that can be mapped to at least one gene set G_k, $i = 1;:::;n$. Then, the moderated t-statistic of gene $g_i \; 2 \; G$ between the two conditions 0 and 1 is calculated as follows (Smyth, 2004):

$$t_i = \frac{K_{i;0}^{\text{norm}} \quad K_{i;1}^{\text{norm}}}{V_i^{\text{limma}} \sqrt{\frac{1}{m_0} + \frac{1}{m_1}}}; \tag{3.5}$$

where $K_{i;0}^{\text{norm}}$ and $K_{i;1}^{\text{norm}}$ are the average expression levels of gene i and m_0, m_1 are the number of samples assigned in conditions 0 and 1, respectively. The interpretation of the moderated t-statistic is analogous to the regular t-statistic with the exception that the usual variance is replaced with posterior variance V_i^{limma}. By borrowing information from all genes in the experiment, V_i^{limma} is shrunk towards a common value, resulting in the t-statistics of an individual gene being more reliable.

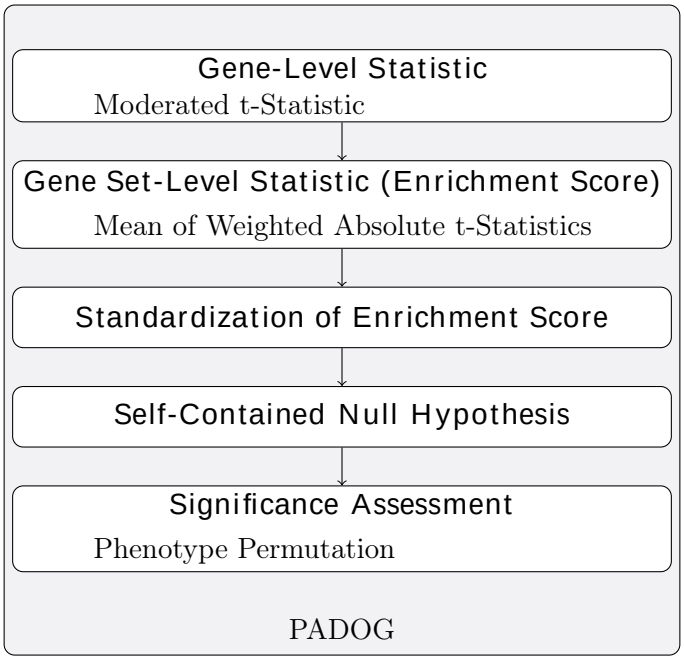

Figure 3.2: Steps Performed in PADOG

In order to compute the weight of each gene to eventually contribute to the gene set statistic, i.e. enrichment score, function $f(g_i)$ is defined as the frequency with which gene g_i can be found across all gene sets under investigation, hence $f(g_i) \; 2 \; f1;:::;ng$ for $g_i \; 2 \; G$. This function is then used to construct function $w(g_i)$ which assigns a weight to each gene:

$$w(g_i) = 1 + \sqrt{\frac{\max(f) \quad f(g_i)}{\max(f) \quad \min(f)}}; \tag{3.6}$$

where $w(g_i)$ is a monotonically decreasing function that is moreover bounded in $[1;2]$. The construction of $w(g_i)$ leads to genes with minimum frequency across all gene sets, i.e. $f(g_i) = \min(f)$, having weight $w = 2$ since

$$w(g_i) = 1 + \sqrt{\frac{\max(f) \quad f(g_i)}{\max(f) \quad \min(f)}} = 1 + \sqrt{\frac{\max(f) \quad \min(f)}{\max(f) \quad \min(f)}} = 1 + 1 = 2:$$

On the other hand, genes with maximum frequency are assigned weight 1 as

$$w(g_i) = 1 + \sqrt{\frac{\max(f) \quad f(g_i)}{\max(f) \quad \min(f)}} = 1 + \sqrt{\frac{\max(f) \quad \max(f)}{\max(f) \quad \min(f)}} = 1 + 0 = 1:$$

The enrichment score of gene set G_k is then computed as the mean of the weighted absolute

moderated t-statistics across the gene set, namely

$$S_0(G_k) = \frac{1}{jG_kj} \sum_{g_i 2 G_k} jt(g_i)j \quad w(g_i): \tag{3.7}$$

In the next step, the standardized enrichment score $S_0^0(G_k)$ of gene set G_k is obtained by performing row randomization on the enrichment score $S_0(G_k)$. Row randomization consists of subtracting the mean and dividing by the standard deviation of enrichment scores obtained by randomly selecting gene sets with an identical size as G_k from the entirety of genes in the experiment (Efron and Tibshirani, 2007). Afterwards, the enrichment score $S_0^0(G_k)$ is again standardized by subtracting the mean and dividing by the standard deviation of the $S_0^0(G_k)$-values from all n gene sets. Consequently, the observed standardized score $S_0(G_k)$ is obtained for each $k\ 2\ f1; :::; ng$.

The probability of observing an enrichment score at least as large as the observed is calculated by permuting the sample labels $N_{\text{ite}} = 1000$ times and computing the resulting observed standardized scores $S_{\text{ite}}(G_k)$, ite $= 1; :::; N_{\text{ite}}$:

$$P_{\text{PADOG}}(G_k) = \frac{\sum_{\text{ite}} \mathbb{1}fS_{\text{ite}}(G_k) \quad S_0(G_k)g}{N_{\text{ite}}}; \tag{3.8}$$

where $\mathbb{1}f\ g$ corresponds to the indicator function. Due to permutation of sample labels, gene-gene correlations are preserved.

3.5 clusterProfiler

`clusterProfiler` is an `R` package that was first introduced in 2012 to perform over-representation analysis for humans, mice and yeast as well as the comparison of functional profiles of various conditions on one level (Wu et al., 2021). The package is part of the `Bioconductor` project in the software `R` and has been updated since its release, now offering the conduct of GSEA and supporting copious additional species. The popularity of `clusterProfiler` is observable in the download statistics provided by `Bioconductor` which indicates a number of 203331 downloads in the year 2021, whereby 91229 downloads were performed from distinct IPs.

`clusterProfiler` offers a multitude of analyses, containing ORA and GSEA for standard gene set databases such as GO (see Section 2.4.1) and KEGG (see Section 2.4.2) but also user-defined gene set databases. It is noted that the analysis performed in this work is restricted to gene sets provided by the former two databases. Outdated versions of `clusterProfiler` additionally offer the conduct of DAVID in alignment with the methodology elaborated in Section 3.1. This utility is assessed in this thesis with regard to the potential for over-optimistic results as well and additionally requires the installation of the `R` package `RDAVIDWebService` (Fresno and Fernandez, 2013).

ORA offered by `clusterProfiler` is based on the hypergeometric distribution as described in Section 2.8.1 to investigate whether the number of genes in a given gene set that are associated with the phenotype of interest is larger than expected. The methodology of GSEA is a variation of Section 3.3 in the sense that it requires a ranking of the genes based on their correlation with the phenotype as input. The enrichment score is then calculated the same way as in GSEA from Section 3.3, namely by running down the ranked list and increasing a running-sum statistic when encountering a gene assigned to the gene set, whereas decreasing when it is not. The enrichment score corresponds to a weighted Kolmogorov-Smirnov Statistic. However, unlike in Section 3.3, the significance of a given enrichment score is assessed by permuting the gene labels in the ranked list and recomputing the enrichment score for the permuted data a fixed number of times, yielding the null distribution of enrichment scores (Yu et al., 2015). The p-value of the enrichment score is eventually computed relative to this null distribution.

4. Analysis Setup

In the following, the steps to assess the potential for over-optimistic results in GSA are described. While in Section 4.1, the general setup of the analysis is elaborated, the steps to pre-process the expression data set for the GSA tools and the components within the respective tools are presented in the Sections 4.2-4.6. This includes, for each step, the set of parameter options utilized to induce over-optimistic results. As Chapter 3 includes GSA tools as well as methods that are implemented in these tools, the term "tools" will be used uniformly throughout the subsequent analysis. It is emphasized that the evaluation of the results of a specific GSA tool also allows conclusions with respect to the potential of over-optimism on the underlying method.

4.1 General Analysis Setup

In order to evaluate the potential of GSA in general and the respective tools under investigation to produce over-optimistic results, a TCGA gene expression data set is utilized (see Section 4.2.1). To be more precise, the entirety of the corresponding samples are randomly labelled such that one half of the samples are labelled with the phenotype "1" and the other half with the phenotype "0". This is performed 5 times, resulting in 5 random phenotype permutations to conduct the analysis with. Then, for each of the 5 phenotype permutations separately, stepwise optimization of the parameter setting is performed to empirically determine the highest number of differentially enriched gene sets that can be obtained from each GSA tool under investigation. The potential of a tool for over-optimistic results is quantified by the maximum difference in the number of differentially enriched gene sets resulting from the default and the optimal configuration of the set of parameters. Furthermore, the potential for over-optimistic results in GSA in general is assessed by a comparison over all tools. Accordingly, a greater difference in the number of differentially enriched gene sets indicates a higher potential for over-optimistic results. It is noted that due to the randomly assigned labels and the resulting lack of biological meaning behind the 5 permutations, one would naturally assume that no gene set is eventually detected as differentially enriched. Therefore, the number of differentially enriched gene sets resulting from a tool, particularly with all parameters in their default setting, provides further insight into a tool's quality.

All methods and tools described in Chapter 3 offer a number of customizations in the parameter choice that can be made in the practical workflow to adapt to the research question. However, in extension to the steps presented in Figure 2.1, a number of additional pre-processing measures have to be performed in practice in order to generate the necessary input for each GSA tool under investigation. Altogether, the pre-processing steps include the pre-filtering of the gene expression data, the removal of duplicated gene IDs caused by gene ID conversion (if necessary), the transformation of the expression data for FCS and differential expression analysis conducted prior to ORA and those FCS tools which require a ranked gene list as input. In this thesis, the entirety of adaptions that can be made in these steps is utilized in the stepwise optimization process.
In this regard, the procedure is as follows: Firstly, all reasonable adaptions that can be made in the preparation process and in the course of the actual GSA tool are identified. It is emphasized that the aim of this thesis is to simulate a well-disposed user and therefore, only those options and adaptions that would be taken into consideration of such operator are put under investigation. Adaptions that would most probably be used willfully to manipulate the results are, on the other hand, explicitly excluded from the analysis. This concerns all possible adaptions in a GSA workflow that might be inappropriate in the given statistical or biological context as e.g. explicitly indicated by the author of the respective tool. After all possible adaptions for a GSA tool are identified, the default choice of each optimization step is set and the steps are ordered in a way to align with the progression of the practical GSA workflow. Whereas for some steps of the optimization process, the default choice is directly indicated by the tool itself, as e.g. presented in the respective user manual, for other steps, the default must be chosen manually. In this context, the default choice is set arbitrarily and deliberately without any knowledge about its effect on the development of the optimization process.
Next, stepwise optimization for each GSA tool is performed to successively find the configuration of parameters that leads to the highest number of differentially enriched gene sets. In this context, each component that is part of the practical GSA workflow remains in its default configuration in the first step and the resulting number of differentially enriched gene sets is declared as the current optimum. Then, the first possible adaption is carried out and in case the resulting number of differentially enriched gene sets increases in comparison to the current optimum, the current optimum is updated and the respective parameter is adapted for all subsequent steps of the optimization process. On the other hand, if this adaption does not lead to an increased number of differentially enriched gene sets, the parameter stays in its default configuration and the current optimum remains unchanged. This process is continued until each parameter is optimized with respect to the resulting number of differentially enriched gene sets.
After this stepwise optimization process is performed for each of the 5 random phenotype permutations and all GSA tools under investigation, the optimal result, i.e. the result with

the highest number of differentially enriched gene sets, is determined for each phenotype permutation, resulting in 5 optimal result tables for each tool.
All parameters that can be adapted in the workflow prior to the actual GSA tools are presented in Figure 4.1. The order of the parameters to be optimized is mainly based on this figure, nevertheless, local changes in the order are made if necessary. For example, for practical programming reasons, the order of the parameters to optimize within differential expression analysis differs between ORA and those FCS tools that require a ranked gene list as input. All adaptions that can be made within the individual GSA tools are then displayed in the respective subsequent sections.

4.2 Data Pre-Processing

In order to generate the required input for a given GSA tool, a number of steps have to be performed prior to the tool. The following sections contain the necessary preparatory steps required for the majority of tools under investigation. If a particular step is not necessary for a certain GSA tool, this is indicated accordingly. Furthermore, in the context of this analysis, not all pre-processing steps contain multiple options, such as the RNA-Seq data set itself and the manner of pre-filtering. These steps are presented nevertheless to impart an overall understanding of the necessary pre-processing steps applied to the gene expression data.

4.2.1 RNA-Seq Data Set

The expression data set utilized in this thesis is the TCGA GBM (Glioblastoma Multiforme) data set which was generated by the TCGA Research Network. The data set consists of 166 samples and prior to pre-filtering and the conversion of gene IDs, it contains gene expression measurements of 56602 genes in the Ensembl gene ID format (Howe et al., 2021). In the course of generating 5 random phenotype permutations, 63 samples are randomly assigned the phenotype "1" while the remaining 63 samples are assigned the phenotype "0".

4.2.2 Pre-Filtering

As suggested by Love et al. (2014), all genes that have less than 10 read counts across all samples are removed. It is noted that `edgeR` offers built-in utilities to filter out genes with low count values but for the purpose of consistency across all tools under consideration, the respective genes are filtered out manually at the beginning of the practical workflow. In addition to the purpose of pre-filtering described in Section 2.5, the reduction of the gene expression data set to genes with higher read counts leads to a simplification of the removal process of duplicated gene IDs resulting from gene ID conversion (see Section 4.2.4). This way, genes that are most probably of low use in the subsequent analysis are

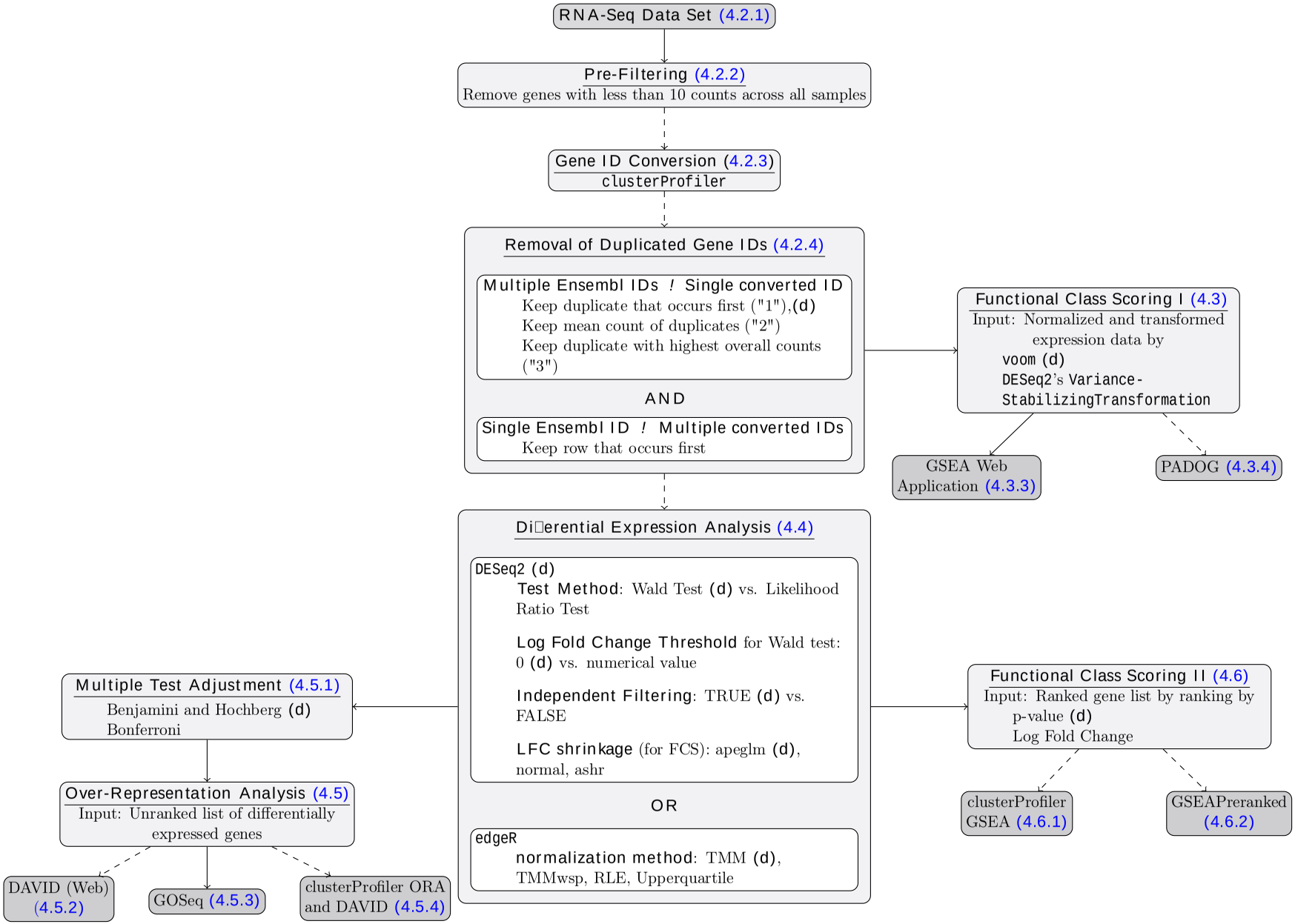

Figure 4.1: Pre-Processing Prior to Application of Gene Set Analysis Tools

Pre-processing steps before running GSA tools under investigation. Bullet points display possible adaptions in the practical workflow to induce an increase in the number of differentially enriched gene sets. Dashed arrows indicate the necessity for gene ID conversion and the resulting removal of duplicated gene IDs for the respective tool. Furthermore, the symbol (d) stipulates the default option.

prevented from generating duplicates to compete with other rows in the gene expression data.

4.2.3 Gene ID Conversion

In the gene expression data set utilized, genes are identified in the Ensembl format (Howe et al., 2021), however, a multitude of tools under investigation, namely `clusterProfiler`, `PADOG`, DAVID and GSEAPreranked, require the gene IDs to be in the Entrez gene ID (Maglott et al., 2010) or HUGO gene symbols (Tweedie et al., 2021) format. Consequently, gene IDs are converted using the functionalities of the `clusterProfiler` package. This conversion scheme is not bijective, i.e. not all genes in the Ensembl format can be transformed into a unique ID of the required format. For some IDs, there is no transformed ID while for others, multiple Ensembl IDs are transformed into the same single converted ID or vice versa. This firstly leads to an immediate reduction of the expression data set and secondly to the occurrence of duplicate IDs. The second aspect is addressed in the following section.

4.2.4 Removal of Duplicated Gene IDs

In the case of multiple Ensembl gene IDs that are mapped to a single Entrez gene ID, the expression data set contains multiple rows of the same corresponding Entrez gene ID whereby each row contains different count data. On the other hand, in the case of a single Ensembl ID being converted to multiple distinct Entrez IDs, the expression data contains a multitude of rows that have different Entrez IDs but identical count data. Both of these matters have to be considered individually. In a real setting, the user has to understand the biological source of such an ambiguous conversion scheme and adjust the removal process of the duplicate IDs accordingly. However, there seems to be no general consensus on a reasonable removal scheme of duplicate IDs. Therefore, three options to remove duplicate gene IDs are utilized in this thesis.

The default approach to removing duplicate gene IDs is, in accordance with Silva et al. (2016), to keep the row in the expression data set which occurs first. This can be applied to the removal of duplicate rows resulting from a single Ensembl gene ID that is converted to multiple Entrez gene IDs as well as rows that have the same Entrez gene ID but different count data. In the second option, duplicates are replaced with a single row that contains the mean count values across all duplicates for each sample in the expression data set. This option is only meaningful in the case when multiple distinct Ensembl IDs are mapped to a single Entrez ID as the count data in each of the rows vary. However, in the case of a single Ensembl gene ID that is converted to a multitude of Entrez gene IDs, each of the resulting rows contains identical count data. Consequently, analogously to option 1, the row that occurs first is kept. Lastly, option 3 consists of keeping the row in the expression data which yields the highest overall count data, i.e. the highest sum of counts across all

samples, among the duplicates. The intuition behind this approach is that a high number of counts is expected to lead to a high power in the detection of differential expression and differential enrichment. As in option 2, this approach is not meaningful in the case of a single Ensembl gene ID that is converted to multiple Entrez IDs and therefore, the row that occurs first is kept.
These three alternative approaches to removing duplicate gene IDs result in 3 distinct, but similar gene expression data sets that are utilized in the stepwise optimization process and referred to as options "1", "2" and "3" respectively. For each phenotype permutation and GSA tool, the gene expression data set yielding the highest number of differentially expressed genes (for ORA tools) or the highest number of differentially enriched gene sets (for FCS tools) is chosen as the optimal expression data set. As described in the above Section 4.2.3, the removal of duplicated gene IDs is only necessary for `clusterProfiler`, `PADOG`, DAVID and GSEAPreranked.

4.3 FCS I (with Gene Expression Data as Input)

As the tools GSEA web application and `PADOG`, which belong to FCS, require the entire expression data set as input and utilize ranking metrics originally developed for microarray data, the expression data must be transformed in accordance with Section 2.7. In this thesis, the two options for RNA-Seq data transformation are `voom` from the `R` package `limma` and `VarianceStabilizingTransformation` from the `R` package `DESeq2`, which are described in further detail in the following. According to Geistlinger et al. (2016), these methods enable the application of microarray methods for the transformed RNA-Seq data. It is noted, however, that there seems to be no general consensus with respect to the method of transformation as well as the necessity itself of performing a transformation in the first place. For example, the documentation of the GSEA web application points out that it was originally developed for microarray data and therefore a proper normalization method, such as trimmed mean of M-values (see Section A.1.2) or median-of-ratios (see Section A.1.1), is necessary. However, a comment is then made that even though the ranking metrics are commonly used for RNA-seq data sets, there has yet to be an evaluation of whether these metrics, initially developed for microarray data, are actually appropriate for RNA-seq data sets.
For each of the two tools GSEA web application and `PADOG`, the optimal transformation method for each of the random phenotype permutations is chosen as the method that yields the highest number of differentially enriched gene sets with the respective GSA tool in its default configuration. It is noted that `voom` is specified as the default transformation method.

4.3.1 voom

voom, which is an acronym for "variance modeling at the observational level", was introduced by Law et al. (2014). The raw counts K_{ij}, $i = 1;:::;N$, $j = 1;:::;m$, are transformed by calculating the log-counts per million (log-cpm) as follows:

$$\kappa_{ij} = \log_2\left(\frac{K_{ij} + 0{:}5}{S_j \quad s_j + 1} \quad 10^6\right); \tag{4.1}$$

where $S_j \quad s_j$ is the library size multiplied by the normalization factor, resulting in the effective library size. The normalization factor s_j is obtained by performing a normalization technique. Law et al. (2014) suggest the use of the Trimmed Mean of M-values method, which is described in further detail in Section A.1.2, however, there again seems to be no general consensus on the choice of the normalization technique. The log-cpm values are by definition normalized for differences in library sizes across samples and by scaling the library size S_j with the Trimmed Mean of M-values method, the transformed values are additionally normalized for compositionality effects. In Equation (4.1), raw count values K_{ij} are offset by $0{:}5$ to reduce the variability of the log-cpm values for low counts and to avoid taking the $\log_2$ of 0. Additionally, the effective library size is offset by 1 to ensure that the fraction in the equation is strictly between 0 and 1.
It is noted that the official voom method presented by Law et al. (2014) further proceeds by estimating the mean-variance trend from the data and incorporating it into precision weights which are calculated for each normalized count observation individually. These precision weights are then fed into the standard limma pipeline (Smyth, 2004). In the context of gene set analyses, however, the term "voom transformation" is solely used to refer to the transformation presented in Equation (4.1). This practice can be observed in Shahjaman et al. (2020) and Zhang et al. (2019).
The transformed count data κ_{ij} in Equation (4.1) have a more stable mean-variance relationship and are closer to the normal distribution compared to the raw counts K_{ij} but are usually more variable for lower magnitudes of transformed counts. This would, in the standard voom workflow, be addressed by estimating the precision weights.

4.3.2 varianceStabilizingTransformation via DESeq2

Another transformation method for RNA-Seq data is integrated in the DESeq2 package and was introduced by Anders and Huber (2010). First, raw counts K_{ij} are normalized using the median-of-ratios method (see Equation (A.1)). The mean-variance relationship is estimated via DESeq2 and the variance stabilizing transformation is performed in a way that for transformed values, the variance is approximately constant throughout the range of the mean. Therefore, this transformation results in approximately homoscedastic values κ_{ij} which are additionally normalized for library size. Furthermore, for large counts, the transformed values are asymptotically equal to the $\log_2$-values of the normalized counts.

4.3.3 GSEA Web Application

GSEA web application utilizes the methodology which is described in Section 3.3. The web application accepts a multitude of gene ID formats as input so that no gene ID conversion (see Section 4.2.4) has to be performed prior to running the application. Consequently, there is no need to optimize over the three expression data sets resulting from gene ID conversion but instead, one pre-filtered gene expression data set is utilized. As the required input is a normalized and transformed expression data set, `voom` is chosen by default and the optimal transformation method is selected in a manner to maximize the number of differentially enriched gene sets with the parameters of the tool in their default configuration. An exception to the default parameter setting is the seed for the generation of the 1000 phenotype permutations in the assessment of gene set significance. In order to facilitate reproducibility of the results and consistency between all runs of the GSEA web application, the seed is set to an identical value across all runs.
The set of adaptions within the GSEA web application that are utilized in this thesis are summarized in Table 4.1. It is noted that in the GSEA web application, the criterion to detect a gene set as differentially enriched is a false discovery rate below 0:25 (FDR < 0:25) and no method for multiple test adjustment has to be chosen in the last step.

Table 4.1: Adaptions within GSEA Web Application

Parameter	Default	Alternative Options
Gene Set Database	GO (MF)	• KEGG • Hallmark
Gene Level-Ranking Metric	Signal2Noise Ratio	• t-Test • Difference of Classes
Exponent	weighted ($p = 1$)	• weighted ($p\ 2\ f1:5; 2g$) • classic ($p = 0$)

The entirety of gene set databases that can be selected in the GSEA web application is part of the Molecular Signature Database ("MSigDB") which is one of the largest and most popular repositories of gene sets (Liberzon et al., 2015). MSigDB is divided into 9 collections of gene set databases, such as "curated databases" or "ontology gene sets". In this context, KEGG is contained by the collection of curated databases and consists of 186 gene sets. Molecular Function, as part of Gene Ontology, contains 1738 gene sets. The default gene set database in the GSEA web application is the Hallmark Gene Set Collection. This collection consists of 50 gene sets that were generated in a way to reduce redundancy across and heteroskedasticity within gene sets by condensing over 4000 gene sets from MSigDB (Liberzon et al., 2015).
These three gene set databases, namely GO, KEGG and Hallmark, are utilized in this thesis and GO with subontology Molecular Function is set as the default gene set database. It

is noted that the actual number of gene sets provided by a gene set database can be lower than indicated above as only those gene sets that fulfill size restrictions are considered by the web application.

Another change that can be made by the user is the gene-level ranking metric which captures, as explained before, the correlation of a gene with the phenotypes. The web application of GSEA offers three options of gene-level ranking metrics for a discrete phenotype and expression data on the log scale. In all of them, a large value of the statistic indicates a more distinct expression of the gene between the phenotypes 0 and 1, compared to a smaller value. In addition to the Signal2Noise Ratio in Equation (3.4), the GSEA web application offers the t-statistic which additionally includes the number of samples m_0 and m_1 of both phenotypes. It is, as commonly used, defined as

$$t = \frac{K_{i;0}^{\text{norm}} \quad K_{i;1}^{\text{norm}}}{\sqrt{\frac{^2_0}{m_0} + \frac{^2_1}{m_1}}}; \tag{4.2}$$

where $K_{i;0}^{\text{norm}}$ and $K_{i;1}^{\text{norm}}$ are the means of the transformed and normalized counts of gene i in the samples ascribed to phenotypes 0 and 1, respectively. Moreover, $_0$ and $_1$ refer to the standard deviations of the transformed and normalized counts in both phenotypes. Furthermore, the user can generate the ranking of the genes based on the average fold change between both phenotypes. Due to the transformation of the expression data using `voom` or `DESeq2`'s `varianceStabilizingTransformation`, the values are consequently on the log scale so that another suitable ranking metric is Difference of Classes, namely

$$\text{DoC} = K_{i;0}^{\text{norm}} \quad K_{i;1}^{\text{norm}}. \tag{4.3}$$

It is noted that the GSEA web application offers additional gene-level ranking metrics but these are only applicable for count data on the natural scale or in the case of a continuous phenotype. In this optimization step, the optimal ranking metric is chosen as the metric that yields the highest number of differentially enriched gene sets based on the optimally transformed expression data set and the optimal gene set database as chosen in the foregone optimization steps.

The last adaption that is utilized within the GSEA web application is the choice of the exponent which corresponds to the parameter p in Equation (3.2). In accordance with the set of optional values offered by the tool, this thesis utilizes the values $p \in \{0; 1; 1.5; 2\}$ to maximize the number of differentially enriched gene sets, whereby value $p = 1$ is the default exponent value. As already elaborated in Section 3.3, a higher exponent leads to an increased weight of the correlation between a gene's expression pattern and the phenotypes in the computation of the enrichment score. In this regard, $p = 0$ leads to equal weights for all genes in the experiment. Subramanian et al. (2005) suggest a choice of $p < 1$ if the user wants to focus on gene sets whose gene members show coherent expression patterns.

On the other hand, they propose to choose $p > 1$ if the gene set database consists of large gene sets and only a small number of genes within the gene sets are expected to show a coherent expression behaviour. In general, however, they recommend $p = 1$ as a reasonable choice. Eventually, the optimal exponent with respect to the number of differentially enriched gene sets is chosen based on the optimal parameter configuration established in the previous optimization steps.
It is noted that the tool additionally offers the option to restrict the inclusion of gene sets in the analysis with respect to gene set size. The default restriction values in the application are 15 and 500, which means that only those gene sets are included whose size is between 15 and 500 genes. The intuition behind this restriction is that the normalization performed for gene set size, as described in Section 3.3, is not accurate for very small or large gene sets. For example, the normalized enrichment score of a small gene set is potentially inflated as a small portion of the genes can generate significant results. Reimand et al. (2019) additionally point out that small gene sets are often contained by larger gene sets and are therefore redundant. This can lead, in turn, to complications in the interpretation and to a more stringent multiple testing adjustment. Furthermore, large gene sets are possibly overly general and thus do not contribute to the interpretability of the results. Due to this reasoning, the minimum and maximum gene set size restrictions are not included in the analysis and therefore remain in their respective default values.
Furthermore, the GSEA web application additionally offers the assessment of differential enrichment of a gene set based on gene set permutation (see Section 2.8.2). This option is, however, omitted from the analysis as the developers of the web application explicitly recommend the use of phenotype permutation. The reason for this is that, in contrast to gene set permutation, phenotype permutation preserves correlation structures between the genes within the gene sets which ultimately leads to results that are more biologically meaningful.

4.3.4 PADOG

As `PADOG` requires the gene IDs in the input data set to be in the Entrez gene ID format, the optimal expression data set is chosen between the three expression data sets resulting from Section 4.2.4 in the first step of the optimization process. In this context, the remaining parameters required to perform `PADOG`, including the method to perform the transformation of the RNA-Seq data set and multiple test adjustment, remain in their default configuration. Based on the resulting optimal expression data set, the optimal RNA-Seq data transformation method is chosen in the next step. Furthermore, the set of utilized parameters in this thesis to induce an increase in the number of differentially enriched gene sets in `PADOG` is presented in Table 4.2.
In order to ensure reproducibility within this thesis, the optional seed to generate the random iterations for the assessment of significance is set. In this context, for the purpose

Table 4.2: Adaptions in PADOG

Parameter	Default	Alternative Options
Multiple Test Adjustment	Benjamini and Hochberg	Bonferroni

of consistency between different runs of `PADOG`, an identical seed is set for each run and therefore, this parameter is explicitly omitted from the optimization process. Furthermore, the minimum gene set size to include gene sets can be adapted and is set to the default value 3. This parameter is explicitly excluded from the analysis based on the reasoning presented in the previous Section 4.3.3. Moreover, the number of random iterations to calculate the p-value of enrichment has a default value of `1000` but can be changed to virtually any value. However, in the context of this thesis, changing this value would correspond to a willful manipulation since the number of iterations can be used to change the results to one's benefit solely by exploiting a random component. Therefore, there are no changes performed within the `PADOG` functionality in this thesis to increase the number of differentially enriched gene sets. Eventually, multiple test adjustment must be performed externally and for this purpose, Benjamini and Hochberg (1995) and Bonferroni (1936) are used in this thesis to maximize the number of differentially enriched gene sets based on the optimal parameter choices established in the previous steps.
It is noted that, in contrast to the other GSA tools under investigation, `PADOG` only offers the use of the KEGG gene set database. Gene sets provided by GO, on the other hand, can only be utilized by manually uploading the corresponding list of gene sets in combination with the annotation between genes and gene sets. Therefore, in this thesis, the analysis of `PADOG` is only performed based on KEGG gene sets.

4.4 Differential Expression Analysis

As all tools that belong to ORA require a list of differentially expressed genes as input, the conduct of differential expression analysis prior to applying the actual GSA tool is necessary. Accordingly, stepwise optimization of this list is performed by adapting the parameters in a way to maximize the number of differentially expressed genes. The intuition behind this is that, in accordance with Fisher's Exact Test displayed in Equation (2.12), one would expect that an increased number of differentially expressed genes leads to a higher overlap between a given gene set and the input list, resulting in a decreased p-value of over-representation. This means that, in a slight deviation from the general analysis setup described in Section 4.1, the set of optimal parameters of differential expression analysis is not found by feeding the input list to the respective ORA tool and maximizing the resulting number of differentially enriched gene sets. After optimization of the unranked list of differentially expressed genes, which is subsequently called "optimized list", the optimal choice between this list and the list of differentially expressed genes resulting

from the set of parameters of differential expression analysis in their default configuration, called "default list", is made with the respective ORA tool in its default configuration.
For tools of FCS that require a ranked list of genes as input, this list is generated based on the results of differential expression analysis and all genes in the experiment are ranked according to a ranking metric presented in Section 4.6. In this context, however, the optimization process of the parameters of differential expression analysis is carried out differently from ORA tools: The parameters of differential expression analysis are adapted step by step and the resulting ranked gene list is fed to the FCS tool in each step, while the tool itself remains in its default configuration. Eventually, the optimal parameters of differential expression analysis are chosen as the set of parameters to generate a ranking that yields the highest number of differentially enriched gene sets with the respective FCS tool in its default configuration.
Both options for differential expression techniques utilized in this thesis are, as already introduced in Chapter 2.6, `DESeq2` and `edgeR` and the corresponding set of parameters to be optimized can be found in Figure 4.1 and are further elaborated in the following Subsections 4.4.1-4.4.2. It is noted that, due to the practical programming workflow, the stepwise optimization process differs between the ORA tools as well as the corresponding FCS tools with respect to the order of the parameters to be optimized.

4.4.1 DESeq2

The first change that can be made in the `DESeq2` workflow only affects the procedure in FCS tools, namely the method to perform log fold change shrinkage. The reason for this is that the calculated p-values to detect differential expression remain unaffected and the log fold change values, which are shrunken in the shrinkage process, are not used to generate an unranked list of differentially expressed genes for ORA. In addition to the standard shrinkage estimator, two alternatives can be chosen.
The first alternative shrinkage estimator was proposed by Love et al. (2014) and starts with the Maximum Likelihood Estimates $\hat{}_i$ based on the Generalized Linear Model (2.2) for the log fold change of gene i between both conditions and applies an Empirical Bayes procedure. A zero centered normal distribution is fit to the log fold changes of all genes, leading to the assumption

$$_i \quad N(0; \ ^2); \quad i = 1; :::; N: \tag{4.4}$$

Based on this distribution as a prior distribution, a generalized linear model is fit a second time to obtain the final estimates of the log fold change of the genes. Additionally, the standard error for each estimate is derived from the posterior distribution's curvature at its maximum. It is noted that this shrinkage estimator was introduced as the original default shrinkage option. However, it has been replaced by the shrinkage estimator `apeglm` (see

Section 2.6.1) since due to the Cauchy distribution as a prior distribution, the latter does not overshrink large log fold change values.
The second alternative shrinkage estimator was introduced by Stephens (2017) and is implemented in the `R` package `ashr`. The method is based on Empirical Bayes and moreover on the assumption that the distribution g of the true log fold changes β_i is unimodal. It uses the vector $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_N)$ of log fold change estimates as well as corresponding estimated standard errors $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_N)$ as input. The overall goal is the estimation of the posterior distribution via Bayes theorem, namely

$$p(\beta_i | \hat{\beta}_i, \hat{s}_i) \propto p(\beta_i | \hat{s}_i) \cdot p(\hat{\beta}_i | \beta_i, \hat{s}_i). \tag{4.5}$$

Distribution g is assumed to be a mixture of a point mass at 0 and an additional mixture of normal distributions centered at 0. Moreover, $p(\hat{\beta}_i | \beta_i, \hat{s}_i)$ is assumed to be approximated by a normal distribution. The general estimation procedure consists of two steps. At first, g is estimated by maximizing a penalized likelihood which is used as a prior distribution in the computation of the posterior distribution (4.5) in the second step. It is noted that, according to Love et al. (2014), the shrinkage methods `apeglm` and `ashr` have shown to have less bias than the shrinkage type `normal`.
Another possible adaption in the `DESeq2` workflow to create a list of genes, ranked or unranked, is the choice of the test method when evaluating the significance of a shrunken log fold change estimate. In addition to Wald test, `DESeq2` offers the likelihood ratio test (Koch, 1999). In the context of this thesis, the likelihood ratio test examines the full model which contains the variable that indicates the phenotype of each sample as well as a reduced model which only consists of the intercept. The likelihood ratio test then tests whether the removed variable explains a significant amount of variation in the data. Consequently, the specification of the alternative hypothesis is necessary in the practical workflow if the likelihood ratio test is chosen as the test method.
As differential expression analysis is usually performed for thousands of genes, multiple test adjustment is inevitable. As this adjustment is, on the downside, associated with a loss of power to detect a gene as differentially expressed, `DESeq2` offers independent filtering to automatically exclude those genes from the analysis that have a marginal possibility of being detected as differentially expressed from the start. This filtering criterion must be independent of the test statistic under the null hypothesis and `DESeq2` therefore proposes the average expression strength across all samples: A gene with a mean normalized read count below the filtering threshold is then omitted from the analysis and the resulting adjusted p-value is set to "NA".
Independent of the test choice discussed above, the default approach of `DESeq2` is to test against the null hypothesis of a log fold change of 0. This would, however, indicate for such a gene that it is completely uninvolved in all processes which is rather unrealistic given the connectedness of the network of genes. Furthermore, as sample size and therefore

statistical power increases, the probability of a gene with a small but non-zero log fold change being detected as differentially expressed increases, leading to a long list of genes consisting of only few that are actually biologically significant. DESeq2's way to address this issue is to include a non-zero log fold change threshold into the statistical testing procedure:

$$H_0 : j \; _i j \tag{4.6}$$

This threshold is set to 0 in the default setting, leading to the standard test procedure but can optionally be set to a non-zero value in the practical workflow. In combination with the intuition elaborated above, the specification of a non-zero log fold change threshold is expected to result in a decreased number of differentially expressed gene sets. It is noted that a log fold change threshold alternative to 0 can only be specified in combination with the Wald test. In the case of a specification of the likelihood ratio test, any log fold change threshold other than 0 leads to an overwriting of the results with the results from the corresponding log fold change threshold and Wald test. Moreover, in the context of the FCS tools that require a ranked list of genes as input, the adaption of the log fold change threshold in combination with log fold change shrinkage is only applicable for the shrinkage method `normal`. For `ashr`, on the one hand, the specification of a log fold change threshold is ignored while for the method `apeglm`, p-values are replaced with s-values which provide for each gene the average of the probability of a wrong sign of the estimated log fold change over all genes with a corresponding equal or lower probability. In the latter case, the resulting s-values are not associated with the research question in this thesis and therefore, the specification of an alternative log fold change threshold is only considered when using the shrinkage method `normal`.

In a real experiment, the user should choose a log fold change threshold that is biologically reasonable for the given research question and the data at hand. In this thesis, the specification of an alternative log fold change threshold is expected to be of low relevance since it can only be utilized in few parameter settings and is expected to lead to a decreased number of differentially expressed genes. Ultimately, an alternative log fold change threshold value of $\log_2(1.05)$ is utilized.

It is noted that the usage of Cook's outlier detection (see Section 2.6.1) is not utilized in the optimization process as a deactivation would naturally lead to an increase in the number of differentially enriched gene sets. The reason for this is that Cook's outlier detection leads to a replacement of outliers with values that comply with the null hypothesis of no differential expression. In this context, a deactivation of Cook's outlier detection is not carried out by simply changing one argument in a single function, but by adapting arguments in two consecutive functions and therefore requires a conscious look into DESeq2's vignette. This means that a user who intends to deactivate Cook's outlier detection is most likely aware of its effects on the results of differential expression analysis. Furthermore,

the developers of DESeq2 recommend a deactivation of Cook's outlier detection explicitly for further exploration of the expression data set in the case of many outliers.

4.4.2 edgeR

The adaptions utilized within edgeR to induce an increase in the number of differentially enriched gene sets is the normalization method and multiple options are offered in addition to TMM described in Section 2.6.2.
The first alternative to normalization via the trimmed mean of M-values is a variant of it, namely TMM with singleton pairing (TMMwsp). The idea behind it is to exclude those genes with zero counts in either sample when comparing samples. The non-zero counts of these genes are then reused to increase the number of features by which the samples are compared. Afterwards, singleton positive counts are paired up between the samples in decreasing order of size and eventually, a slightly modified variant of the trimmed mean of M-values is applied to the re-ordered samples. All in all, TMMWsp is recommended over TMM in the case of many unexpressed genes, i.e. many genes with zero counts.
Another alternative to TMM is the relative log expression (Anders and Huber, 2010), abbreviated with "RLE", which is additionally included in the DESeq2 package. Identically to Equation (A.1), the size factor of sample j is computed as the median over all genes of the ratio of counts to the geometric mean of counts across all samples.
Finally, in the upper-quartile normalization method (Bullard et al., 2010) the size factor of a sample j is calculated by first excluding genes with zero counts across all genes and then taking the upper quartile, i.e. 75th percentile, of counts of all remaining genes in the sample.

4.5 Over-Representation Analysis

In this thesis, the ORA tools chosen to be investigated with respect to the potential for over-optimistic results are DAVID, GOSeq and clusterProfiler. It is noted that DAVID can be performed by the DAVID web application as well as the R package clusterProfiler. Tools that belong to ORA require a list of differentially expressed genes as input and this list is obtained by performing differential expression analysis for each gene in the experiment with the optimal parameters obtained as described in the previous sections. Additionally, adjustment for multiple testing is necessary, which is further elaborated in the following Subsection 4.5.1. The set of possible parameter adaptions for each of the ORA tools is elaborated in the Subsections 4.5.2-4.5.4.

4.5.1 Multiple Test Adjustment

In this thesis, the multiple test adjustment methods utilized to maximize the number of differentially expressed genes in the input list are Benjamini and Hochberg (1995)

as well as Bonferroni (1936). The reason behind this choice is that these two methods are commonly used in many statistical analyses (Reimand et al., 2019). Nevertheless, it is noted that Bonferroni (1936) is known to be more conservative than Benjamini and Hochberg (1995) (Reimand et al., 2019) and therefore is expected to lead to a lower number of differentially expressed genes. Moreover, Reimand et al. (2019) point out that Benjamini and Hochberg (1995) and Bonferroni (1936) assume independence between the statistical tests. However, this assumption can be unrealistic in the context of GSA since gene sets are usually not perfectly disjunct. This might, despite the methods' popularities, eventually lead to inaccurate estimates of the false discovery rate.

4.5.2 DAVID Web Application

The web application DAVID offers several adaptions that are made use of in this thesis to maximize the number of differentially enriched gene sets. These are summarized in Table 4.3. It is noted that the web application technically accepts gene IDs in the Ensembl gene ID format, however, only a marginal fraction of genes can be mapped in the upload of the gene list. As a consequence, the user is recommended by the tool to upload the genes in the Entrez gene ID format, which is complied with in this thesis. Firstly, the user has the

Table 4.3: Adaptions in DAVID Web Application

Parameter	Default	Alternative Options
Gene Set Database	GO (MF)	KEGG
Background Genes	All Genes Annotated to Gene Set Database	All Genes Measured in Experiment
Multiple Test Adjustment	Benjamini and Hochberg	Bonferroni

possibility to choose from a plethora of gene set databases. In alignment with the other tools investigated in this thesis, GO with subontology Molecular Function is chosen as the default configuration and gene set database KEGG is utilized as an alternative option. The optimal choice of a gene set database is therefore made in a way to maximize the number of differentially enriched gene sets with the remaining parameters in their default configuration. Furthermore, in addition to the universe, i.e. set of background genes, comprised of the entirety of genes annotated to the chosen gene set database, the user has the option to provide a customized set of background genes. In accordance with the suggestion made by Geistlinger et al. (2021), the possible choices of background genes are extended to all genes measured in the experiment, i.e. all genes whose differential expression was measured in the preceding differential expression analysis. Consequently, the optimal universe with respect to the number of differentially enriched gene sets is determined with the previously chosen optimal gene set database and remaining parameters in their default configuration.

In alignment with the other GSA tools under investigation in this thesis, the selection of

utilized multiple test adjustment methods consists of Benjamini and Hochberg (1995) and Bonferroni (1936). The choice of the optimal multiple test adjustment method is made in the last step of the stepwise optimization process.
It is noted that the tool offers the calculation of the p-value of over-representation using Fisher's exact method (see Equations (2.11) and (2.12)) in addition to the EASE score. However, this adaption only yields an unadjusted p-value whereas the adjusted p-value remains unchanged and therefore results from the calculation of the EASE score (see Section 3.1). That's why this option is omitted from the analysis in this thesis.

4.5.3 GOSeq

As `GOSeq` is a tool categorized as ORA, the optimal input vector is obtained from the optimal differential expression analysis result table generated as described in the Sections 4.4 and 4.5.1. However, this GSA tool requires as input a variant of the list of differentially expressed genes, namely a named binary vector of all genes in the experiment. In this vector, each differentially expressed gene is assigned the value "1" while all genes not differentially expressed are labelled with "0". `GOSeq` itself offers a variety of changes that can be carried out in the practical workflow and these are presented in Table 4.4.

Table 4.4: Adaptions in GOSeq

Parameter	Default	Alternative Options
Correction for Bias	Length Bias	All Biases
Gene Set Database	GO (MF)	KEGG
Background Genes	All Genes Annotated to at least One Gene Set	All Genes Measured in Experiment
Enrichment Score Calculation	Wallenius Approximation	Resampling
Multiple Test Adjustment	Benjamini and Hochberg	Bonferroni

The first adaption to utilize is the bias to account for in the analysis. The default option, namely length bias, is described in Section 3.2. As already described, length bias refers to the phenomenon that longer genes are more likely to be detected as differentially expressed compared to genes with shorter transcript length. Alternatively, the user can adjust for the total number of counts and therefore all biases present in the expression data instead of only gene length. This adaption facilitates an additional adjustment for a gene's expression level, whereby a higher expression level naturally leads to a higher magnitude of counts and therefore an increased statistical power. This way, a higher weight is assigned to genes with lower counts. A possible downside of this adaption is that accounting for the total number of read counts may lead to the removal of the bias resulting from actual differential expression. In this first step of `GOSeq`, the optimal bias to account for is chosen based on the optimal input vector and with respect to the number of differentially enriched

gene sets, while the remaining parameters remain in their default configuration.
In the next step, the optimal gene set database is chosen. Analogous to the investigation of the other GSA tools, GO with subontology Molecular Function is chosen as the default database. However, it is possible to obtain gene sets from KEGG and based on the previously determined optimal input vector and bias, an optimal gene set database is determined with respect to the number of differentially enriched gene sets.
The third adaption concerns the choice of the universe, i.e. background genes in `GOSeq`. In the default configuration, only those genes which can be assigned to at least a single gene set from the gene set database of choice are used towards the calculation of a p-value of over-representation. Additionally, `GOSeq` offers the possibility to include all genes, whether assigned to a gene set or not, in the assessment of differential enrichment. In the setup of the adapted universe, genes that cannot be assigned to a single gene set database still count towards the entirety of genes outside of the given gene set that is currently tested. It is noted that this was the default option in earlier versions of `GOSeq`. Furthermore, it is underlined that this adaption of the universe differs from the corresponding adaptions in the other ORA tools.
In `GOSeq`, the default method to calculate a p-value of over-representation is Wallenius Approximation, which is an approximation of the computationally expensive resampling technique described in Section 3.2. Nevertheless, resampling is offered as an option to evaluate differential enrichment and is utilized in this thesis with a fixed number of 1000 iterations to maximize the number of differentially enriched gene sets based on the previously established optimal parameter configuration. It is noted that the usage of the regular hypergeometric distribution for the calculation of the p-value is possible, too. But, the user is explicitly discouraged by the authors to use this option since, in the case of no bias being present in the experiment, the probability weighting function reflects this and produces meaningful results nonetheless. Consequently, this option is excluded from the analysis performed in this thesis and the optimal calculation method is chosen between Wallenius approximation and resampling.
As multiple test adjustment is not performed internally in `GOSeq`, the methods by Benjamini and Hochberg (1995) and Bonferroni (1936) are applied using the `R` package `base` (R Core Team, 2021). Eventually, the optimal multiple test adjustment method is chosen based on the optimal parameters as determined in the preceding optimization steps.

4.5.4 clusterProfiler's DAVID and Regular ORA

Since for DAVID and the regular ORA tool, the majority of adaptions that can be made in `clusterProfiler` are identical, these are presented collectively in the following. It is noted that necessary packages to perform DAVID, as this tool is no longer supported in the current version of `clusterProfiler`, are loaded manually into the `R` environment. The regular ORA tool requires the gene IDs to be of the Entrez gene ID format, whereas

DAVID works with the Entrez as well as the Ensembl gene ID format. For the purpose of consistency with the web application of DAVID (see Section 4.5.2), however, DAVID within `clusterProfiler` is analyzed using Entrez gene IDs as well. The following Table 4.5 displays the set of adaptions utilized in `clusterProfiler`'s regular ORA as well as DAVID in order to maximize the number of differentially enriched gene sets. The first possible

Table 4.5: Adaptions in clusterProfiler's Over-Representation Analysis and DAVID

Parameter	Default	Alternative Options
Gene Set Database	GO (MF)	• KEGG • KEGG Modules (regular ORA)
Background Genes	All Genes Annotated to Gene Set Database	All Genes Measured in Experiment
Multiple Test Adjustment	Benjamini and Hochberg	Bonferroni

adaption is the choice of the gene set database. For both tools, GO with subontology Molecular Function is the default choice in the first optimization step. Whereas in DAVID, the gene set database to compare to with respect to the number of differentially enriched gene sets is KEGG, an additional comparison in the regular ORA tool is made with KEGG Modules. After the determination of the optimal gene set database with all remaining parameters in their default configuration, the adaption of the universe, i.e. background set of genes, is utilized in the second step with respect to the number of differentially enriched gene sets. In alignment with the other tools apart from `GOSeq` that implement a form of ORA, the alternative universe is chosen as the entirety of genes measured in the experiment. The last option to consider, with the foregoing parameters in their optimal configuration, is the method of multiple test adjustment and here, Benjamini and Hochberg (1995) and Bonferroni (1936) are again chosen as the two options.

4.6 FCS II (with Ranked Gene List as Input)

In contrast to `PADOG` and the regular GSEA web application, GSEAPreranked and GSEA conducted by `clusterProfiler` require a ranked list of all genes in the experiment as input. The rank of each gene then represents the magnitude of differential expression across the phenotypes and can be generated with a meaningful ranking metric of the user's choice. Reimand et al. (2019) suggest to compute a gene's ranking as

$$\text{rank} = \log_{10}(p) \ \text{sign}(\text{LFC}). \tag{4.7}$$

In this context, p denotes the unadjusted p-value and LFC refers to the log transformed fold change between both phenotypes. Both quantities can be obtained from the results table of differential expression analysis performed with `DESeq2` as well as `edgeR`. This ranking

is subsequently referred to as "ranking by p-value". The intuition behind this ranking metric is that up-regulated genes, i.e. genes with a positive log fold change between both phenotypes, are assigned a positive rank, whereas down-regulated genes are allotted a negative rank. Moreover, a smaller p-value leads to a higher absolute rank compared to a large p-value. This means that, in sum, those genes with a higher absolute and significant fold change between the phenotypes are placed at the top or bottom of the ranked list.
Another ranking metric considered in this work is a ranking generated solely based on the log fold change value of a gene such that

$$\text{rank} = \text{LFC}: \tag{4.8}$$

Love et al. (2014) elaborate that, after performing shrinkage, the resulting shrunken log fold change values are suitable to generate a ranking of genes. This ranking is referred to as "ranking by LFC". In this context, genes that show a stronger up- or down-regulation between both phenotypes are allocated towards the top or bottom of the ranked list, independent of the statistical significance of the respective log fold change.
Since the generation of a ranked gene list requires the conduct of differential expression analysis as well as the actual ranking of the genes, the optimal ranking metric is chosen similarly to the optimal parameters of differential expression analysis. This means that the optimal ranking metric maximizes the number of differentially enriched gene sets resulting from the respective FCS tool in its default configuration.

4.6.1 clusterProfiler's Gene Set Enrichment Analysis

As `clusterProfiler`'s GSEA tool requires a ranked gene list as input, this list is obtained by stepwise optimization of the expression data set with the genes in the Entrez gene ID format, the parameters of differential expression analysis as well as the ranking metric with respect to the number of differentially enriched gene sets. In this context, all parameters within the `clusterProfiler` tool remain in their default configuration. In the next steps, stepwise optimization of the parameters within `clusterProfiler` is performed. The set of parameters under consideration is presented in the following Table 4.6.

Table 4.6: Adaptions in clusterProfiler's Gene Set Enrichment Analysis

Parameter	Default	Alternative Options
Gene Set Database	GO(MF)	• KEGG • KEGG Modules
Exponent	weighted ($p = 1$)	• weighted ($p\ 2\ f1{:}5; 2g$) • classic ($p = 0$)
Multiple Test Adjustment	Benjamini and Hochberg	Bonferroni

It is noted that consistency between different runs of GSEA is ensured by setting an

identical seed in each optimization step. The first parameter to optimize is the gene set database. Analogous to `clusterProfiler`'s regular ORA tool, the considered databases are GO with subontology Molecular Function, KEGG and KEGG Modules. The optimal gene set database is then chosen based on the optimal input ranking and the remaining parameters within the tool in their default configuration. In the next step, the optimal exponent value, which corresponds to the value p in Equation (3.2), is chosen based on the previously established optimal parameters. In accordance with the optimization process of the GSEA web application (see Section 4.3.3), the optimal exponent is chosen among the values $p \in \{0; 1; 1.5; 2\}$, where the default exponent is $p = 1$. In the last step of the optimization process, the optimal multiple test adjustment method is chosen between Benjamini and Hochberg (1995) and Bonferroni (1936).

4.6.2 GSEAPreranked

The developers of GSEAPreranked recommend this tool over the regular GSEA web application (see Section 4.3.3) in the case when the expression data at hand is not conform with the regular GSEA scenario. In contrast to the regular GSEA web application tool, which can convert Ensembl gene IDs to the required gene ID format internally, the user is recommended to convert Ensembl gene IDs to HUGO symbols manually prior to running GSEAPreranked. As GSEAPreranked is part of the web application GSEA, the optimization process has to be performed by feeding the stepwisely optimized ranking of genes into the application by hand. Therefore, for practical reasons, the optimization of the ranked list itself, prior to GSEAPreranked, is limited to the optimal differential analysis method, namely `DESeq2` vs. `edgeR`, and the ranking metric as presented in the parent Section 4.6. In particular, the optimal gene expression data resulting from duplicate gene ID removal is automatically chosen as the one in which the duplicate ID that occurs first it kept (option "1"). Moreover, within the optimal differential expression analysis method, all parameters remain in their default configuration.
After generating the optimal ranked lists of genes, the optimization process proceeds to the parameters within GSEAPreranked. The considered parameters can be found in the following Table 4.7. Analogous to the regular GSEA web application, the first parameter to optimize is the gene set database, where the default option is chosen as GO with subontology Molecular Function and alternative options KEGG and Hallmark gene sets.

Table 4.7: Adaptions in GSEAPreranked

Parameter	Default	Alternative Options
Gene Set Database	GO(MF)	• KEGG • Hallmark
Exponent	weighted ($p = 1$)	• weighted ($p \in \{1.5; 2\}$) • classic ($p = 0$)

Based on the optimal gene set database, the optimal exponent is chosen between values $f0; 1; 1:5; 2g$ in the next step with a default value of $p = 1$. It is noted that the conduct of multiple test adjustment is not necessary in this case since differentially enriched gene sets are indicated by a false discovery rate smaller than $0:25$ (FDR $< 0:25$).

5. Results

In the following, the results of the stepwise optimization process of all GSA tools under investigation are presented. This includes for each of the 5 random phenotype permutations the optimal parameter choice in each step, the resulting number of differentially enriched gene sets and finally, the maximum number of differentially enriched gene sets. As described in Chapter 4, a stronger increase in the number of differentially enriched gene sets induced by optimizing the parameter choice indicates a higher potential of the tool for over-optimistic results. It is noted that the parameter choice in a given step of the optimization process is called optimal if none of the alternative parameter options led to an increase in the number of differentially enriched gene sets. In this regard, an optimal parameter choice might have been the default option even though one or more of the alternatives led to the same number of differentially enriched gene sets.
The complete optimization process for each of the tools under investigation is illustrated in a step diagram to visualize the increase in the number of differentially enriched gene sets obtained by utilizing the parameter adaptions listed and described in Chapter 4. In each step diagram, a dashed line visualizes the effect of the optimal parameter choice on the number of differentially enriched gene sets. In this context, the labels above the dashed line indicate the optimal parameter choice of the corresponding optimization step. Furthermore, additional labels below the dashed line in the step diagram are provided for those GSA tools that generally report a high number of differentially enriched gene sets. These labels serve as an additional measure of visualization and indicate the number of differentially enriched gene sets resulting from the respective optimization step.
It is noted that a complete evaluation of the results would moreover include the assessment of the biological meaningfulness behind the set of gene sets detected as differentially enriched, as a user of GSA tools would only report those results which serve his or her research question. However, a contextual evaluation of the set of differentially enriched gene sets for each phenotype permutation and each tool lies outside of the scope of this thesis and is therefore consigned to the biologically skilled reader. The final set of differentially enriched gene sets for each phenotype permutation and tool is displayed in Appendix A.3.
In the following, the results of the optimization processes of the ORA tools, including the optimization of the input lists of differentially expressed genes, are presented in Section 5.1 and the respective results of the FCS tools are provided in Section 5.2. Furthermore,

a comparison of the results of all tools is given in Section 5.3.

5.1 Over-Representation Analysis

All ORA tools investigated in this thesis require a list of differentially expressed genes as input. A slight variation of this list is required by `GOSeq` whose input is a list of all genes in the experiment labelled according to the presence of differential expression across the phenotypes. As described in Section 4.4, the optimization process of this input list was conducted in two general steps for each tool. In the first step, which is presented in Section 5.1.1, the number of differentially expressed genes was maximized on the single-gene level and the input vector of differentially expressed genes was generated accordingly. This resulted in a default vector and an optimized vector of differentially expressed genes.
In the second step, which is presented in Sections 5.1.2- 5.1.5, the results of the optimization processes of the respective GSA tools are presented. This includes, inter alia, the optimal choice of the input list of differentially expressed genes from step 1 under the component "Optimal Unranked Input List". While the optimal parameter choice "Default" indicates that the input list generated with `DESeq2` in its default configuration led to the highest number of differentially enriched gene sets, the choice "Optimized" signals that the input list with the optimal number of differentially expressed genes resulted in the highest number of differentially enriched gene sets.

5.1.1 Di□erential Expression Analysis

As described in Section 4.4, the optimized input list of differentially expressed genes was obtained by maximizing the number of differentially expressed genes detected by applying the differential expression techniques `DESeq2` or `edgeR` and the criterion of an adjusted p-value $< 0{:}05$. As `GOSeq` accepts genes in the Ensembl gene ID format whereas the remaining ORA tools require gene ID conversion to Entrez gene ID, the optimization process was performed twice, resulting in two optimized unranked lists of differentially expressed genes. In this context, in order to obtain a list of differentially expressed genes in the Entrez gene ID format, the choice of the optimal expression data set resulting from duplicate gene ID removal (see Section 4.2.4) was included in the optimization process. For the expression data set in the Ensembl gene ID format, on the other hand, this step was omitted from the optimization process and the choice was set to default option "1" automatically. As described above, the optimal choice between the default input list of differentially expressed genes and the corresponding optimized list was made in the optimization process of the respective ORA tool (see Sections 5.1.2-5.1.5).
From the illustration of the optimization process of the input lists of differentially expressed genes presented in Figure 5.1, it is visible that for genes in the Entrez gene ID format,

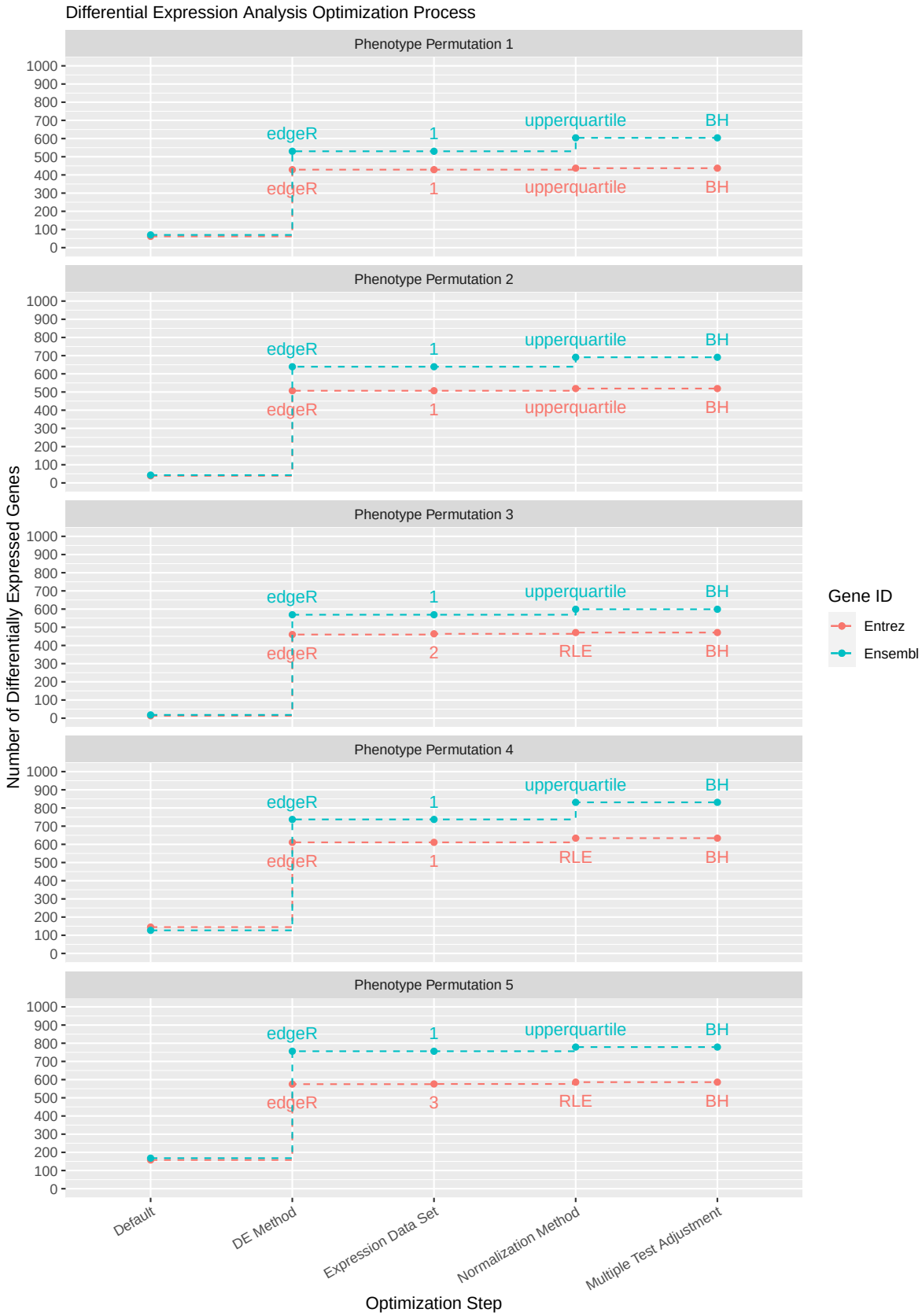

Figure 5.1: Differential Expression Analysis Optimization Process

the number of differentially expressed genes was lower compared to the Ensembl gene ID format across all phenotype permutations and optimization steps. This is a consequence of the gene ID conversion process since not all Ensembl gene IDs could be converted to a corresponding Entrez gene ID.

In the first step of the optimization process, `edgeR` was chosen as the optimal differential expression technique on the single gene-level in all of the 5 random phenotype permutations and for both gene ID formats. This means that, with all parameters within both techniques in their default configuration, `edgeR` led to a higher number of differentially expressed genes compared to `DESeq2` and the difference in the number of differentially expressed genes was considerably high.

In the optimization process of the input list of differentially expressed genes in the Entrez gene ID format, the optimal choice of the duplicate gene ID removal technique depended on the phenotype permutation. Whereas for permutations 1,2 and 4, the default removal technique led to the highest number of differentially expressed genes, taking the mean count value of the respective duplicate rows led to the optimal result in permutation 3. Moreover, keeping the row with the highest overall counts across all samples was chosen as the optimal removal technique for permutation 5. Nevertheless, the increase in the number of differentially expressed genes was minor compared to the effect of the optimal differential expression technique. Consequently, the choice of duplicate gene ID removal technique could be utilized to induce a slight increase in the number of differentially expressed genes.

Since for all 5 phenotype permutations and both gene ID formats, `edgeR` was chosen as the optimal differential expression technique, the only two parameters to consider in the subsequent optimization steps were the normalization technique as well as the method for multiple test adjustment. As displayed in Figure 4.1, in addition to the default option TMM, other normalization techniques utilized in this thesis were TMMwsp, RLE and Upperquartile. While for the expression data in the Ensembl gene ID format, the undisputed normalization technique with respect to the number of differentially expressed genes was the Upperquartile method, the choice for the data in the Entrez gene ID format was split between Upperquartile (for phenotype permutations 1 and 2) and RLE (for phenotype permutations 3, 4 and 5). From Figure 5.1 it becomes apparent that the absolute increase in the number of differentially expressed genes was higher for the gene expression data set in the Ensembl gene ID format. This observation can be substantiated by the fact that the expression data in the Ensembl gene ID contained more genes.

Lastly, the clear optimal choice of the multiple test adjustment technique was Benjamini and Hochberg as it led to the highest number of differentially expressed genes across all gene IDs and random phenotype permutations. Eventually, the number of differentially expressed genes in the 5 phenotype permutations in the Ensembl gene ID format was 604, 691, 599, 831 and 779 respectively. For the genes in the Entrez gene ID format, on the

other hand, the number of differentially expressed genes was 437, 519, 471, 634 and 586 respectively.
To sum up, the utilization of the possible parameter choices in the generation of the input lists led to a considerable increase in the number of differentially expressed genes, particularly caused by the choice of the differential expression technique. A further striking observation in the optimized input lists of differentially expressed genes is that there was a high overlap of the genes across the 5 phenotype permutations despite the fact that the permutations were generated randomly. This is investigated in further detail in Section A.2.

5.1.2 DAVID (Web)

The analysis of the DAVID web application was performed in the "DAVID 2021" version and a gene set was detected as differentially enriched if the respective adjusted p-value was smaller than 0.05. The optimization process is visualized in Figure 5.2 and furthermore, the optimal result tables for each phenotype permutation is provided in Table A.2.
The number of differentially enriched gene sets resulting from the default configuration, including the input lists of differentially expressed genes, amounted to 0 in the phenotype permutations 1-4 and 3 in permutation 5. Eventually, the stepwise optimization process led to an increase in the number of differentially enriched gene sets in all random phenotype permutations except for permutation 1 where the optimal parameter choice was identical to the default parameter configuration and the number of differentially enriched gene sets remained at 0. In the remaining random phenotype permutations, utilizing the optimized input list of differentially expressed genes led to an increase of 1 (in permutations 3 and 5) and 2 (in permutations 2 and 4) differentially enriched gene sets. Furthermore, switching to gene set database KEGG led to a further increase in permutations 2-4, which ranged from 1 to 4 differentially enriched gene sets.
In contrast to the gene set database, an adaption of the universe, i.e. the set of background genes, could not be utilized to surpass the default universe provided by the web application with respect to the number of differentially enriched gene sets. It is noted that the alternative universe resulting from the default input list in permutation 1 differed from the ones utilized in the remaining phenotype permutations. The reason behind this is that the default input list, which was also the optimal list in permutation 1, was generated with `DESeq2` in its default setting. In this setting, a number of adjusted p-values were set to "NA" due to independent filtering (see Section 4.4.1). The respective genes were excluded from the universe since they could neither be classified as "differentially expressed" nor as "not differentially expressed". Eventually, multiple test adjustment using Benjamini and Hochberg led to the highest number of differentially enriched gene sets in all of the 5 phenotype permutations. This result was to be expected since Bonferroni is known to be more conservative.

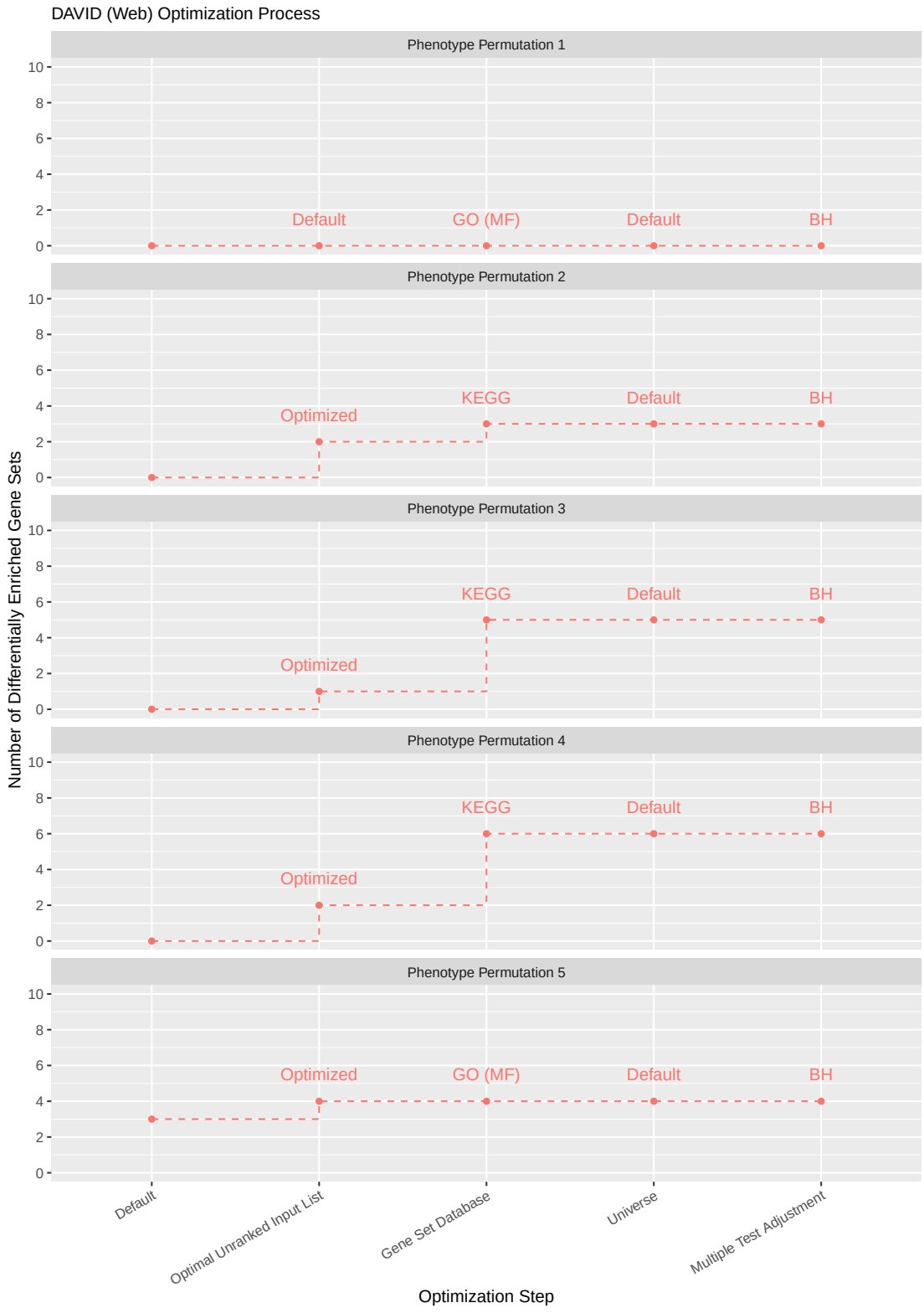

Figure 5.2: DAVID (Web) Optimization Process

To sum up, utilizing the optimized input list of differentially expressed genes led to an increase in the number of differentially enriched gene sets in the majority of the phenotype permutations which indicates that this choice can be utilized to generate over-optimistic results. Moreover, the choice of the gene set database could be used within the tool to obtain a higher number of differentially enriched gene sets, however, not in all cases. It is noted that, in contrast to the other GSA tools under investigation, the DAVID web application offers the option to utilize multiple gene set databases, such as GO and KEGG, simultaneously which would naturally lead to an even higher number of differentially enriched gene sets than in the setting of this thesis.
Analogous to the high overlap of the optimized unranked lists of differentially expressed genes addressed in Section 5.1.1, a multitude of gene sets were detected as differentially enriched across multiple, if not all, of the 3 phenotype permutations in which the chosen input list was generated with `edgeR` and the optimal gene set database was KEGG. (see Table A.2). This observation is investigated in further detail in Section A.2.

5.1.3 GOSeq

The analysis of `GOSeq` was performed under version 1.44.0 and owing to the circumstances that `GOSeq` does not perform multiple test adjustment internally, the adjusted p-values were obtained by using the `R` package `base`. Consequently, the criterion to detect a gene set as differentially enriched was an adjusted p-value < 0.05. The optimization process is visualized in Figure 5.3 and the resulting optimal result tables for each phenotype permutation are presented in Table A.3.
With all parameters in their default configuration, including the input list of differentially expressed genes, the number of differentially enriched gene sets ranged from 0 to 15 across the random phenotype permutations, where the minimum was obtained in permutations 1,2 and 3 and the maximum in permutation 4. In the first step of the optimization process, setting the optimized unranked list of differentially expressed genes as input led to an increase in the number of differentially enriched gene sets for all random phenotype permutations apart from permutation 4. In the remaining permutations 1, 2, 3, and 5 the increase ranged from 4 to 13 differentially enriched gene sets. Analogous to this scheme, accounting for length bias could only be surpassed with respect to the number of differentially enriched gene sets by accounting for all biases in phenotype permutation 4, where the increase amounted to 4 differentially enriched gene sets. This means that in alignment with the possible downside of accounting for all biases described in Section 4.5.3, namely the risk of removing the bias resulting from actual differential expression of genes, accounting for all biases could not induce an increase in the number of differentially enriched gene sets for the majority of the phenotype permutations. In the next optimization step, the optimal gene set database in each phenotype permutation was GO with subontology Molecular Function. Consequently, this parameter did not offer any flexibility with

Figure 5.3: GOSeq Optimization Process

respect to the number of differentially enriched gene sets.
However, setting the universe, i.e. the set of background genes, to all genes measured in the experiment as opposed to all genes that can be annotated to at least one gene set, could yield a notable increase in the number of differentially enriched gene sets. For instance, an increase stronger than twofold was achieved in permutations 3 and 5, namely from 4 to 11 and from 6 to 16 differentially enriched gene sets, respectively. This effect of the adaption of the universe differed from the other ORA tools which can be justified by the fact that the adaption of the universe itself (see Section 4.5.3) and the calculation of the p-value differ. The increase in the number of differentially enriched gene sets can be substantiated by taking a closer look at the calculation of the p-value for a given gene set described in Section 3.2. In the resampling process undergone `1000` times, a random set of differentially expressed genes of the same size as the input list is generated from the universe with each iteration. By additionally including the entirety of genes that do not belong to any gene set, the probability decreases that a gene declared as differentially expressed in a random set is indeed annotated to the given gene set. Consequently, the fraction of genes in the random set of differentially expressed genes that are annotated to the given gene set declines. Thus, the (offset) fraction of resampled sets that contain at least as many genes annotated to the given gene set as the actual input list of differentially expressed genes decreases accordingly. Since in `GOSeq`, this fraction directly corresponds to the p-value (see Equation (3.1)), the p-value of over-representation decreases for the given gene set.
The next parameter, namely the method to calculate the p-value of over-representation, was chosen as Wallenius approximation in all 5 phenotype permutations. This means that using this approximation led to at least as many differentially enriched gene sets, if not more, as calculating the p-value with the exact random sampling technique and the default of `1000` random iterations. Consequently, in the context of this thesis, the approximate calculation method was less conservative than the exact calculation method and therefore, the latter could not be utilized to induce an increase in the number of differentially enriched gene sets. Finally, as expected, Benjamini and Hochberg was the optimal multiple test adjustment method in all phenotype permutations.
To sum up, the set of parameters in the optimization process prior to and within a run of `GOSeq` could be utilized to induce an increase in the number of differentially enriched gene sets in all 5 random phenotype permutations, which ranged from 7 in phenotype permutation 1 to 25 in phenotype permutation 2. Particularly, there was one adaption that posed a definite potential for over-optimistic results, namely the inclusion of all genes in the calculation of the p-value, independent of gene set membership. Furthermore, optimizing the unranked input list of differentially expressed genes led to over-optimistic results in the broad majority of phenotype permutations which consequently indicates a potential for over-optimistic results as well.

5.1.4 clusterProfiler ORA

In this section, the optimization process of the number of differentially enriched gene sets generated with `clusterProfiler`'s regular ORA tool is interpreted. The version of `clusterProfiler` used in this thesis was version 4.0.5 and the criterion to detect a gene set as differentially enriched was firstly an adjusted p-value < 0.05 and secondly an FDR q-value < 0.2. The optimization process is visualized in Figure 5.4 and the resulting sets of differentially enriched gene sets for each phenotype permutation are provided in Table A.4.

From the illustration of the optimization process of `clusterProfiler` ORA, it is visible that in random phenotype permutations 1 and 3, 0 differentially enriched gene sets resulted from the set of parameters in their default configuration, including the input list of differentially expressed genes, while for the remaining permutations, the number of differentially enriched gene sets ranged from 1 to 18 differentially enriched gene sets. It is also observable that the only possibility to generate over-optimistic results lied outside the actual GSA tool and in the generation of the unranked input list of differentially expressed genes. For all phenotype permutations except for permutation 4, an increase could be achieved by optimizing the differential expression technique and its parameters on the single-gene level. The lowest non-zero increase was induced in phenotype permutation 3 and amounted to 6 differentially enriched gene sets, while the highest increase of 17 differentially enriched gene sets was achieved in phenotype permutation 2. To conclude, in 4 out of 5 random phenotype permutations, over-optimistic results could be induced solely by increasing the number of differentially expressed genes in the input vector. This increase could potentially be further intensified by considering each combination of parameters in the differential expression technique instead of the stepwise comparison. Consequently, `clusterProfiler`'s regular ORA tool did not offer any internal parameter optimization to induce over-optimistic results.

Striking about the final result tables, which are presented in Table A.4, is that the overlap of the sets of differentially enriched gene sets across the 4 phenotype permutations whose corresponding optimal results were achieved with the optimized input list was unexpectedly high. For example, gene sets such as "Receptor Ligand Activity" and "Signaling Receptor Activator Activity" appeared for the random phenotype permutations 1-3 and 5. This finding is further investigated in Section A.2.

5.1.5 clusterProfiler DAVID

Analogous to the other GSA tools provided by `clusterProfiler` investigated in this thesis, the analysis of DAVID was performed under version 4.0.5 of the package and similar to the regular ORA tool, those gene sets with an adjusted p-value < 0.05 as well as an FDR q-value < 0.2 were reported as differentially enriched. The optimization process is illustrated in Figure 5.5 and the resulting sets of differentially enriched gene sets for all

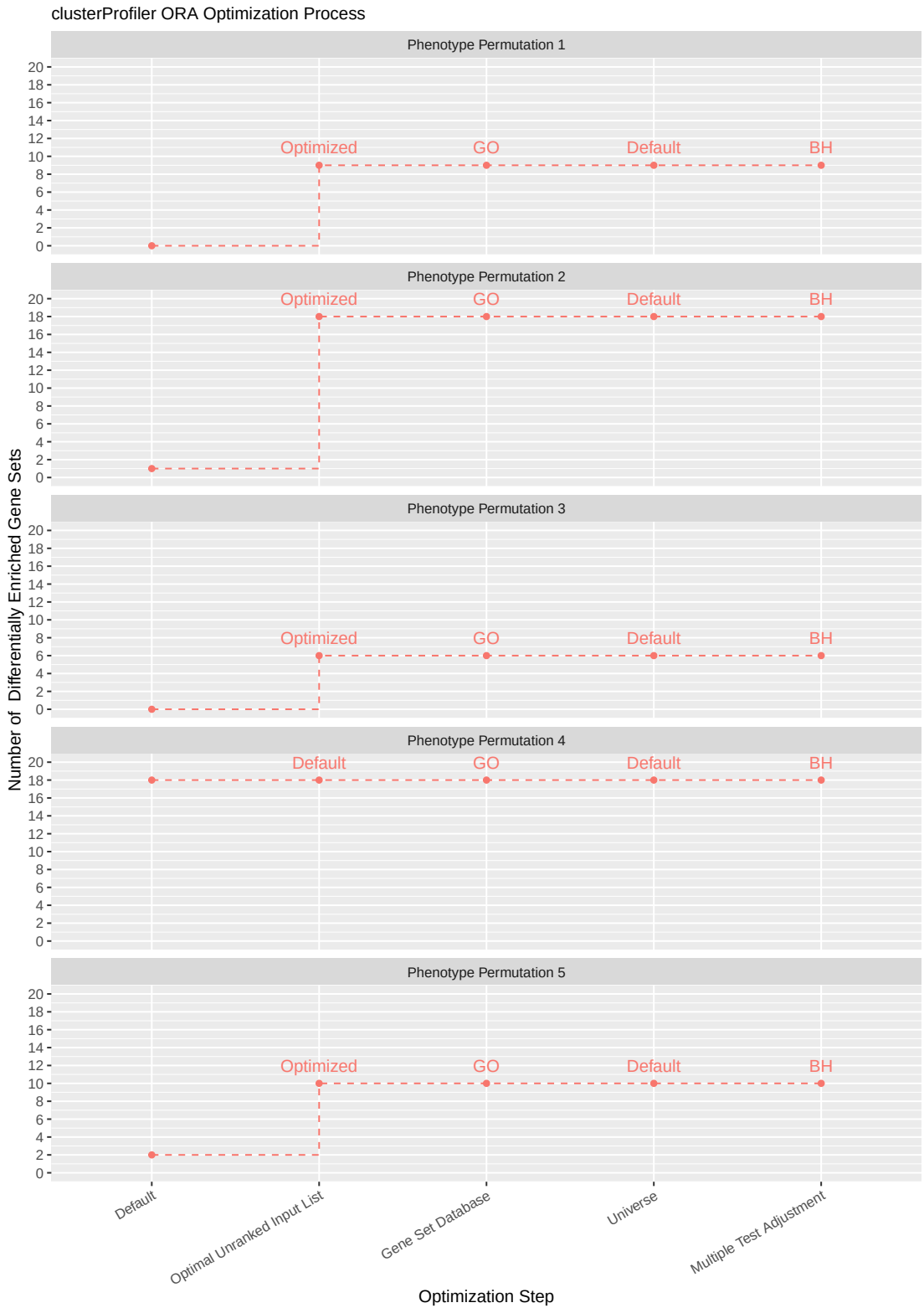

Figure 5.4: clusterProfiler ORA Optimization Process

random phenotype permutations are presented in Table A.5.
With all parameters in their default configuration, including the input list of differentially expressed genes, 0 differentially enriched gene sets resulted in all random phenotype permutations apart from permutation 5. In general, the number ranged from 0 to 2 and was therefore on a similar level compared to the web application of DAVID. Furthermore, identically to the web application of DAVID, the number of differentially enriched gene sets could be increased by using the optimized input list of differentially expressed genes in all phenotype permutations apart from permutation 1. For permutation 1, on the other hand, all of the optimal parameters from the optimization process coincided with the default configuration, meaning that the number of differentially enriched gene sets could not be increased from the value of 0. For the remaining permutations, KEGG was chosen as the optimal gene set database in 3 out of 4 phenotype permutations with a maximum increase of 5 differentially enriched gene sets, while for phenotype permutation 5, GO with subontology Molecular Function could not be surpassed with respect to the number of differentially enriched gene sets. A clear picture presents itself regarding the optimal choice of the universe, namely that the default universe consisting of all genes annotated to the given gene set database was chosen across all phenotype permutations. Finally, the same can be said about the optimal multiple test adjustment method which was Benjamini and Hochberg across all phenotype permutations.
To sum up, over-optimistic results could be achieved in 4 out of the 5 random phenotype permutations for which the choice of the input list offered flexibility with respect to the number of differentially enriched gene sets, while the choice of the gene set database could be utilized in some permutations to induce a further increase.
It is noted that the set of differentially enriched gene sets in the optimal results (see Table A.5) was identical to the one generated with the web application of DAVID across all phenotype permutations (see Table A.2), yet p-values differed slightly.

5.2 Functional Class Scoring

In the following Subsections 5.2.1-5.2.4, the results of the optimization process for each of the FCS tools under investigation are presented. First, the results of those FCS tools are presented that internally generate a ranking of the genes based on the respective magnitude of differential expression. This is followed by the results of the FCS tools that require such ranking as input.

5.2.1 GSEA Web Application

All runs of the GSEA web application were performed under version 4.2.1 and the tool's criterion to detect a gene set as differentially enriched was an FDR q-value < 0.25. The optimization process is displayed in Figure 5.6 and the resulting sets of differentially

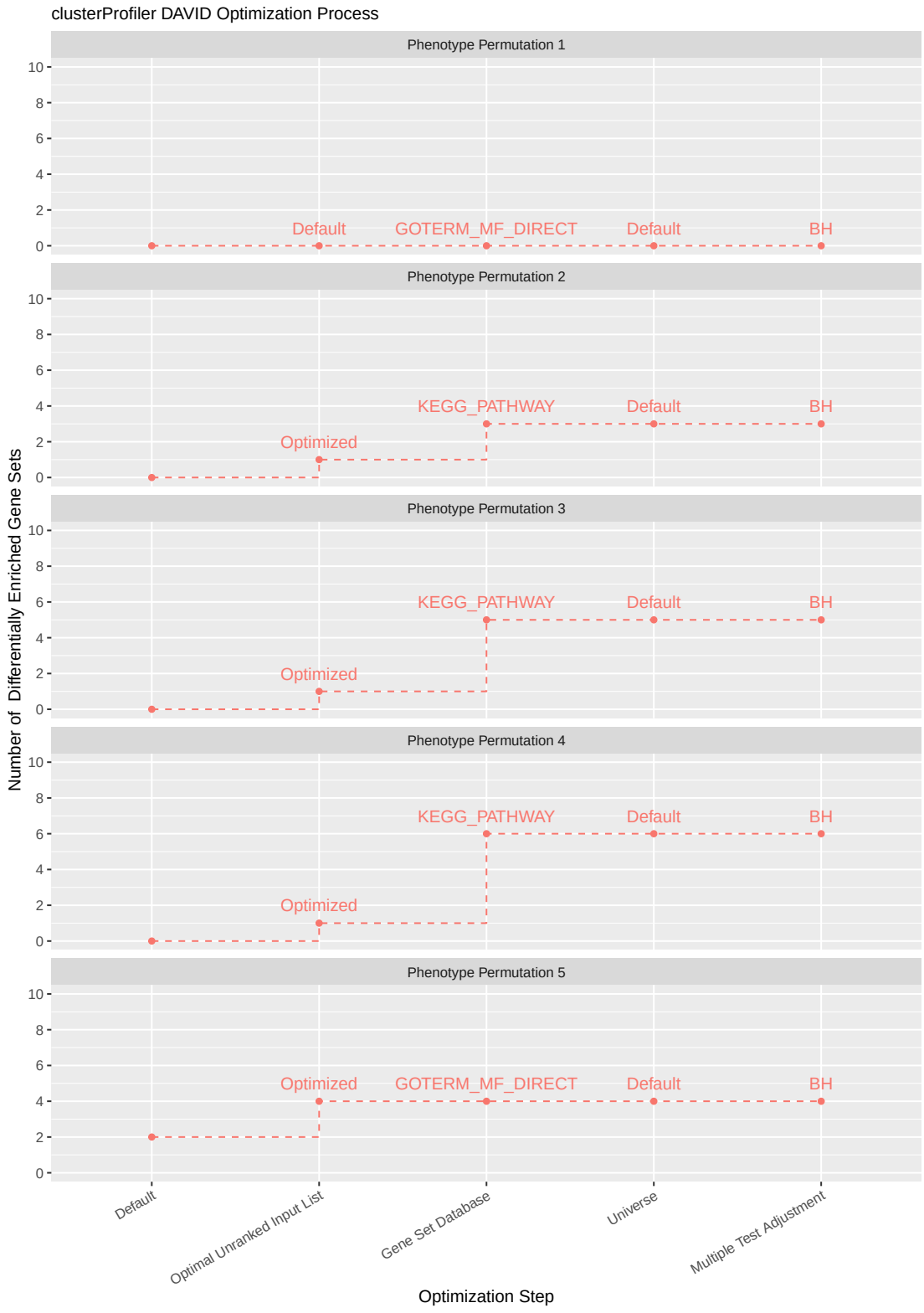

Figure 5.5: clusterProfiler DAVID Optimization Process

enriched gene sets for all random phenotype permutations are provided in Table A.6.
With the default parameter configuration, 0 differentially enriched gene sets resulted for each of the 5 random phenotype permutations. Furthermore, `voom` was determined as the "optimal" transformation method for all random phenotype permutations except for permutation 4, meaning that the number of differentially enriched gene sets did not increase for those. For permutation 4, on the other hand, the choice of the transformation method `varianceStabilizingTransformation` led to an increase to 11 differentially enriched gene sets which was the strongest increase across all optimization steps and phenotype permutations for this tool. Consequently, a clear statement on the effect of the transformation method on the number of differentially enriched gene sets was not possible as for the majority, an increase could not be achieved by the choice of transformation method, whereas for one phenotype permutation, the increase was notable.
Concerning the optimal gene set database, even though GO with subontology Molecular Function generally provides a significantly higher number of gene sets compared to KEGG and Hallmark (see Section 4.3.3), it was not chosen as the optimal option in all 5 random phenotype permutations. On the contrary, by choosing KEGG as the gene set database, the number of differentially enriched gene sets could be increased from 0 to 3 and 0 to 1 in permutations 1 and 3 respectively. Hallmark with its 50 gene sets, however, was never chosen as the optimal gene set database. In contrast to the gene set database, the choice of the optimal gene-level ranking metric was undisputed and laid in the Signal2Noise Ratio, i.e. none of the other metrics led to an increased number of differentially enriched gene sets. Finally, a clear statement on an optimal exponent value cannot be made for the GSEA web application since the optimal choice differed strongly between the phenotype permutations. While the default exponent $p = 1$ was chosen three times, meaning that the number of differentially enriched gene sets could not be increased any further, exponent $p = 0$ was chosen for random phenotype permutation 2 and $p = 1{:}5$ for permutation 3 to induce over-optimistic results. In the latter two permutations, the increase amounted to 5 and 2 differentially enriched gene sets, respectively.
Eventually, in all permutations apart from random phenotype permutation 5, the set of parameters could be utilized to gain a non-zero number of differentially enriched gene sets and therefore an over-optimistic result. However, there was no clear rule as to which adaption definitely led to an increased number of differentially enriched gene sets in all phenotype permutations, since the optimal choice mostly depended on the random phenotype permutation. Moreover, none of the parameter choices led to an equally high increase across the phenotype permutations.

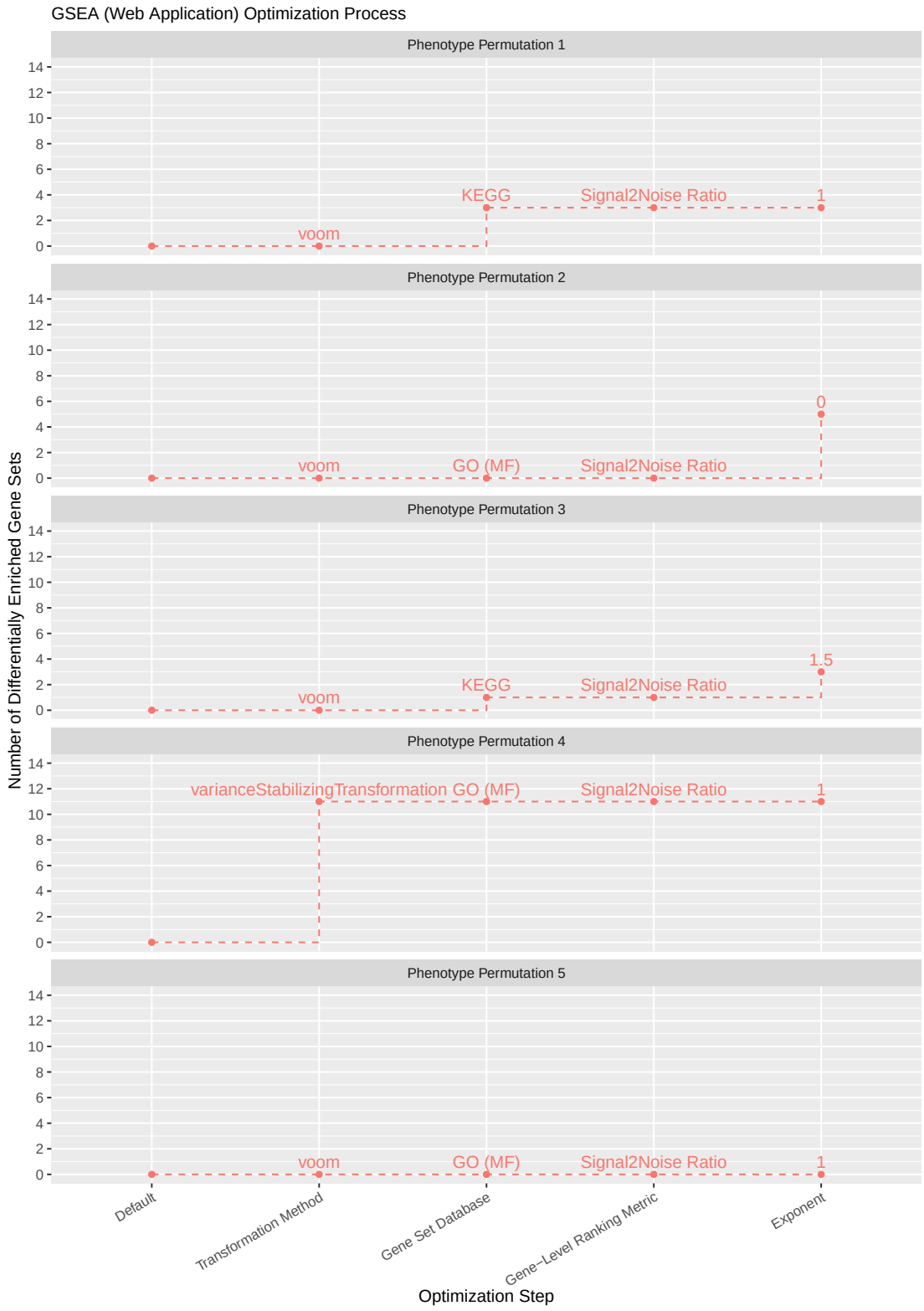

Figure 5.6: GSEA (Web Application) Optimization Process

5.2.2 PADOG

The analysis of `PADOG` was performed under version 1.34.0 and since the tool does not perform multiple test adjustment internally, adjusted p-values were generated using the `R` package `base` and the criterion to detect a gene set as differentially enriched was chosen as an adjusted p-value < 0.05. The optimization process is illustrated in Figure 5.7.

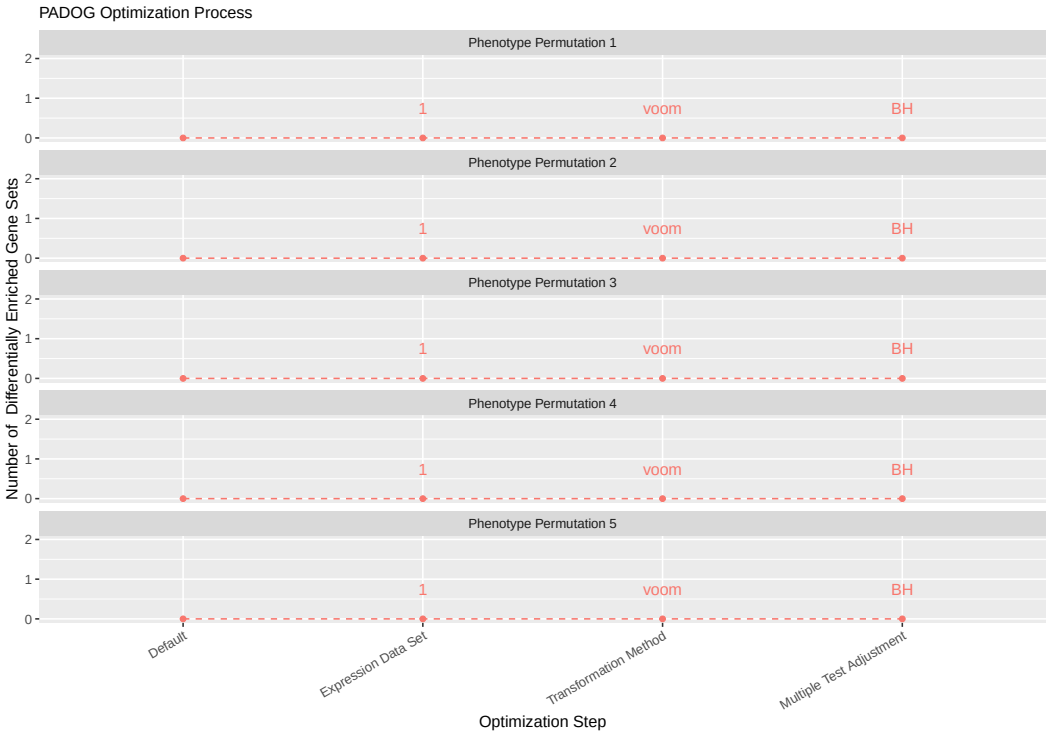

Figure 5.7: PADOG Optimization Process

From Figure 5.7 it is visible that 0 differentially enriched gene sets resulted from the default configuration and that furthermore, this number could not be increased by utilizing any of the parameter adaptions. Consequently, the default results coincided with the optimal results in all random phenotype permutations which in this analysis indicates that there is no potential for over-optimistic results in `PADOG`. This particularly means that the choice of the optimal gene expression data set resulting from the removal of duplicated gene IDs and the RNA-Seq transformation method was immaterial with respect to the resulting number of differentially enriched gene sets in this optimization process since none could induce an increase in the number of differentially enriched gene sets to a non-zero level. Nevertheless, it cannot be ruled out that `varianceStabilizingTransformation` would have induced a higher, non-zero number of differentially enriched gene sets in combination with the gene set database KEGG but due to the optimization process being conducted in a stepwise manner, this comparison was not considered. It is further noted that the choice of the utilized parameters and their set of options are not exclusive. For example,

there might be more possible transformation methods to apply to the RNA-Seq data set. Thus, it cannot be ruled out that over-optimistic results could have been induced through different parameter choices or an alternative optimization process.

5.2.3 GSEAPreranked

Analogous to the regular GSEA web application (see Section 5.2.1), the analysis of GSEA-Preranked was performed under version 4.2.1 and the criterion to detect a differentially enriched gene set was a False Discovery Rate < 0.25 (FDR q-value < 0.25). The optimization process is illustrated in Figure 5.8 and the resulting 10 differentially enriched gene sets with the lowest FDR q-value for each random phenotype permutation are presented in Table A.7.

In all of the random phenotype permutations apart from permutation 3, the number of differentially enriched gene sets was on a higher level compared to those FCS tools which generate a ranked list internally. To be more precise, the number of differentially enriched gene sets resulting from the parameters in their default configuration ranged from 79 to 225, whereas for phenotype permutation 3, the number amounted to 0 differentially enriched gene sets. In the first optimization step, which contained the optimization of the differential expression technique to generate the ranked list of genes, in all phenotype permutations except for permutation 4, the number of differentially enriched gene sets was further increased by generating the ranked list with `edgeR` instead of `DESeq2`. In this regard, the magnitude of the increase ranged between 6 in phenotype permutation 3 and 54 in phenotype permutation 1. In phenotype permutation 4, on the other hand, `DESeq2` could not be surpassed with respect to the number of differentially enriched gene sets. It is noted that it cannot be ruled out that an even stronger increase would have resulted from an additional optimization of the parameters within the differential expression techniques. In all of the 5 phenotype permutations, the ranking generated by p-value, as presented in Equation (4.7), led to the highest number of differentially enriched gene sets with a notable difference compared to the number of differentially enriched gene sets resulting from a ranking by LFC (see Equation (4.8)). In phenotype permutations 1-5, the number of differentially enriched gene sets resulting from a ranking by LFC in this step was 69, 99, 6, 2, 27 respectively.

In contrast to that, the choice of the optimal gene set database was not as clear. Whereas for phenotype permutations 1,2,4 and 5, GO with subontology Molecular Function could not be surpassed with regard to the number of differentially enriched gene sets, it could be increased by 9 with the choice of KEGG as the gene set database in phenotype permutation 3. The fact that KEGG was only chosen as the optimal gene set database in this permutation can be substantiated by the observation that for the majority of the remaining permutations, the magnitude of the number of differentially enriched gene sets was comparable to or even higher than the total number of gene sets offered by KEGG,

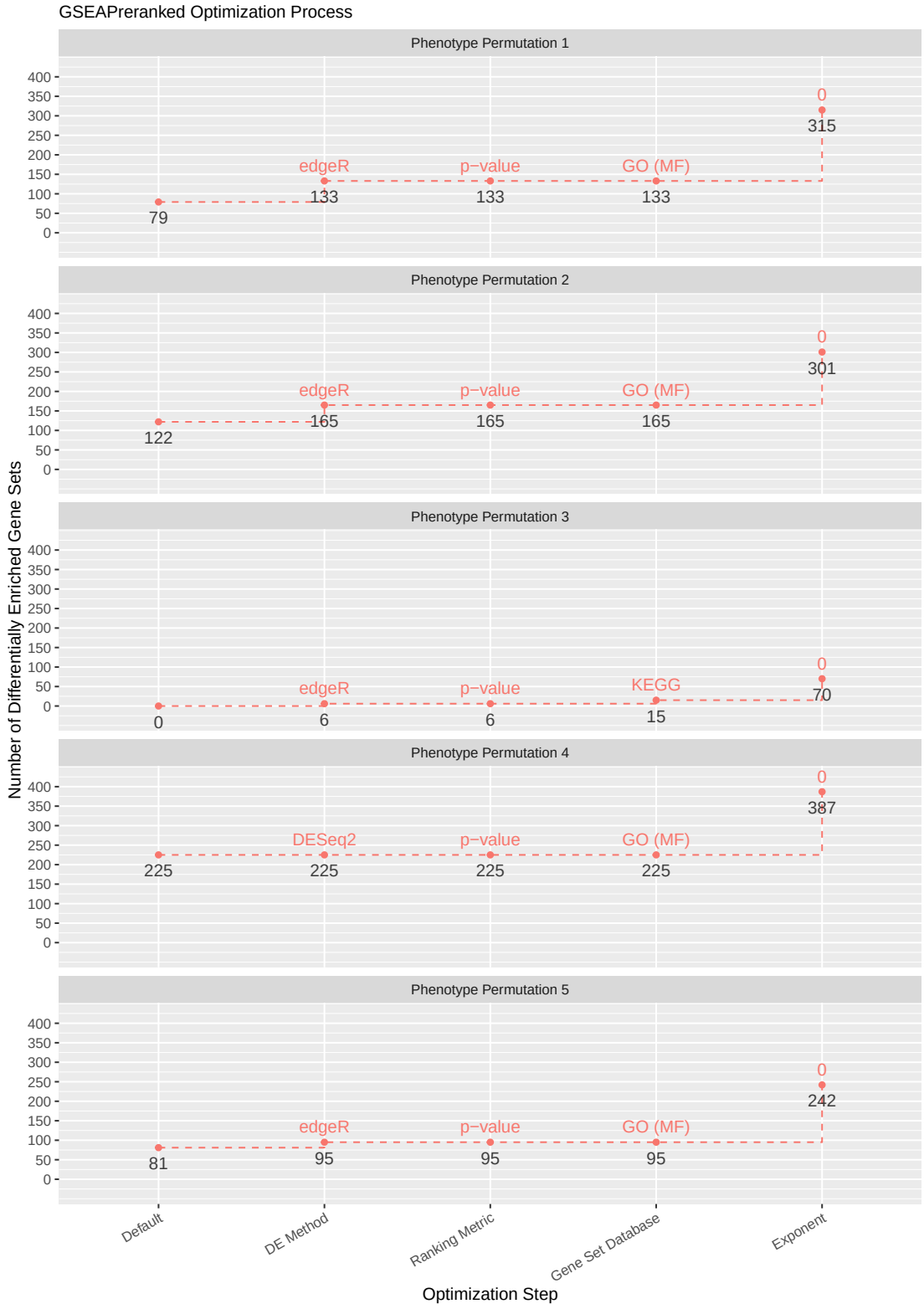

Figure 5.8: GSEAPreranked Optimization Process

Visualization of the optimization process of GSEAPreranked. The labels below the dashed line indicate the number of differentially enriched gene sets resulting from the optimal parameter choice.

namely 178.
Eventually, the choice of exponent $p = 0$ resulted in a notable increase in the number of differentially enriched gene sets in all of the 5 random phenotype permutations, where the minimal increase of 55 was obtained in permutation 3 and the maximum of 182 differentially enriched gene sets in permutation 1. Especially in phenotype permutation 3, the increase was stronger than fourfold.
To sum up, the magnitude of the number of differentially enriched gene sets detected with GSEAPreranked was much higher compared to the number resulting from the regular GSEA web application. Furthermore, the choice of the differential expression technique could be utilized to induce a further increase in the majority of the random phenotype permutations. It is noted that, as already mentioned above, within `DESeq2` and `edgeR`, the parameters were retained in their default configuration and therefore, the values of the increase present a lower bound which could have potentially been further lifted by optimizing the respective parameters within the differential expression techniques. However, the strongest increase in the number of differentially enriched gene sets was achieved by an exponent choice of $p = 0$, which could therefore be used to induce over-optimistic results in each of the random phenotype permutations. This stands in contrast to the regular GSEA web application, in which the choice of the exponent was not as clear since it did not influence the number of differentially enriched gene sets to the same extent.

5.2.4 clusterProfiler GSEA

Analogous to regular ORA and DAVID performed by `clusterProfiler`, all runs of GSEA were performed under version 4.0.5. However, in contrast to these two tools, the only criterion to detect a gene set as differentially enriched was an adjusted p-value < 0.05. The complete optimization process is visualized in Figure 5.9 and for each random phenotype permutation, the resulting 10 differentially enriched gene sets with the lowest adjusted p-values are provided in Table A.8.
With all parameters in their default configuration, the magnitude of the number of differentially enriched gene sets was lower compared to GSEAPreranked but considerably higher than those resulting from the ORA tools, the GSEA web application and `PADOG`. To be more precise, the number of differentially enriched gene sets in `clusterProfiler`'s GSEA ranged from 5 in phenotype permutation 3 and 117 in phenotype permutation 4.
In the first step of the optimization process, it is visible that the choice of the duplicate gene ID removal technique could lead to an increase in the number of differentially enriched gene sets, however, the effect differed greatly between the random phenotype permutations. Whereas in permutation 4, the default option could not be surpassed with respect to the number of differentially enriched gene sets, removing duplicate genes by keeping only the row with the highest overall counts led to an increase of 18 differentially enriched gene sets in phenotype permutation 2. In particular, the increase in the number

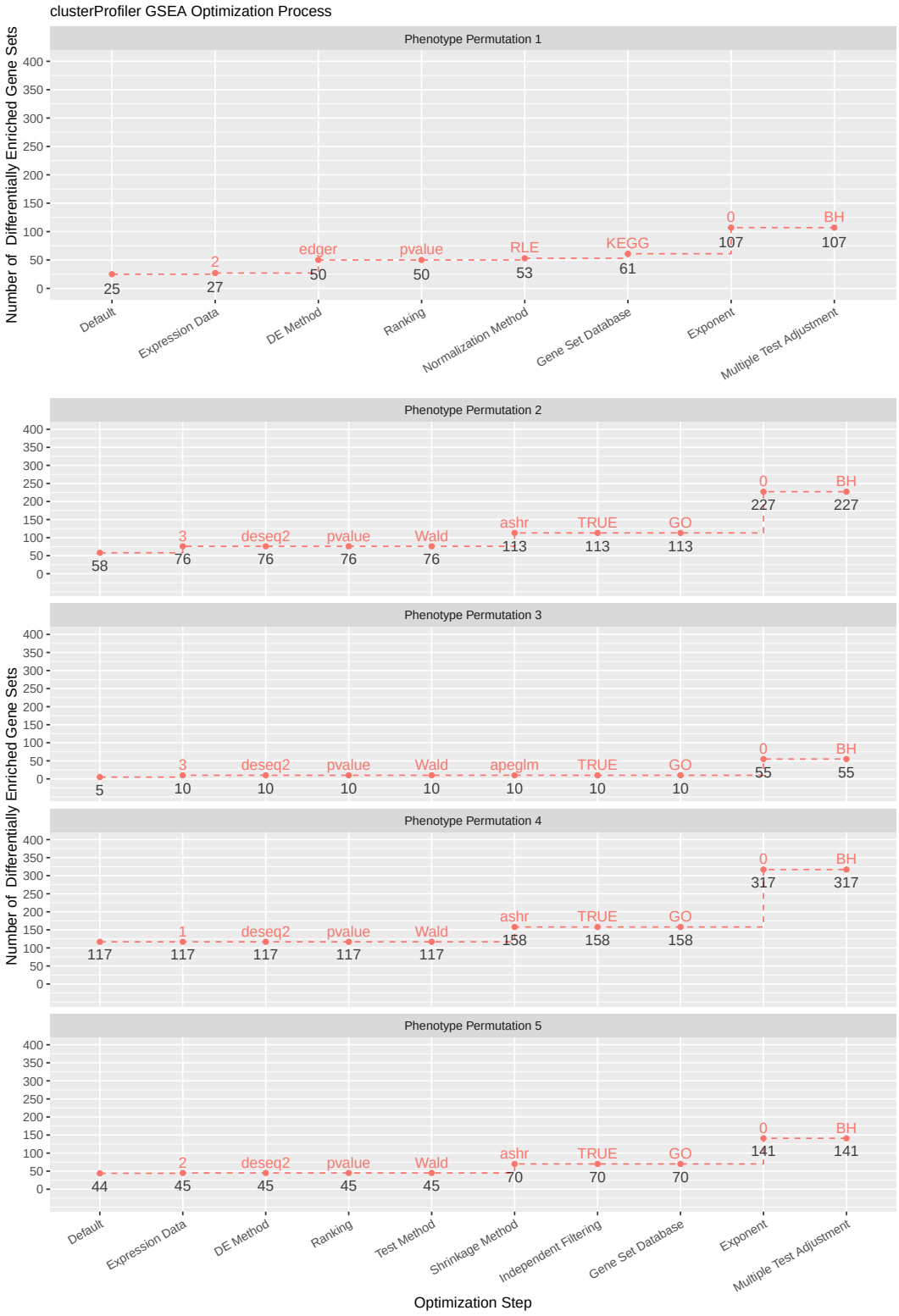

Figure 5.9: clusterProfiler GSEA Optimization Process

Visualization of the optimization process of clusterProfiler GSEA. The labels below the dashed line indicate the number of differentially enriched gene sets resulting from the optimal parameter choice.

of differentially enriched gene sets did not depend on the magnitude of counts prior to this optimization step.
Furthermore, there was no clear choice of the optimal differential expression technique, although in the majority of phenotype permutations, namely 2-5, the default choice of `DESeq2` could not be surpassed with regard to the number of differentially enriched gene sets. This means that the number of differentially enriched gene sets could be increased by the choice of `edgeR` only in phenotype permutation 1. The next optimization steps differed between phenotype permutation 1 and the remaining 4 permutations. For phenotype permutation 1, choosing the normalization technique RLE resulted in a further increase in the number of differentially enriched gene sets while for the remaining permutations, only the shrinkage method could be utilized within `DESeq2` to raise the number of differentially enriched gene sets. In this context, in 3 out of 4 phenotype permutations with `DESeq2` as the optimal differential expression technique, method `ashr` resulted in the highest number of differentially enriched gene sets among the shrinkage methods with an increase between 25 differentially enriched gene sets in permutation 5 and 41 in permutation 4. In phenotype permutation 3, on the other hand, the default shrinkage method `apeglm` could not be surpassed with respect to the number of differentially enriched gene sets.
In alignment with the choice of the optimal differential expression technique, the choice of the optimal gene set database was split between phenotype permutation 1, for which KEGG could be utilized to further raise the number of differentially enriched gene sets, and the remaining 4 phenotype permutations where the default gene set database GO with subontology Molecular Function could not be surpassed with respect to the number of differentially enriched gene sets. Similar to GSEAPreranked, the choice of an exponent value $p = 0$ led to a notable increase in the number of differentially enriched gene sets in all 5 phenotype permutations which ranged between 45 in permutation 3 and 159 in permutation 4. In particular, the increase was approximately twofold in permutations 2, 4 and 5 and stronger than fivefold in permutation 3. Eventually, as expected, multiple test adjustment using Bonferroni did not lead to a higher number of differentially enriched gene sets compared to Benjamini and Hochberg.
To sum up, there was a multitude of parameters that could be adapted to induce overoptimistic results prior to and within `clusterProfiler`'s GSEA tool. Similar to GSEAPreranked (see Subsection 5.2.3), the magnitude of the number of differentially enriched gene sets was among the highest across all GSA tools under investigation and moreover, the choice of the exponent $p = 0$ led to the highest absolute increase.

5.3 Comparison over All Tools

In this section, a comparison of the potential for over-optimistic results of all GSA tools under investigation, quantified by the total increase in the number of differentially enriched gene sets resulting from the respective optimization process, is given in Table 5.1. A further illustration of this table is provided in Figure 5.10 which visualizes the total increase in the number of differentially enriched gene sets for each phenotype permutation across the GSA tools under investigation. Based on this overview, an assessment is made on whether over-optimistic results are induced if the choice of the GSA tool depends on such a comparison between all tools regarding the resulting number of differentially enriched gene sets.

Table 5.1: Total Increase in the Number of Differentially Enriched Gene Sets

Tool	Phenotype Permutation				
	1	2	3	4	5
GOSeq	7	25	11	12	14
DAVID (Web)	0	3	5	6	1
clusterProfiler ORA	9	17	6	0	8
clusterProfiler DAVID	0	3	5	6	2
GSEA (Web)	3	5	3	11	0
PADOG	0	0	0	0	0
GSEAPreranked	236	179	70	162	161
clusterProfiler GSEA	82	169	50	200	97

Total increase in the number of di□erentially enriched gene sets resulting from the stepwise optimization process. The highest increase is indicated in blue.

From Table 5.1 it is visible that GSEAPreranked and `clusterProfiler` GSEA, which are FCS tools that require a ranked list of genes as input, had the highest potential for generating over-optimistic results as the increase in the number of differentially enriched gene sets was significantly higher compared to the other tools under investigation. Particularly, the strongest increase in the two tools was obtained by weighting each gene in the ranked list equally in the computation of the enrichment score (exponent $p = 0$). A less strong but still considerable increase in the number of differentially enriched gene sets was achieved in the process of the generation of the ranked list by optimizing the choice of the differential expression technique and the associated parameters. In the context of these FCS tools, gene set database GO (with subontology Molecular Function) could only be surpassed by KEGG regarding the number of differentially enriched gene sets for one phenotype permutation in each tool. This observation is in alignment with the fact that GO offers a higher number of gene sets compared to KEGG (and Hallmark) and can therefore be used to lead to a higher number of differentially enriched gene sets. Lastly,

the external ranking of the genes based on the p-value led to the highest number of differentially enriched gene sets across all random phenotype permutations and both tools, compared to the alternative ranking metric.
In contrast to that, the remaining FCS tools, namely the regular GSEA web application and `PADOG`, which generate the ranking of the genes internally, had a lower potential for over-optimistic results. Particularly, `PADOG`, which was the only tool chosen for its performance, did not offer any flexibility to increase the number of differentially enriched gene sets by optimizing the set of parameters. Moreover, the number of differentially enriched gene sets remained at a zero level across all of the phenotype permutations which constitutes the best possible scenario given that the phenotype permutations were generated randomly and without any biological meaning. Across the GSEA web application and `PADOG`, there was no clear parameter choice to induce an increase in the number of differentially enriched gene sets. While for the broad majority of the phenotype permutations, none of the RNA-Seq transformation techniques induced an increase in the number of differentially enriched gene sets to a non-zero level, gene set database KEGG could be utilized in some cases in the GSEA web application. The same picture presents itself regarding the choice of the exponent in the GSEA web application, which could be utilized to further increase the number of differentially enriched gene sets for some, but not all, of the random phenotype permutations.

The potential for over-optimistic results of the ORA tools under investigation was comparable to, if not slightly higher than, GSEA web application and higher than `PADOG` since in the majority of phenotype permutations, the number of differentially enriched gene sets could be increased. Across the ORA tools, the highest potential for over-optimistic results was indicated for `GOSeq`. It was shown that the choice of the differential expression technique on the single gene-level and corresponding internal parameters to generate the unranked input list of differentially expressed genes had a substantial effect on the number of differentially enriched gene sets as it induced an increase for the broad majority of the random phenotype permutations. With respect to the gene set database, however, there was no clear optimal choice across all ORA tools. While for `GOSeq` and `clusterProfiler`'s regular ORA tool, GO with subontology Molecular Function could not be surpassed by KEGG with respect to the number of differentially enriched gene sets in any of the phenotype permutations, KEGG led to an increase for the majority of phenotype permutations in the DAVID web application and analogously `clusterProfiler`'s DAVID tool.

To sum up, FCS tools that require the external generation of a ranking based on the magnitude of differential expression on the single-gene level have the highest potential for over-optimistic results which can particularly be induced by setting the exponent parameter to $p = 0$, i.e. by weighting each gene in the ranked list equally in the computation of the enrichment score (see Equation (3.2)). Furthermore, the fact that the different GSA

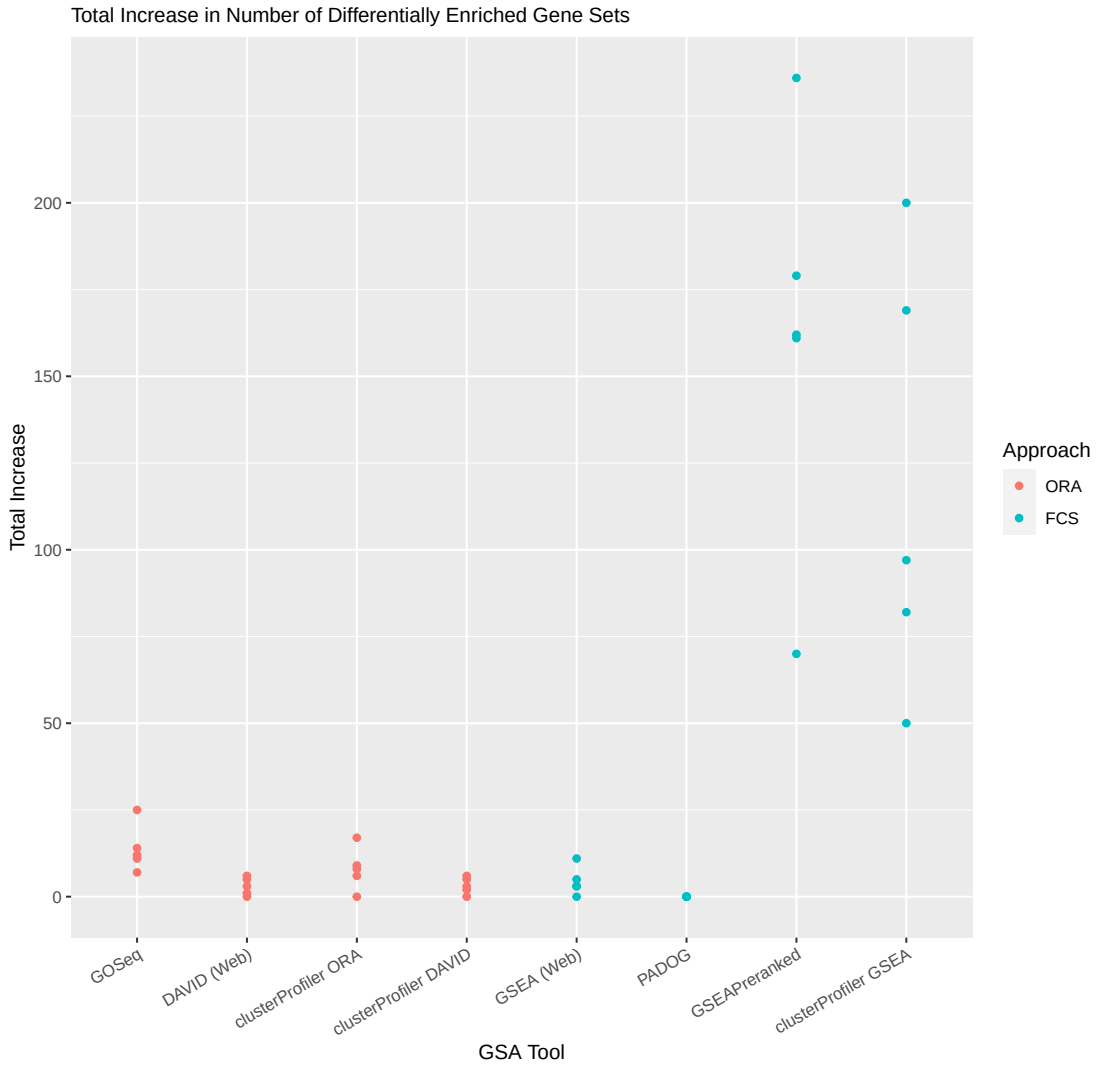

Figure 5.10: Total Increase in Number of Differentially Enriched Gene Sets

tools differ notably in their potential for over-optimistic results indicates that the choice of GSA tool based on the resulting differentially enriched gene sets can in itself lead to over-optimistic results.

It is noted that due to the stepwise optimization process, not all combinations of the parameter options were utilized which means that the values of the absolute increase in Table 5.1 only constitute a lower bound which accordingly also applies to the potential for over-optimistic results.

6. Discussion and Conclusion

The aim of this thesis was to assess GSA and associated methods and tools with respect to the potential for over-optimistic results, i.e. seemingly meaningful but not actually true findings obtained by optimizing certain parameters within the GSA workflow. For this purpose, three GSA methods and eight corresponding tools were introduced for investigation, whereby four tools were classified as ORA and the other four as FCS. The set of possible adaptions in preparation for as well as within the variety of GSA tools was utilized to quantify the potential for over-optimistic results based on a real gene expression data set and with five randomly generated phenotype permutations of the samples. In this context, the potential for over-optimistic results of a given tool was quantified as the maximum difference in the number of differentially enriched gene sets between the default configuration of all parameters and their optimized configuration resulting from a stepwise optimization process.

In this thesis, it was shown that for all tools under investigation except for one, the number of differentially enriched gene sets could be increased for the majority of the random phenotype permutations through stepwise optimization of a variety of parameters. Furthermore, it could be observed that the potential for over-optimistic results generally differed between the types of GSA tools.

It was shown that the FCS tools that require as input a list of genes ranked by their magnitude of differential expression have the highest potential for the generation of over-optimistic results, while also yielding the highest number of differentially enriched gene sets in the default configuration. For both tools, the highest increase for all random phenotype permutations was achieved by weighting each gene from the input list equally in the computation of the enrichment score of each gene set. Moreover, the choice of the differential expression technique on the single gene-level to generate the input list could be utilized to induce an increase in the number of differentially enriched gene sets for the majority of the random phenotype permutations.

For FCS tools that internally generate the ranking of the genes based on their magnitude of differential expression, on the other hand, the potential for over-optimistic results was notably lower. While `PADOG` specifically did not show any potential for over-optimistic results, the optimal parameter choices in the web application of GSEA varied for most of the phenotype permutations.

The potential for over-optimistic results of the ORA tools was found to be comparable, if not slightly higher, to the FCS tools that generate a ranking of genes internally, but significantly lower than those FCS tools that require a ranked list of genes as input. In the context of the ORA tools, an optimization of the input list of differentially expressed genes proved itself as the most reliable measure to increase the number of differentially enriched gene sets.
The fact that these tools, which vary in their approach to conducting GSA, differ notably with respect to the potential for over-optimistic results, indicates that the choice of a GSA tool itself can lead to over-optimistic findings if it is based on which tool provides the most promising results.

As the number of differentially enriched gene sets for each random phenotype permutation and GSA tool was optimized using a stepwise optimization process, not all combinations of the utilized parameters were considered. Therefore, the increase in the corresponding numbers as the result of a single optimization step in each individual tool is to be regarded as a lower bound. This also applies to the overall potential for over-optimistic results of each individual tool which corresponds to the sum of the increase in the number of differentially enriched gene sets of all optimization steps. For example, it cannot be precluded that in the stepwise optimization process of `PADOG`, the combination of the RNA-Seq transformation method `varianceStabilizingTransformation` and the gene set database KEGG would have led to a higher number of differentially enriched gene sets, however, this combination was not included in the stepwise optimization process.
Furthermore, in the context of this thesis, the optimization criterion was chosen solely as the number of differentially enriched gene sets, whereas corresponding adjusted p-values were not included in the criterion. However, taking the respective adjusted p-values into account in each optimization step could have potentially led to a decrease in the adjusted p-values in the subsequent steps. This could have, in turn, resulted in a higher number of differentially enriched gene sets.
Another limitation of the analysis conducted in this thesis is that the selection of parameter values and choices is not exhaustive. For example, differential expression analysis can, in addition to `DESeq2` and `edgeR`, also be conducted with `limma` (Smyth, 2004). Moreover, `voom` and `varianceStabilizingTransformation` are only two among a multitude of RNA-Seq transformation techniques that could potentially lead to a higher number of differentially enriched gene sets.
As there is a lack of common consensus on the conduct of many steps in a GSA workflow and, accordingly, there was no given default option, the default parameter choice was set arbitrarily in this case and deliberately without any knowledge about its effect on the development of the optimization process. Consequently, another default choice could have resulted in a different development of the optimization process. For example, since the

gene set database KEGG generally offers a lower number of gene sets, the choice of KEGG instead of GO as the default gene set database would have led to a notable increase in the number of differentially enriched gene sets in the respective step of the GSEAPreranked optimization process. Moreover, choosing ranking by Log Fold Change as the default ranking metric for those FCS tools that require a ranked list of genes as input would have indicated a stronger inducement of over-optimistic results of the choice of the ranking metric.
Moreover, since the potential for the generation of over-optimistic findings was assessed based on a single gene expression data set, the results are not generalizable for the tools under investigation and GSA in general. This means that in future research, the optimization process could be conducted on a variety of gene expression data sets to form a more reliable conclusion on the potential for over-optimistic results.
Finally, a complete evaluation of the potential for over-optimistic results of each tool under investigation would additionally include a content-related aspect. In particular, an assessment would be made on whether a coherent hypothesis can be extracted from the set of differentially enriched gene sets in the final result table for each random phenotype permutation. The reason for this is that a user would only report those findings which give reasonable answers to his or her research question. Therefore, in future research, collaborations with researchers in the respective biological field could contribute to the meaningfulness of the analysis of the potential for over-optimistic results of GSA.

Overall, the assessment of the potential for over-optimistic results in this thesis generates an awareness of the potential for over-optimistic results in GSA and practices which lead to those. However, to generate more precise statements on the potential with respect to individual tools and GSA in general, further thorough research needs to be conducted and the optimization process extended.

References

M. Ackermann and K. Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):1–20, 2009.

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Nature Precedings*, 2010.

M. Ashburner et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

A.-L. Boulesteix and C. Strobl. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC medical research methodology*, 9(1):85, 2009.

J. H. Bullard, E. Purdom, K. D. Hansen, et al. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):1–13, 2010.

M. Carlson. org.Hs.eg.db: Genome wide annotation for Human. *R package version 3.13.0*, 2021.

S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.

B. Efron and R. Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.

C. Fresno and E. A. Fernandez. *RDAVIDWebService: An R Package for retrieving data from DAVID into R objects using Web Services API.*, 2013. http://www.bdmg.com.ar, http://david.abcc.ncifcrf.gov/.

L. Geistlinger, G. Csaba, and R. Zimmer. Bioconductor's EnrichmentBrowser: seamless navigation through combined results of set- and network-based enrichment analysis. *BMC Bioinformatics*, 17:45, 2016.

L. Geistlinger, G. Csaba, M. Santarelli, M. Ramos, L. Schiffer, N. Turaga, C. Law, S. Davis, V. Carey, M. Morgan, et al. Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in Bioinformatics*, 22(1):545–556, 2021.

D. A. Hosack, G. Dennis, B. T. Sherman, H. C. Lane, and R. A. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(10):1–8, 2003.

K. L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, et al. Ensembl 2021. *Nucleic Acids Research*, 49(D1):884–891, 2021.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009a.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009b.

W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Ole's, H. Pag'es, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):115–121, 2015. URL `http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html`.

M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1):D545–D551, 2021.

P. Khatri, M. Sirota, and A. J. Butte. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, 8(2):e1002375, 2012.

K.-R. Koch. *Parameter Estimation and Hypothesis Testing in Linear Models*. Springer Science & Business Media, 1999.

C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):1–17, 2014.

A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell systems*, 1(6): 417–425, 2015.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.

D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 39(suppl_1):D52–D57, 2010.

F. Maleki, K. L. Ovens, D. J. Hogan, E. Rezaei, A. M. Rosenberg, and A. J. Kusalik. Measuring consistency among gene set analysis methods: A systematic study. *Journal of bioinformatics and computational biology*, 17(5):1940010, 2019.

F. Maleki, K. Ovens, D. J. Hogan, and A. J. Kusalik. Gene Set Analysis: Challenges, Opportunities, and Future Research. *Frontiers in Genetics*, 11:654, 2020.

V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, et al. PGC-1 -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

J. Reimand, R. Isserlin, V. Voisin, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, 14(2):482–517, 2019.

M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):1–9, 2010.

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

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA, 2021. URL `http://www.rstudio.com/`.

M. Shahjaman, M. M. H. Mollah, M. R. Rahman, S. S. Islam, and M. N. H. Mollah. Robust identification of differentially expressed genes from RNA-seq data. *Genomics*, 112(2): 2000–2010, 2020.

T. C. Silva, A. Colaprico, C. Olsen, F. D'Angelo, G. Bontempi, M. Ceccarelli, and H. Noushmehr. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research*, 5, 2016.

J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366, 2011.

G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

A. L. Tarca, S. Draghici, G. Bhatti, and R. Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):1–14, 2012.

A. L. Tarca, G. Bhatti, and R. Romero. A Comparison of Gene Set Analysis Methods in Terms of Sensitivity, Prioritization and Specificity. *PloS One*, 8(11):e79217, 2013.

The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, Jan 2021.

S. Tweedie, B. Braschi, K. Gray, T. E. Jones, R. L. Seal, B. Yates, and E. A. Bruford. Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research*, 49 (D1):D939–D946, 2021.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL `https://CRAN.R-project.org/package=dplyr`. R package version 1.0.7.

T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, T. Feng, L. Zhou, W. Tang, L. Zhan, et al. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141, 2021.

C. Xie, S. Jauhari, and A. Mora. Popularity and performance of bioinformatics software: the case of gene set analysis. *BMC Bioinformatics*, 22(1):1–16, 2021.

M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2):1–12, 2010.

G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609, 2015.

Z. Zhang, D. Yu, M. Seo, C. P. Hersh, S. T. Weiss, and W. Qiu. Novel Data Transformations for RNA-seq Differential Expression Analysis. *Scientific Reports*, 9(1):1–12, 2019.

A. Zhu, J. G. Ibrahim, and M. I. Love. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, 35(12): 2084–2092, 2019.

J. Zyla, M. Marczyk, T. Domaszewska, S. H. Kaufmann, J. Polanska, and J. Weiner 3rd. Gene set enrichment for reproducible science: comparison of CERNO and eight other algorithms. *Bioinformatics*, 35(24):5146–5154, 2019.

A. Appendix

A.1 Normalization

Normalization is performed to remove technical biases from the gene expression data that arise from the sequencing process itself and have sample-specific effects on the count data (Robinson and Oshlack, 2010). These biases would skew the analysis unless accounted for. The first technical bias present when comparing samples of RNA-seq data is the library size or sequencing depth. It refers to the total number of read counts mapped to a sample and arises from the sequencing process, leading to different samples having different library sizes independent of the magnitude of differential expression. Hence, samples with different library sizes are in fact not comparable (Love et al., 2014). The effect of compositionality, on the other hand, is caused by RNA-sequencing data containing only information about relative abundance. For instance, if the majority of counts in a given sample are mapped to a single gene, other genes in this sample will appear to be less expressed in comparison to other samples and therefore may be erroneously detected as down-regulated (Robinson and Oshlack, 2010). Another factor that is addressed by some normalization methods is gene length since genes with a longer transcript length are naturally assigned more reads in the sequencing process. In the context of differential expression analysis, comparisons are made between different samples, but not between genes within a sample. As gene length is a fixed quantity for a given gene across the samples, it, therefore, does not have to be considered by normalization techniques during differential expression analysis. For a comparison of genes within a given sample, on the other hand, normalization for gene length becomes necessary.

A.1.1 Normalization in DESeq2

The normalization factor s_j computed in `DESeq2` accounts for differences in library size between samples as well as compositionality effects. It is computed using the median-of-ratios method (Anders and Huber, 2010), namely

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

The normalized count values can be obtained by dividing the raw read count K_{ij} of sample i in sample j by the normalization factor, i.e.

$$K_{ij}^{\text{norm}} = \frac{K_{ij}}{s_j}: \tag{A.2}$$

It is noted that DESeq2 does not use the normalized counts in the workflow of differential expression analysis but instead includes the normalization factor separately and utilizes the untransformed, raw counts.

A.1.2 Normalization in edgeR

In edgeR, the library size is normalized by finding a normalization factor for each sample such that fold change is minimized for the majority of genes across all samples. The default normalization factor for each sample is calculated using the Trimmed Mean of M-values method (Robinson and Oshlack, 2010), abbreviated by TMM, under the assumption that the majority of the genes in the experiment are not differentially expressed. The idea behind it is to model the expected read counts of gene i in sample j as

$$\mathsf{E}[K_{ij}] = \frac{q_{ij} \quad L_i}{O_j} \quad S_j; \tag{A.3}$$

where L_i is the transcript length of gene i, whereas S_j is the library size of sample j. Quantity O_j, the total RNA output of sample j, is unknown for all samples $j = 1;:::;m$ and cannot be estimated directly. Instead, the relative RNA production

$$\frac{O_r}{O_j} = \frac{\mathsf{E}[K_{ij}]=S_j}{\mathsf{E}[K_{ir}]=S_r} \quad \frac{K_{ij}=S_j}{K_{ir}=S_r} \tag{A.4}$$

is estimated with respect to a fixed reference sample r ($r \quad 2 \; f1;:::;mg$). The first line of Equation (A.4) can be derived using the assumption of most genes not being differential expressed across conditions, i.e. $q_{ij} \quad q_{ir}$. After computing the gene-wise log fold changes

$$M_{ij}^{r} = \log_2 \frac{K_{ij}=S_j}{K_{ir}=S_r} \tag{A.5}$$

and the absolute expression levels

$$A_{ij}^{r} = \frac{1}{2} \left[\log_2 \left(\frac{K_{ij}}{S_j} \right) + \log_2 \left(\frac{K_{ir}}{S_r} \right) \right] \tag{A.6}$$

for sample j relative to sample r, all genes showing extreme expression differences, i.e. large absolute values in M_{ij}^{r} and A_{ij}^{r}, are removed ("trimmed"). The normalization factor of sample j is eventually computed as the weighted mean M_j^{r} of the remaining genes

whereby the weight of a gene i in sample j is computed as the inverse of the approximate asymptotic variance of the gene-wise log fold change M_{ij}^{r} using the delta method. It is noted that the resulting normalization factors for each sample are not used to transform the original count data, as these remain unaffected, but instead to compute the effective library size as a product of the original library size and the respective normalization factor. The effective library then replaces the original library size in all downstream analyses.

A.2 Analysis of TCGA Gene Expression Data Set

As observed multiple times in Chapter 5, several gene sets are detected as differentially enriched for more than one, if not all, of the 5 random phenotype permutations in the ORA tools. Since the phenotype permutations were generated randomly, this finding requires further exploration of the results themselves and the underlying TCGA expression data set. It is noted that this additional analysis is restricted to genes in the Entrez gene ID format since this is the required format for three out of the four ORA tools, namely DAVID (web application) and `clusterProfiler`'s regular ORA and DAVID tool. Furthermore, in contrast to `GOSeq`, these three ORA tools specify which genes from the input list of differentially expressed genes are members of the individual gene sets detected as differentially enriched. This provides further opportunity to fathom the connection between the input lists of differentially expressed genes and the result tables of the respective ORA tool.
Accordingly, a closer look at the optimal result tables of the three ORA tools reveals that a number of genes appear across multiple phenotype permutations but also across the three tools. For example, the gene with ID 3060 can be found as a differentially expressed member of multiple differentially enriched gene sets in the phenotype permutations 1-3 and 5 resulting from `clusterProfiler`'s regular ORA tool. In the optimal results of `clusterProfiler`'s DAVID tool, this gene appears once in phenotype permutations 3 and 4 and the same picture presents itself in the optimal result tables of the DAVID web tool. This means that, in alignment with the overlap of the differentially enriched gene sets across the phenotype permutations, there are several individual genes that appear in multiple input lists of differentially expressed genes. A closer look is therefore taken at the distribution of read count values of the respective genes in the input lists to investigate why they are detected as differentially expressed for several of the phenotype permutations.
An investigation of the distribution of count values of gene 3060, which can be found in Figure A.1, shows that this gene has a single count outlier in sample "TCGA-19-1787" with a significantly higher read count value compared to the remaining samples.
Similar to gene 3060, gene 8360 can be found in the majority of gene sets detected as differentially enriched in the result tables for phenotype permutations 2, 3 and 4 generated with `clusterProfiler`'s DAVID tool as well as the DAVID web tool. In the optimal result tables generated with `clusterProfiler`'s regular ORA tool, on the other hand, gene

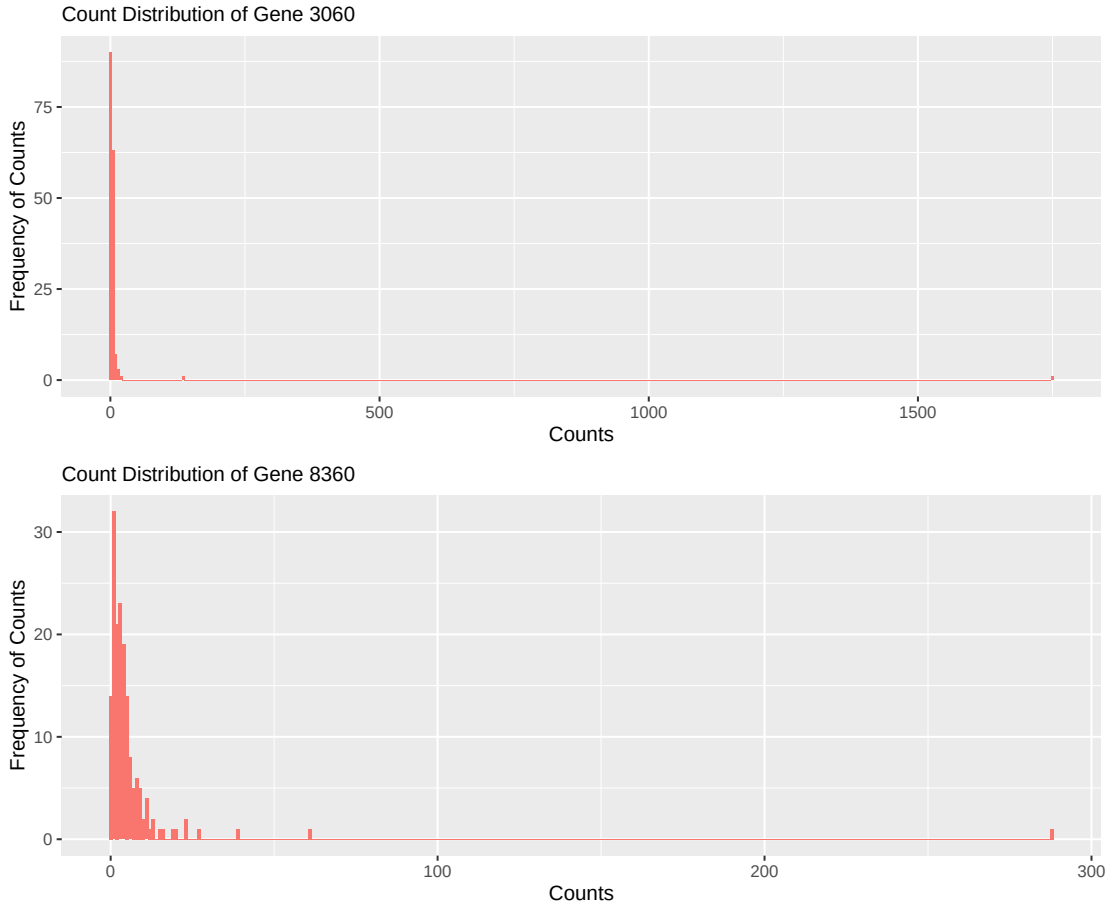

Figure A.1: Count Distribution of Selected Genes

8360 is a member of a single differentially enriched gene set in phenotype permutations 1-3 and 5 each. An investigation of the distribution of count values of this gene (see Figure A.1) reveals that it, too, has a single but notable count outlier. In contrast to gene 3060, however, this count outlier occurs in sample "TCGA-26-5134".

A repetition of this analysis for a number of genes shows a similar pattern, namely a single count outlier value in the count distribution of the respective gene. In particular, the samples in which the respective count outliers occur vary across genes. Consequently, the existence of the outliers cannot be traced to a single sample with a notably higher library size and therefore, the outlier values are not eliminated through normalization in the process of differential expression analysis. The detection as being differentially expressed of a gene with one notable count outlier is therefore the logical consequence since one of the two phenotypes has a considerably higher overall magnitude of count values which, in turn, indicates a higher gene expression level.

In the optimized input lists of differentially expressed genes for the 3 ORA tools, which were generated with `edgeR`, the number of differentially expressed genes amounts to 437, 519, 471, 634, 586 for the respective phenotype permutations (see Section 5.1.1). In alignment with the observations presented above, the overlap between these 5 lists amounts to

178 differentially expressed, which means that there are 178 genes detected as differentially expressed in all of the 5 random phenotype permutations. Consequently, an investigation is carried out on whether this overlap is indeed caused by single count outliers in the gene expression measurement of a number of genes.
In contrast to `edgeR`, `DESeq2` uses Cook's outlier detection by default to replace isolated instances of count outliers with values that comply with the null hypothesis of differential expression (Love et al., 2014). As the consequence, if the TCGA gene expression data set contains a multitude of genes with a single but considerable outlier value each, these outlier values are substituted accordingly in the standard `DESeq2` workflow. In this context, the default input lists for the ORA tools, which were generated with `DESeq2` including Cook's outlier detection, contain 62, 39, 13, 145 and 158 differentially expressed genes for the respective 5 phenotype permutations. Furthermore, there is no overlap between these 5 lists, meaning that there are 0 genes detected as differentially expressed across all 5 phenotype permutations. This is in alignment with the intuition behind Cook's outlier detection and replacement mentioned above. In the same context, it is to be expected that, if the high overlap of differentially expressed genes resulting from `edgeR` is a consequence of single count outliers for a number of genes, a deactivation of Cook's Outlier detection in `DESeq2` leads to a higher number of differentially expressed genes and particularly a higher overlap of the lists of differentially expressed genes across the 5 phenotype permutations.
A differential expression analysis of the TCGA expression data set using `DESeq2`, in which Cook's outlier detection is turned off, results in a notable increase in the number of differentially expressed genes to 383, 449, 420, 561 and 553 for the respective 5 phenotype permutations. These are on a comparable, but slightly lower, level as the number of differentially expressed genes resulting from `edgeR`. In particular, the overlap of differentially expressed genes resulting from `DESeq2` with a deactivated Cook's outlier detection amounts to 151 genes across all 5 phenotype permutations. This is a strong indicator that the high overlap of differentially expressed genes across the random phenotype permutations is caused by single count outliers in the gene expression measurement of a multitude of genes, which in turn leads to an overlap of differentially enriched gene sets across the permutations. Particularly, the intersection between the 151 genes in the overlap across all 5 phenotype permutations resulting from `DESeq2` (with Cook's outlier detection deactivated) and the 178 genes resulting from `edgeR` amounts 148 genes. This means that in this parameter setting, `DESeq2` and `edgeR` detect similar lists of differentially expressed genes.
To illustrate the results of this exploration, `clusterProfiler`'s regular ORA tool is applied with the input lists generated using `DESeq2`, including a disabled Cook's outlier detection. In alignment with the course of the stepwise optimization process presented in Section 5.1.4, the parameters within the ORA tool remain in their default configuration. A special focus is given to the question of whether, in alignment with the overlap of the lists of differentially expressed genes, there is a high overlap of differentially enriched gene sets across the phenotype permutations. The lists of the top 5 differentially enriched gene sets

for each of the 5 random phenotype permutations can be found in Table A.1. From this table, it can be observed that there is an obvious overlap between the result tables across the 5 random phenotype permutations if the input lists are generated with `DESeq2` and a deactivated Cook's outlier detection. Particularly, gene sets "Protein Heterodimerization Activity" and "Receptor Ligand Activity" can be found among the top gene sets. Furthermore, gene 3060 is present in multiple differentially enriched gene sets in all of the 5 random phenotype permutations. Gene 8360, on the other hand, is indicated in one differentially enriched gene set in each of the random phenotype permutations. This is in contrast to the results of the ORA tool with the input list generated with `DESeq2` in its default configuration, i.e. with Cook's outlier detection turned on. As can be observed in Figure 5.4 in step "Default", this set of default input lists leads to 0 differentially enriched gene sets in all random phenotype permutations apart from permutations 4 and 5. In particular, the overlap between the final results of these two permutations amounts to one gene set ("Antigen Binding"), while the overlap across all 5 phenotype permutations is 0.

To sum up, it could be shown that count outliers for a multitude of individual genes cause a high overlap of the results of differentially enriched gene sets across the 5 random phenotype permutations. This finding illustrates the dependence of the results of the ORA tools on the differential expression technique to generate the input lists and the corresponding parameter choice. Furthermore, this finding underlines that the deactivation of Cook's outlier detection can be used to wilfully manipulate the results of the GSA tools and supports the exclusion of this parameter from the stepwise optimization process.

Table A.1: TCGA Expression Data Exploration: Top 5 Differentially Enriched Gene Sets in clusterProfiler ORA with DESeq2 and Deactivated Cook's Outlier Detection

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Adjusted p-Value	q-Value
1	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	0.0008	0.0007
1	GO (MF)	2	GO:0005179 Hormone Activity	0.0031	0.0029
1	GO (MF)	3	GO:0048018 Receptor Ligand Activity	0.0042	0.0040
1	GO (MF)	4	GO:0030546 Signaling Receptor Activator Activity	0.0042	0.0040
1	GO (MF)	5	GO:0005201 Extracellular Matrix Structural Constituent	0.0042	0.0040
2	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	4.2432E-07	3.8326E-07
2	GO (MF)	2	GO:0048018 Receptor Ligand Activity	2.9200E-05	2.6374E-05
2	GO (MF)	3	GO:0030546 Signaling Receptor Activator Activity	2.9200E-05	2.6374E-05
2	GO (MF)	4	GO:0001664 G Protein-Coupled Receptor Binding	0.0030	0.00271
2	GO (MF)	5	GO:0005201 Extracellular Matrix Structural Constituent	0.0099	0.00897
3	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	3.6681E-09	3.5070E-09
3	GO (MF)	2	GO:0048018 Receptor Ligand Activity	0.0046	0.00439
3	GO (MF)	3	GO:0030546 Signaling Receptor Activator Activity	0.0046	0.00439
3	GO (MF)	4	GO:0005201 Extracellular Matrix Structural Constituent	0.0082	0.00783
3	GO (MF)	5	GO:0005104 Fibroblast Growth Factor Receptor Binding	0.0499	0.04775
4	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	5.7452E-13	5.202E-13
4	GO (MF)	2	GO:0048018 Receptor Ligand Activity	0.0047	0.0043
4	GO (MF)	3	GO:0005254 Chloride Channel Activity	0.0047	0.0043

Continued on Next Page

Table A.1: TCGA Expression Data Exploration: Top 5 Differentially Enriched Gene Sets in clusterProfiler ORA with DESeq2 and Deactivated Cook's Outlier Detection

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Adjusted p-Value	q-Value
4	GO (MF)	4	GO:0030546 Signaling Receptor Activator Activity	0.0047	0.0043
4	GO (MF)	5	GO:0015108 Chloride Transmembrane Transporter Activity	0.0067	0.0061
5	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	1.8522E-06	1.8077E-06
5	GO (MF)	2	GO:0034987 Immunoglobulin Receptor Binding	8.0002E-06	7.8079E-06
5	GO (MF)	3	GO:0048018 Receptor Ligand Activity	0.0001	0.0001
5	GO (MF)	4	GO:0030546 Signaling Receptor Activator Activity	0.0001	0.0001
5	GO (MF)	5	GO:0003823 Antigen Binding	0.0018	0.0017

A.3 Optimal Result Tables

In the following, the 5 sets of differentially enriched gene sets resulting from the optimal parameter choice are presented for each tool under investigation except for `PADOG`. If applicable, further relevant quantities are additionally provided for the purpose of interpretability. As the magnitude of the number of differentially enriched gene sets in the optimal result tables of GSEAPreranked and `clusterProfiler`'s GSEA tool are particularly high, only the 10 differentially enriched gene sets with the lowest adjusted p-values are given for each of the 5 random phenotype permutations.

Table A.2: Differentially Enriched Gene Sets in DAVID (Web)

Phenotype Permutation	Gene Set Database	Nr.	Differentially Enriched Gene Set	Count	Fold Enrichment	Adjusted p-Value
1	/	/	/	/	/	/
2	KEGG	1	hsa05322: Systemic Lupus Erythematosus	20	7.9529	8.1697E-10
2	KEGG	2	hsa05034: Alcoholism	21	6.0732	1.5798E-08
2	KEGG	3	hsa04613: Neutrophil Extracellular Trap Formation	20	5.6926	9.9561E-08
3	KEGG	1	Hsa05322: Systemic Lupus Erythematosus	26	10.9213	6.0097E-17
3	KEGG	2	Hsa05034: Alcoholism	27	8.2482	7.3832E-15
3	KEGG	3	Hsa04613: Neutrophil Extracellular Trap Formation	24	7.2160	8.3087E-12
3	KEGG	4	Hsa04080: Neuroactive Ligand-Receptor Interaction	17	2.7511	2.0517E-02
3	KEGG	5	Hsa05203: Viral Carcinogenesis	12	3.3604	3.4181E-02
4	KEGG	1	hsa05322: Systemic Lupus Erythematosus	39	11.8685	1.7872E-28
4	KEGG	2	hsa05034: Alcoholism	39	8.6317	3.1342E-23
4	KEGG	3	hsa04613: Neutrophil Extracellular Trap Formation	37	8.0597	5.5255E-21
4	KEGG	4	hsa04080: Neuroactive Ligand-Receptor Interaction	28	3.2829	4.6880E-06
4	KEGG	5	hsa05203: Viral Carcinogenesis	19	3.8547	8.5931E-05
4	KEGG	6	hsa04217: Necroptosis	13	3.3839	1.7508E-02
5	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	25	3.5826	7.1573E-05
5	GO (MF)	2	GO:0034987 Immunoglobulin Receptor Binding	12	6.7688	3.4898E-04
5	GO (MF)	3	GO:0003823 Antigen Binding	12	4.5126	1.1305E-02
5	GO (MF)	4	GO:0008083 Growth Factor Activity	12	3.8330	3.5836E-02

Table A.3: Differentially Enriched Gene Sets in GOSeq

Phenotype Permutation	Gene Set Database	Nr.	Differentially Enriched Gene Set	Gene Set Size	Adjusted p-Value
1	GO (MF)	1	GO:0030545 Receptor Regulator Activity	517	0.0005
1	GO (MF)	2	GO:0005179 Hormone Activity	116	0.0005
1	GO (MF)	3	GO:0048018 Receptor Ligand Activity	469	0.0005
1	GO (MF)	4	GO:0030546 Signaling Receptor Activator Activity	475	0.0005
1	GO (MF)	5	GO:0005102 Signaling Receptor Binding	1598	0.0006
1	GO (MF)	6	GO:0003674 Molecular Function	17452	0.0100
1	GO (MF)	7	GO:0005184 Neuropeptide Hormone Activity	28	0.0136
2	GO (MF)	1	GO:0048018 Receptor Ligand Activity	469	3.6806E-07
2	GO (MF)	2	GO:0030546 Signaling Receptor Activator Activity	475	3.6806E-07
2	GO (MF)	3	GO:0030545 Receptor Regulator Activity	517	4.0134E-07
2	GO (MF)	4	GO:0046982 Protein Heterodimerization Activity	313	9.5091E-06
2	GO (MF)	5	GO:0005102 Signaling Receptor Binding	1598	3.5163E-05
2	GO (MF)	6	GO:0003674 Molecular$_{F}$*unction*	17452	4.2983E-05
2	GO (MF)	7	GO:0001664 G Protein-Coupled Receptor Binding	278	0.0001
2	GO (MF)	8	GO:0008106 Alcohol Dehydrogenase (NADP+) Activity	21	0.0003
2	GO (MF)	9	GO:0004032 Alditol:NADP+ 1-Oxidoreductase Activity	12	0.0003
2	GO (MF)	10	GO:0004033 Aldo-Keto Reductase (NADP) Activity	27	0.0012
2	GO (MF)	11	GO:0047718 Indanol Dehydrogenase Activity	3	0.0015
2	GO (MF)	12	GO:0008009 Chemokine Activity	48	0.0028

Continued on Next Page

Table A.3: Differentially Enriched Gene Sets in GOSeq

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Gene Set Size	Adjusted p-Value
2	GO (MF)	13	GO:0005179 Hormone Activity	116	0.0036
2	GO (MF)	14	GO:0005488 Binding	15858	0.0043
2	GO (MF)	15	GO:0005104 Fibroblast Growth Factor Receptor Binding	25	0.0074
2	GO (MF)	16	GO:0005201 Extracellular Matrix Structural Constituent	170	0.0074
2	GO (MF)	17	GO:0022804 Active Transmembrane Transporter Activity	298	0.0076
2	GO (MF)	18	GO:0042379 Chemokine Receptor Binding	61	0.0090
2	GO (MF)	19	GO:0008083 Growth Factor Activity	158	0.0320
2	GO (MF)	20	GO:0071855 Neuropeptide Receptor Binding	36	0.0439
2	GO (MF)	21	GO:0015081 Sodium Ion Transmembrane Transporter Activity	146	0.0453
2	GO (MF)	22	GO:0046983 Protein Dimerization Activity	1014	0.0453
2	GO (MF)	23	GO:0031716 Calcitonin Receptor Binding	2	0.0464
2	GO (MF)	24	GO:0030021 Extracellular Matrix Structural Constituent Conferring Compression Resistance	22	0.0464
2	GO (MF)	25	GO:0048020 CCR Chemokine Receptor Binding	38	0.0468
3	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	313	4.8461E-05
3	GO (MF)	2	GO:0048018 Receptor Ligand Activity	469	0.0004
3	GO (MF)	3	GO:0030546 Signaling Receptor Activator Activity	475	0.0004
3	GO (MF)	4	GO:0030545 Receptor Regulator Activity	517	0.0004
3	GO (MF)	5	GO:0003674 Molecular Function	17452	0.0004

Continued on Next Page

Table A.3: Differentially Enriched Gene Sets in GOSeq

Phenotype Permutation	Gene Set Database	Nr.	Differentially Enriched Gene Set	Gene Set Size	Adjusted p-Value
3	GO (MF)	6	GO:0005488 Binding	15858	0.0027
3	GO (MF)	7	GO:0005201 Extracellular Matrix Structural Constituent	170	0.0036
3	GO (MF)	8	GO:0005102 Signaling Receptor Binding	1598	0.0072
3	GO (MF)	9	GO:0046983 Protein Dimerization Activity	1014	0.0303
3	N/A	10	N/A	1609	0.0303
3	GO (MF)	11	GO:0005179 Hormone Activity	116	0.0499
4	GO (MF)	1	GO:0015267 Channel Activity	455	3.7754E-06
4	GO (MF)	2	GO:0022803 Passive Transmembrane Transporter Activity	456	3.7754E-06
4	GO (MF)	3	GO:0022843 Voltage-Gated Cation Channel Activity	134	3.7754E-06
4	GO (MF)	4	GO:0005261 Cation Channel Activity	317	3.8357E-06
4	GO (MF)	5	GO:0022836 Gated Channel Activity	327	4.0739E-06
4	GO (MF)	6	GO:0099094 Ligand-Gated Cation Channel Activity	104	4.0739E-06
4	GO (MF)	7	GO:0005216 Ion Channel Activity	412	4.0739E-06
4	GO (MF)	8	GO:0003674 Molecular$_{F}$*unction*	16649	2.0118E-05
4	GO (MF)	9	GO:0015276 Ligand-Gated Ion Channel Activity	134	2.0118E-05
4	GO (MF)	10	GO:0022834 Ligand-Gated Channel Activity	134	2.0118E-05
4	GO (MF)	11	GO:0005244 Voltage-Gated Ion Channel Activity	191	2.0118E-05
4	GO (MF)	12	GO:0022832 Voltage-Gated Channel Activity	191	2.0118E-05
4	GO (MF)	13	GO:0015318 Inorganic Molecular Entity Transmembrane Transporter Activity	788	0.0001

Continued on Next Page

Table A.3: Differentially Enriched Gene Sets in GOSeq

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Gene Set Size	Adjusted p-Value
4	GO (MF)	14	GO:0008324 Cation Transmembrane Transporter Activity	600	0.0002
4	GO (MF)	15	GO:0005488 Binding	15271	0.0002
4	GO (MF)	16	GO:0022890 Inorganic Cation Transmembrane Transporter Activity	552	0.0004
4	GO (MF)	17	GO:0022857 Transmembrane Transporter Activity	1012	0.0004
4	GO (MF)	18	GO:0015075 Ion Transmembrane Transporter Activity	903	0.0005
4	GO (MF)	19	GO:0005215 Transporter Activity	1121	0.0014
4	GO (MF)	20	GO:0005516 Calmodulin Binding	193	0.0027
4	GO (MF)	21	GO:0005249 Voltage-Gated Potassium Channel Activity	83	0.0037
4	GO (MF)	22	GO:0046873 Metal Ion Transmembrane Transporter Activity	411	0.0059
4	GO (MF)	23	GO:0005242 Inward Rectifier Potassium Channel Activity	21	0.0142
4	GO (MF)	24	GO:0005267 Potassium Channel Activity	114	0.0151
4	GO (MF)	25	GO:0005515 Protein Binding	13160	0.0186
4	GO (MF)	26	GO:0003823 Antigen Binding	140	0.0261
4	GO (MF)	27	GO:0015079 Potassium Ion Transmembrane Transporter Activity	146	0.0418
5	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	313	2.0093E-06
5	GO (MF)	2	GO:0030545 Receptor Regulator Activity	517	1.0121E-05
5	GO (MF)	3	GO:0048018 Receptor Ligand Activity	469	1.0121E-05
5	GO (MF)	4	GO:0034987 Immunoglobulin Receptor Binding	63	1.0121E-05
5	GO (MF)	5	GO:0030546 Signaling Receptor Activator Activity	475	1.0928E-05

Continued on Next Page

Table A.3: Differentially Enriched Gene Sets in GOSeq

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Gene Set Size	Adjusted p-Value
5	GO (MF)	6	GO:0003674 Molecular Function	17452	4.8508E-05
5	GO (MF)	7	GO:0005102 Signaling Receptor Binding	1598	0.0001
5	GO (MF)	8	GO:0005488 Binding	15858	0.0002
5	GO (MF)	9	GO:0046983 Protein Dimerization Activity	1014	0.0012
5	GO (MF)	10	GO:0004967 Glucagon Receptor Activity	3	0.0021
5	GO (MF)	11	GO:0008083 Growth Factor Activity	158	0.0062
5	GO (MF)	12	GO:0005201 Extracellular Matrix Structural Constituent	170	0.0063
5	GO (MF)	13	GO:0003823 Antigen Binding	147	0.0158
5	GO (MF)	14	GO:1905538 Polysome Binding	6	0.0237
5	GO (MF)	15	GO:0005509 Calcium Ion Binding	692	0.0456
5	GO (MF)	16	GO:0005179 Hormone Activity	116	0.0456

Table A.4: Differentially Enriched Gene Sets in clusterProfiler ORA

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Adjusted p-Value	q-Value
1	GO (MF)	1	GO:0048018 Receptor Ligand Activity	0.0005	0.0005
1	GO (MF)	2	GO:0030546 Signaling Receptor Activator Activity	0.0005	0.0005
1	GO (MF)	3	GO:0005179 Hormone Activity	0.0005	0.0005
1	GO (MF)	4	GO:0005184 Neuropeptide Hormone Activity	0.0106	0.0100
1	GO (MF)	5	GO:0046982 Protein Heterodimerization Activity	0.0169	0.0160
1	GO (MF)	6	GO:0008106 Alcohol Dehydrogenase (NADP+) Activity	0.0216	0.0204
1	GO (MF)	7	GO:0005201 Extracellular Matrix Structural Constituent	0.0227	0.0214
1	GO (MF)	8	GO:0052650 Nadp-Retinol Dehydrogenase Activity	0.0392	0.0370
1	GO (MF)	9	GO:0004033 Aldo-Keto Reductase (NADP) Activity	0.0392	0.0370
2	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	5.8347E-06	5.3185E-06
2	GO (MF)	2	GO:0048018 Receptor Ligand Activity	6.1646E-06	5.6192E-06
2	GO (MF)	3	GO:0030546 Signaling Receptor Activator Activity	6.1646E-06	5.6192E-06
2	GO (MF)	4	GO:0008009 Chemokine Activity	0.0024	0.0022
2	GO (MF)	5	GO:0001664 G Protein-Coupled Receptor Binding	0.0045	0.0041
2	GO (MF)	6	GO:0005104 Fibroblast Growth Factor Receptor Binding	0.0048	0.0043
2	GO (MF)	7	GO:0042379 Chemokine Receptor Binding	0.0113	0.0103
2	GO (MF)	8	GO:0005201 Extracellular Matrix Structural Constituent	0.0113	0.0103
2	GO (MF)	9	GO:0022804 Active Transmembrane Transporter Activity	0.0131	0.0119
2	GO (MF)	10	GO:0005179 Hormone Activity	0.0131	0.0119

Continued on Next Page

Table A.4: Differentially Enriched Gene Sets in clusterProfiler ORA

Phenotype Permutation	Gene Set Database	Nr.	Differentially Enriched Gene Set	Adjusted p-Value	q-Value
2	GO (MF)	11	GO:0008106 Alcohol Dehydrogenase (NADP+) Activity	0.0180	0.0164
2	GO (MF)	12	GO:0030021 Extracellular Matrix Structural Constituent Conferring Compression Resistance	0.0196	0.0178
2	GO (MF)	13	GO:0008083 Growth Factor Activity	0.0196	0.0178
2	GO (MF)	14	GO:0004032 Alditol:NADP+ 1-Oxidoreductase Activity	0.0322	0.0293
2	GO (MF)	15	GO:0015081 Sodium Ion Transmembrane Transporter Activity	0.0322	0.0293
2	GO (MF)	16	GO:0004033 Aldo-Keto Reductase (NADP) Activity	0.0322	0.0293
2	GO (MF)	17	GO:0048020 Ccr Chemokine Receptor Binding	0.0322	0.0293
2	GO (MF)	18	GO:0015370 Solute:Sodium Symporter Activity	0.0405	0.0369
3	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	7.3539E-09	7.0261E-09
3	GO (MF)	2	GO:0048018 Receptor Ligand Activity	0.0022	0.0021
3	GO (MF)	3	GO:0030546 Signaling Receptor Activator Activity	0.0022	0.0021
3	GO (MF)	4	GO:0005201 Extracellular Matrix Structural Constituent	0.0022	0.0021
3	GO (MF)	5	GO:0030021 Extracellular Matrix Structural Constituent Conferring Compression Resistance	0.0335	0.0320
3	GO (MF)	6	GO:0005104 Fibroblast Growth Factor Receptor Binding	0.0465	0.0444
4	GO (MF)	1	GO:0022843 Voltage-Gated Cation Channel Activity	5.1931E-07	4.6772E-07
4	GO (MF)	2	GO:0005244 Voltage-Gated Ion Channel Activity	6.5953E-06	5.9402E-06
4	GO (MF)	3	GO:0022832 Voltage-Gated Channel Activity	6.5953E-06	5.9402E-06
4	GO (MF)	4	GO:0005261 Cation Channel Activity	1.8415E-05	1.6586E-05
4	GO (MF)	5	GO:0022836 Gated Channel Activity	1.8415E-05	1.6586E-05

Continued on Next Page

Table A.4: Differentially Enriched Gene Sets in clusterProfiler ORA

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Adjusted p-Value	q-Value
4	GO (MF)	6	GO:0015267 Channel Activity	2.1903E-05	1.9727E-05
4	GO (MF)	7	GO:0022803 Passive Transmembrane Transporter Activity	2.1903E-05	1.9727E-05
4	GO (MF)	8	GO:0005216 Ion Channel Activity	2.9652E-05	2.6707E-05
4	GO (MF)	9	GO:0005249 Voltage-Gated Potassium Channel Activity	5.1544E-05	4.6424E-05
4	GO (MF)	10	GO:0005267 Potassium Channel Activity	0.0004	0.0004
4	GO (MF)	11	GO:0015079 Potassium Ion Transmembrane Transporter Activity	0.0016	0.0015
4	GO (MF)	12	GO:0099094 Ligand-Gated Cation Channel Activity	0.0019	0.0017
4	GO (MF)	13	GO:0046873 Metal Ion Transmembrane Transporter Activity	0.0024	0.0022
4	GO (MF)	14	GO:0015276 Ligand-Gated Ion Channel Activity	0.0058	0.0052
4	GO (MF)	15	GO:0022834 Ligand-Gated Channel Activity	0.0058	0.0052
4	GO (MF)	16	GO:0005516 Calmodulin Binding	0.0058	0.0052
4	GO (MF)	17	GO:0003823 Antigen Binding	0.0127	0.0114
4	GO (MF)	18	GO:0005245 Voltage-Gated Calcium Channel Activity	0.0470	0.0423
5	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	2.7007E-06	2.6480E-06
5	GO (MF)	2	GO:0034987 Immunoglobulin Receptor Binding	9.7092E-06	9.5197E-06
5	GO (MF)	3	GO:0048018 Receptor Ligand Activity	1.6019E-05	1.5707E-05
5	GO (MF)	4	GO:0030546 Signaling Receptor Activator Activity	1.6019E-05	1.5707E-05
5	GO (MF)	5	GO:0008083 Growth Factor Activity	0.0080	0.0078
5	GO (MF)	6	GO:0003823 Antigen Binding	0.0080	0.0078

Continued on Next Page

Table A.4: Differentially Enriched Gene Sets in clusterProfiler ORA

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Adjusted p-Value	q-Value
5	GO (MF)	7	GO:0005179 Hormone Activity	0.0391	0.0384
5	GO (MF)	8	GO:0008528 G Protein-Coupled Peptide Receptor Activity	0.0391	0.0384
5	GO (MF)	9	GO:0071855 Neuropeptide Receptor Binding	0.0391	0.0384
5	GO (MF)	10	GO:0001653 Peptide Receptor Activity	0.0398	0.0390

Table A.5: Differentially Enriched Gene Sets in clusterProfiler DAVID

Phenotype Permutation	Gene Set Database	Nr.	Differentially Enriched Gene Set	Adjusted p-Value	q-Value	Count
1	/	/	/	/	/	/
2	KEGG	1	hsa05322 Systemic Lupus Erythematosus	8.1697E-10	8.3899E-12	20
2	KEGG	2	hsa05034 Alcoholism	1.5798E-08	1.6224E-10	21
2	KEGG	3	hsa04613 Neutrophil Extracellular Trap Formation	9.9561E-08	1.0224E-09	20
3	KEGG	1	hsa05322 Systemic Lupus Erythematosus	6.0097E-17	1.2165E-18	26
3	KEGG	2	hsa05034 Alcoholism	7.3832E-15	1.4946E-16	27
3	KEGG	3	hsa04613 Neutrophil Extracellular Trap Formation	8.3087E-12	1.6819E-13	24
3	KEGG	4	hsa04080 Neuroactive Ligand-Receptor Interaction	2.0517E-02	0.0004	17
3	KEGG	5	hsa05203 Viral Carcinogenesis	3.4181E-02	0.0007	12
4	KEGG	1	hsa05322 Systemic Lupus Erythematosus	1.787E-28	3.0840E-30	39
4	KEGG	2	hsa05034 Alcoholism	3.134E-23	5.4085E-25	39
4	KEGG	3	hsa04613 Neutrophil Extracellular Trap Formation	5.525E-21	9.5349E-23	37
4	KEGG	4	hsa04080 Neuroactive Ligand-Receptor Interaction	4.688E-06	8.0898E-08	28
4	KEGG	5	hsa05203 Viral Carcinogenesis	8.593E-05	1.4829E-06	19
4	KEGG	6	hsa04217 Necroptosis	0.0175	0.0003	13
5	GO (MF)	1	GO:0046982 Protein Heterodimerization Activity	7.1573E-05	1.9783E-06	25
5	GO (MF)	2	GO:0034987 Immunoglobulin Receptor Binding	0.0003	9.6460E-06	12
5	GO (MF)	3	GO:0003823 Antigen Binding	0.0113	0.0003	12
5	GO (MF)	4	GO:0008083 Growth Factor Activity	0.0358	0.0010	12

Table A.6: Differentially Enriched Gene Sets in GSEA Web Application

Phenotype Permutation	Gene Set Database	Nr.	Differentially Enriched Gene Set	Gene Set Size	NES	FDR
1	KEGG	1	Terpenoid Backbone Biosynthesis	15	1.9715	0.1000
1	KEGG	2	Steroid Biosynthesis	17	2.0180	0.1339
1	KEGG	3	Biosynthesis Of Unsaturated Fatty Acids	22	1.8343	0.2264
2	GO (MF)	1	Inorganic Anion Exchanger Activity	15	2.5239	0.1680
2	GO (MF)	2	Dynein Light Intermediate Chain Binding	27	2.1289	0.2376
2	GO (MF)	3	Inorganic Anion Transmembrane Transporter Activity	152	2.1120	0.2380
2	GO (MF)	4	Sodium Ion Transmembrane Transporter Activity	150	2.2685	0.2465
2	GO (MF)	5	Calcium Dependent Cysteine Type Endopeptidase Activity	17	2.1399	0.2485
3	KEGG	1	Prion Diseases	35	1.4059	0.0991
3	KEGG	2	Nicotinate And Nicotinamide Metabolism	24	1.2883	0.1921
3	KEGG	3	Pentose And Glucuronate Interconversions	22	1.3083	0.2067
4	GO (MF)	1	Nucleosomal DNA Binding	40	1.7415	0.1851
4	GO (MF)	2	Phosphatase Activator Activity	23	1.7491	0.1929
4	GO (MF)	3	DNA Polymerase Binding	22	1.6967	0.1978
4	GO (MF)	4	Catalytic Activity Acting On A rRNA	23	1.7187	0.1983
4	GO (MF)	5	Unfolded Protein Binding	118	1.9948	0.1994
4	GO (MF)	6	Protein Folding Chaperone	41	1.7008	0.207
4	GO (MF)	7	Four Way Junction DNA Binding	17	1.6771	0.2135
4	GO (MF)	8	Translation Factor Activity RNA Binding	85	1.75044	0.2146
4	GO (MF)	9	Translation Initiation Factor Activity	51	1.7542	0.2379
4	GO (MF)	10	DNA Secondary Structure Binding	36	1.6199	0.239
4	GO (MF)	11	mRNA 5 Utr Binding	26	1.6233	0.2464
5	/	/	/	/	/	/

Table A.7: Top 10 Differentially Enriched Gene Sets in GSEAPreranked

Phenotype Permutation	Gene Set Database	Nr.	Differentially Enriched Gene Set	Gene Set Size	NES	FDR
1	GO (MF)	1	Cation Channel Activity	329	-6.2520	0.0000
1	GO (MF)	2	Passive Transmembrane Transporter Activity	476	-5.9743	0.0000
1	GO (MF)	3	Gated Channel Activity	339	-5.6448	0.0000
1	GO (MF)	4	Voltage Gated Channel Activity	202	-5.1214	0.0000
1	GO (MF)	5	Metal Ion Transmembrane Transporter Activity	425	-4.8358	0.0000
1	GO (MF)	6	Voltage Gated Cation Channel Activity	143	-4.6671	0.0000
1	GO (MF)	7	Electron Transfer Activity	112	-4.1593	0.0000
1	GO (MF)	8	Oxidoreduction Driven Active Transmembrane Transporter Activity	60	-4.0308	0.0000
1	GO (MF)	9	Potassium Channel Activity	121	-3.8104	0.0000
1	GO (MF)	10	Extracellular Matrix Structural Constituent	172	-3.7962	0.0000
2	GO (MF)	1	Structural Constituent Of Ribosome	163	-6.5915	0
2	GO (MF)	2	Immune Receptor Activity	142	-4.3172	0
2	GO (MF)	3	RRNA Binding	66	-4.1770	0
2	GO (MF)	4	Cytokine Binding	140	-3.2137	0
2	GO (MF)	5	Cytokine Receptor Activity	97	-3.0039	0
2	GO (MF)	6	Gated Channel Activity	339	5.6575	0
2	GO (MF)	7	Metal Ion Transmembrane Transporter Activity	425	5.6348	0
2	GO (MF)	8	Passive Transmembrane Transporter Activity	476	5.4500	0
2	GO (MF)	9	Voltage Gated Channel Activity	202	4.9620	0

Continued on Next Page

Table A.7: Top 10 Differentially Enriched Gene Sets in GSEAPreranked

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Gene Set Size	NES	FDR
2	GO (MF)	10	Voltage Gated Cation Channel Activity	143	4.9096	0
3	KEGG	1	Calcium Signaling Pathway	176	-3.1654	0.0000
3	KEGG	2	Phosphatidylinositol Signaling System	75	-3.0264	0.0000
3	KEGG	3	Systemic Lupus Erythematosus	128	3.2876	0.0000
3	KEGG	4	Erbb Signaling Pathway	87	-2.7294	0.0032
3	KEGG	5	Nod Like Receptor Signaling Pathway	62	-2.5643	0.0046
3	KEGG	6	Fc Epsilon Ri Signaling Pathway	79	-2.5876	0.0047
3	KEGG	7	Type Ii Diabetes Mellitus	47	-2.4408	0.0063
3	KEGG	8	Neuroactive Ligand Receptor Interaction	267	-2.3874	0.0064
3	KEGG	9	Aldosterone Regulated Sodium Reabsorption	42	-2.4028	0.0066
3	KEGG	10	Neurotrophin Signaling Pathway	125	-2.3522	0.0066
4	GO (MF)	1	Protein Heterodimerization Activity	320	-4.6720	0
4	GO (MF)	2	Single Stranded DNA Binding	120	-3.8793	0
4	GO (MF)	3	Structural Constituent Of Ribosome	163	-3.7939	0
4	GO (MF)	4	Ubiquitin Like Protein Ligase Binding	314	-3.5088	0
4	GO (MF)	5	Damaged DNA Binding	68	-3.4345	0
4	GO (MF)	6	Atp Hydrolysis Activity	317	-3.3859	0
4	GO (MF)	7	Ribonucleoprotein Complex Binding	150	-3.3701	0
4	GO (MF)	8	Translation Initiation Factor Activity	51	-3.3125	0

Continued on Next Page

Table A.7: Top 10 Differentially Enriched Gene Sets in GSEAPreranked

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Gene Set Size	NES	FDR
4	GO (MF)	9	Unfolded Protein Binding	113	-3.2716	0
4	GO (MF)	10	Cadherin Binding	332	-3.2372	0
5	GO (MF)	1	Structural Constituent Of Ribosome	163	-6.2912	0
5	GO (MF)	2	R RNA Binding	66	-4.0916	0
5	GO (MF)	3	Catalytic Activity Acting On RNA	389	-3.7734	0
5	GO (MF)	4	Catalytic Activity Acting On A T RNA	124	-3.5942	0
5	GO (MF)	5	Oxidoreduction Driven Active Transmembrane Transporter Activity	60	-3.2590	0
5	GO (MF)	6	Gated Channel Activity	339	4.9070	0
5	GO (MF)	7	Immunoglobulin Receptor Binding	64	4.7984	0
5	GO (MF)	8	Voltage Gated Cation Channel Activity	143	4.5297	0
5	GO (MF)	9	Passive Transmembrane Transporter Activity	476	4.4249	0
5	GO (MF)	10	Voltage Gated Channel Activity	202	4.2316	0

Table A.8: Top 10 Differentially Enriched Gene Sets in clusterProfiler GSEA

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Gene Set Size	NES	adjusted p-Value
1	KEGG	1	hsa00190 Oxidative Phosphorylation	120	-5.1329	2.5692E-09
1	KEGG	2	hsa04020 Calcium Signaling Pathway	238	-4.3841	2.5692E-09
1	KEGG	3	hsa04080 Neuroactive Ligand-Receptor Interaction	355	-4.4548	2.5692E-09
1	KEGG	4	hsa04714 Thermogenesis	218	-4.3786	2.5692E-09
1	KEGG	5	hsa04721 Synaptic Vesicle Cycle	78	-4.5480	2.5692E-09
1	KEGG	6	hsa04723 Retrograde Endocannabinoid Signaling	140	-5.3811	2.5692E-09
1	KEGG	7	hsa05010 Alzheimer Disease	367	-4.6027	2.5692E-09
1	KEGG	8	hsa05012 Parkinson Disease	250	-5.2041	2.5692E-09
1	KEGG	9	hsa05014 Amyotrophic Lateral Sclerosis	349	-4.3484	2.5692E-09
1	KEGG	10	hsa05016 Huntington Disease	292	-5.0897	2.5692E-09
2	GO (MF)	1	GO:0003735 Structural Constituent Of Ribosome	162	-6.9941	6.6833E-09
2	GO (MF)	2	GO:0003823 Antigen Binding	146	-4.1658	6.6833E-09
2	GO (MF)	3	GO:0005216 Ion Channel Activity	424	5.1965	6.6833E-09
2	GO (MF)	4	GO:0005244 Voltage-Gated Ion Channel Activity	196	4.2138	6.6833E-09
2	GO (MF)	5	GO:0005261 Cation Channel Activity	328	4.8489	6.6833E-09
2	GO (MF)	6	GO:0008509 Anion Transmembrane Transporter Activity	456	4.3409	6.6833E-09
2	GO (MF)	7	GO:0015079 Potassium Ion Transmembrane Transporter Activity	151	4.4590	6.6833E-09
2	GO (MF)	8	GO:0015081 Sodium Ion Transmembrane Transporter Activity	146	4.0203	6.6833E-09
2	GO (MF)	9	GO:0015267 Channel Activity	470	5.1929	6.6833E-09

Continued on Next Page

Table A.8: Top 10 Differentially Enriched Gene Sets in clusterProfiler GSEA

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Gene Set Size	NES	adjusted p-Value
2	GO (MF)	10	GO:0015276 Ligand-Gated Ion Channel Activity	139	4.0913	6.6833E-09
3	GO (MF)	1	GO:0005216 Ion Channel Activity	424	-4.0543	6.0150E-08
3	GO (MF)	2	GO:0046873 Metal Ion Transmembrane Transporter Activity	423	-3.9495	6.0150E-08
3	GO (MF)	3	GO:0022803 Passive Transmembrane Transporter Activity	471	-3.7478	6.2841E-07
3	GO (MF)	4	GO:0015267 Channel Activity	470	-3.7349	8.2359E-07
3	GO (MF)	5	GO:0004984 Olfactory Receptor Activity	168	3.5478	2.1765E-06
3	GO (MF)	6	GO:0005261 Cation Channel Activity	328	-3.5354	4.8678E-06
3	GO (MF)	7	GO:0022836 Gated Channel Activity	336	-3.5611	4.8678E-06
3	GO (MF)	8	GO:0015079 Potassium Ion Transmembrane Transporter Activity	151	-3.2723	4.1215E-05
3	GO (MF)	9	GO:0015081 Sodium Ion Transmembrane Transporter Activity	146	-3.1767	9.2551E-05
3	GO (MF)	10	GO:0000149 Snare Binding	113	-3.0744	0.0002
4	GO (MF)	1	GO:0003823 Antigen Binding	146	4.8277	8.5929E-09
4	GO (MF)	2	GO:0005216 Ion Channel Activity	424	4.5561	8.5929E-09
4	GO (MF)	3	GO:0005244 Voltage-Gated Ion Channel Activity	196	4.7392	8.5929E-09
4	GO (MF)	4	GO:0005261 Cation Channel Activity	328	4.4481	8.5929E-09
4	GO (MF)	5	GO:0015079 Potassium Ion Transmembrane Transporter Activity	151	4.1813	8.5929E-09
4	GO (MF)	6	GO:0015267 Channel Activity	470	4.5137	8.5929E-09
4	GO (MF)	7	GO:0022803 Passive Transmembrane Transporter Activity	471	4.5273	8.5929E-09
4	GO (MF)	8	GO:0022832 Voltage-Gated Channel Activity	196	4.7392	8.5929E-09

Continued on Next Page

Table A.8: Top 10 Differentially Enriched Gene Sets in clusterProfiler GSEA

Phenotype Permutation	Gene Set Database	Nr.	Di□erentially Enriched Gene Set	Gene Set Size	NES	adjusted p-Value
4	GO (MF)	9	GO:0022836 Gated Channel Activity	336	4.9937	8.5929E-09
4	GO (MF)	10	GO:0022843 Voltage-Gated Cation Channel Activity	138	5.3255	8.5929E-09
5	GO (MF)	1	GO:0003735 Structural Constituent Of Ribosome	162	-6.5631	1.003E-08
5	GO (MF)	2	GO:0003823 Antigen Binding	146	4.3554	1.003E-08
5	GO (MF)	3	GO:0005216 Ion Channel Activity	424	4.8178	1.003E-08
5	GO (MF)	4	GO:0005244 Voltage-Gated Ion Channel Activity	196	4.2153	1.003E-08
5	GO (MF)	5	GO:0015267 Channel Activity	470	4.5222	1.003E-08
5	GO (MF)	6	GO:0019843 RRNA Binding	64	-4.0857	1.003E-08
5	GO (MF)	7	GO:0022803 Passive Transmembrane Transporter Activity	471	4.5282	1.003E-08
5	GO (MF)	8	GO:0022832 Voltage-Gated Channel Activity	196	4.2153	1.003E-08
5	GO (MF)	9	GO:0022836 Gated Channel Activity	336	4.9131	1.003E-08
5	GO (MF)	10	GO:0022843 Voltage-Gated Cation Channel Activity	138	4.2398	1.003E-08

B. Electronic Appendix

The electronic appendix contains an electronic version of this thesis (`MA_Wuensch.pdf`) as well as three folders. The folder `Code` consists of the `R` codes and the screenshots (for the web tools) to reproduce the results in Chapter 5. Furthermore, the folder `Results` comprises the complete and final result tables of all GSA tools under investigation. Finally, the folder `TCGA Investigation` comprises the `R` code and result tables to reproduce the analysis performed in Section A.2. A detailed description of the electronic appendix is given in the document `README`.

Declaration of authorship

I hereby confirm that I have authored this Master's thesis independently and without use of other resources other than those indicated. The ideas taken directly or indirectly from external sources are duly acknowledged in the text. The material, either full or in part, has not been previously submitted for grading at this or any other academic institution.

Munich, April 14, 2022

Milena Wünsch