Sensitivity And Specificity Of Gene Set Analysis

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Farhad Maleki

Permission to Use

Abstract

High-throughput technologies are widely used for understanding biological processes. Gene set analysis is a well-established computational approach for providing a concise biological interpretation of high-throughput gene expression data. Gene set analysis utilizes the available knowledge about the groups of genes involved in cellular processes or functions. Large collections of such groups of genes, referred to as gene set databases, are available through online repositories to facilitate gene set analysis. There are a large number of gene set analysis methods available, and current recommendations and guidelines about the method of choice for a given experiment are often inconsistent and contradictory. It has also been reported that some gene set analysis methods suffer from a lack of specificity. Furthermore, the sheer size of gene set databases makes it difficult to study these databases and their effect on gene set analysis.

In this thesis, we propose quantitative approaches for the study of reproducibility, sensitivity, and specificity of gene set analysis methods; characterize gene set databases; and offer guidelines for choosing an appropriate gene set database for a given experiment. We review commonly used gene set analysis methods; classify these methods based on their components; describe the underlying requirements and assumptions for each class; suggest the appropriate method to be used for a given experiment; and explain the challenges and pitfalls in interpreting results for each class of methods. We propose a methodology and use it for evaluating the effect of sample size on the results of thirteen gene set analysis methods utilizing real datasets. Further, to investigate the effect of method choice on the results of gene set analysis, we develop a quantitative approach and use it to evaluate ten commonly used gene set analysis methods. We also quantify and visualize gene set overlap and study its effect on the specificity of over-representation analysis. We propose Silver, a quantitative framework for simulating gene expression datasets and evaluating gene set analysis methods without relying on oversimplifying assumptions commonly made when evaluating gene set analysis methods. Finally, we propose a systematic approach to select appropriate gene set databases for conducting gene set analysis for a given experiment. Using this approach, we highlight the drawbacks of meta-databases such as MSigDB, a well-established gene set database made by extracting gene sets from several sources including GO, KEGG, Reactome, and BioCarta.

Our findings suggest that the results of most gene set analysis methods are not reproducible for small sample sizes. In addition, the results of gene set analysis significantly vary depending on the method used, with little to no commonality between the 20 most significant results. We show that there is a significant negative correlation between gene set overlap and the specificity of over-representation analysis. This suggests that gene set overlap should be taken into account when developing and evaluating gene set analysis methods. We show that the datasets synthesized using Silver preserve complex gene-gene correlations and the distribution of expression values. Using Silver provides unbiased insight about how gene set analysis methods behave when applied on real datasets and real gene set databases. Our quantitative study of several well-established gene set databases reveals that commonly used gene set databases fall short in representing some phenotypes.

The proposed methodologies and achieved results in this research reveal the main challenges facing gene set analysis. We identify key factors that contribute to the lack of specificity and reproducibility of gene set analysis methods, establishing the direction for future research. Also, the quantitative methodologies proposed in this thesis facilitate the design and development of gene set analysis methods as well as gene set databases and benefit a wide range of researchers utilizing high-throughput technologies.

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Tony Kusalik for supporting me for the duration of my Ph.D research. His guidance and support in conducting this research have dramatically improved my ability to communicate my research to others effectively.

I would also like to thank my committee: Dr. Ian McQuillan, Dr. Longhai Li, and Dr. Matthew Links for their insightful feedback on my research. They have all played a role in polishing my research and presentation skills. Also, I would like to thank my external examiner who took the time out of their schedule to participate in my defence.

I would also like to express my gratitude to Dr. Alan Rosenberg and Dr. Elham Rezaei, who gave me the opportunity to work with them and utilize their expertise in rheumatology research. In doing so, I was able to demonstrate the value of my Ph.D. research from a biomedical perspective.

My labmates have also been valuable collaborators. In particular, I would like to thank Daniel Hogan and Conor Lazarou for providing feedback on my paper entitled ”Gene Set Overlap: An Impediment to Achieving High Specificity in Over-representation Analysis”. Katie Ovens also contributed to many of the papers resulting from my Ph.D. research and provided help with any evaluation that required the analysis of gene expression studies. I could not ask for better friends and colleagues than those I have been working with during my Ph.D. years in the Bioinformatics lab.

Contents

List of Tables

List of Figures

List of Abbreviations

DNA	Deoxyribonucleic Acid
ES	Enrichment Score
FDR	False Discovery Rate
FSC	Functional Scoring
GEO	Gene Expression Omnibus
GSEA	Gene Set Enrichment Analysis
JIA	Juevenile Idiopathic Arthritis
KEGG	Kyoto Encyclopedia of Genes and Genomes
MDS	Multi-dimensional Scaling
NES	Normalized Enrichment Score
ORA	Over-representation Analysis
PCA	Principal Component Analysis
SEA	Simple Enrichment Analysis
SNR	Signal-to-noise Ratio
TPR	True Positive Rate
TNR	True Negative Rate

1 Introduction

High-throughput technologies are widely used for understanding the biology of an organism. The primary challenge facing the deployment of these technologies is gaining insight from the data generated. Gene expression is often studied using high-throughput technologies. A common approach in gene expression analysis is comparing the gene expression levels for treatment(s) versus control(s) for each gene to find genes with different expression patterns across control and case samples. For each gene, a p-value—which is the probability of obtaining a random gene expression difference as extreme as the observed difference, assuming that the gene is not differentially expressed—is calculated. Then, in order to reduce the number of false positives resulting from multiple comparisons, an adjustment is made. Finally, genes with an adjusted p-value smaller than a given threshold value α are predicted as differentially expressed. This approach is also known as single gene analysis. The main shortcomings of this approach are as follows:

- High-throughput technologies make monitoring a large number of genes in a single experiment possible but introduces the “curse of dimensionality”. In a typical gene expression study, the differential expression of thousands of genes are tested; therefore, the adjustment for multiple comparisons are made for a large number of comparisons. This may result in false negatives [14, 16, 11].

- Selecting a cutoff value in single gene analysis is often challenging. It has been shown that the biological interpretation of gene expression experiments may change depending of the cutoff value being used [13]. Conservative cutoff values might lead to false negatives and relaxed cutoff values might result in false positives [4, 3].

- Cellular processes and functions are often the result of groups of genes working in concert. Single gene analysis ignores this fact and leaves all interpretation to the researcher. The information about genes that correspond to some biological functions or processes is known and available through public knowledge bases such as GO [7], KEGG [9], and OMIM [1]. Single gene analysis does not use such information even though incorporating this information might facilitate gaining biological insight.

- There are many factors that may confound the data resulting from a high-throughput study [17]. These errors may affect the observed change in expression of a single gene and therefore confound the results of single gene analysis. Expression patterns of a set of genes, rather than a single gene, are more robust in the presence of such noise [10, 6].

Gene set analysis, also known as enrichment analysis, is an alternative approach that can alleviate the shortcomings of single gene analysis [11, 15]. In gene set analysis the difference in expression patterns of a group of genes that share a common biological attribute are tested. Such groups of genes are referred to as gene sets and a collection of such gene sets are referred to as a gene set database.

Although gene set analysis aims at alleviating the shortcomings of single gene analysis, it is not a panacea to solve all challenges in high-throughput data analysis, and it comes with its own complications and limitations. In this thesis, we study the shortcomings and limitations of gene set analysis and offer guidelines for choosing a proper gene set analysis method for a given experiment. In addition, we propose a quantitative approach for choosing a sample size for achieving reproducible results for a given gene set analysis method. Further, we develop a quantitative framework for the evaluation of gene set analysis methods using real expression datasets. Our findings establish a direction for further research on developing gene set analysis methods.

This thesis is composed of several papers each tailored to focus on one research goal. Chapter 2 reviews the literature on gene set analysis, lays out the main shortcomings of commonly used gene set analysis methods and sets the direction for future research. Chapter 3 proposes a quantitative methodology to study the effect of sample size on the reproducibility of gene set analysis. Using the proposed method, we investigate the impact of sample size on the reproducibility of thirteen widely used gene set analysis methods. In Chapter 4, we propose an approach for evaluating gene set analysis methods using real expression datasets and compare ten well-established gene set analysis methods. We quantify and visualize gene set overlap in several gene set databases, which are specifically designed for gene set analysis methods, in Chapter 5. We also assess the hypothesis that there is a negative correlation between gene set overlap and the specificity of over-representation analysis. In Chapter 6, we discuss that the main reason behind the lack of consensus in the research community about the gene set analysis method to be used for a given experiment is the absence of gold standard datasets, i.e. expression datasets with their list of differentially enriched gene sets known *a priori*. We propose Silver, a framework for synthesizing expression datasets without relying on oversimplifying assumptions commonly used when evaluating gene set analysis methods. Silver also provides a methodology for evaluating a gene set analysis method using quantitative measures such as sensitivity, specificity, and accuracy. Gene set databases are a crucial component of gene set analysis. In Chapter 7, we propose a systematic approach to select appropriate gene set databases for conducting gene set analysis for a given experiment. Using this approach, we study well-established gene set databases such as Gene Ontology [2], KEGG [8], Reactome [5], and BioCarta [12]. Finally, Chapter 8 contains a summary and suggestions for future research. Each chapter of this thesis consists of a journal or conference paper. Some of these papers have been published (Chapters 3, 4, and 5), and some are in the final stages of preparation for submission (Chapters 2, 6, and 7). For the chapters consisting of published papers, there is an introductory section added linking the paper to the bulk of the thesis. Appendix F describes the copyright licensing for the papers and includes the corresponding documents. Also for some of the published papers, page limitations meant that there were more research results than could be described in the paper. Those additional research results appear as appendices to the thesis, and are described in the relevant chapter.

References

[1] Joanna Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. McKusick's online mendelian inheritance in man (OMIM®). *Nucleic Acids Research*, 37(suppl 1):D793–D796, 2009.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[3] Yoram Ben-Shaul, Hagai Bergman, and Hermona Soreq. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, 21(7):1129–1137, 2005.

[4] Thomas Breslin, Patrik Edén, and Morten Krogh. Comparing functional annotation analyses with Catmap. *BMC Bioinformatics*, 5(1):193, 2004.

[5] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, 2013.

[6] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, pages 107–129, 2007.

[7] Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.

[8] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[9] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2015.

[10] Seon-Young Kim and David J Volsky. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(1):144, 2005.

[11] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003.

[12] Darryl Nishimura. Biocarta. *Biotech Software & Internet Report*, 2(3):117–120, 2001.

[13] Kuang-Hung Pan, Chih-Jian Lih, and Stanley N Cohen. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 102(25):8961–8965, 2005.

[14] Raghavakaimal Sreekumar, Panagiotis Halvatsiotis, Jill Coenen Schimke, and K Sreekumaran Nair. Gene expression profile in skeletal muscle of type 2 diabetes and the effect of insulin treatment. *Diabetes*, 51(6):1913–1920, 2002.

[15] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric 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 of the United States of America*, 102(43):15545–15550, 2005.

[16] X Yang, R Pratley, S Tokraks, C Bogardus, and P Permana. Microarray profiling of skeletal muscle tissues from equally obese, non-diabetic insulin-sensitive and insulin-resistant Pima indians. *Diabetologia*, 45(11):1584–1593, 2002.

[17] Wei Zhang, Ilya Shmulevich, and Jaakko Astola. *Microarray Quality Control.* John Wiley & Sons, 2005.

2 Background

Gene set analysis methods differ in their various components such as gene score, gene set score, the null hypothesis, and the method of significance assessment. This chapter reviews and classifies gene set analysis methods based on their components. We discuss the strengths and shortcomings of each class of gene set analysis methods and establish the direction for further research in gene set analysis. An extended version of this chapter will be submitted to *Frontiers in Genetics*.

Abstract

Gene set analysis methods are widely used to provide insight into high-throughput gene expression data. These methods rely on various assumptions and have different requirements, strengths and weaknesses. In this paper, we classify gene set analysis methods based on their components, describe the underlying requirements and assumptions for each class, suggest the appropriate method to be used for a given experiment, and describe the challenges and pitfalls in interpreting results for each class of method.

2.1 Introduction

High-throughput technologies such as DNA microarrays and RNA-Seq are widely used to monitor the activity of thousands of genes in a single experiment. The primary challenge facing the deployment of these technologies is to gain biological insight from the generated data.

The early approach for analysing gene expression data is single gene analysis, where expression measures of each gene for case and control samples are compared using a statistical test such as t-test or Wilcoxon rank-sum test and a p-value is calculated. Then, in order to reduce the number of false positives resulting from multiple comparisons, an adjustment for multiple comparisons is made. Next, genes with an adjusted p-value smaller than a given threshold are predicted as being differentially expressed. Finally, a biological interpretation is attempted using these genes.

Single gene analysis suffers from several shortcomings. In a high-throughput gene expression study, many single-gene tests are typically performed. Consequently, adjustment for multiple comparisons is performed for a large number of genes. Such adjustments may lead to many false negatives by detecting very few or even no gene as being differentially expressed [45, 55, 41]. This issue is more pronounced when using conservative methods, such as Bonferroni and Šídák for multiple comparisons adjustment [11].

In the single-gene approach often researchers use arbitrary cutoff values to choose a reasonable number of genes for further study and interpretation. Different choices of threshold value may lead to different biological interpretations [43]. Conservative threshold values may cause false negatives and relaxed thresholds may cause false positives [8, 6].

Further, cellular processes are often associated with changes in the expression patterns of groups of genes that share common biological functions or attributes. A meaningful change in a group of these genes is more biologically reliable and interpretable than a change in a single gene. *A priori* knowledge about some of these sets of genes is available through public online databases such as GO [9], KEGG [28], and OMIM [3]. The single-gene approach simply disregards this information. Incorporating this information in the data analysis may provide valuable insight about underlying biological processes or functions.

Although high-throughput technologies make the monitoring of expression of thousands of genes in a single experiment possible, they introduce a challenge of dealing with high dimensional data, often referred to as the "curse of dimensionality" [7]. To deal with high dimensional data dimensionality reduction methods are used for downstream analysis and visualization. Relying on sets of biologically related genes is the most intuitive and biologically relevant approach to dimensionality reduction in high-throughput gene expression studies.

When differences in measured values for a single gene across treatments are subtle, the single-gene approach makes it difficult to differentiate the true difference in gene expression from the difference due to biological variability of samples [41, 46]. Gene set analysis, on the other hand, might be able to detect such subtle but concordant changes in expression patterns of genes within a gene set.

Further, the single-gene approach may report several hundred to a few thousand genes as being differentially expressed. Interpreting a long list of differentially expressed genes is a cumbersome task prone to investigator bias toward a hypothesis of interest.

Gene set analysis, also known as enrichment analysis, is an attempt to resolve these shortcomings and to gain insight from gene expression data. The primary aim of gene set analysis is to identify enrichment or depletion of expression levels of a given set of genes of interest, referred to as a gene set.

There are a large number of gene set analysis methods available [21]. These methods differ in their various components such as their underlying assumptions, notion of differential enrichment, null hypothesis, and their significance assessment procedure. Study of gene set analysis methods based on their components helps to understand the strengths and weaknesses of each category of methods, select an appropriate method for a given experiment, facilitate the interpretation of the outcomes of gene set analysis, and develop new methods with higher sensitivity and specificity.

In this paper, we review gene set analysis methods based on their various components, classify different approaches based on their components, characterize the strengths and shortcomings of each approach, and describe the underlying requirements and assumptions for each class of methods. We also suggest the appropriate method to be used for a given experiment, and describe the challenges and pitfalls in interpreting results for each class of gene set analysis methods. In addition, we set direction for further research in gene set analysis.

2.2 Gene set analysis

Data from a high-throughput case-control experiment can be organised in an expression matrix. This matrix is generated by joining the corresponding expression values for all samples in the experiment. Each column of the matrix corresponds to the expression measures for one sample and each row of the matrix corresponds to the expression measures for one gene across all samples. This expression matrix is the base for expression analysis including single gene and gene set analysis. Figure 2.1 shows an expression matrix with $\|C\|$ control samples and $\|T\|$ case samples.

There are many gene set analysis methods available. Over-representation analysis (ORA), functional scor-

$$\begin{array}{c}
\begin{array}{cccccc}
A^{(c_1)} & \dots & A^{(c_{\|C\|})} & A^{(t_1)} & \dots & A^{(t_{\|T\|})}
\end{array} \\
\begin{bmatrix}
g_1^{(c_1)} & \dots & g_1^{(c_{\|C\|})} & g_1^{(t_1)} & \dots & g_1^{(t_{\|T\|})} \\
g_2^{(c_1)} & \dots & g_2^{(c_{\|C\|})} & g_2^{(t_1)} & \dots & g_2^{(t_{\|T\|})} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
g_{m-1}^{(c_1)} & \dots & g_{m-1}^{(c_{\|C\|})} & g_{m-1}^{(t_1)} & \dots & g_{m-1}^{(t_{\|T\|})} \\
g_m^{(c_1)} & \dots & g_m^{(c_{\|C\|})} & g_m^{(t_1)} & \dots & g_m^{(t_{\|T\|})}
\end{bmatrix}
\end{array}$$

Figure 2.1: Expression matrix for a pairwise comparison where $A^{(c_1)}, \dots, A^{(c_{\|C\|})}$ columns represent control samples and $A^{(t_1)}, \dots, A^{(t_{\|T\|})}$ columns represent case samples. c_i represents the index of a sample in the control group, and t_j represents the index of a sample in the case/treatment group.

ing (FCS), and pathway topology-based methods are three main categories of gene set analysis methods [31]. In this paper, we focus on ORA and FCS methods that comprise the main body of gene set analysis methods used by researchers. For a review and a comparison of topology-based methods refer to Mitrea et al. [40] and Ihnatova et al. [25].

2.2.1 Over-representation analysis

ORA is the natural extension of single gene analysis and one of the most widely used gene set analysis methods. Due to its simplicity, well-established underlying statistical model, and ease of implementation, ORA is available through many bioinformatics toolkits. Huang et al. listed 68 gene set analysis methods and tools of which 40 are ORA-based. These tools differ in their various components such as gene set databases, data visualization, and user interfaces [21]. ORA uses a list L of genes predicted as being differentially expressed by a single gene analysis method.

Given L and a gene set G_i that has n_i' genes in common with L, ORA considers G_i as being differentially enriched if the occurrence of n_i' differentially expressed genes in G_i is unlikely to be due to chance. Table 2.1 illustrates the contingency table representation for the over-representation of differentially expressed genes in G_i given L and U. The set of n genes under study is called the reference set or background set and depicted by U, and $\overline{G_i}$ is the complement of G_i with respect to U.

Under the null hypothesis that there is no association between differential expression and membership in G_i, we can assume that G_i is the result of a simple random sampling of $\|G_i\|$ genes from U; therefore, the probability of having n_i' differentially expressed genes within G_i can be calculated using the hypergeometric distribution as follows [11]:

$$f(n_i'; n, \|G_i\|, \|L\|) = \frac{\binom{\|G_i\|}{n_i'} \times \binom{n-\|G_i\|}{\|L\|-n_i'}}{\binom{n}{\|L\|}} \tag{2.1}$$

Table 2.1: Representation of ORA as a contingency table. Each cell contains a count of genes satisfying the condition associated with the row and column.

	Genes in L	Genes not in L	Total
Genes in G_i	n'_i	$\|G_i\| - n'_i$	$\|G_i\|$
Genes in $\overline{G_i}$	$\|L\| - n'_i$	$n - \|G_i\| - (\|L\| - n'_i)$	$n - \|G_i\|$
Total	$\|L\|$	$n - \|L\|$	n

The significance of the association between genes in G_i and genes in L can be assessed using Fisher's exact test, as follows:

$$p = \sum_{j=n'_i}^{\|G_i\|} f(j; n, \|G_i\|, \|L\|) = 1 - \sum_{j=0}^{n'_i - 1} f(j; n, \|G_i\|, \|L\|) \tag{2.2}$$

Fisher's exact test gives the exact p-value for both small and large cell counts in Table 2.1. Less computationally demanding alternatives can also be used to calculate an approximation for the p-value.

Binomial distribution can be used to estimate the p-value for Fisher's exact test [12]. For large values of n, the hypergeometric distribution tends to a binomial distribution. The binomial estimation of Equation 5.1 is as follows:

$$f_b(n'_i; \|L\|, \frac{\|G\|}{n}) = \binom{\|L\|}{n'_i} \times \left(\frac{\|G_i\|}{n}\right)^{n'_i} \times \left(1 - \frac{\|G_i\|}{n}\right)^{\|L\| - n'_i} \tag{2.3}$$

Therefore, Equation 5.2 can be estimated as follows:

$$p = 1 - \sum_{j=0}^{i-1} f_b(j; \|L\|, \frac{\|G_i\|}{n}) \tag{2.4}$$

where f_b in Equation 2.3 and 2.4 represents the binomial distribution density function.

Another alternative to estimate the p-value is the χ^2 test for equality of proportions [53]. This test has also been used in the context of gene set enrichment analysis [12, 56, 30].

2.2.2 Functional scoring methods

The main assumptions of ORA are that genes are independent and equally effective in biological processes. Although these assumptions simplify the problem modelling, they are not biologically valid. It is well-established that genes, proteins, and other biomolecules often act in concert [51]. In addition, ORA disregards quantitative information for all genes that are not predicted as differentially expressed. Usually, genes predicted as differentially expressed are the result of applying a p-value cutoff, and all the quantitative measures for genes with a p-value greater than the cut-off value are disregarded. However, a consistent change even with a p-value slightly greater than the cutoff value may contribute to the detection of pathway activities.

In contrast to ORA, the main goal of FCS methods is to use all information from an expression matrix to address the enrichment problem without relying on the aforementioned biologically invalid assumptions.

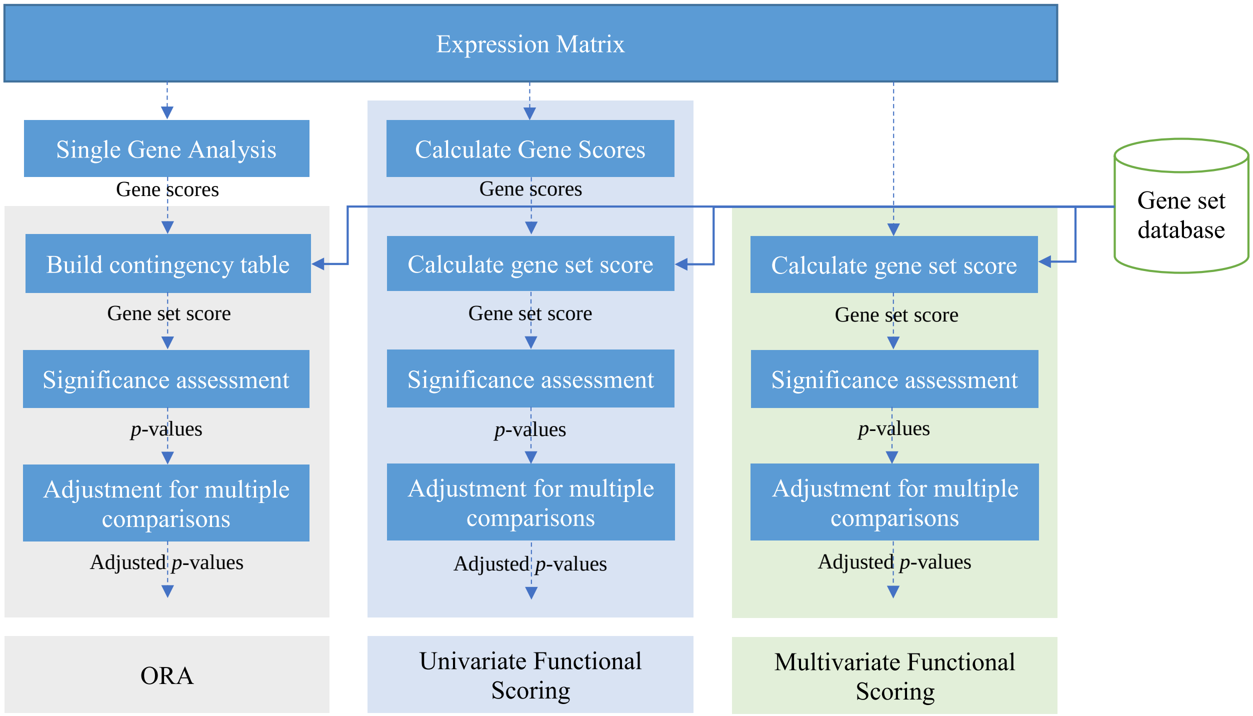


Figure 2.2: A schematic view of over-representation analysis (ORA) and univariate and multivariate FCS methods.

Therefore, FCS methods—instead of working with a list of differentially expressed genes—take advantage of an expression matrix of gene expression measures for all genes to discern differential enrichment of gene sets.

There are many FCS methods available [41, 19, 46, 32, 52, 5, 50, 39, 35, 4, 54, 49, 20, 44]. These methods can be categorized into two classes: univariate and multivariate methods. In univariate FCS methods, usually a per gene score is calculated for each gene using each row of the expression matrix. Then these per gene scores are used to calculate a gene set score for each gene set. Finally, the significance of the gene set scores is assessed and differentially enriched gene sets are reported. Multivariate methods skip the step for calculating gene scores and directly calculate gene set scores from the expression matrix. Figure 2.2 illustrates a schematic view of univariate and multivariate FCS methods and also ORA methods.

An FCS method often consists of a set of common components such as a per gene score that is a statistic that summarizes the expression level of a gene across control and case samples; a per gene set score that summarizes the expression level of genes within a gene set as a single statistic; a procedure for significance assessment; and an adjustment for multiple comparisons.

Univariate functional scoring methods

GSEA [41] is the most widely used univariate FCS method. GSEA uses a gene score such as signal-to-noise ratio (SNR) difference between gene expression measures in control and case samples to calculate a per gene score. The signal-to-noise ratio difference used in GSEA is as follows [47]:

$$SNR(g_i) = \frac{\frac{\sum_{j=1}^{\|C\|} g_i^{(c_j)}}{\|C\|} - \frac{\sum_{j=1}^{\|T\|} g_i^{(t_j)}}{\|T\|}}{\sigma'_{c,i} + \sigma'_{t,i}} \tag{2.5}$$

$$\sigma'_{c,i} = \max\left(\sigma\left(g_i^{(c_1)}, \dots, g_i^{(c_{\|C\|})}\right), 0.2 \times \frac{\sum_{j=1}^{\|C\|} g_i^{(c_j)}}{\|C\|}\right)$$

where $g_i^{(c_j)}$ is the gene expression level for gene g_i in sample $A^{(c_j)}$ (see Figure. 2.1); $\sigma'_{c,i}$ is the standard deviation of expression levels for gene g_i among control samples; $g_i^{(t_j)}$ and $\sigma'_{t,i}$ are defined analogously using case samples.

GSEA ranks all genes according to their scores. Then to measure the association between members of a given gene set G_i and treatments, it calculates a gene set score—also referred to as the enrichment score (*ES*) in GSEA terminology—using a Kolmogorov-Smirnov statistic. The *ES* value for G_i, denoted as $ES(G_i)$, is calculated using a running sum initialized as 0. Assume $g_1, \dots, g_n$ is the sorted list of all genes under study according to SNR difference in decreasing order. For each gene in the sorted list starting with the first one the running sum (enrichment score) is updated by adding a value of $+\sqrt{\frac{n-\|G_i\|}{\|G_i\|}}$ when the gene belongs to G_i and by subtracting a value of $\sqrt{\frac{\|G_i\|}{n-\|G_i\|}}$ when the gene does not belong to G_i [41]. The *ES* value is calculated "as the maximum observed positive deviation of the running sum" [41], as shown in Equation 2.6.

$$ES(G_i) = \max_{1 \le l \le n} \sum_{k=1}^{l} x_k \tag{2.6}$$

$$x_k = \begin{cases} +\sqrt{\frac{n-\|G_i\|}{\|G_i\|}} & g_k \in G_i \\ \\ -\sqrt{\frac{\|G_i\|}{n-\|G_i\|}} & g_k \notin G_i \end{cases}$$

After calculation of the actual *ES* value for all gene sets, the method determines the maximum *ES*, denoted as *MES*. The significance of the calculated *MES* value is assessed using a permutation test. The sample labels are permuted 1000 times, and for each permutation a *MES* value is calculated. Finally, the significance of *MES* of the actual data is calculated as the fraction of permutations that lead to an *MES* higher than the *MES* of the actual data.

It should be mentioned that the significance of the *MES* does not provide any insight about the significance of the enrichment score of a given gene set G_i, although this is the main purpose of enrichment analysis. In fact, assessing the significance of the *MES* tests the null hypothesis that "no gene set is associated with the class distinction" [41], where the rank ordering is used as the measure of association. Therefore, rejection of this null hypothesis only suggests that there is at least one gene set for which the rank ordering of its members is associated with the sample classes, i.e. phenotypes.

The method suggested by Mootha et al. [41] as inferred from Equation 2.6 calculates *ES* as the "maximum observed positive deviation of the running sum" [41] and, therefore, it does not detect differential enrichment of gene sets that have the majority of their genes up-regulated unless the phenotypes are swapped and the GSEA procedure is run again. Hence, GSEA should be considered as a one-sided test [50]. In addition, in order to be able to rely on the enrichment scores, the significance of each ES should be assessed. However, the method proposed by Mootha et al. [41] tested the null hypothesis that "no gene set is associated with the class distinction" which is not extendable to the ES for each gene set.

Damian et al. [10] raised concern about the capabilities of GSEA by suggesting a synthesized example. They showed that GSEA may ignore highly enriched gene sets solely due to the size of gene sets. In their hypothetical example they assumed that there is a given dataset of gene expression values for genes in three gene sets G_1, G_2, and G_3 of size n, $5n$, and $4n$, respectively, where—after calculation of per gene scores and sorting them—genes in G_1 ranked higher than genes in G_2, and genes in G_2 ranked higher than genes in G_3. In order to better demonstrate the problem, assume that G_1 is the only enriched gene set with all genes being down-regulated, and also G_2 and G_3 are not differentially enriched. GSEA assigns enrichment scores of $3n$, $4n$, and 0, respectively to G_1, G_2, and G_3. Therefore, G_2 is preferred to G_1, although G_1 is the only enriched gene set. Furthermore, Subramanian et al. [46] reported that GSEA leads to high enrichment scores for gene sets clustered around the middle of the sorted list of all genes. These gene sets are often not associated with the phenotypes under study [46].

Considering these shortcomings, Tian et al. [50] suggested using the t-test or Wilcoxon rank-sum test statistics as alternative gene set scores instead of the Kolmogorov-Smirnov statistic in GSEA. They suggested that these scores are able to detect moderate but coordinated shift from the background distribution. In order to generate the background distribution, they used both gene sampling and phenotype permutation (see Section 2.3). In fact, instead of testing differences in distribution of per gene scores across treatments, they tested a location change, i.e. shift in mean or median. The shortcoming of the method is its lack of sensitivity in detecting a differentially enriched gene set where some of its genes are up-regulated and some down-regulated [26]. This is due to the inherent inability of the average to detect those effects.

PAGE, a parametric method for gene set enrichment analysis, was proposed as a statistically more sensitive and computationally less demanding alternative for gene set enrichment analysis [32]. PAGE tests the null hypothesis that "all genes in a given microarray dataset are independent of each other and identically distributed, that is they are not co-regulated" [32]. It uses fold change between sample groups, i.e. treatments, to calculate a Z-score for a given gene set G_i. The significance of this Z-score is then calculated using a normal distribution. PAGE starts with calculating the fold change value of each gene as the per gene score. Next, it calculates mean (μ) and standard deviation (σ) of all fold change values. Then, for a given gene set G_i, it calculates μ_i as the average fold change value of genes in G_i. After that, a score Z_i is calculated as follows:

$$Z_i = \frac{\mu_i - \mu}{\frac{\sigma}{\|G_i\|}} \tag{2.7}$$

Finally, the significance of Z_i is assessed using the standard normal distribution. The rationale behind using the normal distribution is that according to the Central Limit Theorem [14], the sampling distribution of the average of an independent random variable for large sample sizes is normal, regardless of the distribution of the sampling population. Therefore, the distribution of average fold change values for gene sets should be normal. This method has been reported to achieve a high sensitivity while suffering from a low specificity.

In another attempt to address the aforementioned shortcomings of GSEA, Subramanian et al. [46]—almost the same group who proposed GSEA—adjusted the method by using a weighted Kolmogorov-Smirnov statistic as the gene set score. They also used False Discovery Rate (FDR) to adjust for multiple comparisons [46]. The outline of the method is as follows. First, the per gene score for each gene is calculated. Assuming $g_1, \ldots, g_n$ is the list of all genes sorted according to the per gene score, the gene set score is calculated as follows:

$$\begin{aligned} ES(G_i) &= \max_{1 \le k \le n} \left(P_{hit}(G_i, k) - P_{miss}(G_i, k)\right) \\ P_{hit}(G_i, k) &= \sum_{\substack{g_t \in G_i \\ t \le k}} \frac{|r_t|^p}{R(G_i)} \\ R(G_i) &= \sum_{g_t \in G_i} |r_t|^p \\ P_{miss}(G_i, k) &= \sum_{\substack{g_t \notin G_i \\ t \le k}} \frac{1}{n - \|G_i\|} \end{aligned} \tag{2.8}$$

where p is a positive constant and a parameter of the method; r_t is the per gene score for the t^{th} gene in the sorted list. The enrichment scores are normalized to account for gene set size. The significance of the normalized enrichment scores (NES) is assessed using gene sampling or phenotype permutation (see Section 2.3). Finally, an adjustment for multiple comparisons is made.

It should be mentioned that the enrichment score in the adjusted GSEA is similar to, but not the same as, the enrichment score in GSEA. To calculate the enrichment scores, both methods calculate a running sum by traversing the list of all genes ranked according to their per gene scores. For each gene in the list, the original GSEA method updates the running sum by a constant value, while the adjusted GSEA increases the running sum with a value of $\frac{|r_t|^p}{\sum_{g_t \in G_i} |r_t|^p}$ for genes in G_i. This increases the effect of genes with higher absolute value of the per gene score ($|r_t|^p$), i.e. genes at the beginning or at the end of the ordered list, and to decrease the effect of genes in the middle of the ordered list on gene set scores. Hereafter, we use GSEA to refer to the updated GSEA, unless stated otherwise. GSEA is still a one-sided test. In addition, it is not obvious how GSEA addresses the effect of gene set size, as it was reported to affect the results of the original GSEA [10]. Further, an ad hoc choice of 1 for p has been used in the updated version of GSEA.

Irizarry et al.[26] proposed the use of a simple parametric method as an alternative to GSEA. They mentioned that GSEA is based on a Kolmogorov-Smirnov test which is known for its lack of sensitivity. In order to avoid using a Kolmogorov-Smirnov test statistic and also a permutation test, which is computationally demanding, they suggested using a parametric method that employs standard normal distribution to assess the significance of each enrichment score. They used the two-sample t-test statistic as the per gene score to measure the degree of association between each gene and phenotypes. For a given gene g, this value is denoted by $t(g)$. They evaluated the assumption of normality of $t(g)$ values for all genes using a Q-Q plot for 8 datasets—all datasets used by Subramanian et al. [46] and Mootha et al. [41]. Based on the observed Q-Q plots, they suggested that assuming standard normal distribution for the distribution of $t(g)$ values in practice is valid. For a given gene set G_i, they suggested a Z-score as follows:

$$\begin{aligned} \textit{Z-Score}(G_i) &= \sqrt{\|G_i\|} \times \bar{t}(G_i) \qquad (2.9) \\ \bar{t}(G_i) &= \frac{\sum_{g \in G_i} t(g)}{\|G_i\|} \end{aligned}$$

By accepting the assumption that the t-test statistic has standard normal distribution and also ignoring the correlation between gene set members, they inferred that the Z-score has standard normal distribution as well. Therefore, they assessed the significance of Z-scores using standard normal distribution. Hereafter, we refer to this method as SEA.

Irizarry et al. [26] admitted that a limitation of the proposed Z-score is that it may not be able to detect gene sets where almost half of the genes are up-regulated and the rest are down-regulated. To deal with this issue, they suggested a standardized χ^2-test score as follows:

$$\chi^2\textit{-score}(G_i) = \frac{\sum_{g \in G_i} \left(t(g) - \bar{t}\,(G_i)\right)^2 - \left(\|G_i\| - 1\right)}{2\left(\|G_i\| - 1\right)} \qquad (2.10)$$

They approximated the distribution of χ^2-score for a gene set of large size, e.g. 20 or higher, using the standard normal distribution to calculate the significance of the gene set score.

Tamayo et al. [47] refuted the claim made by Irizarry et al. [26] that their simple enrichment analysis method outperforms GSEA [46]. They focused on the assumption made by Irizarry et al. [26] to ignore gene-gene correlation, questioning its practicality and whether it is realistic. Comparing the results of SEA and GSEA, they reported that SEA uniformly produces more significant gene sets. For example, they reported that in a pancreas dataset [1], SEA predicted 42% of gene sets as significantly differentially enriched, a number almost 5 times more than that from GSEA.

In addition, Tamayo et al. [47], using the approach of Gatti et al. [16], tested the effect of gene-gene correlation on the results of GSEA and SEA, where there is no significant correlation structure between gene profiles and phenotypes. In this regard, for each dataset, they produced results for both SEA and GSEA for 1000 datasets with random permutation of phenotype labels in the expression profile. Since after random permutations of gene profile labels there is almost no relation between gene profiles and phenotypes, we expect almost no significant gene set to be reported as differentially enriched by gene set enrichment analysis methods. Tamayo et al. reported that while GSEA predicted almost 0% of gene sets as differentially enriched, SEA predicted many gene sets as differentially enriched.

Jiang and Gentleman [27] suggested several per gene and per gene-set scores as extensions to gene set enrichment analysis. They suggested a linear model for calculating a per gene score. Equation 2.11 shows the linear model.

$$Y_{g,i} = \mu_g + \beta_g X_i + \epsilon_{g,i} \qquad (2.11)$$

where $Y_{g,i}$ is the measured expression value for gene g from the i^{th} sample; for a given gene g and i $(1 \leq i \leq n)$, $\epsilon_{g,i}$ is assumed to be a independent normally distributed variable with mean of zero; X_i is a binary variable showing phenotype, i.e. class, of the i^{th} sample. For a given gene g, μ_g represents the mean of expression measures for the phenotype corresponding to $X_i = 0$, and β_g represents the difference between mean of expression measures of g for the phenotype corresponding to $X_i = 1$ and μ_g. They used $\frac{\hat{\beta}_g}{s_g}$ as the per gene score, where $\hat{\beta}_g$ is the estimate of β and s_g is an estimate for standard deviation of expression measurements for gene g. In addition, they suggested using median and also a signed-test, which is a non-parametric test to assess consistent differences in paired samples, as alternatives to the per gene set statistic. The sign-test was used to assess the prevalence of up- or down-regulation of genes within a gene set, regardless of the magnitude of this regulation. They found a lack of sensitivity when using the sign test as gene set score. Also, they suggested that median is less susceptible to outlier effects in comparison to using mean as a gene set score.

Multivariate functional scoring methods

Multivariate FCS methods, unlike univariate FCS methods, directly calculate gene set scores from expression data and skip the intermediate step of calculating gene scores (see Figure 2.2).

Goeman et al. [19] proposed the GlobalTest method, based on a generalized linear model, to address the question whether the global expression patterns of genes in a given gene set G_i is significantly associated with a biological outcome of interest. The outcome of interest may be a binary group label representing two experimental conditions or a continuous variable. The idea behind the GlobalTest method is that if genes in a given gene set G_i can be used to correctly predict a biological outcome, then genes in G_i should have different expression patterns for different outcomes. In GlobalTest, the expression profile of genes in G_i across samples is represented using a matrix X, where $X_{k,j}$ is the expression value of the j^{th} gene of G_i in the k^{th} sample; the biological outcome of interest is represented as an $n \times 1$ vector, where Y_k is the outcome of interest for the k^{th} sample. In a pairwise comparison of phenotypes, Y_k is a binary value representing the phenotype of the k^{th} sample. In order to model the relation between X and Y, GlobalTest uses the following generalized linear model:

$$E(Y_i \mid \beta) = h^{-1}(\alpha + \sum_{i=1}^{m} \beta_j x_{i,j}) \tag{2.12}$$

where β_j $(1 \leq j \leq m)$ is the regression coefficient for the expression value of gene g_j; α is an intercept value; h is a link function that can be the identity function resulting in a linear regression model, or *logit* function resulting in a logistic regression model. In order to test if genes in G_i are able to predict the biological outcome, the following null hypothesis should be tested:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

Considering the fact that the number of samples is usually less than the number of variables, i.e. gene set size $\|G_i\|$, this null hypothesis cannot be tested in a classical way. In order to address this problem, Goeman et al. accepted the simplifying assumption that the regression coefficients all come from the same distribution with a mean of zero and an unknown variance of τ^2. In this case, the aforementioned null hypothesis is equivalent to the following null hypothesis.

$$H_0 : \tau^2 = 0$$

An implementation of the GlobalTest method is available as an R-package from Bioconductor [17]. The implementation of GlobalTest uses a diagonal covariance matrix, which means that the correlation between genes in a given gene set is ignored [2].

Analysis of covariance, another multivariate statistical method, has been used in gene set enrichment analysis [38, 22]. In addition, Kong et al. [33] used Hotelling's T^2-test for enrichment analysis. This test is the natural generalization of the t-test for testing the difference between multivariate means of two populations. The test statistic for a given gene set G_i is as follows:

$$T^2 = (\bar{X}_C - \bar{X}_T)^{tr}(S\frac{n_1 + n_2}{n_1 n_2})^{-1}(\bar{X}_C - \bar{X}_T) \tag{2.13}$$

where $\bar{X}_C$ and $\bar{X}_T$ are the mean expression vectors of genes in the gene set for control and treatment samples, respectively; n_1 and n_2 are the number of control and treatment samples, respectively; tr denotes the matrix transpose operator. Under the null hypothesis and when $n > m + 1$, the following statistic has an F-distribution with m and $n - m - 1$ degrees of freedom, where m is the size of the gene set and $n = n_1 + n_2$:

$$\frac{n - m - 1}{(n - 2)m} T^2 \tag{2.14}$$

Since m, i.e. gene set size, is often bigger than n, i.e. sample size, Kong et al. [33] employed singular value decomposition for dimension reduction to be able to use this approach.

2.3 Significance assessment of gene set score

Gene sampling, phenotype permutation, parametric methods, and dynamic programming are four main approaches that have been used to assess the significance of gene set scores. In this section we study these approaches.

In gene sampling the significance of a gene set score for a given gene set G_i is assessed by comparing it to the scores of randomly assembled sets of $\|G_i\|$ genes from the reference set U, i.e. all genes under study. In this method, a large number of random gene sets are assembled, and their scores are calculated. Then the significance value of the gene set score of G_i is calculated as the fraction of assembled gene sets that lead to stronger scores than the score of G_i, where a score in comparison to another is considered stronger if it is more in favour of rejecting the null hypothesis of interest. Since gene sampling does not depend on the number of samples, it has been widely used for gene set analysis [46, 50, 2]. The main shortcoming of gene sampling is that it relies on the unrealistic assumption of independence between genes within a gene set. Usually genes within a gene set show a highly correlated behaviour; therefore, the gene sampling method may incorrectly predict a gene set as differentially enriched only because of high correlation between its genes. In this regard, it may cause false positive predictions. Another shortcoming of gene sampling is being computationally demanding. For each gene set G_i, the whole process of gene set score calculation should be repeated for a large number of randomly assembled gene sets. In implementations of gene-sampling approaches, usually the number of assembled gene sets is an order of magnitude of 1000. This number of repetitions makes the significance evaluation computationally demanding. Moreover, gene sampling may lead to a lack of statistical reliability of the significance values for large gene sets [29]. Even using an order of magnitude of 1000 assembled gene sets may not be big enough to represent the background distribution; therefore, the significance value for large gene sets may not be statistically reliable.

Phenotype permutation, also known as sample permutation, assesses the significance of a gene set score of a given gene set G_i by permuting sample labels. First, the gene set score of G_i is calculated. Let S_{G_i} denote the gene set score of G_i according to the actual gene expression profile. Then a large number of expression profiles are synthesized by permuting the sample labels, i.e. the column labels of the actual expression profile. For a synthesized expression profile, we expect no association between the expression patterns of genes in G_i and the phenotypes. Next, for each synthesized expression profile, the gene set score of G_i is calculated. Finally, the significance of S_{G_i} is calculated as the fraction of the synthesized expression profiles that lead to a stronger score than S_{G_i}, where a score in comparison to another is considered stronger if it is more in favour of rejecting the null hypothesis of interest.

Phenotype permutation, unlike gene sampling, does not rely on the unrealistic assumption of gene independence, but it requires a large number of samples with almost equal number of samples for each phenotype. This condition most often is not satisfied. Instead, due to limited budgets, lack of technicians, or ethical conduct in animal and human research, having a large number of samples is not a choice for many researchers. In some cases, like for rare diseases, having a large sample size is not possible at all. Therefore, phenotype permutation is not applicable, and some gene set analysis tools provide gene sampling as an alternative to phenotype permutation [46].

The parametric approach is another way to assess the significance of gene set scores [32, 26]. In this approach, first, a gene set score is proposed. Then, under the null hypothesis and by accepting some simplifying assumptions, a parametric distribution for the gene set statistic is proposed. Finally, the parametric distribution is used to assess the significance of gene set statistics. Although this approach, unlike gene sampling and phenotype permutation, is not computationally demanding, it has been criticized as being too simplistic and unable to detect truly differentially enriched gene sets [47].

Keller et al. [29] used a dynamic programming approach to assess the significance of the enrichment score used in the method proposed by Mootha et al. [41]. Their dynamic programming approach assessed the significance of the gene set scores derived from the unweighted Kolmogorov-Smirnov statistic. For a given array containing n genes and a given gene set G_i, first, they calculated the gene set score RS_{G_i}. Then they calculated its p-values as the probability of obtaining a gene set score equal to or greater than RS_{G_i}, assuming that there is no association between the distribution of genes in G_i and the phenotypes. Since there are $\binom{n}{\|G_i\|}$ enrichment scores possible [29], they calculated the number of enrichment scores less than RS_{G_i} and then used the following formula to calculate the p-values:

$$\textit{p-value}(RS_{G_i}) = 1 - \frac{\text{The number of enrichment scores that are less than } RS_{G_i}}{\binom{n}{\|G_i\|}} \tag{2.15}$$

In order to calculate the number of enrichment scores that are less than RS_{G_i} using a dynamic programming approach, they initialized a $2\|G_i\|(n-\|G_i\|+1) \times (n+1)$ matrix M, where the different rows—indexed from $-(n - \|G_i\|) \times \|G_i\|$ to $(n - \|G_i\|) \times \|G_i\|$—represent all possible running sum scores. They initialized $M(0,0) = 1$ and the rest of the elements of M as 0. Starting from the second column ($k = 1$), they updated all elements of the matrix, column by column, according to Equation 2.16.

$$M(j,k) = \begin{cases} M(j - n + \|G_i\|, k-1) + M(j + \|G_i\|, k-1) & \text{if } -|RS| < j < |RS_{G_i}| \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

Finally, $M(0,n)$ was reported as the number of enrichment scores with a maximum deviation smaller than RS_{G_i}. Keller et al. [29] suggested that their proposed dynamic programming approach is more efficient than the permutation approach and that their method does not suffer from the statistically unreliable results produced by the permutation method, when the number of permutations is not large enough. They claimed that their approach is almost 10 times faster than phenotype permutation and gene sampling. It should be mentioned that the main shortcoming of this approach is that, unlike permutation approaches, it is not extendable to other gene set scores such as weighted Kolmogorov-Smirnov statistic in GSEA.

2.4 Null hypothesis in gene set enrichment analysis

Defining a null hypothesis is an essential step in conducting any statistical inference. Different null hypotheses have been used in gene set enrichment analysis: Competitive null hypothesis [18], self-contained null hypothesis [18], and hybrid null hypothesis [2]. Understanding the implications of these null hypotheses is essential for having a valid interpretation of the the result of gene set analysis. In this section, we discuss the limitations and requirements of each class of hypothesis.

For a given gene set G_i, a competitive null hypothesis states that genes in G_i do not have a different expression pattern in comparison to the rest of the genes under study. The competitive approach, for detecting differentially enriched gene sets, compares the expression patterns of genes in G_i with the rest of the genes under study. To conduct the hypothesis test, a gene sampling approach is used. Consequently, the competitive approach has been criticized for using genes as sampling units, whereas the purpose of the experiment is to detect changes across phenotypes [18, 2]. It also has been criticized for ignoring the correlation between genes within a gene set. Therefore, methods based on the competitive approach may detect a gene set as being differentially enriched just because of the correlation between its genes [18, 2].

For a given gene set G_i, a self-contained null hypothesis states that genes in G_i do not have a different expression pattern across phenotypes. To test this hypothesis, methods based on the self-contained approach use phenotype permutation. Consequently, testing a self-contained null hypothesis leads to preserving the complex correlation of genes within a gene set. However, it requires a large and almost equal number of samples for each phenotype. This condition may not be met by many experiments.

The hybrid null hypotheses are a subclass of the self-contained null hypotheses. Hybrid null hypotheses state that genes in G_i are not differentially associated with a phenotype of interest, where the measure of association is defined according to all genes under study. Methods based on this approach calculate a gene set score for a given gene set G_i using all genes, i.e. genes in $G_i \cup \bar{G}_i$; then they assess the significance of this score in a similar way to the self-contained approach. Therefore, testing these hypotheses requires a large and almost equal number of samples for each phenotype. GSEA and its variants, which are based on Kolmogorov-Smirnov statistic, use the hybrid null hypotheses [41, 46, 23].

2.5 Discussion

Despite having a well-established underlying statistical model, ORA suffers from several shortcomings. ORA relies on the gene-gene independence assumption that is known to be biologically invalid [16, 48]. Also, ORA uses a list of differentially expressed genes as input and treats all genes equally regardless of their magnitude of differential expression. Differentially expressed genes are determined using a single gene analysis method, where the use of arbitrary thresholds is often a common practice. It has been shown that the choice of these thresholds might affect the result of the downstream analysis [43]. ORA is also incapable of detecting low but concordant signals, i.e. below the used threshold, from genes within a gene set. These concordant signals are believed to be biologically important.

FCS methods aim at solving some of these problems. There are many FCS methods available, but there is no consensus among researchers about the method of choice for a given experiment [18, 34, 2, 26, 15, 24, 47]. Therefore, a systematic methodology for evaluation of gene set analysis is of great value for the research community.

A gold standard dataset for evaluation of gene set analysis methods requires the enrichment status of all gene sets in a given gene set database to be known *a priori*. Currently, such a dataset is not available; therefore, simulated datasets have been widely used to evaluate gene set analysis methods [13, 42, 2]. These simulated datasets often rely on oversimplifying assumptions that are not capable of capturing the true nature of gene expression and gene set collections. Most often expression data have been simulated by ignoring the complex correlations between genes within gene sets [13, 42, 2]. Moreover, gene set collections have usually been simulated to be a small number of non-overlapping sets of equal size, a situation that is substantially different from the real gene set databases. Due to oversimplifying assumptions, evaluation of gene set analysis using these datasets has led to inconsistent and contradictory results [36]. Synthesizing datasets that preserve the true nature of gene expression data and gene set databases is an essential step in the evaluation of new and existing gene set analysis methods.

Competitive gene set analysis methods rely on gene sampling for the significance assessment. Gene sampling is based on the assumption that genes are independent. This assumption is known to be biologically invalid and may cause some gene sets to be predicted as being differentially enriched solely due to the correlation between its genes. This issue introduces false positives and decreases the specificity of these methods. Self-contained gene set analysis methods also suffer from a lack of specificity.

Lack of specificity of gene set analysis methods is one of the main drawbacks to gaining insight from the results of gene set analysis. One of the main reasons behind this issue is gene set overlap which is an integral part of gene set databases [37]. For example, if the null hypothesis for a self-contained method is that there is no difference in the average expression of genes in a gene set between case and control samples, a significant change in the expression of a gene within a gene set might result in differential enrichment of that gene set. The problem arises in the presence of gene set overlap. In the presence of gene set overlap, such gene(s) may occur in several gene sets; therefore, differential expression of a gene or a few genes may cause a large number of gene sets to be predicted as being differentially enriched. Hence, gene set overlap seems to be the main challenge facing these methods [37]. There have been several attempts to alleviate the effect of gene set overlap [49, 44]. Although these methods lead to a higher specificity, they suffer from low sensitivity. Further research in developing gene set analysis methods that increase specificity without sacrificing sensitivity is required.

Self-contained gene set analysis methods rely on phenotype permutation for significance assessment; therefore, in an experiment where the number of samples is small, the calculated significance score by phenotype permutation is not reliable; therefore, the use of these methods is highly discouraged. For example, in a case-control study where there are 3 control and 3 case samples, the total number of unique ways that one can divide these 6 samples into two groups each of size 3 is 20; therefore, the smallest non-zero p-values achievable is 0.5. Competitive methods should be preferred for such experiments.

2.6 Conclusion

In this paper, we reviewed a set of well-established gene set analysis methods. We discussed the shortcoming and strengths of gene set analysis method based on their various components such as their gene set score, null hypothesis, and their methods of significance assessment. We also set the direction for conducting further research in gene set analysis.

A systematic methodology for evaluating gene set analysis methods can help to resolve the lack of consensus about the method of choice for a given experiment. A method for synthesizing expression datasets that represent the characteristics of real expression data and developing benchmark datasets based on this method will prove useful. In addition, developing benchmark datasets using such methods will facilitate the evaluation of gene set analysis methods. A quantitative study of characteristics of gene set databases, such as gene set overlap, can help researchers better understand the effect of these characteristics on the results of a given gene set analysis method and develop new methods that achieve higher specificity without losing sensitivity.

References

[1] Amir Abdollahi, Christian Schwager, Jörg Kleeff, Irene Esposito, Sophie Domhan, Peter Peschke, Kai Hauser, Philip Hahnfeldt, Lynn Hlatky, Jürgen Debus, et al. Transcriptional network governing the angiogenic switch in human pancreatic cancer. *Proceedings of the National Academy of Sciences*, 104(31):12890–12895, 2007.

[2] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, 2009.

[3] Joanna Amberger, Carol A Bocchini, Alan F Scott, and Ada Hamosh. Mckusick's online mendelian inheritance in man (OMIM®). *Nucleic Acids Research*, 37(suppl 1):D793–D796, 2009.

[4] David A Barbie et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108, 2009.

[5] William T Barry, Andrew B Nobel, and Fred A Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.

[6] Yoram Ben-Shaul, Hagai Bergman, and Hermona Soreq. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, 21(7):1129–1137, 2005.

[7] Daniel P Berrar, Werner Dubitzky, and Martin Granzow. *A practical approach to microarray data analysis*. Springer, 2003.

[8] Thomas Breslin, Patrik Edén, and Morten Krogh. Comparing functional annotation analyses with Catmap. *BMC Bioinformatics*, 5(1):193, 2004.

[9] Gene Ontology Consortium et al. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.

[10] Doris Damian and Malka Gorfine. Statistical concerns about the GSEA procedure. *Nature Genetics*, 36(7):663–663, 2004.

[11] Sorin Drăghici. *Statistics and data analysis for microarrays using R and bioconductor*. CRC Press, 2016.

[12] Sorin Drăghici, Purvesh Khatri, Rui P Martins, G Charles Ostermeier, and Stephen A Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.

[13] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.

[14] John E Freund, Irwin Miller, and Marylees Miller. *John E. Freund's Mathematical Statistics: With Applications*. Pearson Education India, 2004.

[15] Brooke L Fridley, Gregory D Jenkins, and Joanna M Biernacka. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One*, 5(9):e12693, 2010.

[16] Daniel M Gatti, William T Barry, Andrew B Nobel, Ivan Rusyn, and Fred A Wright. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics*, 11(1):574, 2010.

[17] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):1–16, 2004.

[18] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[19] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[20] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14(1):7, 2013.

[21] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[22] Manuela Hummel, Reinhard Meister, and Ulrich Mansmann. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, 24(1):78–85, 2008.

[23] Jui-Hung Hung, Troy W Whitfield, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, Charles DeLisi, et al. Identification of functional modules that correlate with phenotypic difference: the influence of network topology. *Genome Biology*, 11(2):R23, 2010.

[24] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, 13(3):281–291, 2011.

[25] Ivana Ihnatova, Vlad Popovici, and Eva Budinska. A critical comparison of topology-based pathway analysis methods. *PLOS One*, 13(1):e0191154, 2018.

[26] Rafael A Irizarry, Chi Wang, Yun Zhou, and Terence P Speed. Gene set enrichment analysis made simple. *Statistical Methods in Medical Research*, 18(6):565–575, 2009.

[27] Zhen Jiang and Robert Gentleman. Extensions to gene set enrichment. *Bioinformatics*, 23(3):306–313, 2007.

[28] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 2015.

[29] Andreas Keller, Christina Backes, and Hans-Peter Lenhof. Computation of significance scores of unweighted gene set enrichment analyses. *BMC Bioinformatics*, 8(1):290, 2007.

[30] Purvesh Khatri, Sorin Draghici, G Charles Ostermeier, and Stephen A Krawetz. Profiling gene expression using onto-express. *Genomics*, 79(2):266–270, 2002.

[31] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Computational Biology*, 8(2):e1002375, 2012.

[32] Seon-Young Kim and David J Volsky. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(1):144, 2005.

[33] Sek Won Kong, William T Pu, and Peter J Park. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, 22(19):2373–2380, 2006.

[34] Qi Liu, Irina Dinu, Adeniyi J Adewale, John D Potter, and Yutaka Yasui. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, 8(1):431, 2007.

[35] Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson, and Peter J Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1):161, 2009.

[36] Henryk Maciejewski. Gene set analysis methods: statistical models and methodological differences. *Briefings in Bioinformatics*, 15(4):504–518, 2013.

[37] Farhad Maleki and Anthony J. Kusalik. Gene set overlap: An impediment to achieving high specificity in over-representation analysis. In *12th International joint conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*, Prague, Czech Republic, February 2019.

[38] U Mansmann, R Meister, et al. Testing differential gene expression in functional groups goeman's global test versus an ancova approach. *Methods of Information in Medicine*, 44(3):449–453, 2005.

[39] Joëlle Michaud, Ken M Simpson, Robert Escher, Karine Buchet-Poyau, Tim Beissbarth, Catherine Carmichael, Matthew E Ritchie, Frédéric Schütz, Ping Cannon, Marjorie Liu, et al. Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics*, 9(1):363, 2008.

[40] Cristina Mitrea, Zeinab Taghavi, Behzad Bokanizad, Samer Hanoudi, Rebecca Tagett, Michele Donato, Calin Voichita, and Sorin Draghici. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, 4:278, 2013.

[41] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003.

[42] Dougu Nam and Seon-Young Kim. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9(3):189–197, 2008.

[43] Kuang-Hung Pan, Chih-Jian Lih, and Stanley N Cohen. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 102(25):8961–8965, 2005.

[44] Cedric Simillion, Robin Liechti, Heidi E.L. Lischer, Vassilios Ioannidis, and Rémy Bruggmann. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics*, 18(1):151, 2017.

[45] Raghavakaimal Sreekumar, Panagiotis Halvatsiotis, Jill Coenen Schimke, and K Sreekumaran Nair. Gene expression profile in skeletal muscle of type 2 diabetes and the effect of insulin treatment. *Diabetes*, 51(6):1913–1920, 2002.

[46] Aravind Subramanian 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.

[47] Pablo Tamayo, George Steinhardt, Arthur Liberzon, and Jill P Mesirov. The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, 25(1):472–487, 2016.

[48] Adi L Tarca, Gaurav Bhatti, and Roberto Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLOS One*, 8(11):e79217, 2013.

[49] Adi Laurentiu Tarca, Sorin Draghici, Gaurav Bhatti, and Roberto Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):136, 2012.

[50] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–13549, 2005.

[51] Charles A Tilford and Nathan O Siemers. Gene set enrichment analysis. In *Protein Networks and Pathway Analysis*, pages 99–121. Springer, 2009.

[52] John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(1):225, 2005.

[53] Gerald Van Belle, Lloyd D Fisher, Patrick J Heagerty, and Thomas Lumley. *Biostatistics: a methodology for the health sciences*, volume 519. John Wiley & Sons, 2004.

[54] Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133–e133, 2012.

[55] X Yang, R Pratley, S Tokraks, C Bogardus, and P Permana. Microarray profiling of skeletal muscle tissues from equally obese, non-diabetic insulin-sensitive and insulin-resistant pima indians. *Diabetologia*, 45(11):1584–1593, 2002.

[56] Sheng Zhong, Kai-Florian Storch, Ovidiu Lipan, Ming-Chih J Kao, Charles J Weitz, and Wing H Wong. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Applied Bioinformatics*, 3(4):261–264, 2004.

3 Size Matters: How Sample Size Affects the Reproducibility and Specificity of Gene Set Analysis

Irreproducibility of high-throughput experiments has been a major challenge for the research community. Small sample size is known to be one of the major causes behind the irreproducibility of high-throughput experiments. Despite the existence of general guidelines, there is no quantitative approach for suggesting a sufficient sample size to be used for a given gene set analysis method. Therefore, gene set analysis methods have often been proposed without providing specific guidance regarding the required number of samples. This chapter presents a quantitative approach for such an evaluation.

An initial version of this work was accepted as a regular paper at the “2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM18)” with 534 full paper submissions and an acceptance rate of 19.6%. An extended version of this paper has been subsequently invited for consideration by “BMC Human Genomics”. In this Chapter, we present the extended version submitted to “BMC Human Genomics”.

Citation

Maleki, F., Ovens, K., McQuillan, I., & Kusalik, A. J. (Dec. 6, 2018). Sample Size and Reproducibility of Gene Set Analysis. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 122-129). IEEE. Madrid, Spain. doi:10.1109/BIBM.2018.8621462

Maleki, F., Ovens, K., McQuillan, I., & Kusalik, A. J. Size Matters: How Sample Size Affects the Reproducibility and Specificity of Gene Set Analysis. BMC Human Genomics (Submitted on April 28, 2018).

Author contributions

Farhad Maleki designed and developed the methodology for assessing the effect of sample size on the result of gene set analysis methods, implemented the pipeline for performing gene set analysis using 13 methods, performed the statistical analysis and data visualization, and wrote the paper (except “Data” Subsection). Katie Ovens preprocessed microarray datasets and performed single gene analysis and wrote the “Data” subsection. The automation of the pipeline for running all experiments was done by Katie Ovens. Ian

McQuillan helped with review and revision of the paper. Anthony Kusalik supervised the research and helped edit and revise the paper.

Abstract

Gene set analysis is a well-established approach for interpretation of data from high-throughput gene expression studies. Achieving reproducible results is an essential requirement in such studies. One factor of a gene expression experiment that can affect reproducibility is the choice of sample size. However, choosing an appropriate sample size can be difficult, especially because the choice may be method-dependent. Further, sample size choice can have unexpected effects on specificity.

In this paper, we report on a systematic, quantitative approach to study the effect of sample size on the reproducibility of the results from 13 gene set analysis methods. We also investigate the impact of sample size on the specificity of these methods. Rather than relying on synthetic data, the proposed approach uses real expression datasets to offer an accurate and reliable evaluation.

Our findings show that, as a general pattern, the results of gene set analysis become more reproducible as sample size increases. However, the extent of reproducibility and the rate at which it increases vary from method to method. In addition, even in the absence of differential expression, some gene set analysis methods report a large number of false positives, and increasing sample size does not lead to reducing these false positives. The results of this research can be used when selecting a gene set analysis method from those available.

3.1 Introduction

The choice of sample size is an important decision to make when designing a gene expression experiment. Choosing an appropriate sample size for obtaining a desired statistical power is feasible for basic statistical procedures; however, making such choices for complex procedures such as gene set analysis is not straightforward. To the best of our knowledge, there is no methodological approach to determine the optimal sample size for reaching a predetermined statistical power or achieving reproducible results in gene set analysis, where sample size refers to the number of biological replicates per treatment, tissue, or condition. Consequently, researchers either use the largest possible number of samples considering available resources—such as funding, specimens, and technicians—for conducting the experiments, or they use an arbitrary sample size—as small as two or three samples per treatment. Using unnecessarily large sample sizes wastes resources and might involve ethical concerns. On the other hand, using small sample sizes may yield unreliable and irreproducible results.

The impact of the number of used samples on the results of differential expression analyses has been studied. Tsai et al. [27] suggested a methodology for sample size estimation. They assumed an equal standardized effect size and a constant gene-gene correlation for differentially expressed genes. Relying on these assumptions, they estimated an appropriate sample size as that which led to the highest number of true positives using a beta-binomial distribution for the two-sample z-test. Their proposed approach, as they reported, might underestimate the number of required samples when the gene-gene correlation is not constant. Stretch et al. [20] reported that using a small number of samples may lead to irreproducible results in differential expression studies. Schurch et al. [19] evaluated 11 tools for differential expression analysis using a dataset with 48 controls and 48 cases. The results of these methods when using subsets of 3 controls and 3 cases were compared to the results when using all samples. They reported that for 8 methods only 20 to 40% of the differentially expressed genes were among the genes reported when using all samples. Furthermore, they suggested that to increase this percentage to a value larger than 85%, at least 20 samples per treatment are required.

Gene expression analysis typically reports several hundred genes as differentially expressed. Biological interpretation of such a large number of genes is laborious and prone to investigator bias(es) in favour of, or against the hypothesis under study. The main aim of gene set analysis—also known as enrichment analysis—is to alleviate these problems. Many gene set analysis methods are available. These methods, unlike basic statistical tests, are complex procedures; therefore, estimating sample size for obtaining a predetermined statistical power or reproducible results is challenging.

In this paper, we extend an earlier work on sample size and reproducibility in the context of gene set analysis [15]. We study a comprehensive list of 13 gene set analysis methods: PAGE [11], GAGE [13], Camera [30], ROAST [29], FRY (from the R package *limma*) [17], GSEA (both gene permutation (GSEA-G) and sample permutation (GSEA-S) versions) [21], ssGSEA [3], GSVA [10], PLAGE [26], GlobalTest [9], PADOG [25], and over-representation analysis (ORA) [6]. All of the methods can be used for pairwise comparison of phenotypes or treatments (e.g. case versus control). Using real datasets, we evaluate the reproducibility of the results of these methods across sample sizes that are commonly used in gene expression studies.

In addition, we assess the specificity of gene set analysis methods across sample sizes. Tarca et al. [24] evaluated the specificity of gene set analysis methods by calculating the number of false positives for datasets generated from permutations of sample/phenotype labels of actual expression datasets. Since after permutation of sample labels there should be no association between differential enrichment of gene sets and phenotypes, all gene sets predicted as differentially enriched by a method were considered as false positives. However, each group of samples corresponding to a condition (case or control) in their generated datasets contained a mixture of both case and control samples from the original dataset; therefore, the characteristics of the generated dataset could be different from that of the original dataset. Furthermore, they did not evaluate the specificity of gene set analysis methods across sample sizes. To avoid the shortcoming of the approach used by Tarca et al., we conduct an experiment by generating datasets of various sizes, where case and control samples for each generated dataset are the result of sampling without replacement from control samples of an actual expression dataset. Since in the generated datasets control and case samples are from actual controls, any gene set predicted as differentially enriched by a method can be considered a false positive.

3.2 Data and methodology

3.2.1 Data

Repositories such as Gene Expression Omnibus (GEO) [7] and ArrayExpress [18] make large scale expression datasets publicly available. In this research, three such case-control datasets with unrelated phenotypes from the *Affymetrix GeneChip Human Genome U133 Plus 2.0* microarray platform were obtained from GEO: 1) renal cell carcinoma tissue (77 controls and 77 cases, GSE53757) [28], 2) gingival tissues (64 controls and 183 cases, GSE10334) [5], and 3) skin tissue in psoriasis patients (64 controls and 58 cases, GSE13355) [22]. The raw data were preprocessed with the *GEOquery v2.46.15* R package and normalized with *justRMA* normalization from the *affy v1.56.0* package.

MDS (multidimensional scaling) plots in Figures B.1, B.2, and B.3 in Supplementary Materials visualize the similarity between samples, respectively, for datasets GSE53757, GSE10334, and GSE13355. These plots illustrate that dataset GSE10334 has the lowest intra-class similarity and less distinction between control and case samples, while dataset GSE13355 has highest intra-class similarity and shows a clear distinction between case and control samples. All visualizations in this paper are produced for the dataset with intermediate characteristics, GSE53757, unless otherwise noted.

Probe IDs were converted to their corresponding Entrez gene identifiers using the *hgu133plus2.db v3.2.3* R package. To avoid over-emphasizing genes with a large number of probes on the arrays, it is a common practice in gene set analysis to collapse duplicate IDs. This was accomplished by using the *collapseRows* function from *WGCNA v1.61* with the *MaxMean* method that selects the probe that has the maximum average value across samples when multiple probes map to the same gene. Collapsing the probes resulted in 20,514 genes in each experiment from an initial 54,675 probes.

3.2.2 Methodology

A proper study of the effect of sample size on the results of gene set analysis methods requires conducting expression studies with datasets of various sizes. These studies must be conducted while all potentially confounding factors such as phenotype under study, the platform for measuring gene expression, experiment protocol, laboratory technician skill level, and environmental conditions stay constant. Datasets of various sizes for which these factors are constant is not currently available. In this research, we utilize a systematic, quantitative approach to study the effect of sample size on the reproducibility and specificity of gene set analysis methods using large scale publicly available gene expression datasets.

Assume that D is a dataset containing n_C control samples and n_T case samples, where both n_C and n_T are relatively large numbers (> 50). Given an integer n, where $n < n_C$, and $n < n_T$, we generate a balanced case-control dataset by randomly selecting n samples from the n_C controls of D and n samples from the n_T case samples of D. The random sampling is performed without replacement; therefore, the chosen samples are unique within the generated dataset. Hereafter, we refer to such a balanced case-control dataset as a replicate dataset of size $2 \times n$ and the entire process for assembling a replicate dataset as the data generation procedure. Also, to avoid confusion, we refer to D as the original dataset.

To obtain results that do not depend on a specific composition of samples, for each given n, we repeat the dataset generation procedure m times to construct m replicate datasets of size $2 \times n$. These replicate datasets are then used for downstream analysis. Due to the nature of random sampling, these replicate datasets are different.

The dataset generation procedure assembles replicate datasets from an original dataset; therefore, confounding factors remain almost invariable. For instance, all these replicate datasets have the same platform and use the same experiment protocol; they also have been made by the same technician(s). This makes it possible to study the effect of sample size on the result of gene set analysis methods while keeping the confounding factors nearly constant.

Different gene set analysis methods then are applied to replicate datasets of various sample sizes and their results are used to investigate the effect of sample size on the reproducibility and specificity of gene set analysis.

In this research, 13 widely used gene set analysis methods are studied—PAGE [11], GAGE [13], Camera [30], ROAST [29], FRY [17], GSEA [21], ssGSEA [3], GSVA [10], PLAGE [26], PADOG [25], GlobalTest [9], and ORA [6]. The following R packages are used for conducting gene set analysis. PLAGE, GSVA, and ssGSEA are obtained from *GSVA* package version *1.18.0*; the *phyper* method from the *stats* package version *3.4.4* is used to implement ORA; the *GSEA.1.0.R* script downloaded from the Broad Institute software page for GSEA are used to run GSEA-S and GSEA-G; ROAST, FRY, and Camera are run using the *limma* package version *3.34.9*; GAGE and PAGE are obtained from the *gage* package version *2.20.1*. For each sample size n ($n \in \{3, \ldots, 20\}$), $m = 10$ replicate datasets of size $2 \times n$ are generated. Then a gene set analysis method is applied to each replicate dataset. For all gene set analysis methods the default parameters are used. A Benjamini-Hochberg correction [4] with a false discovery rate of 0.05 is performed for a fair comparison across methods. Also, the GO gene sets are extracted from *MSigDB* version *6.1* [21] and used as the gene set database for all experiments. Hereafter, this database is referred to as $\mathbb{G}$.

After generating $D_1^{(2\times n)}, \ldots, D_m^{(2\times n)}$, each replicate dataset $D_i^{(2\times n)}$ $(1 \leq i \leq m)$ and the gene set database $\mathbb{G}$ are used as inputs to a gene set analysis method ψ, and the results after correction for multiple comparisons are stored in a vector $R^{\psi}_{D_i^{(2\times n)}}$. The k^{th} component of this vector is the adjusted p-value resulting from testing the differential enrichment of the k^{th} gene set of $\mathbb{G}$; therefore, the length of $R^{\psi}_{D_i^{(2\times n)}}$ is equal to the number of gene sets in $\mathbb{G}$.

The differential enrichment status of the k^{th} gene set in $\mathbb{G}$ is determined by comparing the k^{th} component of $R^{\psi}_{D_i^{(2\times n)}}$ against a significance level $\alpha = 0.05$. For each element of $R^{\psi}_{D_i^{(2\times n)}}$ that is less than α, the corresponding gene set in $\mathbb{G}$ is considered as differentially enriched, and non-differentially enriched otherwise. We denote the set of all gene sets predicted as differentially enriched by $S^{\psi}_{D_i^{(2\times n)}}$.

We use the Jaccard similarity coefficient, also known as Jaccard index [2], to quantify the reproducibility of the results of a gene set analysis method ψ when applied to replicate datasets $D_i^{(2\times n)}$ and $D_j^{(2\times n)}$. The Jaccard similarity coefficient is defined as follows:

$$J(S^{\psi}_{D_i^{(2\times n)}}, S^{\psi}_{D_j^{(2\times n)}}) = \frac{S^{\psi}_{D_i^{(2\times n)}} \cap S^{\psi}_{D_j^{(2\times n)}}}{S^{\psi}_{D_i^{(2\times n)}} \cup S^{\psi}_{D_j^{(2\times n)}}} \tag{3.1}$$

A Jaccard similarity coefficient of 0 indicates no overlap, i.e. no agreement, between $S^{\psi}_{D_i^{(2\times n)}}$ and $S^{\psi}_{D_j^{(2\times n)}}$, and a value of 1 indicates complete overlap, i.e. $S^{\psi}_{D_i^{(2\times n)}} = S^{\psi}_{D_j^{(2\times n)}}$. Hereafter, we refer to the Jaccard similarity coefficient as the overlap score.

For each pair of replicate datasets $D_i^{(2\times n)}$ and $D_j^{(2\times n)}$ $(1 \leq i, j \leq m)$, we calculate $J(S^{\psi}_{D_i^{(2\times n)}}, S^{\psi}_{D_j^{(2\times n)}})$, the overlap between the sets of gene sets predicted as differentially enriched by method ψ when analysing $D_i^{(2\times n)}$ and $D_j^{(2\times n)}$. The resulting overlap score is stored in position (i, j) of an upper triangular matrix, which is called an overlap matrix and visualized in Section 3.3. Since $J(S^{\psi}_{D_i^{(2\times n)}}, S^{\psi}_{D_j^{(2\times n)}}) = J(S^{\psi}_{D_j^{(2\times n)}}, S^{\psi}_{D_i^{(2\times n)}})$, for each sample size $2 \times n$ we need to calculate $\frac{m\times(m-1)}{2}$ overlap scores. Overlap scores $J(S^{\psi}_{D_i^{(2\times n)}}, S^{\psi}_{D_j^{(2\times n)}})$ $(1 \leq i, j \leq m)$ indicate the extent to which the results of a gene set analysis method is reproducible when analyzing replicate datasets of size $2 \times n$. High overlap indicates that method ψ using datasets of size $2 \times n$ yield reproducible results. For each method ψ, we conduct a Kruskal-Wallis test to statistically assess if there is a significant difference between the overlap scores across sample sizes $(3 \leq n \leq 20)$. The overlap scores for each sample size $2 \times n$ is represented as a multiset $P^{\psi}_{(2\times n)}$, which is a set but with repetition allowed. $P^{\psi}_{(2\times n)}$ is defined as follows:

$$P^{\psi}_{(2\times n)} = \{J(S^{\psi}_{D_i^{(2\times n)}}, S^{\psi}_{D_j^{(2\times n)}}) \mid 1 \leq i < j \leq m\} \tag{3.2}$$

The adjusted p-values resulting from the gene set analysis of a given expression dataset are often sorted based on their adjusted p-values. Then gene sets with the smallest adjusted p-values are considered for further investigation and interpretation. Therefore, not only is the differential enrichment status of gene sets important but also the order of their significance. To assess the agreement in the order of gene sets reported as differentially enriched when analyzing replicate datasets of the same size using a method ψ, we use Kendall's coefficient of concordance [2]. The Kendall's coefficient of concordance ranges between 0 and 1, with 0 indicating no agreement and 1 indicating complete agreement, i.e. the same order of gene sets when sorted by their adjusted p-values. Since we aim to quantify the agreement in the order of gene sets predicted as differentially enriched across replicate datasets of the same size, we only consider gene sets that are predicted as differentially enriched for at least one replicate dataset.

Further, we compare the overlap between the results of a gene set analysis method ψ when applied to dataset $D_{2\times n}$ and D. This is done to evaluate if the dataset with the smaller sample size is enough to reproduce the results when using the whole dataset D. Therefore, we construct a multiset $W^{\psi}_{(2\times n)}$ of m overlap scores, as follows:

$$W^{\psi}_{(2\times n)} = \{J(S^{\psi}_{D_i^{(2\times n)}}, S^{\psi}_{D}) \mid 1 \leq i \leq m\} \tag{3.3}$$

High overlap scores suggest that a sample size of $2 \times n$ might be enough for obtaining equivalent results as those achieved using the whole dataset.

Also, to evaluate the specificity of gene set analysis methods, we conduct an additional experiment, referred to as the control-control experiment. The procedure for generating control-control replicate datasets in this experiment is similar to that of the balanced case-control replicate datasets with the sole difference being that only actual control samples are used for constructing the replicate datasets. In other words, each replicate dataset for the control-control experiment is made by random sampling (without replacement) of n samples from the n_C controls of D (where $n < \frac{n_C}{2}$) and another n samples from the n_C control samples of D. The former group of n samples are considered controls in the replicated dataset and the latter group are considered cases.

3.3 Experimental results

To visualize the change in reproducibility of a gene set analysis method across sample sizes, we utilize an arrangement of modified heat maps, hereafter referred to as a pine plot. Each triangular heat map in a pine plot—referred to as a layer—represents values above the diagonal in an overlap matrix (as described in Methodology). Each layer visualizes the overlap scores calculated based on the results of the gene set analysis of replicate datasets of the same sample size. The colour intensity of a cell (i, j) in each layer represents $J(S^{\psi}_{D_i^{(2\times n)}}, S^{\psi}_{D_j^{(2\times n)}})$, which is the overlap between the result of gene set analysis of two replicate datasets ($S^{\psi}_{D_i^{(2\times n)}}$ and $S^{\psi}_{D_j^{(2\times n)}}$). When $i = j$, a value of 1 is assigned to the cell (i, i) of the overlap matrix, which is represented with a red colour and serves as a visual reference point.

The pine plots in Figures 3.1 and 3.2 depict the change in reproducibility of results from GSEA-S, GSEA-G, ORA, and GAGE for replicate datasets of size 2×3, 2×5, 2×10, 2×15, and 2×20, where $2 \times n$ represents the size of a dataset with n controls and n cases. These methods were chosen as they represent the range of results shown by all 13 methods. The pine plots for the remaining methods for dataset GSE53757 are provided as Supplementary Material. Visualizations and tables corresponding to datasets GSE10334 and GSE13355 are available from the authors.

The plots in Figures 3.1 and 3.2 show that reproducibility increases as sample size increases. However, the extent of the increase in the overlap scores is not the same across all methods. GSEA-S, as depicted in Figure 3.1, has lower overlap scores overall compared to GSEA-G. In Figure 3.2, ORA and GAGE have higher overlap scores when using sample sizes larger than 2×10 compared to GSEA-S and GSEA-G.

The pine plots for all 13 methods illustrate the following results. PADOG, GSEA-S, and Camera show the lowest overlap scores. Pine plots for ROAST, PAGE, ORA, GSVA, and GSEA-G show a distinct transition from low overlap (for sample sizes less than or equal to 2×5) to high overlap scores (for sample sizes more than 2×10) as sample size increases. ssGSEA, GlobalTest, GAGE, and PLAGE report a large number of gene sets as differentially enriched. These methods tend to achieve high overlap scores as well.

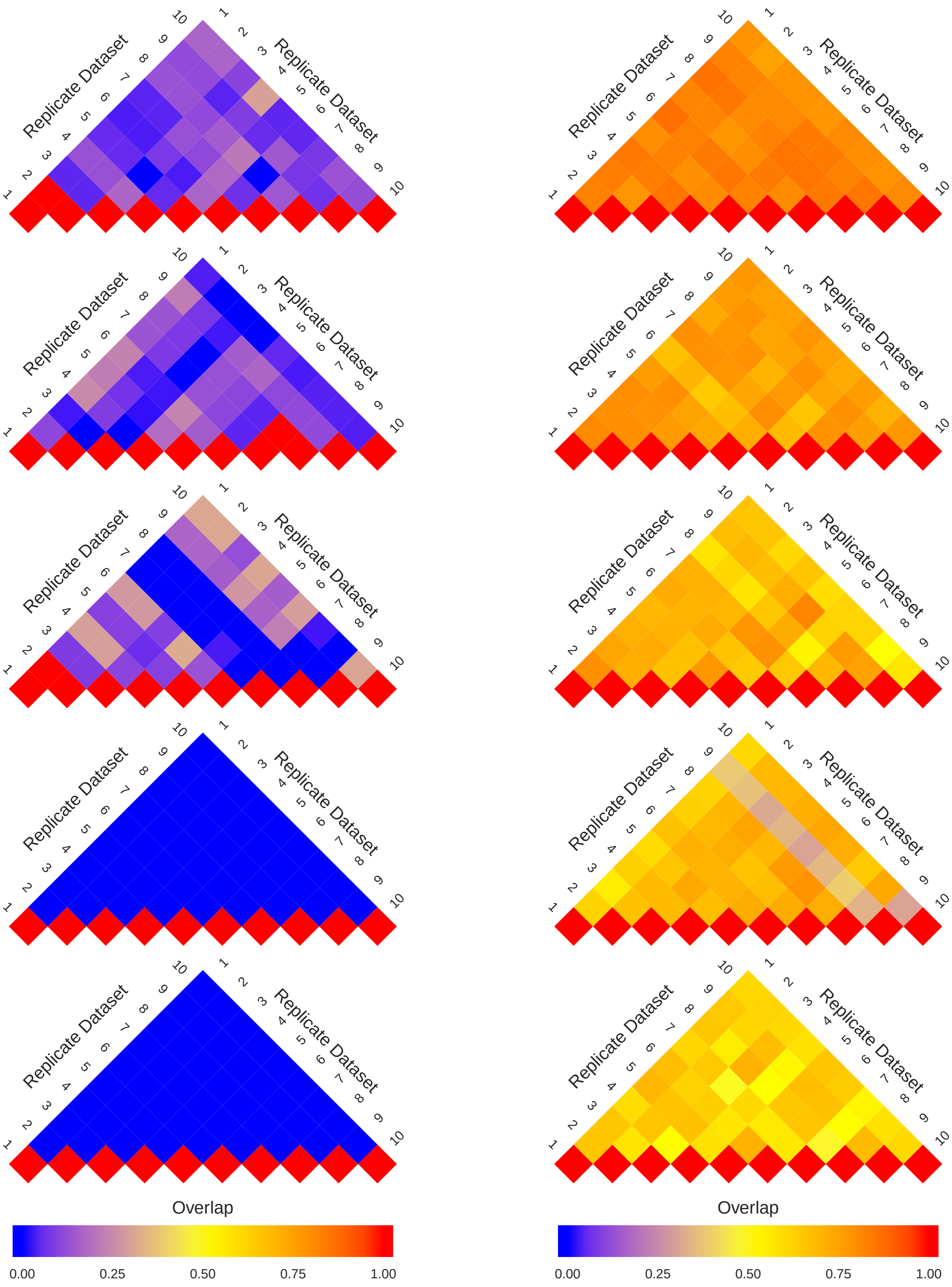


Figure 3.1: Pine plots for dataset GSE53757 showing reproducibility of the results from GSEA-S (left) and GSEA-G (right) across sample sizes. Reproducibility is quantified by overlap score (Equation 3.1). Each layer of the pine plot illustrates the overlap score of the results of a method for 10 replicate datasets with the same sample size. From top to bottom, the pine plots show replicates with sample size 2 × 20, 2 × 15, 2 × 10, 2 × 5, and 2 × 3. The overlap score ranges from 0 to 1 represented by a gradient from blue to red, respectively, separated by yellow in the middle (overlap of 0.5). The overlap score increases as sample size increases; however, the rate of increase for GSEA-S is lower than that of GSEA-G.

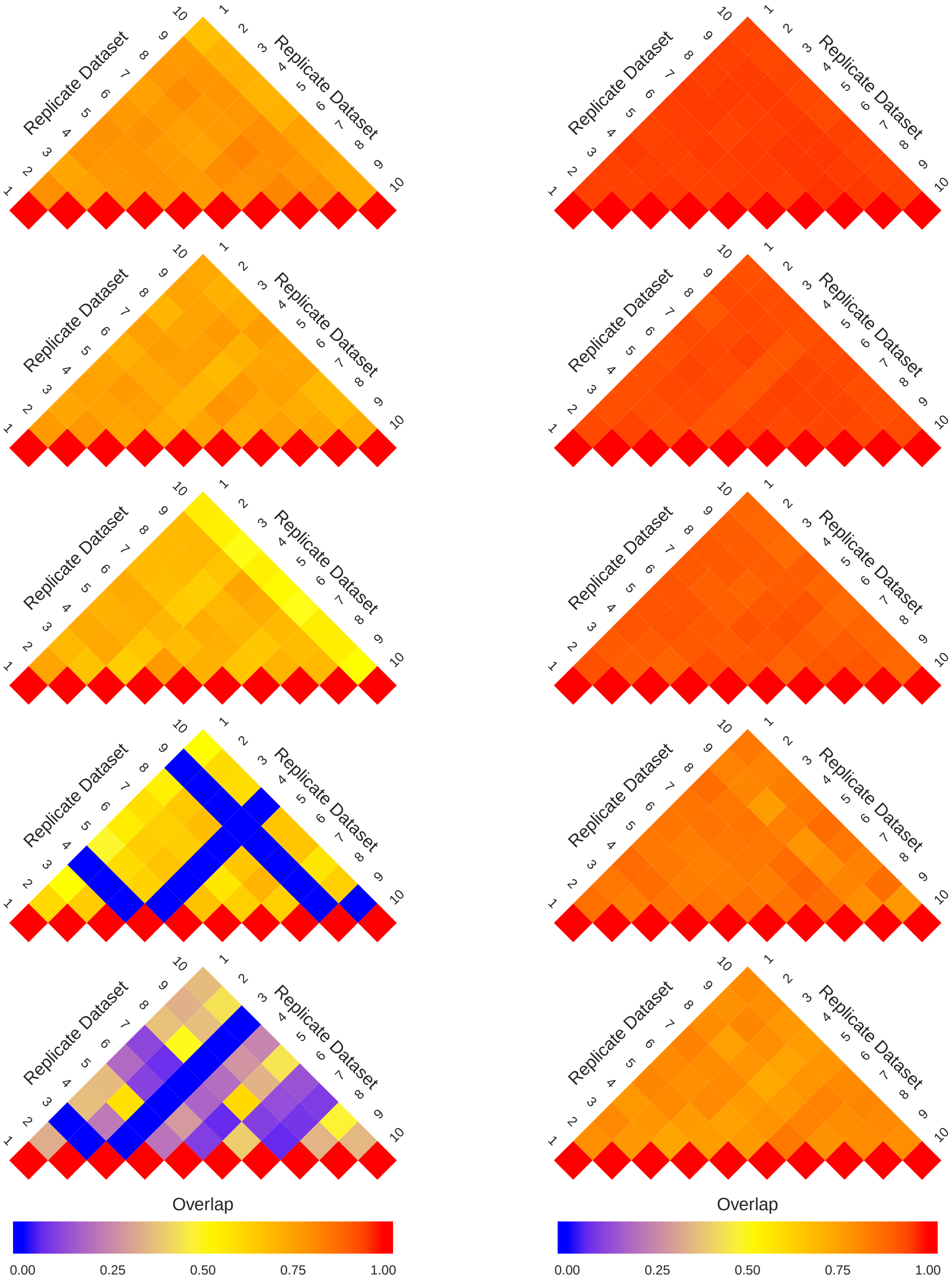


Figure 3.2: Pine plots for dataset GSE53757 showing reproducibility of the results from ORA (left) and GAGE (right). See Figure 3.1 caption for more information. The pine plots suggest that the overlap between replicates is larger in comparison to that of GSEA-S and GSEA-G. GAGE has more agreement between replicates when using lower sample sizes such as 3 compared to the other methods shown (including in Figure 3.1), and the overlap scores continue to improve for higher numbers of samples.

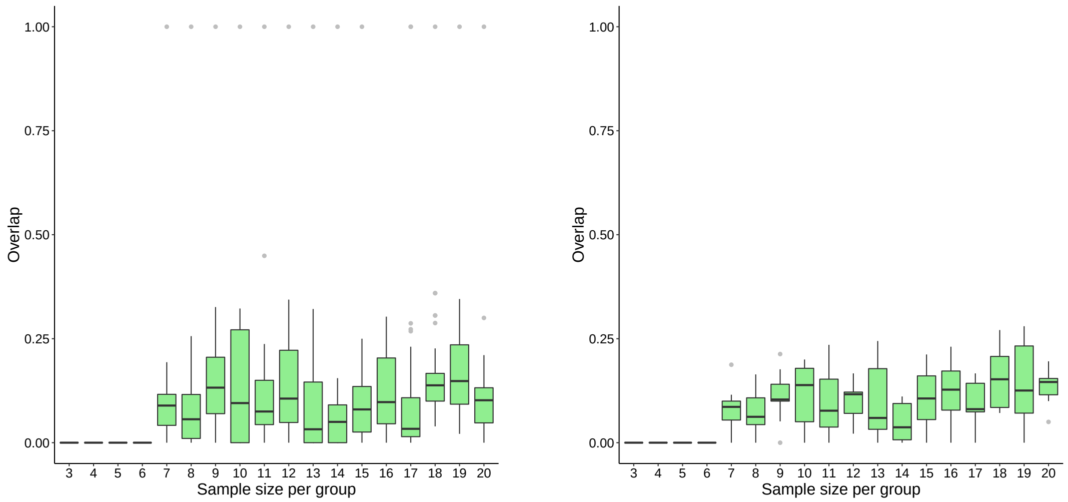


Figure 3.3: Box plots showing the distribution of overlap scores resulting from gene set analysis utilizing GSEA-S when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. Dataset sample sizes are $2 \times n$ ($3 \leq n \leq 20$), where n is the sample size per group. The x-axis shows n, the sample size per group, and the y-axis shows the overlap scores.

The box plots in Figures 3.3, and 3.4, and 3.5 illustrate the distribution of the overlap scores across sample sizes when the results of the replicate datasets are compared to each other (calculated using Equation 3.1), as well as when the results of replicate datasets are compared to the results using the entire dataset, i.e. original dataset (calculated using Equation 3.3). Figures 3.3, 3.4, and 3.5 show representative results from three methods; plots for the remaining methods are in Supplementary Materials. For all methods except GSEA-S and PADOG, the agreement between overlap scores increases as sample size increases.

To statistically assess if there is a significant difference between the overlap scores across sample sizes, i.e. $P^{\psi}_{(2\times3)}, \ldots, P^{\psi}_{(2\times20)}$, for each method, a Kruskal-Wallis test was used. The p-values resulting from these tests, shown in Table B.1 in Supplementary Materials, suggest that the overlap scores significantly vary across sample sizes irrespective of the original dataset being used.

Figure 3.6 depicts Kendall's concordance coefficients for replicate datasets across sample sizes for all methods under study. The figure illustrates that concordance coefficient increases as the sample size increases.

Figure 3.7 illustrates the average number of gene sets reported as differentially enriched across sample sizes. PADOG, GSEA-S, and Camera tend to report a small number of differentially enriched gene sets compared to the other methods. The number of gene sets reported as differentially enriched by GAGE, GSVA, ROAST, and FRY substantially increases as sample size increases, while the rest of the methods reach an almost constant number of gene sets predicted as being differentially enriched. Visualizations analogous to Figures 3.6 and 3.7 for datasets GSE10334 and GSE13355 are provided in Supplementary Materials. These Figures, as well as the pine plots and box plots for these datasets, showed patterns consistent with those for dataset GSE53757.

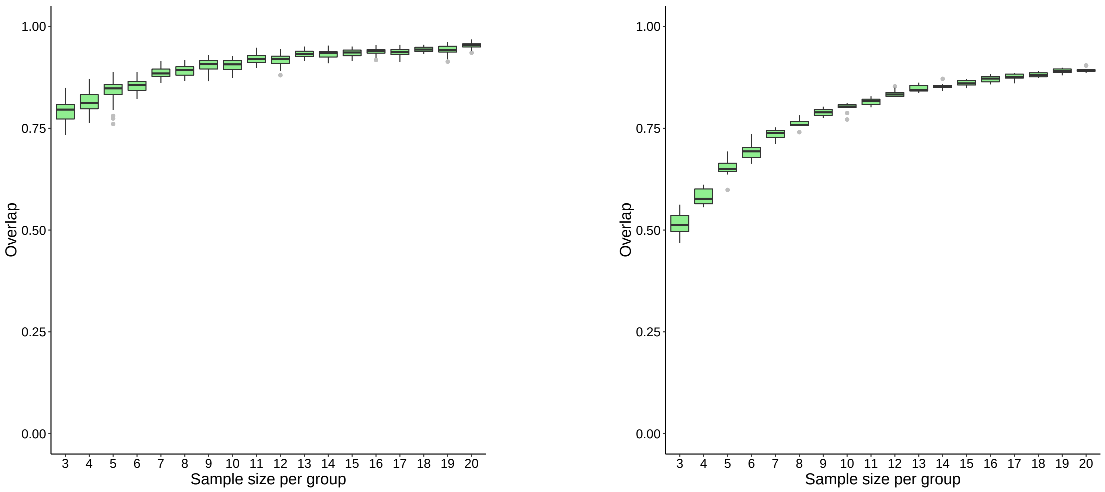


Figure 3.4: Box plots showing the distribution of overlap scores resulting from gene set analysis utilizing GAGE when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

The control-control experiment was conducted using dataset GSE53757. Table B.2 in Supplementary Materials shows the average number of differentially enriched gene sets across sample sizes in the control-control experiment for all of the methods. Since there is no true differential enrichment expected in the control-control experiment, the average values in this table represent the average number of false positives. ssGSEA, GAGE, PAGE, and PADOG are methods with non-zero false positive counts across sample sizes. ssGSEA results in the largest number of false positives followed by GAGE, and then PAGE. Also, increasing sample size for these methods does not reduce the frequency of false positives. Surprisingly, an increase in sample size leads to an increase in false positives when using GAGE. The number of false positives reported by PADOG decreases rapidly as sample size increases, and for sample sizes larger than 2×5, it only reports a small number of false positives. Camera reports almost no false positives for sample sizes less than 2×9, but it reports a small number of false positives as sample size increases. ORA, GSVA, GlobalTest, PLAGE, ROAST, and FRY rarely, if ever, report any false positives.

3.4 Discussion

Reproducibility of the results of gene set analysis methods, as in any other scientific context, is an essential condition for having confidence in the methods. In this research, we applied a systematic approach for quantitatively assessing the reproducibility of gene set analysis methods. By using real expression datasets, the proposed approach strives for a realistic assessment of the reproducibility of gene set analysis methods across sample sizes. We also measure the specificity of gene set analysis methods as a complement to the reproducibility assessment.

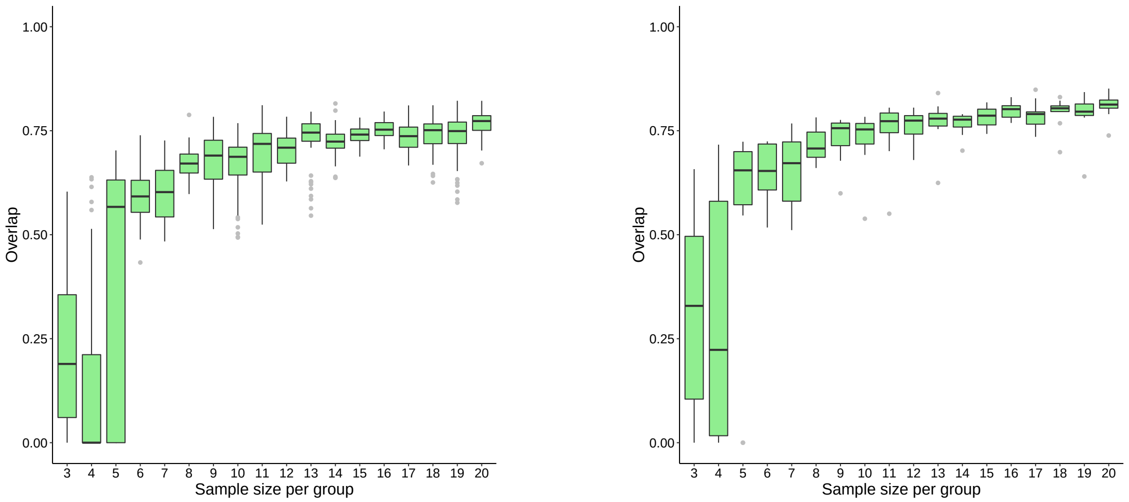


Figure 3.5: Box plots showing the distribution of overlap scores resulting from gene set analysis utilizing ORA when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

To visualize overlap between the results of a gene set analysis method across sample sizes, we introduced and used pine plots. The utility of pine plots, however, is not limited to this application. In general, pine plots can be used to visualize the interaction between several variables defined using a symmetric function while controlling for potentially confounding factors. In practice, most functions for measuring the interaction between variables are symmetric functions—for example, Pearson correlation and Spearman's rank correlation coefficients. Also, symmetry is a necessary condition for any well-defined metric or distance function [12]. Therefore, the symmetric condition does not limit the usability of pine plots. The pine plots in this paper illustrated a general increase in reproducibility as sample size increases. While boxplots can only show the distribution of overlap scores, pine plots are capable of showing the extent of the overlap between each pair of replicate datasets, further highlighting the reproducibility of replicate datasets of the same size.

The reproducibility of gene set analysis across replicate datasets is a necessary condition for obtaining biologically meaningful results, but not a sufficient one. To illustrate, consider a hypothetical method that always predicts all of the gene sets in a gene set database as differentially enriched regardless of the phenotype being examined. Obviously, the results of this method are perfectly reproducible (always an overlap score of 1). However, such a method produces a large number of false positives—i.e., suffers from a lack of specificity—and, as a consequence, does not provide any biological insight. Therefore, we also investigated the specificity of gene set analysis methods. Gene set analysis methods that tend to predict very few gene sets as differentially enriched achieve high specificity. However, these methods often suffer from a lack of sensitivity. In this research, we used the number of differentially enriched gene sets predicted by each method to reveal such scenarios.

The pine plots and box plots showed an increase in reproducibility as sample size increases. However, the

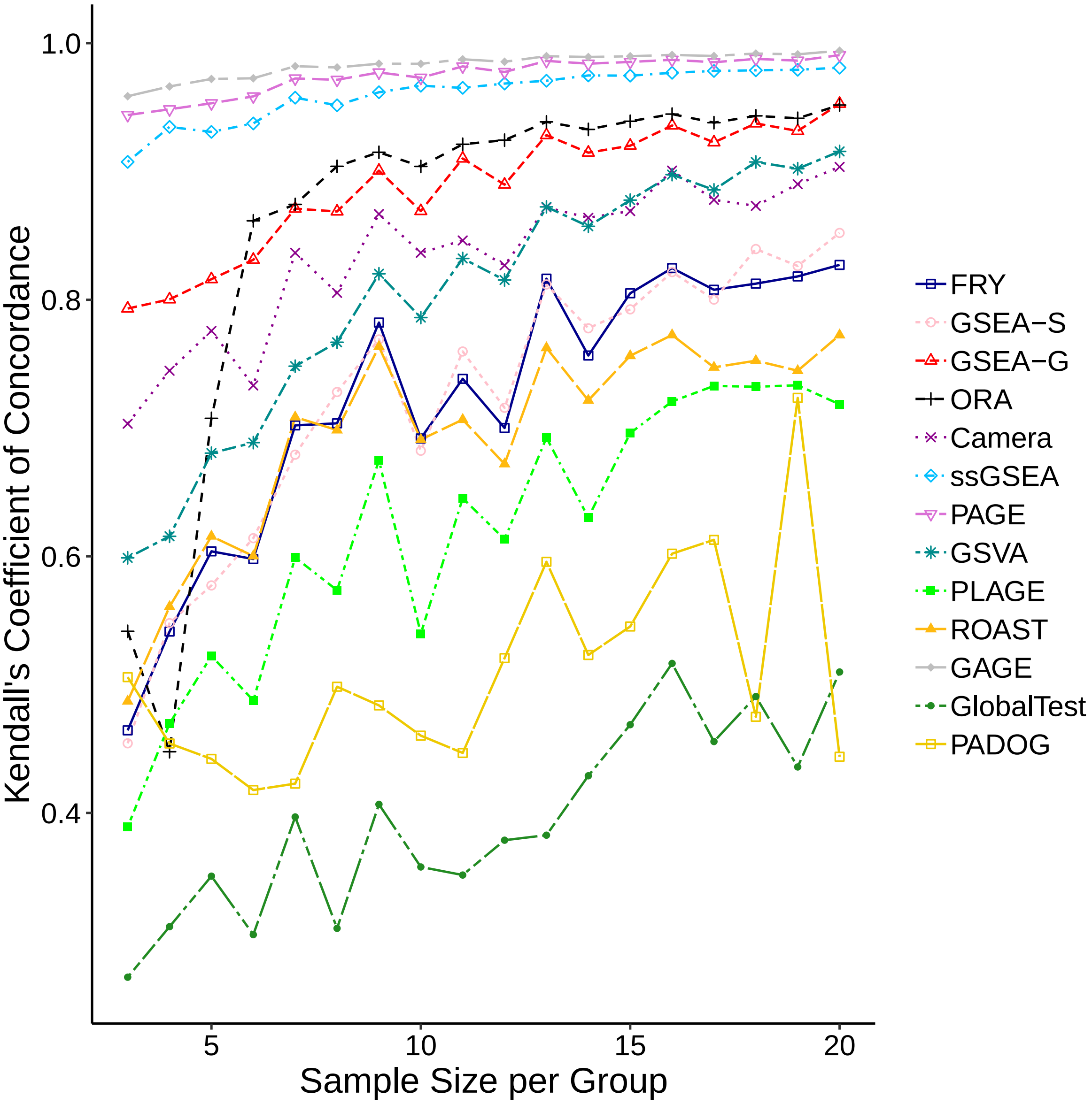


Figure 3.6: Kendall's coefficient of concordance for each method under study when using the original dataset GSE53757 for generating replicate datasets. The x-axis shows the sample size. The y-axis shows concordance coefficients of the results of gene set analysis of 10 replicate datasets of the same size.

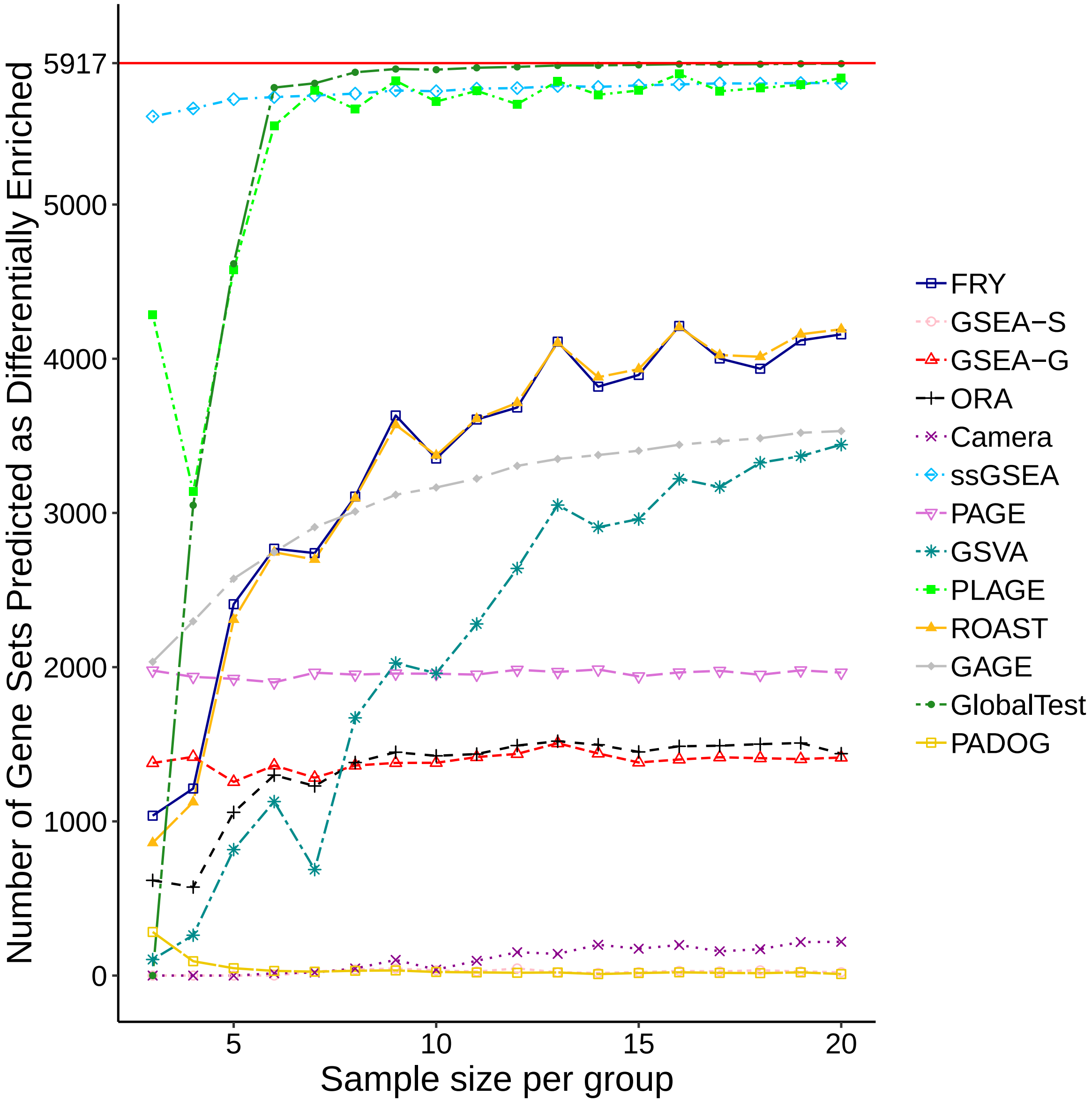


Figure 3.7: The number of gene sets predicted as differentially enriched for each method under study when using the original dataset GSE53757 for generating replicate datasets. The x-axis shows the sample size per group. The y-axis shows the average number of gene sets predicted as differentially enriched across 10 replicate datasets of the same size. The red line parallel to the x-axis shows the size of the gene set database being used, i.e. the maximum possible number of gene sets that could be predicted as being differentially enriched.

extent of reproducibility and the rate at which it grows by sample size was different across methods.

We evaluated reproducibility not only based on the differential enrichment of gene sets but also on the order in which they are predicted, where the order is defined using significance values of gene sets. Kendall concordance coefficients were used to determine if the gene sets predicted as significantly differentially enriched by each method are consistently reported in the same order across replicated datasets of the same sample size.

This research provides insights into the behaviour of specific gene set analysis methods. For instance, GSEA-S achieves low overlap scores across replicate datasets, even for larger sample sizes such as 2 × 20. Meanwhile no differentially enriched gene sets are predicted by this method using small sample sizes (see Figures 3.1, 3.3, and 3.7). GSEA-S was expected to predict few differentially enriched gene sets using small sample sizes since large sample sizes are required for this method to assess significance based on sample permutation. For example, for an experiment with 3 control and 3 case samples, 20 distinct sample permutations exist—the combination of 3 out of 6. Therefore, the smallest non-zero p-value is 0.05, which we did not consider significant. For a sample size of 2×10 or higher, GSEA-S predicts an average of 15 gene sets as differentially enriched (Figure 3.7). This number remains steady while Kendall's concordance increases (Figure 3.6), which suggests that 2 × 10 samples might be a reasonable lower bound for using GSEA-S.

PADOG, which has been designed to take gene set overlap into account and increase specificity, has low overlap scores between replicate datasets even for the largest sample sizes considered. Like GSEA-S and Camera, PADOG also predicts few gene sets as differentially enriched. Lack of reproducibility across large sample sizes for both methods may suggest that the gene sets predicted as differentially enriched are false positives, i.e. gene sets incorrectly predicted as being differentially enriched.

PLAGE, ssGSEA, and GlobalTest tend to report nearly all gene sets as differentially enriched. This is always the case for ssGSEA regardless of the sample size of the replicate datasets used. For small sample sizes, PLAGE and GlobalTest report fewer gene sets as differentially enriched, but this number rapidly increases for larger sample sizes. Furthermore, PLAGE appears to be more sensitive to the dataset used since it predicts far fewer gene sets as differentially enriched for the dataset with higher variability across case and control samples, as with GSE10334 (see Figures B.2 and B.21 in Supplementary Materials). The number of gene sets predicted explains why these methods achieve high overlap scores. Since it is unlikely that such a large number of gene sets are differentially enriched in a living organism, we assume that these methods also predict a large number of false positives.

Given the above, relying on gene sets predicted as being differentially enriched by GlobalTest, PLAGE or ssGSEA may lead to interpretations that are incorrect or biased towards a hypothesis of interest. However, the most statistically significant gene sets, i.e. gene sets with the lowest adjusted p-value, suggested by these methods may still be biologically relevant. Therefore, we suggest further research be conducted to evaluate how best to use these methods and interpret their results.

GlobalTest's and PLAGE's relatively low Kendall concordance coefficients depicted in Figure 3.6 show that the order in which they predict the differentially enriched gene sets is not conserved across replicate datasets. For PLAGE this may be explained by considering that for each gene set, the method defines “the activity level in terms of the first eigenvector, ‘metagene’, in the singular value decomposition” [26]. By ignoring other eigenvectors of an expression profile for a gene set, it cannot completely capture the variability of expression of genes within a gene set. This means that the gene sets predicted as being differentially enriched by PLAGE could show variation in statistical significance across replicate datasets, and therefore, be ranked differently for each replicate dataset.

For sample sizes larger than 2 × 5, ORA remains consistent in the number of gene sets reported as differentially enriched suggesting that 2 × 5 is a reasonable lower bound for sample sizes when using ORA. PAGE also shows a behaviour similar to that of ORA. Since PAGE and ORA are parametric methods and their calculated gene set scores are a function of the number of differentially expressed genes, this behaviour was expected.

When replicate datasets of various sizes are generated from one original dataset, it is expected that a gene set analysis method analyzing these replicate datasets will report approximately the same number of differentially enriched gene sets. However, this is not the case with GAGE, GSVA, FRY, and ROAST. A dramatic increase in the number of gene sets predicted as being differentially enriched was observed for these methods as sample size increases. This increase may also be partially responsible for the observed increase in the overlap scores of these methods as sample size increases.

FRY and ROAST closely mirror each other in the number of gene sets predicted as being differentially enriched as well as their Kendall concordance coefficients across sample sizes. As FRY was designed to be a fast approximation of ROAST, this behaviour is understandable. However, since these methods, as well as GAGE and GSVA, report large numbers of gene sets as being differentially enriched, we assume that these methods may lead to more false positives as sample size increases.

Measuring sensitivity using real datasets is challenging, if not impossible, as the differential enrichment status of gene sets for real datasets are not known. Simulated data, on the other hand, often suffer from oversimplified assumptions such as constant or zero gene-gene correlation or normally distributed expression values [8, 16, 1] leading to biased evaluations of gene set analysis methods [23]. We suggest development of standard synthesized datasets without relying on such assumptions as future research in the community. This would alleviate the challenges caused by the lack of gold standard datasets for the evaluation of gene set analysis methods. A methodology such as the one utilized in this paper could be used for such standard datasets.

The results of the control-control experiment indicate that some methods suffer from a lack of specificity, even in the absence of differential expression. It would be expected that as the number of samples increases, the number of false positives reported would decrease, i.e. specificity increases; however, it is not the case for some methods. As depicted in Table B.2, GAGE and ssGSEA report the highest number of false positives. Almost all gene sets are predicted by ssGSEA as differentially enriched in both the control-control experiment and case-control experiments (see Figures 3.7, B.21, and B.22), regardless of the sample size. GAGE reports a large number of false positives as sample size increases. PAGE and GSEA-G also report a large number of false positives, but this number remains relatively consistent and is not affected by the sample size used. Therefore, increasing the sample size is not a viable solution for improving the results of these methods. PADOG reports fewer false positives as sample size increases; however, PADOG also reports a small number of differentially enriched gene sets in the case-control experiments as well. This shows that although PADOG achieves a high specificity, it may suffer from a lack of sensitivity, as the reported false positives might overwhelm the results. GSEA-S and Camera have a similar issue with being overwhelmed with false positives since they also report a small number of gene sets as differentially enriched. GSVA, Globaltest, FRY, and ROAST do not appear to report false positives in the control-control experiment regardless of sample size being used. However, this control-control experiment measures the number of false positive in the absence of differential expression and does not say anything about the specificity of these methods in the presence of differential expression. The large number of gene sets predicted as differentially enriched in the case-control experiments, as depicted in Figures 3.7, B.21, and B.22 suggests that these methods suffer from a lack of specificity in the presence of differential expression, as such a drastic change in gene expression is unlikely for a living organism. ORA, as expected, did not report any false positive in the absence of differentially expressed genes. It also reported a smaller number of differentially enriched gene sets in the case-control experiments—in comparison to GSVA, Globaltest, FRY, and ROAST. However, it still reported a substantially large number of differentially enriched gene sets for each case-control experiment (almost 20% of gene sets in $\mathbb{G}$). This suggests that ORA could still suffer from lack of specificity in the presence of differential expression though not to the degree of the other methods such as GSVA, Globaltest, FRY, and ROAST. This can be attributed to the presence of gene set overlap [14].

In this paper, we evaluated the results of gene set analysis methods using balanced datasets, i.e. datasets with the same number of cases and controls. While we expect consistent results for datasets that are not drastically imbalanced, we suggest investigating the effect of dataset imbalance on the results of gene set analysis methods.

3.5 Conclusion

The systematic methodology described in this paper can be successfully used to evaluate the reproducibility of results from gene set analysis methods, allowing comparison across methods and sample sizes. The methodology was employed to evaluate the reproducibility of 13 widely used gene set analysis methods. The proposed methodology also made it possible to measure the specificity of these methods using real datasets. From the results we conclude that, as a general pattern, reproducibility, as measured by an overlap score, increases with sample size. However, the rate of increase is method-dependent. Our findings suggest that for all methods in the study, achieving reproducible results using small sample sizes—such as 3, 4, or 5 samples per group—is unlikely. However, we observed that increasing sample size is not a panacea for achieving biologically reliable results, as for some methods it decreased the specificity, i.e. introduced more false positives.

Results from this paper can aid researchers in making a choice among common gene set analysis methods for their work.

References

[1] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, 2009.

[2] Gerald J Bakus. *Quantitative Analysis of Marine Biological Communities: Field Biology and Environment*. John Wiley & Sons, Hoboken, New Jersey, 2007.

[3] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108, 2009.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 289–300, 1995.

[5] Ryan T Demmer, Jan H Behle, Dana L Wolf, Martin Handfield, Moritz Kebschull, Romanita Celenti, Paul Pavlidis, and Panos N Papapanou. Transcriptomes in healthy and diseased gingival tissues. *Journal of Periodontology*, 79(11):2112–2124, 2008.

[6] Sorin Drăghici. *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. CRC Press, Boca Raton, FL, 2016.

[7] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[8] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, pages 107–129, 2007.

[9] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[10] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14(1):7, 2013.

[11] Seon-Young Kim and David J Volsky. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(1):144, 2005.

[12] Nicholas Loehr. *Advanced Linear Algebra*. Chapman and Hall/CRC, Boca Raton, FL, 1 edition, 2014.

[13] Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson, and Peter J Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1):161, 2009.

[14] Farhad Maleki and Anthony J. Kusalik. Gene set overlap: An impediment to achieving high specificity in over-representation analysis. In *10th International Conference on Bioinformatics Models, Methods, and Algorithms*, pages 182–193, Prague, Czech Republic, February 2019.

[15] Farhad Maleki, Katie Ovens, Ian McQuillan, and Anthony J Kusalik. Sample size and reproducibility of gene set analysis. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 122–129. IEEE, December 2018.

[16] Dougu Nam and Seon-Young Kim. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9(3):189–197, 2008.

[17] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[18] Philippe Rocca-Serra, Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Sergio Contrino, Jaak Vilo, Niran Abeygunawardena, Gaurab Mukherjee, Ele Holloway, et al. Arrayexpress: a public database of gene expression data at EBI. *Comptes Rendus Biologies*, 326(10):1075–1078, 2003.

[19] Nicholas J Schurch, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G Simpson, Tom Owen-Hughes, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, 22(6):839–851, 2016.

[20] Cynthia Stretch, Sheehan Khan, Nasimeh Asgarian, Roman Eisner, Saman Vaisipour, Sambasivarao Damaraju, Kathryn Graham, Oliver F Bathe, Helen Steed, Russell Greiner, et al. Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. *PloS One*, 8(6):e65380, 2013.

[21] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[22] William R Swindell, Andrew Johnston, Steve Carbajal, Gangwen Han, Christian Wohn, Jun Lu, Xianying Xing, Rajan P Nair, John J Voorhees, James T Elder, et al. Genome-wide expression profiling of five mouse models identifies similarities and differences with human psoriasis. *PloS One*, 6(4):e18266, 2011.

[23] Pablo Tamayo, George Steinhardt, Arthur Liberzon, and Jill P Mesirov. The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, 25(1):472–487, 2016.

[24] Adi L Tarca, Gaurav Bhatti, and Roberto Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS One*, 8(11):e79217, 2013.

[25] Adi Laurentiu Tarca, Sorin Draghici, Gaurav Bhatti, and Roberto Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):136, 2012.

[26] John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(1):225, 2005.

[27] Chen An Tsai, Sue Jane Wang, Dung Tsa Chen, and James J Chen. Sample size for gene expression microarray experiments. *Bioinformatics*, 21(8):1502–1508, 2004.

[28] Christina A Von Roemeling, Derek C Radisky, Laura A Marlow, Simon J Cooper, Stefan K Grebe, Panagiotis Z Anastasiadis, Han W Tun, and John A Copland. Neuronal pentraxin 2 supports clear cell renal cell carcinoma by activating the ampa-selective glutamate receptor-4. *Cancer Research*, 74(17):4796–4810, 2014.

[29] Di Wu, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E Visvader, and Gordon K Smyth. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010.

[30] Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133–e133, 2012.

4 Method Choice in Gene Set Analysis has Important Consequences for Analysis Outcome

There are a large number of gene set analysis methods available. These methods vary in their components such as their underlying null hypothesis. This raises a question about the choice of gene set analysis method for a given experiment and to what extent that choice may affect the outcome of the analysis. This chapter presents the evaluation of 10 widely used gene set analysis methods.

This paper was accepted as a regular paper at the "12^{th} International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)" with 271 paper submissions from 47 countries, of which 12.5% were accepted as full papers. The paper also received the "Best Student Paper Award".

Citation

Maleki, F., Ovens, K., Rezaei, E., Rosenberg, A. & Kusalik, A. (Feb. 22, 2019). Method Choice in Gene Set Analysis has Important Consequences for Analysis Outcome. In Proceedings of the 12th International joint conference on Biomedical Engineering Systems and Technologies. (pp. 43-54). Prague, Czech Republic. doi:10.5220/0007375000430054

Author contributions

Farhad Maleki designed and developed the methodology for conducting the experiments, performed the statistical analysis and data visualization, and wrote the paper (except "Data" Subsection and "Biomedical evaluation" Section). Katie Ovens preprocessed microarray datasets, performed single gene analysis, and wrote the "Data" Subsection. Also, "Biomedical evaluation" was conducted by Katie Ovens and Elham Rezaei under the supervision of Alan Rosenberg. This Section was written by Katie Ovens and Elham Rezaei, and edited by all authors. Anthony Kusalik supervised the research and helped edit and revise the paper.

Abstract

Gene set enrichment analysis is a well-established approach for gaining biological insight from expression data. With many gene set analysis methods available, a question is raised about the consistency of the results of these methods. In this paper, we answer this question with a systematic analysis of ten commonly used gene set analysis methods when applied to microarray data. The statistical analysis suggests that there is a significant difference between the results of these methods. Comparison of the 20 most statistically significant gene sets reported by these methods showed little to no agreement regardless of the dataset being used. This observation suggests that the outcome of a study can be highly dependent on the choice of the gene set analysis method. Comparing the 100 most statistically significant gene sets also led to the same conclusion. Furthermore, biological evaluation using a juvenile idiopathic arthritis dataset agreed with the results of the statistical analysis. The 20 most statistically significant gene sets for some methods showed relevance to the biology of juvenile arthritis, supporting their utility, while most methods led to results that were irrelevant or marginally relevant to the known biology of the disease.

4.1 Introduction

High-throughput technologies have made it possible to study the expression activity of a large number of genes in a single experiment. These technologies are commonly used to investigate the effect of different stimuli on the expression activity of genes and detect differential expression. A typical gene expression study may lead to reporting several hundred genes as being differentially expressed. Biological interpretation of such an extensive list of genes is difficult. Gene set analysis, also referred to as gene set enrichment analysis, has been widely used to alleviate this problem by detecting a concordant change in the expression pattern of groups of genes that are known to be related to particular functions, processes, or cellular components. Such groups of genes are known as gene sets.

Due to the lack of gold standard datasets where the enrichment status of gene sets are *a priori* known, evaluation of gene set analysis methods is challenging. In the absence of such gold standard datasets, researchers have used artificial datasets to evaluate the sensitivity and specificity of gene set analysis methods. These datasets often rely on simplifying assumptions about the distribution of gene expression measures. Also, they either ignore the complex gene-gene correlation pattern among genes within gene sets or model it using a constant value [7, 16, 1], even though gene-gene correlation has been reported to have a profound impact on the results of enrichment analysis methods [28]. Real expression datasets have also been used to evaluate the sensitivity and specificity of gene set analysis methods [29]. Since the true enrichment status of gene sets are not *a priori* known in real datasets, relying on unverified assumptions about the differential enrichment of gene sets in these datasets does not provide an authentic framework for the evaluation of gene set analysis methods [15]. Consequently, there is no consensus among researchers about the method to use for a given experimental design.

Many gene set enrichment analysis methods are available. These methods vary in their underlying statistical model and the way they quantify a change in the expression pattern of genes within a gene set. A natural question that arises is whether the results of gene set analysis are comparable across methods. In this research, we compare the results of 10 widely used gene set analysis methods to test if the choice of gene set analysis method significantly affects the result of a gene expression study. In addition, since the most statistically significant gene sets are of more value to researchers, we statistically and biologically assess the agreement of the most significant gene sets for all methods under study.

In the rest of the paper, Section 4.2 describes the data and methodology used. Section 4.3 presents the experimental results. The biological evaluation of the results of gene set analysis methods are presented in Section 4.4. Section 6.4 offers insight gained from the experiments and provides suggestions for further research. Finally, Section 6.5 ends the paper with a short summary and conclusions.

4.2 Data and methodology

4.2.1 Data

In this study, four large case-control experiments in humans from the Affymetrix GeneChip Human Genome U133 Plus 2.0 microarray platform were selected for evaluation of gene set analysis methods. These datasets originated from 1) renal cell carcinoma tissue and healthy controls (77 controls and 77 cases, GSE53757) [32], 2) skin from patients with psoriasis and healthy control tissue (64 controls and 58 cases, GSE13355) [27], 3) gingival tissues from healthy and diseased individuals (64 controls and 183 cases, GSE10334) [5], and 4) blood samples from individuals with rheumatoid factor (RF)-negative polyarthritis and healthy individuals (23 controls and 35 cases, GSE26554) [30].

The raw data were preprocessed by first reading the CEL files into R using the *GEOquery* version 2.46.15 R package, and generating the normalized expression table using the *affy* version 1.56.0 package and *justRMA* normalization [10], which have been widely used for normalizing Affymetrix data [17, 33, 37]. Probe IDs were converted to their corresponding Entrez gene identifiers using the *hgu133plus2.db* version 3.2.3 R package. To avoid over-emphasizing genes with a large number of probes on the arrays, it is a common practice in gene set analysis to collapse duplicate IDs. This was accomplished by using *collapseRows* from *WGCNA* version 1.61 with the *MaxMean* method. *MaxMean* selects the probe that has the maximum average value across samples when multiple probes map to the same gene. Collapsing the probes resulted in 20,514 genes in each experiment from an initial 54,675 probes.

The multidimentional scaling (MDS) plots visualizing the case and control samples from each dataset are shown in Figure 4.1. These plots were produced using *cmdscale* from the *stats* R package version 3.4.4 with default parameters.

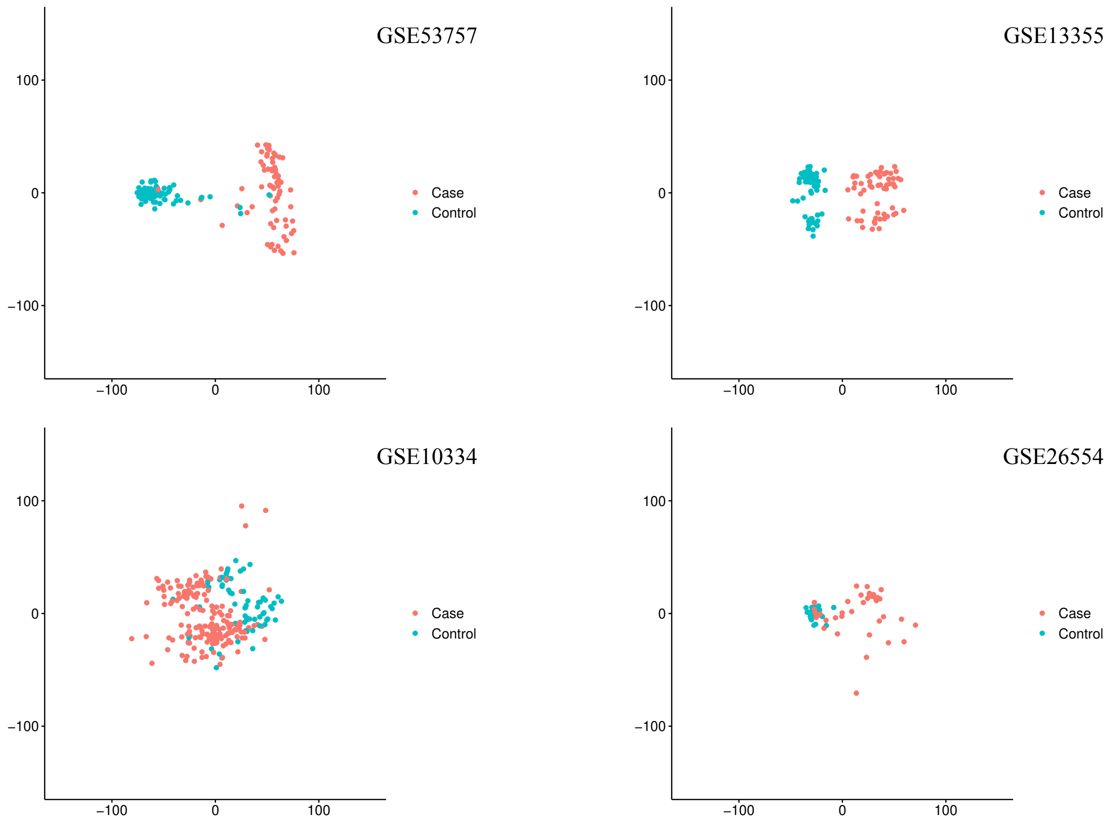


Figure 4.1: MDS plots for samples from the datasets under study. The MDS plots from top to bottom are for datasets with GEO IDs GSE35757, GSE13355, GSE10334, and GSE26554, respectively.

4.2.2 Methodology

In this research, we compare 10 gene set analysis methods: PAGE [12], GAGE [13], Camera [35], ROAST [34], FRY (from the *limma* package) [25], GSEA [26], ssGSEA [3], GSVA [9], PLAGE [31], and over-representation analysis (ORA) [6].

The following R packages are utilized in this study: *GSVA* package version 1.18.0 is used for GSVA, PLAGE, and ssGSEA; the *phyper* method from the *stats* package version 3.4.4 is utilized to implement ORA; the *GSEA.1.0.R* script downloaded from the Broad Institute software page for GSEA provides GSEA; the *limma* package version 3.34.9 is used to run Camera, ROAST, and FRY; the *gage* package version 2.20.1 is used for PAGE and GAGE.

In addition to a gene expression dataset, gene set analysis requires a database of gene sets as input. In this research, we used the GO gene sets—hereafter referred to as $\mathbb{G}$—extracted from *MSigDB* version 6.1 [26]. The GO database is widely used for gene set analysis.

For each gene expression dataset D_i and method ψ_j, gene set analysis is conducted using the default parameters proposed by the authors of ψ_j. To adjust for multiple comparisons, the Benjamini-Hochberg adjustment [4] with a false discovery rate of 0.05 is applied. The resulting adjusted p-values are denoted by a vector $R_{D_i}^{\psi_j}$, where $R_{D_i}^{\psi_j}(n)$—the n^{th} element of this vector—represents the adjusted p-value resulting from gene set analysis of the n^{th} gene set in the gene set database $\mathbb{G}$ using method ψ_j.

For a significance level $\alpha = 0.05$, we define a vector $E_{D_i}^{\psi_j}$ as follows:

$$E_{D_i}^{\psi_j}(n) = \begin{cases} 1, & \text{if } R_{D_i}^{\psi_j}(n) < \alpha \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

where $E_{D_i}^{\psi_j}$ represents the predicted differential enrichment status of gene sets in $\mathbb{G}$—1 for differentially enriched and 0 for non-differentially enriched—and $E_{D_i}^{\psi_j}(n)$ is the n^{th} element of $E_{D_i}^{\psi_j}$. This is accomplished using *cochran.qtest* method from the *RVAideMemoire* R package version 0.9.69.3.

For a given dataset D_i, we statistically assess whether there is a significant difference between these predictions across different methods or not. Since the enrichment status is a dichotomous variable and there are paired data for each gene set (enrichment status for the same gene set across methods), we conduct Cochran's Q test for $E_{D_i}^{\psi_j}$ across all values of ψ_j, i.e. all methods. As post hoc analysis, Wilcoxon sign test is conducted for pairwise comparisons if the Cochran's Q test suggests a significant difference across methods.

Moreover, given the result of two gene set analysis methods ψ_j and ψ_k, we compare the similarity of their results when analyzing dataset D_i using the Jaccard index [2] as follows:

$$J(S_{D_i}^{\psi_j}, S_{D_i}^{\psi_k}) = \frac{S_{D_i}^{\psi_j} \cap S_{D_i}^{\psi_k}}{S_{D_i}^{\psi_j} \cup S_{D_i}^{\psi_k}} \tag{4.2}$$

where $S_{D_i}^{\psi_j}$ is the set of all statistically significant gene sets—i.e. gene sets with an adjusted p-value less than α—when analyzing dataset D_i using ψ_j. A Jaccard index of 1 corresponds to the highest similarity, i.e. $S_{D_i}^{\psi_j} = S_{D_i}^{\psi_k}$, while a Jaccard index of 0 represents no similarity. Also, we define the Jaccard index to be 1 if $S_{D_i}^{\psi_j}$ and $S_{D_i}^{\psi_k}$ both are empty sets. In this paper, we interchangeably refer to the Jaccard index as overlap score.

In addition, since the most statistically significant results—i.e. gene sets predicted as being differentially enriched with the lowest p-values—are of the most interest to researchers, we investigate the agreement among the methods regarding their most significant results. In this regard, we define $S(D_i, \psi_j, t)$ to be the set of up to t statistically most significant gene sets predicted as being differentially enriched—with an adjusted p-value less than α—when analyzing dataset D_i using ψ_j. It should be noted that in cases where the number of differentially enriched gene sets is less than t, $S(D_i, \psi_j, t)$ is equal to the entire set of differentially enriched gene sets resulting from analysis of D_i using ψ_j. After determining $S(D_i, \psi_j, t)$ for each method ψ_j, we quantify the agreement of different methods for their most significant results using an overlap score of $J(S(D_i, \psi_j, t), S(D_i, \psi_k, t))$. In this research, we investigate agreement between the top 20 (and also the top 100) most significant results reported by each method.

4.3 Experimental results

First, each of the four datasets was analyzed using the ten methods under study. Next, $E_{D_i}^{\psi_j}$, i.e. the differential enrichment status of gene sets in $\mathbb{G}$, were determined. For each dataset D_i a Cochran's Q test with a significance level $\alpha = 0.05$ was used to statistically assess if there is a significant difference between the differential enrichment status of gene sets in $\mathbb{G}$ across the methods under study. The Cochran's Q test for all datasets showed a statistically significant difference between the results of the methods under study (see Tables C.1 to C.5 in the Appendix C for test results and the post hoc analysis).

Figure 4.2, using a series of triangular heat maps, illustrates the extent of overlap between the results of the 10 gene set analysis methods for the four datasets and three different scenarios: 1) when overlap is measured from the top 20 most significant gene sets predicted by each method, 2) when overlap is measured from the top 100 most significant gene sets predicted by each method, and 3) when overlap is measured from all the significant gene sets predicted by each method. Each cell in these heat maps represents the overlap score between the results of two methods. A blue hue of a cell indicates low overlap and a red hue indicates high overlap in enriched gene sets between two methods. The heat maps in Figure 4.2 show that, regardless of the datasets being used, the consistency—as measured by overlap score—between the results of different gene set analysis methods is generally low. However, as we move from scenario 1 to 3, the overlap between the results of some of the methods increases. In some instances, such as ROAST and FRY, the amount of overlap remains consistently high across scenarios. The consistency among methods when considering the top 20 most statistically significant results is much lower than the consistency when considering all significant results. This pattern is also observed when comparing the top 100 most statistically significant gene sets to all gene sets predicted as being differentially enriched. Also, Camera and GSEA have little consistency with all other methods under study.

Table 4.1 shows the total number of differentially enriched gene sets reported for all the datasets and all ten methods. GSEA, Camera, and ORA predict a smaller number of gene sets as differentially enriched compared to the other methods.

Figure 4.3 visualizes the distribution of the size of the top 20 and the top 100 most significant gene sets predicted as being differentially enriched for each method. These box plots further highlight the difference between the results of the gene set analysis methods. GAGE, ORA, and ssGSEA tend to report larger gene sets, i.e. gene sets that contain higher numbers of genes, in comparison to the other methods regardless of the dataset being analyzed.

4.4 Biological evaluation

Juvenile idiopathic arthritis (JIA) is a class of childhood arthritis with unknown cause developing before the age of 16 years and persisting for at least 6 weeks. JIA comprises seven categories including: 1) systemic

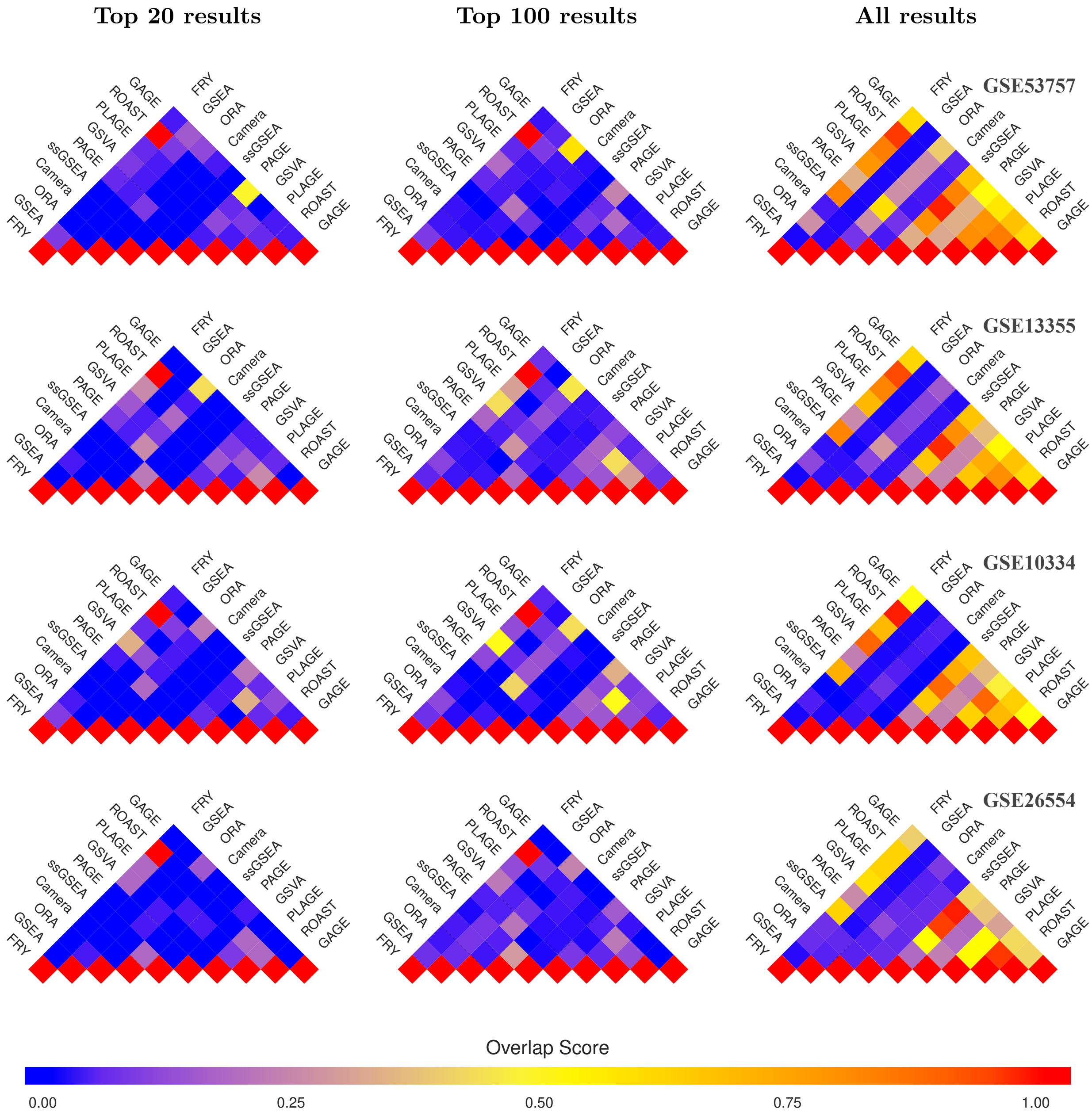
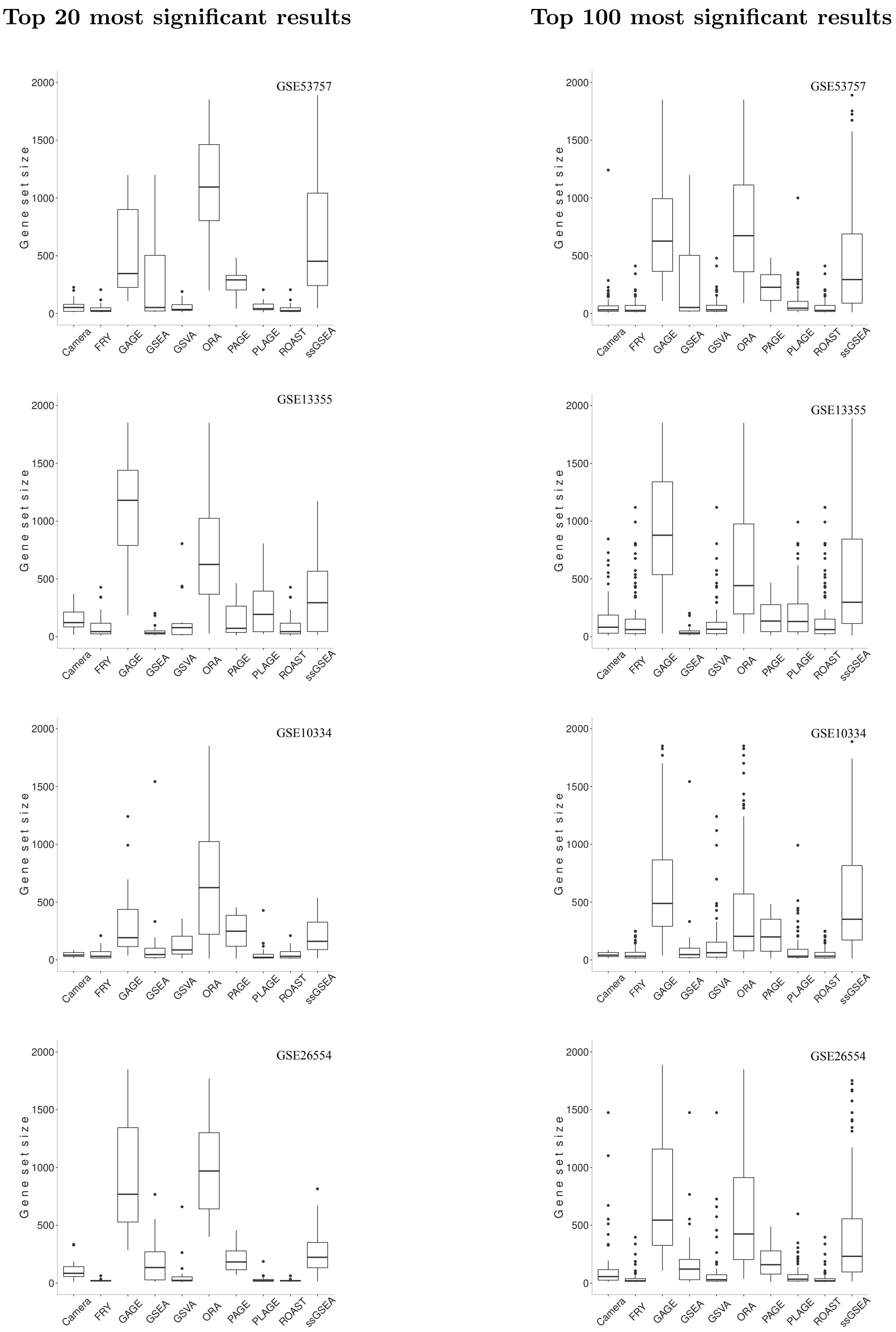


Figure 4.2: A set of triangular heat maps depicting the consistency of the results of gene set analysis methods—as measured by overlap score—across databases. Each triangular heat map illustrates the overlap score of the results of gene set analysis methods when analyzing a gene expression dataset. The layers in the plot, from top to bottom, correspond to datasets with GEO id of GSE53757, GSE13355, GSE10334, and GSE26554, respectively. Ranging from 0 to 1, the overlap score is represented by color hues from blue to red, separated by yellow in the middle (overlap of 0.5). The plot suggests that there is little consistency between the results of the gene set analysis methods under study. This lack of consistency is more pronounced among the top 20 (left column) and top 100 (middle column) most statistically significant results compared to all differentially enriched gene sets (right column).



Top 20 most significant results

Top 100 most significant results

Figure 4.3: Box plots visualizing the distribution of gene set size among the top 20 (left) and top 100 (right) most statistically significant gene sets reported by each method. The plots, from top to bottom, correspond to datasets with GEO ID of GSE53757, GSE13355, GSE10334, and GSE26554 respectively.

Table 4.1: Number of gene sets predicted as being differentially enriched by each method for each dataset.

	GSE53757	GSE13355	GSE10334	GSE26554
FRY	4937	4876	4241	3660
GSEA	19	17	17	33
ORA	1547	573	130	222
Camera	155	73	3	313
ssGSEA	5844	5869	5862	5846
PAGE	1967	1400	1375	1054
GSVA	4730	3847	3819	2988
PLAGE	5900	5830	5242	5698
ROAST	4949	4737	4256	3380
GAGE	3951	3899	3887	2441

arthritis, 2) oligoarthritis, 3) polyarthritis rheumatoid factor (RF)-negative, 4) polyarthritis RF-positive, 5) psoriatic arthritis, 6) enthesitis-related arthritis (ERA), and 7) undifferentiated [19]. For biological validation of methods under study, a JIA dataset containing RF-negative polyarthritis samples and healthy controls was obtained from the same Affymetrix GeneChip Human Genome U133 Plus 2.0 microarray platform as the other datasets (23 controls and 35 cases, GSE26554).

Expression profiles tend to be distinguishable among JIA categories. Gene expression and genome-wide genotyping have identified genes associated with different JIA subtypes, particularly *HLA* gene complex, *PTPN22*, *PTPN2*, *STAT4*, *ANKRD55*, Interleukin *(IL)2-IL21*, *IL-2RA*, *IL-6*, *SH2B3-ATXN2*, *MIF*, *SLC11A1* (*NRAMP1*), *TNFA*, *TNFAIP3*, *TRAF1/C5*, *VTCN1*, *CCL5*, *CD14*, and *WISP3* [21, 20, 23, 14, 36, 8]. The functions of these genes are chiefly regulating production and function of inflammatory biomarkers and their receptors. For instance, *PTPN2* modulates the expression of *IL-2*, *IL-4*, *IL-6*, and *IFN*. Variants of this gene can cause impairment in the regulation of inflammatory pathways, including joint inflammation [11, 22, 24]. The inflammatory process is mediated by an array of innate regulators including interleukins, chemokines, growth factors, and matrix metalloproteinases (MMPs)[18]. There has been increasing interest in identifying molecules involved in regulating immune responses related to susceptibility to, and outcome of, JIA.

Biological evaluation of the 10 gene set enrichment analysis methods under study was performed based on the gene sets/pathways that are known to play a role in JIA using the dataset GSE26554 to determine the biological relevance of the gene sets predicted as being differentially enriched.

All of the top 20 gene sets predicted as being differentially enriched by GAGE showed general relevance to JIA. For example, the top 3 gene sets were “immune response” (GO:0006955), “regulation of immune system process” (GO:0002682), and “immune system process” (GO:0002376). All these gene sets are relatively large and nonspecific to the actual disease or related pathways, with sizes approaching or exceeding 1000 genes. Several gene sets contained fewer genes, while still potentially relating to JIA, including "response to cytokine" (GO:0034097) and "inflammatory response" (GO:0006954). Furthermore, many of the gene sets in the top 20 predicted using GAGE had many terms relating to a general immune process or response (8 of the top 20 gene sets predicted as being differentially enriched).

GSEA also predicted a moderate number of gene sets that are thought to play a role in JIA, but unlike GAGE, these gene sets were much smaller and related to more specific processes. The small predicted gene sets related to JIA included the HLA complex in the gene set "trans-Golgi network" (GO:0005802). Other gene sets predicted as differentially enriched included "positive regulation of antigen receptor-mediated signaling pathway" (GO:0050857), which involves the cross-linking of antigen receptors of immune cells, and "cellular response to interferon-beta" (GO:0035458), which involves responses to a particular cytokine.

ORA also predicted a moderate number of gene sets related to cytokines, while PAGE predicted gene sets related to immune response. The few gene sets relevant to JIA reported by ORA and PAGE were also reported by GAGE. This agrees with our observations in Section 4.3, as the results of these three methods moderately overlap. The other six methods produced few gene sets—among their top results—associated with immune response or inflammation, as shown by the overlap scores (in the triangular heat maps for dataset GSE26554) for the top 20 and top 100 most significant results reported by these methods.

4.5 Discussion

In this research, we showed that there is a significant difference between the results of ten commonly used gene set analysis methods. We quantified the similarity between the results of the methods using Jaccard index. Since researchers value the most statistically significant results, we studied the distribution of gene set size for the top 20 and top 100 most significant results of each method.

The results showed that ROAST and FRY share the same top 20 and top 100 significantly enriched gene sets. This is expected as FRY was designed to be a computationally efficient approximation of ROAST. Also, there are moderate overlaps between the top 20 (and top 100) most significant results reported by ORA, GAGE, and PAGE. This similarity can be explained as all three of these methods are parametric gene set analysis methods; ORA and GAGE are based on two-sample t-tests, and PAGE is based on a z-score. These observations support the validity of the experimental design in this research.

When considering all gene sets predicted as being differentially enriched, there are moderate to high overlaps between the results of all methods, with the exception of ORA, Camera, and GSEA. These high overlaps appear to be a consequence of the high number of gene sets reported as being differentially enriched. As seen by combining the results in Figure 4.2 and Table 4.1, two methods with high overlap between their results also report a high number of gene sets as being differentially enriched. This happens when some methods report a large proportion of gene sets as being differentially enriched. At the extreme, if two methods report all gene sets, the overlap will be its maximum value, i.e. 1. However, as also depicted in Figure 4.2, there is no or very small overlap between the top 20 (and top 100) differentially enriched gene sets reported by methods that achieve high overlap scores when considering all of their significant results. These observations suggest that the methods under study generally do not agree in the gene sets they reported as most statistically significant.

The high numbers of reported differentially enriched gene sets (for some methods) are not an artifact of the choice of the expression datasets or the preprocessing steps. Single gene expression analysis reported that GSE53757 had 571 differentially expressed genes, GSE13355 had 121, GSE10334 had 12, and GSE26554 had 5 differentially expressed genes. These results were produced using the *limma* package with a log fold change cutoff of ±2, a Benjamini-Hochberg correction for multiple comparisons, and a significant level $\alpha = 0.05$.

The number of differentially enriched gene sets reported by ORA and Camera, compared to all other methods under study, seem to be more sensitive to the variation between the case and control groups of each dataset (see Figure 4.1). For the datasets that have more distinct groups, more gene sets are reported as differentially enriched. When the variation is low, ORA—for example—predicts fewer differentially enriched gene sets. One explanation is that the list of genes predicted as being differentially expressed, an input to ORA, is based on a t-test. The t-test statistic denominator represents the variance of expression measures for a gene, and when the sample variation is high, as with GSE10334, the statistic value decreases and the derived p-value increases; this results in a small number of differentially expressed genes. This, in turn, decreases the number of differentially enriched gene sets reported by ORA.

GSEA and Camera typically report a small number of differentially enriched gene sets for each data set. Since the number of reported differentially enriched gene sets is small, these methods have a very small overlap with the results of other methods and also to each other.

Gene sets extracted from GO are associated with GO terms, where the more general terms usually correspond to larger gene sets, and specific terms correspond to smaller gene sets. As depicted in Figure 4.3, for the 20—and also top 100—most statistically significant gene sets reported, ORA, GAGE, PAGE, and ssGSEA tend to report gene sets with larger sizes. Although there could be cases where a gene set with a large size is associated with a phenotype of interest, these three methods consistently report larger gene sets across the datasets compared to the other methods. This may be a sign of systematic bias in favour of large gene set sizes, which are usually less informative. With ORA in particular, the amount of variation between gene set sizes tends to be high as well; also, the median gene set size is typically higher than all other methods. On the other hand, reported gene sets by GSEA have a low median size, although variation between sizes is larger than some of the other methods such as Camera, FRY, ROAST, and GSVA. This is because GSEA reports a mixture of small gene sets followed by large gene sets in the top 20 and 100 results. Camera also reports a small number of gene sets, usually with small sizes. PLAGE, FRY, ROAST, and GSVA—on the other hand—report a large number of gene sets as differentially enriched but, like Camera, their most significant gene sets have small sizes. This could suggest that the reported gene sets are very specific to a particular biological process, molecular function, or cellular component. However, this does not necessarily mean these gene sets are biologically informative to the phenotype under study. To assist with interpreting these results, the relevancy of the most significant genes sets for the JIA dataset was explored by biological interpretation.

The biological evaluation in Section 4.4 suggests that GAGE performed the best followed by GSEA, ORA, and PAGE at predicting the most gene sets that were relevant to the phenotype of interest. This is on par with the overlap scores where GAGE achieved moderate overlap scores with ORA and PAGE. GSEA reports a small number of gene sets as being differentially enriched and therefore achieves low overlap scores with other methods. However, some of its reported gene sets showed relevance to specific immune system processes, which could be more informative compared to some of the more general gene sets reported by GAGE, ORA, and PAGE. We suggest that these results be confirmed further with validation performed on a wide variety of datasets to ensure the results are not dataset or phenotype dependent.

These observations further highlight the lack of agreement between the results of gene set analysis methods. Our results support the utility of methods such as GAGE, GSEA, ORA, and PAGE in gaining biological insight. Drawing a conclusion based on the results of the other methods, even their most significant results, is more challenging and prone to investigator bias toward a hypothesis of interest. This is even a more serious problem for methods that report a large number of gene sets as being differentially enriched. Since it is unlikely for a living organism to undergo such a dramatic change involving several thousand gene sets, this can be interpreted as the lack of specificity, i.e. incorrectly reporting a large number of gene sets as being differentially enriched. We suggest developing methods with higher specificity without sacrificing sensitivity as future research.

Often, it is the case that researchers studying the same phenomenon come up with different results (e.g. different implicated gene sets) even though they appear to have each followed a valid methodology. We are left searching for an explanation for the difference in results. Since our study shows that there is a lack of consistency between the results of gene set analysis methods, part of the explanation could be using different gene set analysis methods, if different gene set analysis methods were used.

4.6 Conclusion

In this paper, we studied the consistency of the results of ten commonly used gene set analysis methods when applied to real expression datasets. The data analysis showed that there is a significant difference between the results of these methods. Our study suggests that not only do these methods differ in the gene sets reported as being differentially enriched, but they also differ in the distribution of the size of the reported gene sets. Further, there is little to no overlap between the results of top 20 (and top 100) most statistically significant gene sets reported, except between FRY and ROAST.

The biological validation of the most significant results using a JIA dataset revealed that GAGE performs the best followed by GSEA, ORA, and PAGE at predicting the most gene sets relevant to the phenotype of interest. The biological evaluation of the most significant results reported by the gene set analysis methods revealed that the majority of the methods reported gene sets that are not related to the known biology of JIA. GAGE was the only method with all of its top 20 gene sets relevant to the biology of juvenile arthritis. In addition, GSEA, ORA, and PAGE reported relevant gene sets, with GSEA reporting fewer but more specific gene sets. This supports the utility of these methods for gene set analysis. However, any more general conclusion would require a broader study incorporating datasets from various phenotypes.

References

[1] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, 2009.

[2] Gerald J Bakus. *Quantitative Analysis of Marine Biological Communities: Field Biology and Environment*. John Wiley & Sons, 2007.

[3] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108, 2009.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995.

[5] Ryan T Demmer, Jan H Behle, Dana L Wolf, Martin Handfield, Moritz Kebschull, Romanita Celenti, Paul Pavlidis, and Panos N Papapanou. Transcriptomes in healthy and diseased gingival tissues. *Journal of Periodontology*, 79(11):2112–2124, 2008.

[6] Sorin Drăghici. *Statistics and Data Analysis for Microarrays using R and Bioconductor*. CRC Press, 2016.

[7] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.

[8] Yi-Man E Fung, Deborah J Smyth, Joanna MM Howson, James D Cooper, Neil M Walker, H Stevens, Linda S Wicker, and John A Todd. Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus. *Genes and Immunity*, 10(2):188, 2009.

[9] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14(1):7, 2013.

[10] Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15–e15, 2003.

[11] L. B Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome Research*, 10(10):1435–1444, 2000.

[12] Seon-Young Kim and David J Volsky. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(1):144, 2005.

[13] Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson, and Peter J Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1):161, 2009.

[14] A Martinez, J Varade, A Marquez, MC Cenit, L Espino, N Perdigones, JL Santiago, M Fernández-Arquero, H De La Calle, R Arroyo, et al. Association of the STAT4 gene with increased susceptibility for some immune-mediated diseases. *Arthritis & Rheumatism*, 58(9):2598–2602, 2008.

[15] Ravi Mathur, Daniel Rotroff, Jun Ma, Ali Shojaie, and Alison Motsinger-Reif. Gene set analysis methods: a systematic comparison. *BioData Mining*, 11(1):8, 2018.

[16] Dougu Nam and Seon-Young Kim. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9(3):189–197, 2008.

[17] Benjamin A. Neely and Paul E. Anderson. Complementary domain prioritization: A method to improve biologically relevant detection in multi-omic data sets. In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS, (BIOSTEC 2017)*, pages 68–80. INSTICC, SciTePress, 2017.

[18] Ross E Petty, Ronald M Laxer, Carol B Lindsley, and Lucy Wedderburn. *Textbook of Pediatric Rheumatology*. Elsevier Health Sciences, 2015.

[19] Ross E Petty, Taunton R Southwood, Prudence Manners, John Baum, David N Glass, Jose Goldenberg, Xiaohu He, Jose Maldonado-Cocco, Javier Orozco-Alcala, Anne-Marie Prieur, et al. International league of associations for rheumatology classification of juvenile idiopathic arthritis: second revision, edmonton, 2001. *The Journal of Rheumatology*, 31(2):390, 2004.

[20] JD Phelan, SD Thompson, and DN Glass. Susceptibility to jra/jia: complementing general autoimmune and arthritis traits. *Genes and Immunity*, 7(1):1, 2006.

[21] Sampath Prahalad. Genetics of juvenile idiopathic arthritis: an update. *Current Opinion in Rheumatology*, 16(5):588–594, 2004.

[22] Sampath Prahalad. Genetic analysis of juvenile rheumatoid arthritis: approaches to complex traits. *Current problems in pediatric and adolescent health care*, 36(3):83, 2006.

[23] Sampath Prahalad and David N Glass. A comprehensive review of the genetics of juvenile idiopathic arthritis. *Pediatric Rheumatology*, 6(1):11, 2008.

[24] Sampath Prahalad, Mary H Ryan, Edith S Shear, Susan D Thompson, Edward H Giannini, and David N Glass. Juvenile rheumatoid arthritis: linkage to hla demonstrated by allele sharing in affected sibpairs. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 43(10):2335–2338, 2000.

[25] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[26] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric 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 of the United States of America*, 102(43):15545–15550, 2005.

[27] William R Swindell, Andrew Johnston, Steve Carbajal, Gangwen Han, Christian Wohn, Jun Lu, Xianying Xing, Rajan P Nair, John J Voorhees, James T Elder, et al. Genome-wide expression profiling of five mouse models identifies similarities and differences with human psoriasis. *PloS One*, 6(4):e18266, 2011.

[28] Pablo Tamayo, George Steinhardt, Arthur Liberzon, and Jill P Mesirov. The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, 25(1):1–16, 2012.

[29] Adi L Tarca, Gaurav Bhatti, and Roberto Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS One*, 8(11):e79217, 2013.

[30] Susan D Thompson, Miranda C Marion, Marc Sudman, Mary Ryan, Monica Tsoras, Timothy D Howard, Michael G Barnes, Paula S Ramos, Wendy Thomson, Anne Hinks, et al. Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13. *Arthritis & Rheumatism*, 64(8):2781–2791, 2012.

[31] John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(1):225, 2005.

[32] Christina A Von Roemeling, Derek C Radisky, Laura A Marlow, Simon J Cooper, Stefan K Grebe, Panagiotis Z Anastasiadis, Han W Tun, and John A Copland. Neuronal pentraxin 2 supports clear cell renal cell carcinoma by activating the ampa-selective glutamate receptor-4. *Cancer Research*, 74(17):4796–4810, 2014.

[33] Sean West and Hesham Ali. Sensitivity analysis of granularity levels in complex biological networks. In Ana Fred and Hugo Gamboa, editors, *Biomedical Engineering Systems and Technologies*, pages 167–188. Springer International Publishing, 2017.

[34] Di Wu, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E Visvader, and Gordon K Smyth. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010.

[35] Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133–e133, 2012.

[36] Tsung-Chieh Yao, Yi-Chan Tsai, and Jing-Long Huang. Association of rantes promoter polymorphism with juvenile rheumatoid arthritis. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 60(4):1173–1178, 2009.

[37] Joanna Zyla, Michal Marczyk, and Joanna Polanska. Sensitivity, specificity and prioritization of gene set analysis when applying different ranking metrics. In *10th International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 61–69. Springer, 2016.

5 Gene Set Overlap: An Impediment to Achieving High Specificity in Over-representation Analysis

Often, gene set analysis leads to predicting a large number of biologically irrelevant or uninformative gene sets as being differentially enriched [40]. This makes gaining biological insight from the results of gene set analysis difficult. It also may lead to a biased interpretation of the results in favour of a hypothesis of interest. This chapter presents a study of gene set overlap and its effect on over-representation analysis. We quantify gene set overlap, show that it is a ubiquitous phenomenon across gene set databases, and assess its effect on the specificity of over-representation analysis. Particularly, we assess the hypothesis that there is a significant correlation between gene set overlap and specificity of over-representation analysis.

This paper was accepted as a regular paper at the "12^{th} International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)" with 271 paper submissions from 47 countries, of which 12.5% were accepted as regular papers.

Citation

Maleki, F. & Kusalik, A. (Feb. 22, 2019). Gene Set Overlap: An Impediment to Achieving High Specificity in Over-representation Analysis. In Proceedings of the 12th International joint conference on Biomedical Engineering Systems and Technologies. (pp. 182-193). Prague, Czech Republic. doi:10.5220/0007376901820193

Author contributions

Farhad Maleki conducted the research and wrote the paper. Anthony Kusalik supervised the research and helped edit and revise the paper.

Abstract

Gene set analysis methods are widely used to analyze data from high-throughput “omics” technologies. One drawback of these methods is their low specificity or high false positive rate. Over-representation analysis is one of the most commonly used gene set analysis methods. In this paper, we propose a systematic approach to investigate the hypothesis that gene set overlap is an underlying cause of low specificity in over-representation analysis. We quantify gene set overlap and show that it is a ubiquitous phenomenon across gene set databases. Statistical analysis indicates a strong negative correlation between gene set overlap and the specificity of over-representation analysis. We conclude that gene set overlap is an underlying cause of the low specificity. This result highlights the importance of considering gene set overlap in gene set analysis and explains the lack of specificity of methods that ignore gene set overlap. This research also establishes the direction for developing new gene set analysis methods.

5.1 Introduction

High-throughput “omics” technologies have been widely used to investigate biological questions that require screening of a large number of biomolecules. The main challenge facing these technologies is analyzing the generated data to gain biological insight. An RNA-Seq experiment, for example, may suggest several hundred genes as being differentially expressed. Manual interpretation of such a large set of genes is impractical and susceptible to investigator bias toward a hypothesis of interest.

Gene set analysis is a well-established computational approach to gain biological insight from data resulting from high-throughput gene expression experiments [22]. It relies on the assumption that most biological processes are the consequence of a coordinated activity of a group of genes. Therefore, the primary goal of gene set analysis is to detect concordant changes in expression patterns of predefined groups of genes, referred to as gene sets. Members of a given gene set often share a common biological function or attribute. MSigDB [29], GeneSigDB [11], GeneSetDB [3], Go-Elite [45], and Enrichr [27] are among the most widely used gene set databases. These databases have been generated from various sources including GO [4], KEGG [26], Reactome [25], and BioCarta [34].

Often gene set analysis methods report a large number of gene sets as being differentially enriched, where the majority of the reported gene sets are biologically irrelevant or uninformative [40]. The rapid growth of the size of gene set databases is intensifying this issue. Consequently, gaining biological insight from the results of gene set analysis is becoming more challenging and prone to investigator biases in favour of a hypothesis of interest. For example, Araki et al. used GeneSetDB to analyze a list of 79 differentially expressed Affymetrix probe sets [3] resulting from an experiment where endothelial cells were induced to undergo apoptosis [24]. After correction for multiple hypothesis testing, they reported 1694 gene sets as statistically significant, i.e. differentially enriched. Interpreting this large number of gene sets is challenging.

Understanding the factors contributing to low specificity in gene set analysis helps in choosing methods that are more robust against these factors. Such an understanding also facilitates interpreting the results of gene set analysis methods and accelerates the development of new methods that address these contributing factors to achieve higher specificity without sacrificing sensitivity and accuracy.

Specificity of gene set analysis methods in the absence of differential expression of genes has been studied. Tarca et al. [40] investigated the specificity of sixteen gene set analysis methods in the absence of differential expression and showed that even when there is no differential expression, some gene set analysis methods produce a large number of false positives. However, their approach cannot be used to assess the specificity of a gene set analysis method in the presence of differentially expressed genes.

Overlap between gene sets has been suggested as being responsible for the low specificity of gene set analysis methods. To deal with overlap between gene sets, PADOG [41] assigns lower weights to genes that belong to more than one gene set. For a given gene g, this weight is negatively correlated with the number of gene sets containing g. TopGO [2] is another attempt to deal with gene set overlap. It considers that Gene Ontology (GO) terms are organized as a directed acyclic graph encoding a hierarchy of general-to-more-specific terms. This structure leads to commonality between the genes corresponding to a child node and those of its parent(s). TopGO proposes a gene elimination and a gene down-weighting procedure to decorrelate the GO graph structure resulting from these relations. MGSA [6] utilises a Bayesian approach that considers the overlap between GO categories to reduce the number of false positives. SetRank [38] is another attempt at reducing the number of false positives by considering the overlap between gene sets.

Parallel to the development of gene set analysis methods, various gene set databases have been developed. The prevailing trend in developing gene set databases has been introducing more gene sets and increasing database size. Figure 5.1 illustrates the growth of MSigDB across its versions. This gene set database has been designed for gene set analysis in human, and its current version includes gene sets from various sources such as GO, KEGG, Reactome, and BioCarta. This gene set database has undergone a 13-fold increase in the number of gene sets compared to its first version. Given the limited number of known genes for human, this steep growth leads to an increase in the number of gene sets overlapping with each other.

The MSigDB Hallmark gene set collection is an attempt to reduce gene set overlap by computational and manual curation of gene sets from MSigDB. However, this collection currently (version 6.2) only includes 50 gene sets [28]. "Enrichment Map" is another approach of dealing with gene set overlap using a network-based visualization approach [31]. To the best of our knowledge, there is no systematic study of the effect of gene set overlap on the results of gene set analysis. In this paper, we investigate the hypothesis that gene set overlap plays a prominent role in the lack of specificity of over-representation analysis (ORA), which is one of the most widely used gene set analysis methods [15].

The rest of the paper is organised as follows. In Section 5.2, we briefly describe ORA. In Section 5.3, we show that gene set overlap is a ubiquitous phenomenon in gene set databases; we use quantitative measures to visualize gene set overlap in GeneSetDB [3], GeneSigDB [12], and MSigDB [29], which are well-established

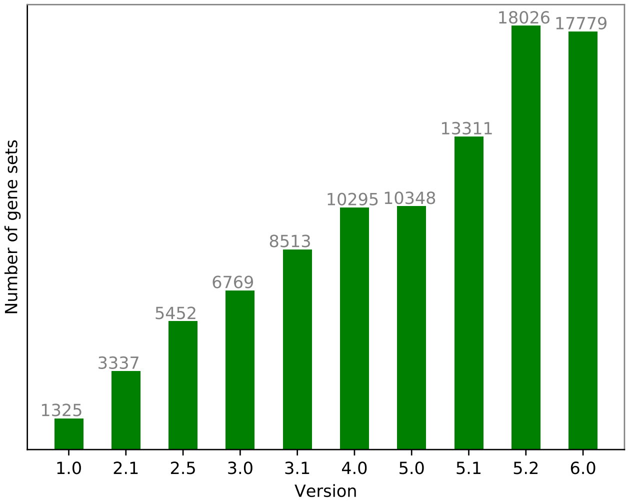


Figure 5.1: The number of gene sets in different versions of MSigDB

gene set databases. In Section 5.4, using these quantitative measures, we introduce a methodology to study the effect of gene set overlap on the specificity of ORA. In Section 5.5, we describe the experimental results; using the methodology introduced in Section 5.4, we statistically investigate the effect of gene set overlap on the specificity of ORA by assessing the correlation between gene set overlap and specificity. In Section 6.4, we discuss the implication of gene set overlap and the challenges it entails. We also provide suggestions for developing and evaluating gene set analysis methods. Finally, Section 6.5 offers a short summary and conclusion.

5.2 Over-representation analysis

Many algorithms have been proposed and used for gene set analysis, of which ORA is one of the most widely used. Due to its simplicity, well-established underlying statistical model, and ease of implementation, ORA is available through many tools [7], [8], [10], [23], [30], [42], [43], [44], [46], [47], [48]. This method defines a concordant change in expression pattern of members of a given gene set as a change that is unlikely to happen by chance. It also quantifies the concept of change as the number of differentially expressed genes in a pairwise comparison of phenotypes, e.g. “cancerous” versus “non-cancerous”.

ORA can be outlined as follows [16]. Suppose that data analysis for an experiment using a high-throughput technology predicts a set of differentially expressed genes L, and that the intersection of L and a given gene set G_i contains n'_i genes. In addition, assume that the set of background genes, i.e. all genes with a non-zero probability of being differentially expressed, contains n genes. For example, the background genes in a microarray study can be the set of all genes represented on the arrays. Denote the background set as U. Let $\overline{G_i}$ refer to the complement of G_i with respect to U, i.e. all genes in U but not in G_i. Given L, G_i, and U, ORA assesses whether the number of differentially expressed genes in G_i is more than what it should be just by chance, i.e. it is over-represented. Table 5.1 represents ORA as a contingency table, where $\| \bullet \|$ is the

Table 5.1: Representation of ORA as a contingency table. Each cell contains a count of genes satisfying the condition given by the row and column.

	Genes $\in L$	Genes $\notin L$	Total
Genes $\in G_i$	n'_i	$\|G_i\| - n'_i$	$\|G_i\|$
Genes $\in \overline{G_i}$	$\|L\| - n'_i$	$(n - \|G_i\|)$ - $(\|L\| - n'_i)$	$n - \|G_i\|$
Total	$\|L\|$	$n - \|L\|$	n

cardinality operator.

Assuming that genes are selected using a simple random sampling approach, ORA can be modeled using a hypergeometric distribution [16]. Accordingly, the probability of having n'_i genes from G_i among differentially expressed genes, i.e. L, is as follows:

$$f(n'_i; n, \|G_i\|, \|L\|) = \frac{\binom{\|G_i\|}{n'_i} \times \binom{n-\|G_i\|}{\|L\|-n'_i}}{\binom{n}{\|L\|}} \tag{5.1}$$

In addition, Fisher's exact test can be used to examine the significance of the association between genes in G_i and genes in L. The p-value can be calculated for over-representation of G_i based on Equation 5.2.

$$\begin{aligned} p &= \sum_{j=n'_i}^{\|G_i\|} f(j; n, \|G_i\|, \|L\|) \\ &= 1 - \sum_{j=0}^{n'_i-1} f(j; n, \|G_i\|, \|L\|) \end{aligned} \tag{5.2}$$

5.3 Overlap in gene set database

ORA, as with other gene set analysis methods, relies on availability of a gene set database. Gene set databases are developed by collecting genes that are manually or computationally inferred to share a common biological function or attribute. The availability of *a priori* knowledge through public repositories such as GO [4], KEGG [26], and OMIM [21] makes it possible to develop gene set databases. There are many publicly available gene set databases including L2L [33], SignatureDB [36], CCancer [13], GeneSigDB [12], GeneSetDB [3], and MSigDB [29]. The latter three are widely used for gene set analysis.

MSigDB is the gene set database integrated with GSEA [39]. MSigDB acquires gene sets through manual curation and computational methods [29]. As a meta-database, MSigDB extracts gene sets from several sources including GO [4], KEGG [26], Reactome [25], and BioCarta [34].

GeneSigDB is another database of gene sets extracted from published experimental expression studies of genes, proteins, or miRNAs. GeneSigDB relied on PubMed searches to collect papers relevant to a set of search terms mainly focused on cancer, lung disease, development, immune cells, and stem cells. To develop the database, the authors downloaded the relevant papers and then manually transcribed gene sets from them or their supplementary documents.

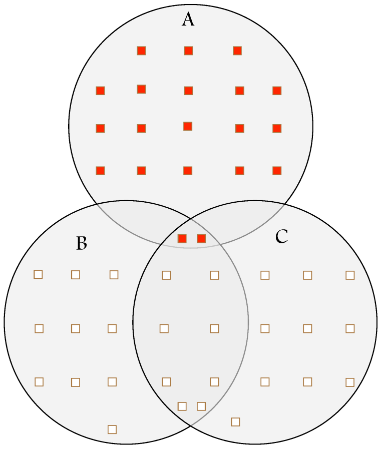


Figure 5.2: A hypothetical example: gene set overlap leading to lack of specificity of ORA. Each circle represents a gene set. Rectangles coloured in red and white represent differentially expressed and non-differentially expressed genes, respectively. Gene set B (and also C) is predicted as being differentially enriched by ORA solely due to partial overlap with A, a truly differentially enriched gene set.

GeneSetDB, as another meta-database, is a collection of 26 public databases focused on pathways, phenotypes, drugs, gene regulation, or Gene Ontology. The primary focus of GeneSetDB is human, although it supports mouse and rat using computationally inferred homology [3].

5.3.1 Gene set overlap and ORA: A hypothetical example

To show how overlap of gene sets can affect the results of ORA, in this section we present a hypothetical example. Suppose that in a high-throughput experiment, the expression activity of 10000 genes has been measured. After conducting the experiment and performing single gene analysis, 100 genes have been predicted as being differentially expressed. Consider gene sets A, B, and C as illustrated in Figure 5.2, where gene sets are depicted as circles, and genes belonging to each gene set are depicted as rectangles. In each gene set, genes predicted as being differentially expressed are coloured in red and the rest of the genes are coloured in white. As shown in Figure 5.2, all genes in A have been predicted as being differentially expressed. Table 5.2 illustrates the contingency table for over-representation of B.

Table 5.2: The contingency table for the over-representation of B. DE stands for differentially expressed and Non-DE stands for non-differentially expressed.

	DE Genes	Non-DE Genes	Total
Genes in B	2	18	20
Genes not in B	98	9882	9980
Total	100	9900	10000

According to Fisher's exact test, B is predicted as being differentially enriched with a 95% confidence level ($p\text{-}value = 0.0167$). This result is primarily due to the overlap between A and B. This example suggests that gene set overlap can lead to a lack of specificity in gene set analysis methods.

In this paper, we use GeneSigDB version 4, GeneSetDB for Human (downloaded on February 2, 2018), and MSigDB version 6.0, unless stated otherwise.

5.3.2 Measuring gene set overlap

To study gene set overlap and its effect on the specificity of ORA, we use the Jaccard coefficient to quantify the overlap between two gene sets. We then use this quantitative measure to visualize gene set overlap in MSigDB, GeneSetDB, and GeneSigDB.

For given sets A and B, the Jaccard coefficient is defined as follows:

$$J(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|} \tag{5.3}$$

The Jaccard coefficient is a value between 0 and 1, where $J(A,B) = 0$ means that there is no overlap between A and B; $J(A,B) = 1$ means that there is a complete overlap between A and B, i.e. $A = B$; and other values ($0 < J(A,B) < 1$) represent partial overlaps between A and B. The Jaccard index can be used to quantify the overlap between two sets; for example, it can be used to measure the overlap between two gene sets from a gene set database or the overlap between a gene set and a set of differentially expressed genes resulting from a gene expression study. Hereafter, we refer to Jaccard index as overlap score.

For a given set of genes L_i and a gene set database $\mathbb{G}$, we define the overlap coefficient, or overlap score, of L_i with respect to $\mathbb{G}$ as follows:

$$O(L_i, \mathbb{G}) = \sum_{G_j \in \mathbb{G}} J(L_i, G_j) \tag{5.4}$$

This measure is representative of the cumulative overlap of L_i with all gene sets in the gene set database $\mathbb{G}$. For the sake of brevity, whenever gene set database $\mathbb{G}$ can be inferred from the context, we use the phrase "overlap score of L_i" to refer to $O(L_i, \mathbb{G})$. Note that $O(L_i, \mathbb{G})$, which is the summation of overlap between L_i and each gene set in the gene set database $\mathbb{G}$, should not be mistaken with overlap between two sets of genes. The latter is calculated using the Jaccard index (Equation 5.3).

5.3.3 Visualization of gene set overlap

Graph-based visualizations have been used to facilitate the interpretation of gene set analysis results [9, 31]. We also visualize a gene set database as a graph, where each gene set G_i is represented as a vertex v_i, and there is an edge between two vertices v_i and v_j if $J(G_i, G_j) > 0$; the value of $J(G_i, G_j)$ is used as the weight for this edge. Since the Jaccard coefficient is symmetric, the graph defined using this measure is an undirected graph. Due to the sheer number of overlapping gene sets, such a graph has a large number of edges. To visualize substantial overlaps between gene sets, we only show overlap scores greater than or equal to 0.5, while retaining all vertices. In other words, in all graph visualizations in this paper, an edge between two vertices v_i and v_j indicates that their corresponding gene sets, i.e. G_i and G_j, share at least half of

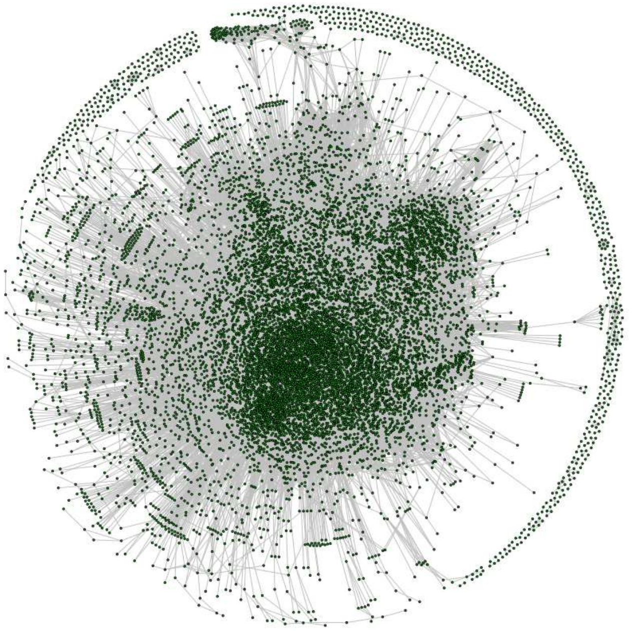


Figure 5.3: The graph representing the overlap between gene sets in MSigDB. In this graph, each vertex represents a gene set in MSigDB, and each edge represents an overlap with Jaccard coefficient greater than or equal to 0.5 between two gene sets (see Equation 5.3). The "hairball" is the result of a large number of gene sets with a substantial overlap (≥ 0.5) with each other.

their genes. The "hairballs" in Figure 5.3 and also Figures D.1 and D.2 (in the Appendix D) are due to the existence of a large number of edges, i.e. pairs of gene sets with a substantial amount of overlap. These graphs highlight the existence of gene set overlap as a ubiquitous phenomenon in gene set databases. The graph visualization can be generated using Fruchterman Reingold layout [19] in Gephi (version 0.9.2) [5].

To further visually inspect the gene set overlap in a given gene set database $\mathbb{G}$, we use a frequency plot. For each gene set G_i in $\mathbb{G}$, we calculate $f_i = \|\{G_j \mid J(G_i, G_j) > 0 \ (j \neq i) \ and \ G_j \in \mathbb{G}\}\|$. f_i is the number of gene sets G_j $(j \neq i)$ in $\mathbb{G}$ with a non-zero overlap with G_i. After calculating f_i values for all G_i in $\mathbb{G}$, we use a frequency plot to show the distribution of f_i values. Figure 5.4 and also Figures D.3, and D.4 (in the Appendix D) illustrate the distribution of f_i values for MSigDB, GeneSetDB, and GeneSigDB, respectively. These figures are in agreement with Figure 5.3, D.1, and D.2 and show the prevalence of gene set overlap in the aforementioned gene set databases.

Figure 5.4 suggests that overlap scores in MSigDB follow a multimodal distribution. This can be attributed to the fact that MSigDB is a meta-database that extracts gene sets from several sources including GO, KEGG, Reactome, and BioCarta. A compelling result revealed by Figure 5.4 is that majority of gene sets in MSigDB have at least a non-zero overlap with more than 1000 other gene sets in MSigDB (out of a total of 17778 gene sets). Also, there is no gene set in MSigDB without overlap with some other gene set(s). Finally, there are gene sets that overlap with the majority of gene sets in MSigDB. For example, the gene set associated with the "cellular response to organic substance" GO term (GO:0071310) has one non-zero overlap with 17292 gene sets. This gene set is associated with a general GO term and therefore overlaps a large number of gene

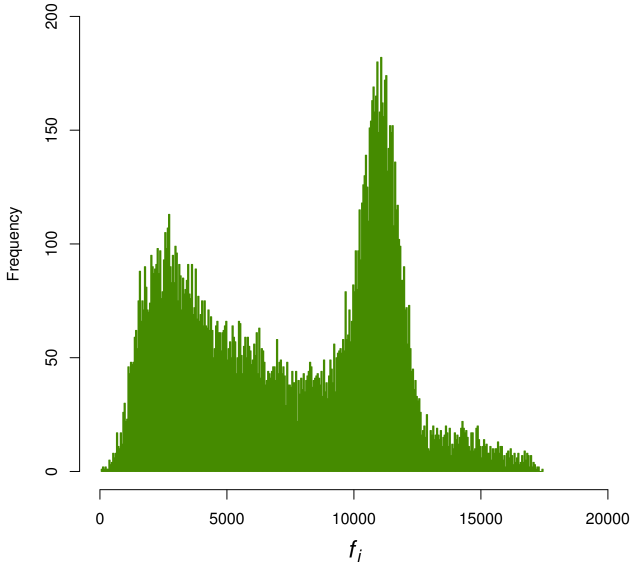


Figure 5.4: A frequency plot for f_i values in MSigDB illustrates the prevalence of gene set overlap. For each gene set G_i in a gene set database $\mathbb{G}$ (MSigDB here), f_i is the number of gene sets G_j ($j \neq i$) in $\mathbb{G}$ with $J(G_i, G_j) > 0$.

sets including the gene sets defined using relatively more specific GO terms.

5.4 Methodology

Evaluation of ORA using a quantitative measure such as specificity requires a gold standard dataset for which the differentially enriched gene sets are *a priori* known. Such a gold standard does not exist. In this section, we propose a methodology for a quantitative evaluation of the effect of gene set overlap on the specificity of ORA in the absence of such a gold standard dataset.

To perform ORA, a single gene analysis method must be conducted to predict the set of differentially expressed genes. This set serves as one of the inputs to ORA. In practice, often noise and biological variability introduce errors—i.e. false positives and false negatives—in the result of single gene analysis. In the context of single gene analysis, false positives are genes that are not differentially expressed but predicted as being so, and false negatives are genes that are differentially expressed but predicted as not being such. False negatives in single gene analysis may reduce the sensitivity of ORA, while false positives may reduce the specificity. To avoid the interference of the single gene analysis errors in the study of gene set overlap and its effect on the specificity of ORA, we assume that differentially expressed genes have been identified correctly; also, this is the same assumption that ORA relies on. Therefore, to perform the quantitative evaluation, a scenario in which all genes in a given gene set have been accurately detected as being differentially expressed is considered.

To deal with the absence of a gold standard dataset, in this paper the following procedure is used to identify the true enrichment status of gene sets. Given a gene set database $\mathbb{G} = \{G_j \mid 1 \leq j \leq m\}$ and L_i, a set of differentially expressed genes, and a fixed parameter γ, for each gene set $G_j \in \mathbb{G}$ we consider G_j as being truly differentially enriched if at least $100 \times \gamma$ percent of its members are differentially expressed genes, i.e. $\frac{\|G_j \cap L_i\|}{\|G_j\|} \geq \gamma$. Otherwise, G_j is considered as not being truly differentially enriched. γ serves as a threshold; since there is no consensus about such a threshold value, we repeat the main experiments for a wide range of values for γ, and we show that regardless of the value chosen for γ the results are consistent. In the rest of the paper, the set of truly differentially enriched and truly nondifferentially enriched gene sets are denoted by $T_i^+(\gamma)$ and $T_i^-(\gamma)$, respectively, and are defined as follows:

$$T_i^+(\gamma) = \{G_j \in \mathbb{G} \mid \frac{\|G_j \cap L_i\|}{\|G_j\|} \geq \gamma\} \tag{5.5}$$

$$T_i^-(\gamma) = \{G_j \in \mathbb{G} \mid \frac{\|G_j \cap L_i\|}{\|G_j\|} < \gamma\} \tag{5.6}$$

Hereafter, for the sake of brevity, we avoid writing the parameter γ; for example, we refer to $T_i^+(\gamma)$ and $T_i^-(\gamma)$ as T_i^+ and T_i^- respectively.

Given γ and L_i, Equations 5.5 and 5.6 determine the true enrichment status of all gene sets in $\mathbb{G}$. Knowing the true enrichment status of gene sets, we run ORA. The parameters (inputs) for running ORA are: a list of differentially expressed genes L_i, a significance level α, a background set U, and a gene set database $\mathbb{G} = \{G_j : 1 \leq j \leq m\}$.

In this research, the experiments were conducted using *python* version 3.6.2. To implement ORA, the *fisher_exact* method from the *stats* module of *scipy* version 0.19.1 was used. Also, the Benjamini-Hochberg FDR adjustment for multiple comparisons was performed using the *multipletests* method (with method parameter equal to *fdr_bh*) from *statsmodels* version 0.8.0.

For each gene set G_j in $\mathbb{G}$, ORA calculates a p-value p_j. After calculating $p_1, \ldots, p_m$—the p-values corresponding to the over-representation of gene sets $G_1, \ldots, G_m$ in $\mathbb{G}$ according to Equation 5.2—the Benjamini-Hochberg FDR adjustment [14] for multiple comparisons is applied. All gene sets with an adjusted p-value less than α are predicted as significant, i.e. being differentially enriched. $\mathbb{G}_i^+$ is defined as the set of all such significant gene sets. $\mathbb{G}_i^+$ includes both true positives and false positives. $\mathbb{G}_i^-$ is defined as the set of all nonsignificant gene sets, i.e. $\mathbb{G}_i^- = \mathbb{G} - \mathbb{G}_i^+$. $\mathbb{G}_i^-$ includes both true negatives and false negatives. For the given value of γ, true positives (TP_i), false positives (FP_i), true negatives (TN_i), and false negatives (FN_i) are identified based on Equations 5.7, 5.8, 5.9, and 5.10.

$$TP_i = \quad T_i^+ \cap \mathbb{G}_i^+ \tag{5.7}$$

$$FP_i = \quad \mathbb{G}_i^+ - T_i^+ \tag{5.8}$$

$$TN_i = \quad T_i^- \cap \mathbb{G}_i^- \tag{5.9}$$

$$FN_i = \quad \mathbb{G}_i^- - T_i^- \tag{5.10}$$

Using these values, specificity (SPC_i) is calculated according to Equation 5.11.

$$SPC_i = \quad \frac{\|TN_i\|}{\|TN_i\| + \|FP_i\|} \tag{5.11}$$

To be able to gain insight that is unbiased toward a single set L_i, this process is repeated many times, each time with a different L_i. We denote the set of all L_i as $\mathbb{L} = \{L_i \mid 1 \leq i \leq l\}$.

Algorithm 1 (see the Appendix D) illustrates the methodology for conducting the experiment. In each iteration of the algorithm, i.e. the outer loop, a gene set L_i from $\mathbb{L}$ is used, and the process is repeated for all gene sets in $\mathbb{L}$. In addition, for each set $L_i \in \mathbb{L}$, the overlap score of L_i with respect to gene set database $\mathbb{G}$, i.e. $O(L_i, \mathbb{G})$, is calculated according to Equation 5.4. Having overlap score and specificity measure for each $L_i \in \mathbb{L}$, the relationship between overlap and the specificity of ORA can be assessed using statistical methods (see Section 5.5).

5.5 Experimental results

To study the effect of gene set overlap on the specificity of ORA using Algorithm 1, MSigDB—one of the most widely used gene set databases devoted to gene set analysis—was used as the gene set database $\mathbb{G}$. Since ORA requires a list (set) of differentially expressed genes as input, Algorithm 1 requires a collection of such lists (denoted as $\mathbb{L}$ in the algorithm). ImmuneSigDB [20] version 6.0 was used to provide such a collection. ImmuneSigDB contains lists of differentially expressed genes, each created by identifying differentially expressed genes in a dataset extracted from Gene Expression Omnibus (GEO) [17]. Therefore, each list in ImmuneSigDB represents a set of differentially expressed genes derived from a high-throughput study.

To investigate the association between gene set overlap and the specificity of ORA results, first the overlap score $O(L_i, \mathbb{G})$ was calculated for each list L_i in ImmuneSigDB. In this experiment, a significance level $\alpha = 0.05$ and γ values equal to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99 were used. For each value of γ, Algorithm 1 was run to calculate SPC_i corresponding to each $L_i \in \mathbb{L}$. Figure 5.5 illustrates the relationship between gene set overlap and the number of false positives for $\gamma = 0.5$. As overlap score increases, we observe an increase in the number of false positives and therefore a decline in the specificity. We observed the same pattern for all the aforementioned values of γ.

To study the relationship between gene set overlap and the specificity of ORA, we used a statistical test of correlation. Choosing a proper test of correlation requires assessment of the normality assumption. To test the null hypothesis that specificity values are normally distributed, we used the Shapiro-Wilk test [37]. Table D.1 shows the test results for the aforementioned values of γ. Considering those results, as confirmed by the histogram in Figure D.5 (see Appendix D), we concluded that specificity values are not normally distributed. Therefore, a Spearman's rank correlation coefficient test, a nonparametric test, was conducted for each value of γ to test the null hypothesis that there is no correlation between specificity and overlap scores. Table 5.3 shows the result of this test for various values of γ. Considering these results, we concluded that there is a strong negative correlation between gene set overlap and specificity of ORA.

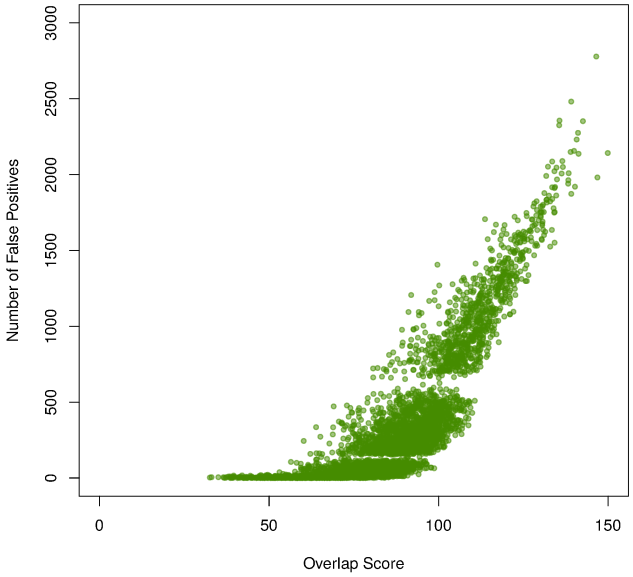


Figure 5.5: Number of false positives increases as overlap score increases ($\gamma = 0.5$). A similar pattern was observed for other values of γ.

Table 5.3: The result of Spearman rank correlation tests for different values of γ. All p-values are less than 0.0000001.

γ	r_s	p-value
0.10	-0.884064	<0.0000001
0.20	-0.880628	<0.0000001
0.30	-0.879913	<0.0000001
0.40	-0.879589	<0.0000001
0.50	-0.879366	<0.0000001
0.60	-0.879301	<0.0000001
0.70	-0.879302	<0.0000001
0.80	-0.879301	<0.0000001
0.90	-0.879307	<0.0000001
0.99	-0.879307	<0.0000001

5.6 Discussion

In this research, we proposed a systematic approach for evaluating the specificity of over-representation analysis. Using the proposed method, we demonstrated that there is a significant negative correlation between the specificity of ORA and gene set overlap. In other words, gene set overlap increases the number of false positives, i.e. gene sets incorrectly predicted as being differentially enriched. The increase in the number of false positives makes interpreting the results of ORA difficult and prone to investigator biases toward a hypothesis of interest. It also hinders reproducibility of gene set analysis results.

We also showed that gene set overlap is a ubiquitous phenomenon across gene set databases. The existence of multifunctional genes is one contributor to this phenomenon. Multifunctional genes are genes associated with several molecular functions or biological processes; therefore, they appear in several gene sets, contributing to gene set overlap. Multifunctional genes are commonplace; for example, Pritykin et al. [35] identified 2517 multifunctional genes in the human genome. As a consequence, gene set overlap is an integral characteristic of gene set databases. Another factor contributing to the prevalence of gene set overlap in databases that define some (or all) of their gene sets based on GO is the child-parent relationship between GO terms. GO terms are organized as a directed acyclic graph; each node represents a GO term; and each edge between two nodes represents a parent-child relationship between terms, with the child term being more specific than its parent term(s). Therefore, gene sets derived from GO terms that are involved in such child-parent relationships share common genes; this, in turn, contributes to the existence of gene set overlap.

Being an integral part of gene set databases, gene set overlap should be considered in the design and evaluation of gene set analysis methods. However, many gene set analysis studies have used simulated collections of non-overlapping gene sets for method evaluation and comparison [1], [18], [32]. Therefore, gene set overlap and its effect on the outcome of gene set analysis methods have been overlooked. We suggest using datasets that account for overlap as a requirement in the evaluation of gene set analysis methods.

Although many gene set analysis methods and tools have been developed, there are very few methods that consider gene set overlap. For example, PADOG is an attempt for addressing gene set overlap that leads to a small number of false positives (has high specificity), but its sensitivity has been reported to be lower than that of other gene set analysis methods (see Table S2 from the work by Tarca et al., 2013). SetRank is another gene set analysis method designed with gene set overlap in mind to increase specificity [38]. The authors of SetRank claimed that due to a lower number of false positives, the significant results reported by this method are more reliable than other methods. Therefore, it may be a viable solution for the lack of specificity of gene set analysis methods. A rigorous evaluation of the specificity and sensitivity of this method is suggested as future research.

The existence of gene set databases that accurately represent biological processes and functions is essential to the success of gene set analysis. Increasing the size of gene set databases by depositing more gene sets has been the common trend in developing gene set databases. The increase in the number of gene sets has introduced more gene set overlap, which in turn leads to a higher false positive rate. There is a need to focus on quality rather than sheer quantity in developing gene set databases. We suggest further research on the quality control of gene set databases.

Another suggestion for improving the specificity of current methods is to exclude irrelevant or uninformative gene sets before conducting gene set analysis. Considering the size of gene set databases, filtering these gene sets is laborious and, if done manually, prone to investigator bias toward gene sets considered “relevant”. Developing a computational approach for filtering irrelevant or uninformative gene sets would be worthwhile.

In the proposed method for evaluating ORA, we considered scenarios with only one differentially enriched gene set. In practice, a specific phenotype may be the result of altering several biological processes or functions, i.e. multiple gene sets. We expect that the differential enrichment of several gene sets intensifies the extent to which gene set overlap reduces specificity. In other words, we expect to see a larger number of false positives compared to the situation considered in this work. The proposed method is capable of handling scenarios with several differentially enriched gene sets. Also, Algorithm 1 can be used seamlessly with sensitivity or accuracy instead of specificity.

Since the input to ORA is a list of differentially expressed genes, we utilized ImmuneSigDB [20] for evaluating ORA. However, some gene set analysis methods require an expression matrix that represents expression level of genes under study across control and case samples. The proposed methodology is capable of evaluating such gene set analysis methods. To do so, the only requirement is developing expression profiles with the differentially enriched gene set(s) encoded in expression values. Therefore, our methodology can be used as a systematic approach to study specificity, sensitivity, and accuracy of other gene set analysis methods. For example, we suggest the study of the relationship between gene set overlap and the specificity of GSEA [39], which is another well-established gene set analysis method, as future work.

In the absence of gene set overlap, gene set analysis is a trivial problem, as many methods have achieved high specificity when being evaluated (by their authors) using simulated gene set databases with non-overlapping gene sets. If gene set overlap was considered in the evaluation of these methods, the lack of specificity of many gene set analysis methods would be obvious. For example, assume a gene set analysis method that uses average expression value of genes within a gene set (in control versus case samples) to predict the enrichment status of a gene set. Also assume that there is a single differentially expressed gene that appears in 100 gene sets. Such a method would report all 100 gene sets as being differentially enriched, while most of these gene sets might be biologically irrelevant. Therefore, we strongly recommend considering gene set overlap in any attempt for evaluating gene set analysis using simulated data.

5.7 Conclusion

In this paper, we proposed a systematic approach to study the effect of gene set overlap on the result of ORA (over-representation analysis). Using the proposed method and statistical analysis, we showed that there is a significant negative correlation between gene set overlap and specificity of ORA. We quantified gene set overlap and showed that it is a ubiquitous phenomenon across gene set databases. The proposed approach for the study of the relationship between gene set overlap and specificity of ORA can easily be used to investigate the effect of gene set overlap on different gene set analysis methods using quantitative measures such as specificity, sensitivity, and accuracy.

Considering the effect of gene set overlap on the results of ORA, it is essential to develop and use methods that address gene set overlap and achieve higher specificity without sacrificing sensitivity in the prediction of differentially enriched gene sets. Due to the lack of gold standard datasets, where the differentially enriched gene sets are known *a priori*, simulated datasets have been widely used for evaluation of gene set analysis methods. The databases used in these studies are often a collection of non-overlapping gene sets of the same size. This setting is substantially different from a real gene set database where gene set overlap is common. By completely ignoring gene set overlap, some methods achieve high specificity on simulated data but behave inadequately when working in real settings. We strongly recommend that the use of non-overlapping datasets be avoided for evaluation of gene set analysis methods.

References

[1] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, 2009.

[2] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.

[3] Hiromitsu Araki, Christoph Knapp, Peter Tsai, and Cristin Print. Genesetdb: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, 2:76 – 82, 2012.

[4] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[5] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Proceedings Of The Third International Conference On Weblogs And Social Media (ICWSM)*, volume 8, pages 361–362, 2009.

[6] Sebastian Bauer, Julien Gagneur, and Peter N. Robinson. Going bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Research*, 38(11):3523–3532, 2010.

[7] Tim Beißbarth and Terence P Speed. GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[8] Gabriel F Berriz, Oliver D King, Barbara Bryant, Chris Sander, and Frederick P Roth. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–2504, 2003.

[9] Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. Cluego: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093, 2009.

[10] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. GO:: Termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.

[11] Aedín C Culhane, Markus S Schröder, Razvan Sultana, Shaita C Picard, Enzo N Martinelli, Caroline Kelly, Benjamin Haibe-Kains, Misha Kapushesky, Anne-Alyssa St Pierre, William Flahive, et al. GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Research*, page D1060–D1066, 2011.

[12] Aedín C. Culhane, Thomas Schwarzl, Razvan Sultana, Kermshlise C. Picard, Shaita C. Picard, Tim H. Lu, Katherine R. Franklin, Simon J. French, Gerald Papenhausen, Mick Correll, and John Quackenbush. Genesigdb—a curated database of gene expression signatures. *Nucleic Acids Research*, 38(suppl 1):D716–D725, 2010.

[13] Sabine Dietmann, Wanseon Lee, Philip Wong, Igor Rodchenkov, and Alexey V Antonov. CCancer: a bird's eye view on gene lists reported in cancer-related studies. *Nucleic Acids Research*, 38(suppl 2):W118–W123, 2010.

[14] Sorin Drăghici. *Statistics and data analysis for microarrays using R and bioconductor*. CRC Press, 2016.

[15] Sorin Draghici, Purvesh Khatri, Rui P Martins, G Charles Ostermeier, and Stephen A Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, Feb 2003.

[16] Sorin Drăghici, Purvesh Khatri, Rui P Martins, G Charles Ostermeier, and Stephen A Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, 2003.

[17] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[18] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, pages 107–129, 2007.

[19] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[20] Jernej Godec, Yan Tan, Arthur Liberzon, Pablo Tamayo, Sanchita Bhattacharya, Atul J Butte, Jill P Mesirov, and W Nicholas Haining. Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. *Immunity*, 44(1):194–206, 2016.

[21] Ada Hamosh, Alan F Scott, Joanna Amberger, Carol Bocchini, David Valle, and Victor A McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–55, Jan 2002.

[22] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[23] Xiaoli Jiao, Brad T Sherman, Da Wei Huang, Robert Stephens, Michael W Baseler, H Clifford Lane, and Richard A Lempicki. David-ws: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13):1805–1806, 2012.

[24] Nicola A Johnson, Shiladitya Sengupta, Samir A Saidi, Khashayar Lessan, Stephen D Charnock-Jones, Laurie Scott, Richard Stephens, Tom C Freeman, Brian DM Tom, Michael Harris, et al. Endothelial cells preparing to die by apoptosis initiate a program of transcriptome and glycome regulation. *The FASEB Journal*, 18(1):188–190, 2004.

[25] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl_1):D428–D432, 2005.

[26] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–62, Jan 2016.

[27] Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L. Jenkins, Kathleen M. Jagodnik, Alexander Lachmann, Michael G. McDermott, Caroline D. Monteiro, Gregory W. Gundersen, and Avi Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 2016.

[28] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

[29] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[30] Steven Maere, Karel Heymans, and Martin Kuiper. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, 2005.

[31] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili, and Gary D Bader. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS One*, 5(11):e13984, 2010.

[32] Dougu Nam and Seon-Young Kim. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9(3):189–197, 2008.

[33] John C Newman and Alan M Weiner. L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biology*, 6(9):R81, 2005.

[34] Darryl Nishimura. Biocarta. *Biotech Software & Internet Report: The Computer Software Journal for Scientists*, 2(3):117–120, 2001.

[35] Yuri Pritykin, Dario Ghersi, and Mona Singh. Genome-wide detection and analysis of multifunctional genes. *PLOS Computational Biology*, 11(10):e1004467, 2015.

[36] Arthur L Shaffer, George Wright, Liming Yang, John Powell, Vu Ngo, Laurence Lamy, Lloyd T Lam, R Eric Davis, and Louis M Staudt. A library of gene expression signatures to illuminate normal and pathological lymphoid biology. *Immunological Reviews*, 210(1):67–85, 2006.

[37] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[38] Cedric Simillion, Robin Liechti, Heidi E.L. Lischer, Vassilios Ioannidis, and Rémy Bruggmann. Avoiding the pitfalls of gene set enrichment analysis with setrank. *BMC Bioinformatics*, 18(1):151, 2017.

[39] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric 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.

[40] Adi L Tarca, Gaurav Bhatti, and Roberto Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS One*, 8(11):e79217, 2013.

[41] Adi Laurentiu Tarca, Sorin Draghici, Gaurav Bhatti, and Roberto Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):136, 2012.

[42] Jing Wang, Suhas Vasaikar, Zhiao Shi, Michael Greer, and Bing Zhang. Webgestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research*, pages W130–W137, 2017.

[43] Gunnar Wrobel, Frédéric Chalmel, and Michael Primig. goCluster integrates statistical analysis and functional interpretation of microarray expression data. *Bioinformatics*, 21(17):3575–3577, 2005.

[44] Andrew Young, Nathan Whitehouse, J Cho, and C Shaw. Ontologytraverser: an R package for GO analysis. *Bioinformatics*, 21(2):275–276, 2005.

[45] Alexander C Zambon, Stan Gaj, Isaac Ho, Kristina Hanspers, Karen Vranizan, Chris T Evelo, Bruce R Conklin, Alexander R Pico, and Nathan Salomonis. Go-elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics*, 28(16):2209–2210, 2012.

[46] Barry R Zeeberg, Weimin Feng, Geoffrey Wang, May D Wang, Anthony T Fojo, Margot Sunshine, Sudarshan Narasimhan, David W Kane, William C Reinhold, Samir Lababidi, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4):R28, 2003.

[47] Barry R Zeeberg, Haiying Qin, Sudarshan Narasimhan, Margot Sunshine, Hong Cao, David W Kane, Mark Reimers, Robert M Stephens, David Bryant, Stanley K Burt, et al. High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of common variable immune deficiency (CVID). *BMC Bioinformatics*, 6(1):1, 2005.

[48] Bing Zhang, Stefan Kirov, and Jay Snoddy. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, 33(suppl 2):W741–W748, 2005.

6 Silver: Forging Almost Gold Standard Datasets for Evaluation of Gene Set Analysis Methods

Despite the existence of many studies comparing gene set analysis methods, there is no consensus regarding the method of choice for a given experiment, and existing guidelines and suggestions are most often contradictory [12] (See Appendix A for examples of such inconsistent and contradictory guidelines and recommendations). In view of these inconsistencies, it is necessary to identify and tackle obstacles leading to such disagreements. The lack of gold standard datasets is one of the main factors that prevents researchers from a sound and unbiased evaluation of gene set analysis methods. In the absence of gold standard datasets, where the differential enrichment status of gene sets are known prior to conducting gene set analysis, both simulated and real datasets have been used. In this paper, we discuss the shortcomings of the available approaches for evaluation of gene set analysis methods and propose a framework for evaluation of gene set analysis methods. This paper built on top of the methodology we published earlier (see reference below) for simulating kinome microarray data.

Maleki, F., & Kusalik, A. (2015). A Synthetic Kinome Microarray Data Generator. Microarrays, 4(4), 432-453.

Author contributions

Farhad Maleki designed and developed the framework, performed the statistical analysis and data visualization, and wrote the paper. Katie Ovens preprocessed microarray and RNA-Seq datasets, performed single gene analysis, and wrote the corresponding portion of the paper. Ian McQuillan helped with revising and clarifying the "Framework for Evaluating Gene Set Analysis Methods" Section. Anthony Kusalik supervised the research and helped edit and revise the paper.

Abstract

Gene set analysis has been widely used to gain insight from high-throughput expression studies. Although various tools and methods have been developed for gene set analysis, there is no consensus among researchers regarding best practice(s). Most often, evaluation studies have reported contradictory recommendations of which methods are superior. Therefore, an unbiased quantitative framework for evaluation of gene set analysis methods is valuable. Such a framework requires gene expression datasets where enrichment status of gene sets is *a priori* known. In the absence of such gold standard datasets, artificial datasets are commonly used for evaluation of gene set analysis methods; however, they often rely on oversimplifying assumptions that make them biased in favour of or against a given method. In this paper, we propose a quantitative framework for evaluation of gene set analysis methods by synthesising expression datasets using real data without relying on oversimplifying and unrealistic assumptions while preserving complex gene-gene correlations and retaining the distribution of expression values. The utility of the quantitative approach is shown by evaluating ten widely used gene set analysis methods. An implementation of the proposed method is publicly available.

6.1 Introduction

High-throughput technologies are widely used to monitor the expression activity of many genes in a single experiment. Analyzing high dimensional data resulting from these technologies is challenging. Gene set analysis, also known as enrichment analysis, is widely used to address this challenge and to gain insight from the resulting data. Gene set analysis employs *a priori* knowledge of groups of genes that are known to be associated with cellular components, processes, or functions. Such groups of genes, also referred to as gene sets or pathways, can be extracted from knowledgebases such as GO [2] and KEGG [9]. Hereafter, we refer to such a collection of gene sets as a gene set database. Given a gene set database and a case-control gene expression dataset, gene set analysis aims to find gene sets from the database that are differentially enriched when comparing case to control samples; for example, a pathway that is activated (or deactivated) in case samples when compared to controls.

Many gene set analysis methods have been developed [3, 8, 10, 11, 20, 23, 25] and it has been shown that different gene set analysis methods may lead to significantly different results in terms of gene sets reported as differentially enriched [16]. Considering the large number of available methods, it is natural to wonder which method should be used. Answering this question in a quantitative manner requires a gold standard expression dataset where differentially enriched gene sets are *a priori* known. Due to the absence of such gold standard datasets, real datasets with presumed enrichment status of gene sets [22, 26] and synthesized datasets [1, 7, 18] have been used. Unfortunately, these have had shortcomings and evaluating gene set analysis methods remains a challenge.

Evaluations using real datasets are often based on questionable assumptions about the differential enrichment status of the gene sets. For example, Tarca et al. [22] used 42 microarray datasets from GEO. For each dataset, they defined a particular KEGG or Metacore disease gene set to be differentially enriched whenever the dataset was from a similar phenotype. Such assumptions make this kind of evaluation unreliable.

Evaluations using synthesized datasets have also been conducted relying on oversimplifying assumptions that may not represent the true nature of real data [1, 7, 18]. Such evaluations have often used normally distributed expression values with no gene-gene correlations [7, 18] or constant correlations [1], even though more complex gene-gene correlations exist and profoundly impact the results of gene set analysis [21]. Moreover, normally distributed values with constant mean and standard deviation ignores heterogeneity of variance, which is a common phenomenon in high-throughput data [13]. Hence, the resulting datasets may be biased in favour or against a specific method or class of methods. Furthermore, these evaluations often utilize gene sets of equal size, as opposed to gene sets of varying sizes (the typical situation), which have been reported to affect some methods [5]. Another shortcoming of these evaluations is that they only consider non-overlapping gene sets. Therefore, overlap of gene sets and its effect on the specificity of gene set analysis [14] have been overlooked.

Recently Mathur et al. created artificial gene expression datasets in a more realistic manner [17]. They conducted a systematic comparison of 4 gene set analysis methods using real expression datasets by sampling with replacement from control and case samples; i.e. simulating controls (and also cases) from both actual controls and cases. They simulated differential expression “by shifting the gene expression in the [simulated] case groups according to a range of values”. The authors followed a bootstrap approach, repeating this procedure 100 times and reported the average behaviour of each gene set analysis method across the datasets. However, in each generated dataset, the simulated controls (and also cases) contained a heterogeneous mixture of both actual control and case samples; therefore, the simulated datasets were not representative of the actual data. This approach is also computationally demanding, and almost impractical for evaluating computationally expensive methods such as SetRank [19].

Despite the existence of many studies comparing gene set analysis methods, there is no consensus regarding the method of choice for a given experiment, and existing guidelines and suggestions are often contradictory [12]. In this research, we propose Silver, a framework for evaluating gene set analysis methods. The framework synthesizes gene expression datasets without relying on oversimplifying assumptions such as normally distributed expression values and zero or constant gene-gene correlations. The synthesized expression datasets preserve the true distribution of gene expression values and retain complex gene-gene correlation patterns. This approach incorporates gene set overlap, which has been shown to have a significant impact on the results of gene set analysis methods [14]. Also, it is computationally affordable as it does not rely on bootstrapping. We showcase the utility of Silver by providing a comprehensive evaluation of nine commonly used gene set analysis methods together with a recent method aimed at increasing specificity.

6.2 Framework for evaluating gene set analysis methods

Silver, the proposed framework for evaluation of gene set analysis methods, is presented in this section. The framework consists of a methodology for synthesizing "almost gold" standard expression datasets and a quantitative approach for comparing gene set analysis methods. Silver is available as a GitHub repository.

6.2.1 Synthesizing expression datasets

To synthesize an expression profile with n_C controls and n_T cases, first we identify an actual expression dataset $\Lambda = (\Lambda^C, \Lambda^T)$ where $\Lambda^C = \{A^{(C_1)}, \ldots, A^{(C_n)}\}$ are n control samples and $\Lambda^T = \{A^{(T_1)}, \ldots, A^{(T_{n'})}\}$ are n' case samples. Each $A^{(C_i)}$ and $A^{(T_j)}$ is a vector of the expression levels of m genes for some positive integer m. Also, it is required that $n \geq n_C + n_T$ and $n' \geq n_T$. Then, $\overline{\Lambda^C} = \{A^{(C_{i_1})}, \ldots, A^{(C_{i_{n_C}})}\}$ and $\overline{\Lambda^T} = \{A^{(C_{j_1})}, \ldots, A^{(C_{j_{n_T}})}\}$ are created through a random sampling without replacement so that $\overline{\Lambda^C}$ and $\overline{\Lambda^T}$ are disjoint subsets of Λ^C. $\overline{\Lambda^C}$ and $\overline{\Lambda^T}$ together form an expression matrix, where each column corresponds to a member of $\overline{\Lambda^C}$ or $\overline{\Lambda^T}$ and each row corresponds to the expression values for a gene g_k $(1 \leq k \leq m)$ across samples in $\overline{\Lambda^C}$ and $\overline{\Lambda^T}$. In other words, the generated expression matrix contains $n_C + n_T$ columns and m rows, where m is the number of genes in the original case and control samples.

Given a set $X \subset \{g_1, \ldots, g_m\}$, where X is a user input, for each gene g_t in X we adjust the expression levels of g_t in $\overline{\Lambda^T}$ by simulating differential expression through the following process. We first create $\Re$, which is a table of expression values with m rows and n_T columns; each column is selected from the actual cases Λ^T through random sampling without replacement. The columns of $\Re$ are $A^{(T_{\ell_1})}, \ldots, A^{(T_{\ell_{n_T}})}$, where $1 \leq \ell_1 < \cdots < \ell_{n_T} \leq n'$. Each row of $\Re$ represents the expression values for a gene across $A^{(T_{\ell_1})}, \ldots, A^{(T_{\ell_{n_T}})}$. This table is used to simulate differential expression of each gene in X according to some specified criterion. To simulate differential expression of a gene g_t in X by a given fold change $FC(g_t)$, we randomly choose a row e from rows of $\Re$ that satisfies the differential expression criterion considering the simulated control expression values for g_t (from $\overline{\Lambda^C}$) and $FC(g_t)$. Among criteria one can use are t-test, Wilcoxon Rank-Sum test, and median fold change. The current expression values for gene g_t in $\overline{\Lambda^T}$ (a vector of size n_T) are replaced with the vector of expression measures from row e of $\Re$. The choice of genes selected for differential expression (X) depends on the purpose of simulation. Note that given the initial expression level of g_t, if the intended fold change value—which is a user input—is unrealistically high or low considering the distribution of expression values in the original dataset, a row e that meets the criterion might not exist. In such cases, optionally, expression values for g_t can be shifted in a manner similar to Mathur et al. [17]. After updating expression values for genes in X, $(\overline{\Lambda^C}, \overline{\Lambda^T})$ is returned as the synthesized dataset. As the proposed method inherits the characteristics of an input dataset, care should be taken when choosing input datasets. Having a MDS (multidimensional scaling) plot of the input datasets where controls and cases cluster separately might be a good rule of thumb for choosing a dataset of reasonably high quality [16].

Another required component for gene set analysis is a gene set database. Instead of following the common approach of generating a small number of non-overlapping gene sets of equal size, we use real gene set databases to evaluate gene set analysis methods. This is possible as we synthesize expression profiles using real data and therefore retain real gene identifiers along with their gene expression characteristics from an actual dataset.

6.2.2 Quantitative approach

We utilize the aforementioned procedure to synthesize expression datasets where the enrichment status of given gene sets are known *a priori*. To achieve this goal, we select a group of gene sets $G_1, \ldots, G_q$—from a gene set database $\mathbb{G}$—and synthesize an expression dataset with genes in these gene sets being differentially expressed. However, not only do we need to consider $G_1, \ldots, G_q$ as being differentially enriched, but we also need to consider gene sets that "substantially overlap" with $G_1, \ldots, G_q$ as being differentially enriched. However, there is no consensus about what should be considered as a "substantial overlap". We use a methodology similar to that proposed by Maleki and Kusalik [14] to address this ambiguity and to determine the enrichment status of gene sets.

Assuming that L is the list of all genes that are differentially expressed in the synthetic dataset $(\overline{\Lambda^C}, \overline{\Lambda^T})$, we consider a gene set G_i as truly differentially enriched if the following inequality holds:

$$f(G_i, L) = \frac{\|G_i \cap L\|}{\|G_i\|} > \gamma$$

where γ is a value between 0 and 1 and $\| \bullet \|$ is set cardinality. $f(G_i, L)$ represents the proportion of genes in G_i that are differentially expressed. Hereafter, we refer to f as the coverage score of G_i given L or simply as the coverage score of G_i in situations where L can be inferred from the context. Since there is no consensus in the research community about an appropriate value of γ, we use a wide range of values for γ and evaluate a gene set analysis method for each value. Knowing the truly enriched gene sets in a simulated dataset and results of a gene set analysis method for that dataset, we can then quantitatively evaluate the result of a gene set analysis method in terms of sensitivity and specificity.

6.3 Results

Using the proposed framework, we evaluate nine commonly used gene set analysis methods: PAGE [10], GSEA (both gene permutation and phenotype permutation versions) [20], PLAGE [23], GAGE [11], ssGSEA [3], ROAST [25], GSVA [8], over-representation analysis (ORA) [6], as well as SetRank [19], a more recent method claiming to increase specificity. We use the following R packages in this study: *GSVA* package version 1.18.0 for GSVA, PLAGE, and ssGSEA; the *limma* package version 3.34.9 for ROAST; the *gage* package version 2.20.1 for PAGE and GAGE; the *SetRank* version 1.1.0 for SetRank. ORA was run using the WebGestalt online service [24]. GSEA was obtained from the Java-based application v3.0 (build 0160) at the Broad Institute software page for GSEA.

The gene set analysis methods were evaluated using data simulated from 2 microarray datasets and 1 RNA-seq dataset, downloaded from GEO. The microarray experiments were case-control experiments in humans from the Affymetrix GeneChip Human Genome U133 Plus 2.0 microarray platform from studies of renal cell carcinoma tissue (77 controls and 77 cases, GSE53757) and skin tissue in psoriasis patients (64 controls and 58 cases, GSE13355). These datasets were normalized as described in a previous work [15].

The RNA-seq dataset originated from normal and lesional psoriatic skin (82 controls and 92 cases, GSE54456). The 80-base single-stranded reads were trimmed with Trimmomatic 0.36 and mapped to the GRCh38 human genome using STAR 2.2.51 to obtain raw counts. The dataset was normalized using TMM normalization from the *edgeR* R package. The Ensembl gene IDs were translated to human Entrez gene IDs using *biomaRt*. Ensembl IDs (and also Probe IDs for microarrays) were collapsed to obtain a unique set of Entrez gene identifiers using methods described in a previous work [15]. GO gene sets (a total of 5917) were extracted from MSigDB version 6.1. For each expression dataset, genes not represented in the dataset were removed from these gene sets.

From each original dataset, we simulated a dataset containing 20 controls and 20 cases. Ten gene sets—(1) GO:0003823, (2) GO:0019724, (3) GO:0060070, (4) GO:0005126, (5) GO:0008009, (6) GO:0030851, (7) GO:0002544, (8) GO:0045087, (9) GO:0002253, and (10) GO:0006954—of various sizes (see Table 6.1) associated with immune system processes were selected for being differentially enriched in each simulated dataset. The genes in the ten gene sets were differentially expressed with mixed proportions of up- and down-regulated genes (see Table 6.1) and absolute log2 fold change values between 1 and 3. Hereafter, we refer to these ten gene sets as the target gene sets. Also, independent two-sample t-test was used as the differential expression criterion.

Figure 6.1 illustrates the volcano plot and also a Q-Q plot of the average expression value of cases versus controls for a synthesized dataset generated from GSE53757. This volcano plot resembles a typical volcano plot resulting from differential expression analysis. Further, a two-sample Kolmogorov-Smirnov test was used to assess if the average expression levels in a simulated dataset follows the same distribution as the real dataset it was generated from. As indicated in Table 6.2, the null hypothesis cannot be rejected for any of the datasets, suggesting that the distributions are the same.

Table 6.1: Information about the target gene sets in this study.

	Gene sets as numbered above									
	1	2	3	4	5	6	7	8	9	10
Up-regulated	24	27	65	206	24	3	7	245	155	218
Down-regulated	13	21	2	44	17	5	8	229	142	210
Size	76	67	92	250	41	15	15	538	383	428

Table 6.2: The results of KS-test for all datasets.

	statistic	p-value
GSE53757	0.010	0.28
GSE13355	0.006	0.79
GSE54456	0.011	0.26

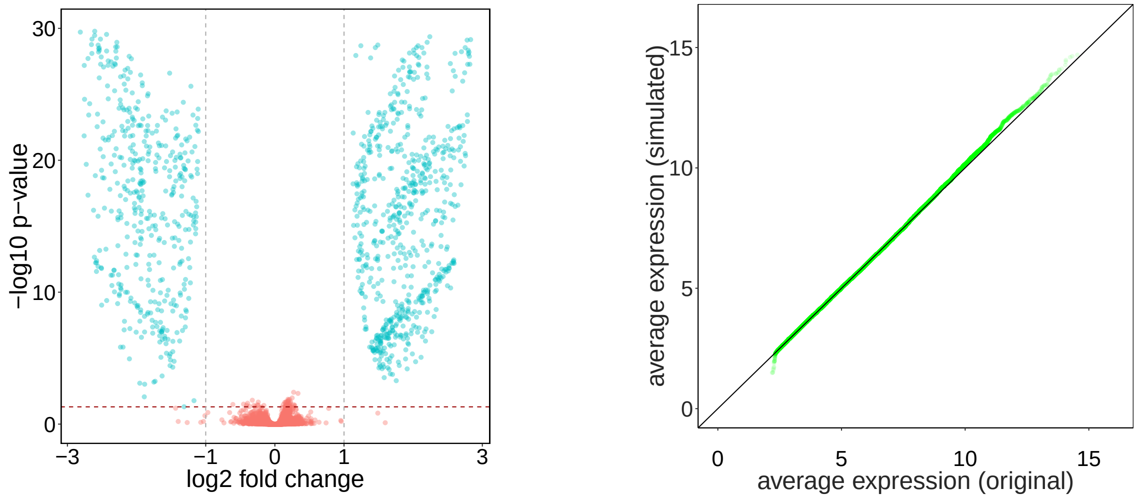


Figure 6.1: (Left) Volcano plot shows differentially expressed genes resulting from simulated data (20 control and 20 case samples) using dataset GSE53757. The blue points represent genes that were differentially expressed and the red points represent non-differentially expressed genes. The vertical dotted lines indicate the log fold change thresholds that were considered significant. The red horizontal line indicates the p-value cutoff 0.05. The p-values were obtained by performing differential expression analysis using the *limma* R package. (Right) A Q-Q plot of the average expression value of cases versus controls for a synthesized dataset generated from GSE53757.

We now demonstrate the utility of Silver as a means to evaluate the ten gene set analysis methods. For each method, the default parameters—as suggested by its author(s)—were used. To achieve comparable results, the Benjamini-Hochberg adjustment [4] for multiple comparison with a false discovery rate of 0.05 was applied to the reported p-values for each method.

Each plot in Figure 6.2 illustrates the reported p-values—resulting from running a method—for each gene set versus its coverage score given the list of differentially expressed genes for the dataset synthesized from GSE53757. These plots show the lack of specificity of the methods under study. Almost all methods reported a large number of gene sets as being differentially enriched regardless of the coverage scores. As depicted in Figure 6.2, ORA, GAGE, PAGE, and PLAGE reported gene sets with high coverage as being differentially enriched, while other methods reported some gene sets with high coverage—i.e. gene sets with a large proportion of their genes being differentially expressed—as non-enriched.

Figure 6.3 shows the rank of the target gene sets based on adjusted p-values reported by each method. The heat map shows that the ranking of the target gene sets substantially differs across methods, with some methods not being able to report some of the target gene set as differentially enriched. Also, GSEA-G and GSVA only ranked gene sets highly when most of its genes were up-regulated. GSEA-S and ssGSEA reported the majority of target gene sets near the bottom of their results. GAGE, PAGE, PLAGE, and ORA ranked the target gene sets higher in comparison to other methods. SetRank, while ranking six of the ten target gene sets highly, failed to report the other four target gene sets.

Figure 6.4 shows the Receiver Operator Characteristic (ROC) curves for the results of each method for all three synthetic datasets. The ROC curves suggest that GSEA (both gene permutation and phenotype

Table 6.3: The sensitivity (TPR) and specificity (TNR) of gene set analysis methods for data simulated from GSE53757.

	$\gamma = 0.1$		$\gamma = 0.3$		$\gamma = 0.5$		$\gamma = 0.9$		$\gamma = 0.99$	
Method	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR
GAGE	0.73	0.98	0.40	1.00	0.35	1.00	0.32	1.00	0.32	1.00
GSEA-G	1.00	0.05	0.98	0.07	0.97	0.03	0.97	0.14	0.97	0.16
GSEA-S	0.68	0.39	0.59	0.13	0.61	0.01	0.64	0.00	0.64	0.00
GSVA	1.00	0.07	0.99	0.16	0.98	0.19	0.96	0.42	0.96	0.44
PAGE	0.97	0.90	0.59	1.00	0.52	1.00	0.48	1.00	0.47	1.00
PLAGE	1.00	0.49	0.89	0.99	0.78	1.00	0.72	1.00	0.72	1.00
Roast	0.76	0.67	0.57	0.72	0.53	0.74	0.51	0.92	0.51	0.92
ssGSEA	0.01	0.98	0.01	0.96	0.01	0.94	0.02	0.97	0.02	0.96
SetRank	1.00	0.01	1.00	0.03	1.00	0.05	0.99	0.22	0.99	0.28
ORA	0.99	0.69	0.60	1.00	0.50	1.00	0.44	1.00	0.44	1.00

permutation versions) and ssGSEA performed poorly regardless of the value of γ. Also, GSVA performed moderately better than these methods. ORA, ROAST, PAGE, GAGE, and PLAGE achieved a relatively higher area under the curve. This supports the reliability of the most statistically significant results reported by these methods.

Table 6.3 shows the sensitivity and specificity of the methods under study when analyzing the dataset synthesized from GSE53757 across γ values. As depicted in Table 6.3, GSEA-G, GSVA, and SetRank achieve high specificity with the consequence of having low sensitivity. GAGE, PAGE, and ssGSEA achieve high sensitivity while sacrificing specificity, with ssGSEA being the least specific. These results are consistent across all γ values. PLAGE, while achieving high sensitivity (across $\gamma > 0.1$), also achieves 0.7 or higher specificity. However, due to the sheer size of gene set databases (5,000+ for GO gene sets, and 16,000+ for MSigDB), such a specificity is not high in absolute terms and leads to hundreds to thousands of false positives, depending on the database size.

Results using the other two datasets were consistent with the observations reported in Figures 6.2 and 6.3 and Table 6.3 (data not shown due to limited space).

6.4 Discussion

We proposed Silver, a framework for evaluating gene set analysis methods consisting of a method for synthesizing expression datasets and a quantitative approach for evaluating gene set analysis methods. While the proposed methodology does not generate gold standard datasets, it is capable of generating expression datasets without relying on common oversimplifying assumptions and preserves the characteristics of real (input) datasets; therefore, it is not limited to a specific gene expression platform. The synthesized datasets inherit the distribution of expression values and complex gene-gene correlations from real data, preserving technical and biological variability. This was expected as the proposed method incorporates real data and was confirmed by Kolmogorov–Smirnov tests and visualizations. Although we used the methodology for simulating expression datasets as part of an evaluation of gene set analysis methods, its utility is not limited to this role, and it can be used in any context where one needs expression datasets with control over differentially expressed genes.

Moreover, Silver utilizes real gene set databases to avoid using artificial databases of non-overlapping gene sets of equal size that are unrealistic and to also substantially affect the results of gene set analysis methods [14]. Using Silver, we evaluated a comprehensive list of gene set analysis methods providing key insights into weaknesses and strengths of these methods. A compelling observation revealed by Figures 6.2 and 6.3 is that some methods—such as ROAST, GSVA, SetRank, GSEA-G, and GSEA-S—did not report some gene sets as being differentially enriched even when all the genes in those gene sets were differentially expressed. For example, gene set GO:0002544—with 15 genes of which 7 were up-regulated and 8 were down-regulated—was not reported as differentially enriched by the aforementioned methods. This suggests an inadequacy of these methods in detecting gene sets with both up- and down-regulated genes. This leads to under-reporting of pathways in which by definition some genes must be up- and some down-regulated during a biological process or function.

The experiments showed that ssGSEA suffers from a lack of specificity, although it should be mentioned that the available R package implementing this method [8] was not from the authors of ssGSEA. We strongly recommend against using the current implementation in GSVA package version 1.18.0 with default parameters.

SetRank was designed with the goal of increasing specificity. The available implementation does not report the non-significant gene sets; as such, a fair comparison of it with other methods via ROC curves is not possible. However, its scatter plot in Figure 6.2 and its sensitivity and specificity in Table 6.3 reveal a lack of sensitivity. As illustrated by Figure 6.3, four out of the ten target gene sets were not detected by SetRank. This might be due to SetRank explicitly removing some gene sets based on a level of overlap between gene sets. However, excluding the small gene sets where most or all genes are differentially expressed suggests that SetRank sacrifices sensitivity in favour of increasing specificity.

In this study, we showcased the utility of Silver using a scenario of differentially enriching ten gene sets of processes related to immune system. However, this by no means is a comprehensive evaluation of these methods. We suggest studying various scenarios using gene sets from different phenotypes, only up-regulated (or down-regulated) gene sets of various sizes, and different levels of fold change gene expression values. In addition, we only evaluated gene set analysis methods using balanced datasets of 20 case and 20 control samples. The study of gene set analysis methods with different samples sizes and with unequal numbers of cases and controls is suggested as future research.

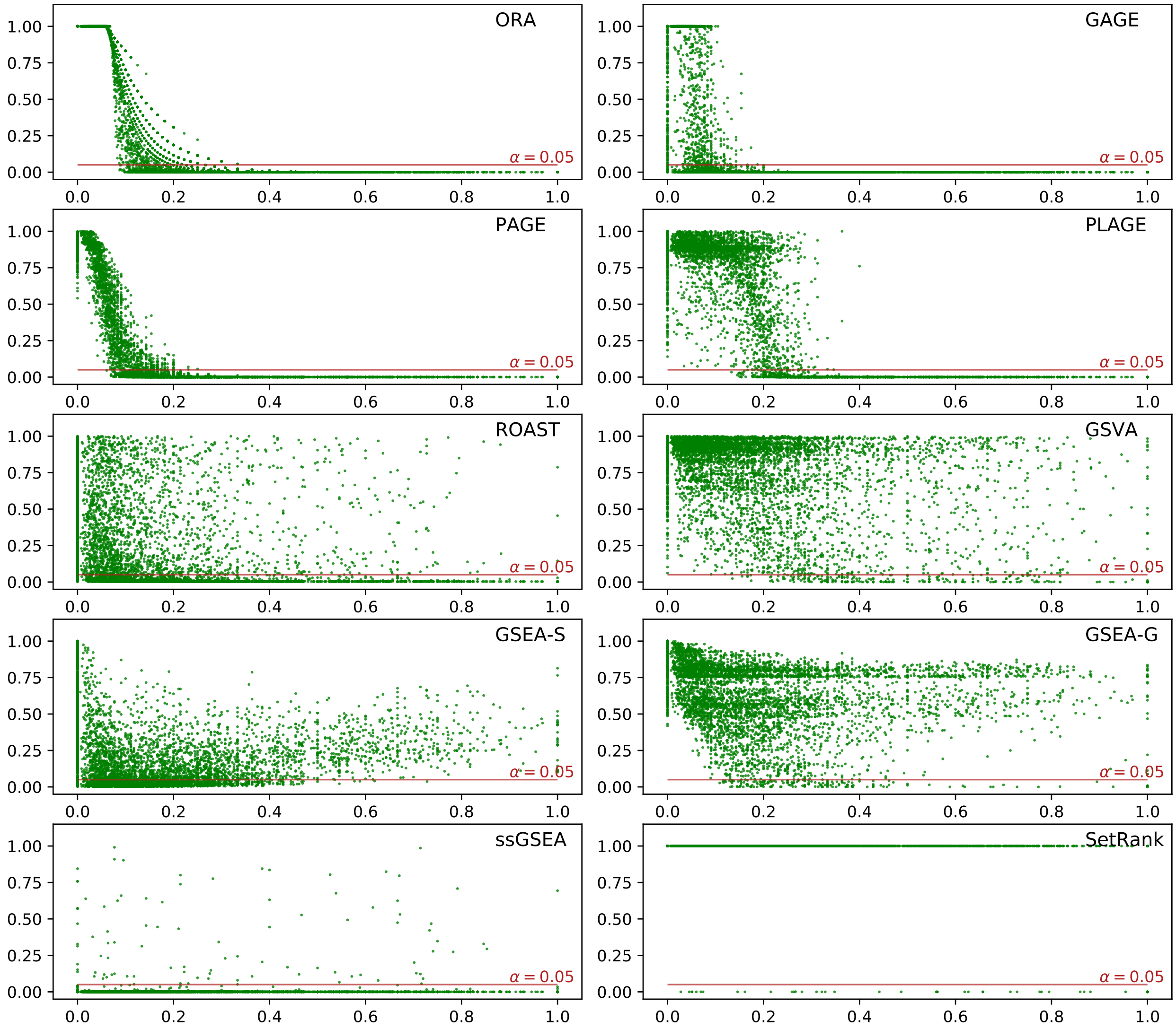
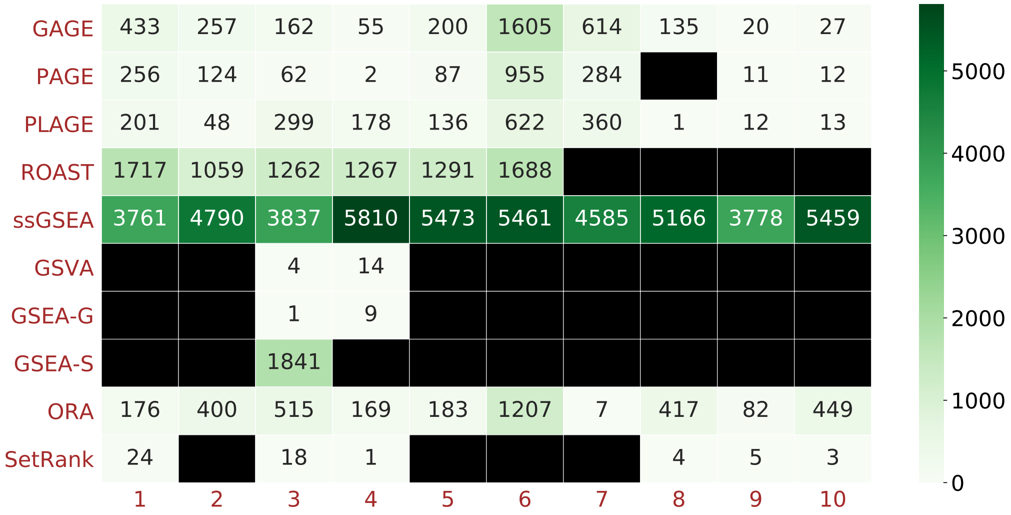


Figure 6.2: Scatter plots of the relationship between gene set coverage (x-axis) and the statistical significance (adjusted p-value) of the results of each method (y-axis). Each point in green represents a gene set. The red line shows a p-value cutoff of $\alpha = 0.05$. Since SetRank only returned statistically significant results (points under the red line), we assign a p-value of 1 to visualize the coverage scores for non-significant results.



	1	2	3	4	5	6	7	8	9	10
GAGE	433	257	162	55	200	1605	614	135	20	27
PAGE	256	124	62	2	87	955	284		11	12
PLAGE	201	48	299	178	136	622	360	1	12	13
ROAST	1717	1059	1262	1267	1291	1688				
ssGSEA	3761	4790	3837	5810	5473	5461	4585	5166	3778	5459
GSVA			4	14						
GSEA-G			1	9						
GSEA-S			1841							
ORA	176	400	515	169	183	1207	7	417	82	449
SetRank	24		18	1				4	5	3

Figure 6.3: Heat map of the rank of the 10 target gene sets as reported by each method. The results of each method were sorted based on the adjusted p-values (smallest to largest); the rank of each target gene set was determined as its rank in the sorted list. A black cell with no number shows that the adjusted p-value was not less than $\alpha = 0.05$.

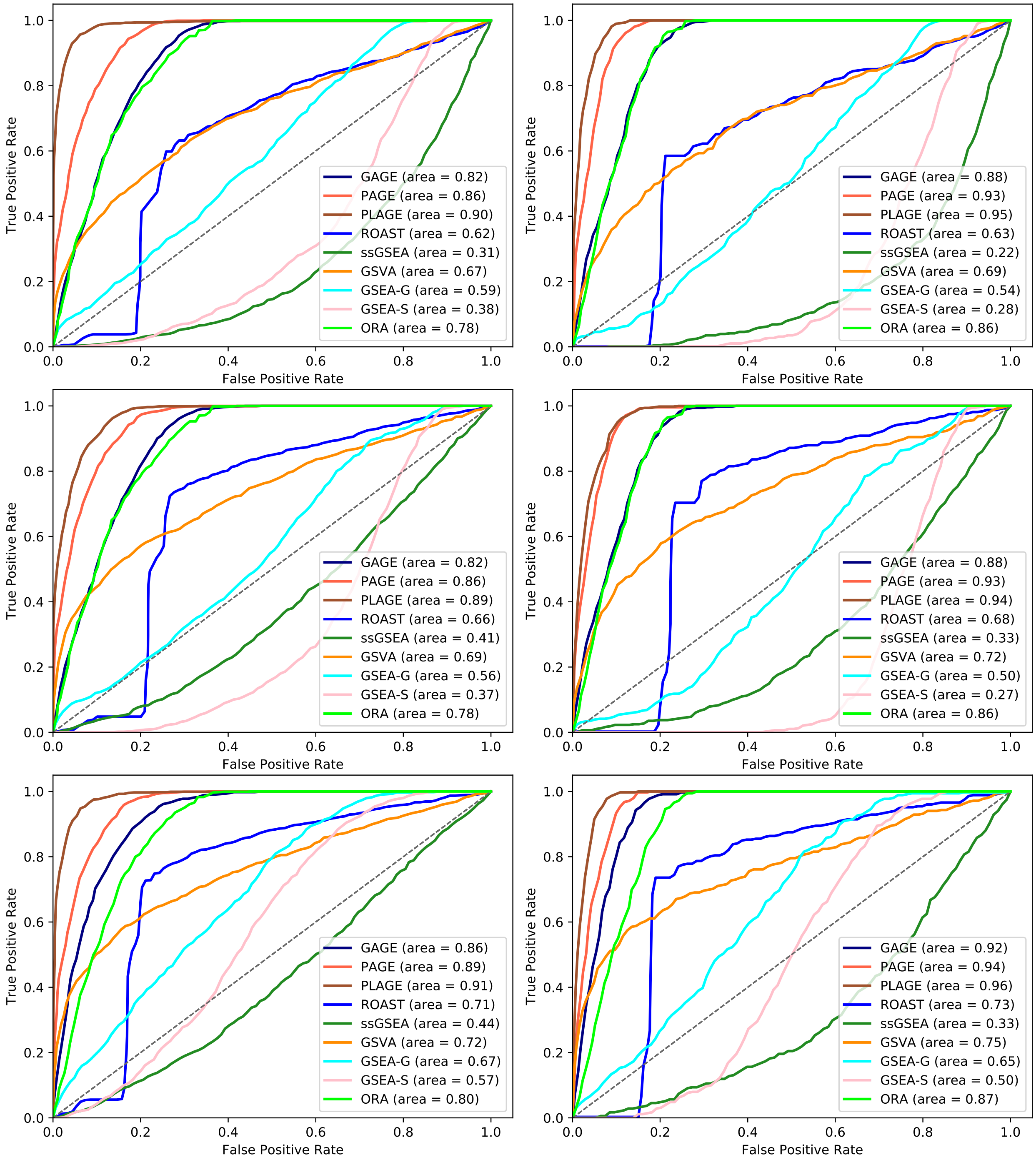


Figure 6.4: Receiver Operator Characteristic curves (ROC) for each method using $\gamma = 0.3$ (left column) and 0.5 (right column) for the dataset synthesized based on microarray dataset GSE53757, GSE13355, and RNA-Seq dataset GSE54456 (from top to bottom). The plots show the relationship between the true positive rate (y-axis) and the false positive rate (x-axis). A method with higher area under the curve (shown for each method) is considered better. The black dotted diagonal line (y=x) represents a method with random ordering of significance values.

6.5 Conclusion

In this paper, we proposed Silver, a framework for generating synthetic data that avoids common oversimplifying assumptions. We showed the utility of this framework by evaluating a comprehensive list of gene set analysis methods. The evaluation revealed key insights about these methods. It showed a lack of specificity as the main challenge facing these gene set analysis methods. Moreover, we found that some methods lack sensitivity when dealing with gene sets/pathways that are a mixture of up- and down-regulated genes.

Considering the key insights revealed using Silver, we strongly discourage using artificial datasets that rely on oversimplifying assumptions such as normally distributed expression values or non-overlapping gene sets of the same size, as they are not realistic and do not provide an accurate evaluation of gene set analysis methods. We anticipate that using Silver as a means for evaluation of existing and new gene set analysis methods will provide a better understanding of these methods and lead to development of gene set analysis methods that achieve a high specificity without sacrificing sensitivity.

References

[1] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, 2009.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[3] David A Barbie et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108, 2009.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995.

[5] Doris Damian and Malka Gorfine. Statistical concerns about the GSEA procedure. *Nature Genetics*, 36(7):663–663, 2004.

[6] Sorin Drăghici. *Statistics and data analysis for microarrays using R and bioconductor*. CRC Press, 2016.

[7] Bradley Efron and Robert Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.

[8] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14(1):7, 2013.

[9] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[10] Seon-Young Kim and David J Volsky. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(1):144, 2005.

[11] Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson, and Peter J Woolf. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1):161, 2009.

[12] Henryk Maciejewski. Gene set analysis methods: statistical models and methodological differences. *Briefings in Bioinformatics*, 15(4):504–518, 2013.

[13] Farhad Maleki and Anthony Kusalik. A synthetic kinome microarray data generator. *Microarrays*, 4(4):432–453, 2015.

[14] Farhad Maleki and Anthony J. Kusalik. Gene set overlap: An impediment to achieving high specificity in over-representation analysis. In *12th International joint conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*, Prague, Czech Republic, February 2019. (to appear).

[15] Farhad Maleki, Katie Ovens, Ian McQuillan, and Anthony J Kusalik. Sample size and reproducibility of gene set analysis. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 122–129. IEEE, 2018.

[16] Farhad Maleki, Katie L. Ovens, Elham Rezaei, Alan M. Rosenberg, and Anthony J. Kusalik. Method choice in gene set analysis has important consequences for analysis outcome. In *12th International joint conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*, Prague, Czech Republic, February 2019. (to appear).

[17] Ravi Mathur, Daniel Rotroff, Jun Ma, Ali Shojaie, and Alison Motsinger-Reif. Gene set analysis methods: a systematic comparison. *BioData Mining*, 11(1):8, 2018.

[18] Dougu Nam and Seon-Young Kim. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9(3):189–197, 2008.

[19] Cedric Simillion, Robin Liechti, Heidi E.L. Lischer, Vassilios Ioannidis, and Rémy Bruggmann. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics*, 18(1):151, 2017.

[20] Aravind Subramanian 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.

[21] Pablo Tamayo, George Steinhardt, Arthur Liberzon, and Jill P Mesirov. The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, 25(1):1–16, 2012.

[22] Adi L Tarca, Gaurav Bhatti, and Roberto Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One*, 8(11):e79217, 2013.

[23] John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(1):225, 2005.

[24] Jing Wang, Suhas Vasaikar, Zhiao Shi, Michael Greer, and Bing Zhang. Webgestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Research*, 45(W1):W130–W137, 2017.

[25] Di Wu, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E Visvader, and Gordon K Smyth. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010.

[26] Joanna Zyla, Michal Marczyk, and Joanna Polanska. Sensitivity, specificity and prioritization of gene set analysis when applying different ranking metrics. In *10th International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 61–69. Springer, 2016.

7 Gene Set Databases

Gene set databases, as an input to gene set analysis, have the potential to considerably impact the outcome of the analysis. Therefore, understanding the characteristics of these databases is vital to the success of gene set analysis. In this chapter, we propose and use quantitative measures for assessing the similarity between gene set databases and calculating the degree to which a given gene is represented in the gene sets of a given gene set database. We also propose a methodology to statistically determine whether a phenotype of interest is well-represented in a given gene set database. This paper has been accepted as a regular paper in the 10^{th} ACM Conference on Bioinformatics, Computational Biology, and Health Informatics.

Citation

Maleki, F., Ovens, K., McQuillan, I., Rezaei, E., Rosenberg, A. & Kusalik, A. (Sep. 2019). Gene Set Databases: A Fountain of Knowledge or a Siren Call? 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB 2019). Niagara falls, NY, USA. (to appear)

Author contributions

Farhad Maleki designed and developed the methodology for conducting the experiments, performed the statistical analysis and data visualization, and wrote the paper (except the “Data preprocessing” subsection and “Biomedical case study” section). Katie Ovens preprocessed the microarray datasets, performed single gene analysis, and wrote the “Data” Subsection. Also, the biomedical case study was conducted by Katie Ovens under the supervision of Alan Rosenberg. The “Biomedical case study” section was written by Katie Ovens. Elham Rezaei provided a list of genes associated with JIA based on literature evidence and provided comments on the gene sets potentially relevant to JIA. Ian McQuillan provided insightful feedback on the methodology and helped with editing the paper. Anthony Kusalik supervised the research and helped edit and revise the paper.

Abstract

Gene set analysis is a well-established approach for analyzing high-throughput gene expression data. Gene set databases used for gene set analysis may affect the outcome of the analysis. Therefore, understanding characteristics of these databases is vital to the success of gene set analysis. Due to the sheer size of the gene set databases, a comprehensive qualitative evaluation of these databases is impractical. In this paper, we quantitatively study several well-established gene set databases. We propose and use a quantitative measure for assessing the similarity between gene set databases. Also, we introduce a permeability score to quantify the degree to which a group of genes of interest co-occur in gene sets of a database. Using the permeability score, we propose a methodology to statistically determine whether a phenotype of interest is well-represented in a given gene set database. To study the effect of the choice of gene set database on the result of gene set analysis and show the utility of the permeability score, we conduct an experiment using two widely used gene set analysis methods and three expression datasets. The results suggest that the choice of gene set database might profoundly affect the outcome of the analysis. Also, our findings show that the permeability score can be used as a guideline for selecting an appropriate gene set database for a given study.

7.1 Introduction

Gene set enrichment analysis is a common methodology in high-throughput gene expression studies. For a case-control dataset, gene set analysis attempts to determine if a group of related genes shows different expression patterns in case samples versus controls. Conducting gene set analysis requires a collection of gene sets, hereafter refered to as a gene set database, containing groups of genes that are known to be associated with biological pathways, functions, processes, or cellular components. Gene sets can be extracted from knowledge bases such as Gene Ontology [4], KEGG [12], Reactome [8], and BioCarta [19].

MSigDb is a well-established gene set database designed for gene set analysis [14]. As a meta-database, it includes several gene set databases, each extracted from a different source, such as one of the aforementioned knowledge bases. MSigDB also contains gene sets constructed using differentially expressed genes resulting from expression studies of various phenotypes.

Gene set databases vary regarding the gene sets/pathways they represent. Adriaens et al. [1] compared pathways involved in fatty acid metabolism from several pathway databases including KEGG, BioCarta, and Reactome. Based on their observations, they concluded that "biological pathway content varies greatly in quality and completeness" [1]. In a different study of MSigDB, Tragante et al. reported a predominance of immunological and cancer-related pathways among curated gene sets extracted from canonical pathways (CP) such as BioCarta, KEGG, and Reactome, as well as gene sets from chemical and genetic perturbations (CGP) [26]. It is currently not known to what extent using different gene set databases impacts the results of gene set analysis. There are only a few general recommendations regarding gene set databases and no quantitative measure derived in a systematic way for choosing a gene set database for a given study. Among these general recommendations are: choosing databases consistent with the hypothesis of interest [18], using an up-to-date database [27], filtering out small genesets (with less than 15 genes) and large gene sets (more then 500 genes) [25], and combining databases to take advantage of different knowledge bases [26]. However, even these general recommendations are often method and data specific and might not be applicable to other gene set analysis methods. Moreover, small gene sets often are associated with more specific pathways, functions, processes, or cellular components; therefore, they can be more biologically informative, and filtering out all gene sets with less than 15 genes might lead to information loss.

Due to the sheer size of gene set databases, a fine-grained qualitative evaluation of these datasets by an expert is impractical. In this study, we propose new quantitative measures—similarity score, presence score, and permeability score—to compare four well-established gene set databases: KEGG, GO, Reactome, and BioCarta. We use similarity score to quantify the similarity between two gene set databases; presence score to measure the degree to which a gene is represented in a database; and permeability score to assess the co-occurrence of genes within gene sets of a database. Further, we statistically assess the significance of a permeability score for a given list of genes and a gene set database. For this purpose, we use lists of genes of interest from Human Phenotype Ontology (HPO) [13] as well as a list of genes compiled from literature. Also, we study the impact of the choice of gene set databases on the results of two widely used gene set analysis methods utilizing three microarray datasets.

The rest of the paper is organized as follows. Section 7.2 presents data acquisition and preprocessing as well as the quantitative measures and the methodology for the significance assessment of permeability scores. Section 7.3 presents the results of the quantitative study of the selected gene set databases as well as a biomedical case study of three juvenile idiopathic arthritis (JIA) datasets using two widely used gene set analysis methods across several gene set databases. The utility of the quantitative approach proposed in this paper is shown in Section 7.4 with a focus on discussion of the relationship between permeability score and gene set analysis results. Section 7.5 concludes the paper with recommendations for improving current gene set databases and a method to identify an appropriate gene set database for a given experiment.

7.2 Materials and Methods

7.2.1 Data preprocessing

The gene set databases used for evaluation included gene sets from GO, KEGG, Reactome, and BioCarta, which are subsets of MSigDB version 6.2 [14]. We also conducted the analyses using the entire MSigDB version 6.2. Further, two additional gene set databases were derived to investigate how gene set analysis results differ when using more current versions of the databases included in MSigDB. These databases were constructed as follows. A current version of KEGG was downloaded using the KEGGREST v1.22.0 package from R. All the pathways for human (KEGG organism ID: hsa) were downloaded (a total of 321 gene sets).

Hereafter, we refer to this database as KEGG*.

A GO gene set database was extracted using the GO.db R package, which contains a snapshot of GO gene sets from October 10, 2018. A gene set database was constructed by extracting all GO terms for human and associating them with Entrez gene identifiers annotated with each GO term using the org.Hs.eg.db R package. For each gene set associated with a term, genes with only "IEA" (electronically inferred) as their evidence of association were removed from the gene set. We refer to this database, which contains 5,509 gene sets, as GO*.

Three JIA case-control datasets collected using the Affymetrix GeneChip Human Genome U133 Plus 2.0 microarray platform were obtained from Gene Expression Omnibus (GEO). The three datasets were from peripheral blood mononuclear cells (PBMCs): 1) GSE13501 (59 controls, 21 systemic JIA samples), 2) GSE26554 (23 controls and 36 polyarthritis samples), and 3) GSE7753 (30 controls, 17 systemic JIA samples). The GEOquery v2.46.15 R package was used for reading CEL files. The data were normalized using RMA normalization from the affy v1.56.0 R package. Probe IDs were converted to Entrez gene identifiers using the hgu133plus2.db v3.2.3 R package. Finally, the *collapseRows* function from WGCNA v1.61 with the *MaxMean* method was used to collapse the expression measures based on Entrez gene identifiers.

7.2.2 Gene set database similarity

We measure the overlap between two genes sets G_i and G_j using the Jaccard index [5], which is defined as follows:

$$J(G_i, G_j) = \frac{\|G_i \cap G_j\|}{\|G_i \cup G_j\|} \tag{7.1}$$

For given sets G_i and G_j, $J(G_i, G_j)$ is a number between 0 and 1, with 0 representing no overlap between two sets (disjoint sets) and 1 representing complete overlap between G_i and G_j (set equality). We use the terms Jaccard index and overlap score interchangeably.

To assess the similarity between gene set databases $\mathbb{G}$ and $\mathbb{G}'$, where $\mathbb{G} = \{G_1, \ldots, G_n\}$, $\mathbb{G}' = \{G'_1, \ldots, G'_m\}$ and G_i $(1 \leq i \leq n)$ and G'_j $(1 \leq j \leq m)$ are gene sets, we use the following equation.

$$\mathbb{S}(\mathbb{G}, \mathbb{G}') = \frac{\sum_{i=1}^{n} \max_{1 \leq j \leq m} J(G_i, G'_j)}{2 \times \|\mathbb{G}\|} + \frac{\sum_{j=1}^{m} \max_{1 \leq i \leq n} J(G'_j, G_i)}{2 \times \|\mathbb{G}'\|} \tag{7.2}$$

Note that $\mathbb{S}$ is the summation of two components. In the first component, for each gene set G_i in $\mathbb{G}$, we find a gene set from $\mathbb{G}'$ that is the most similar to G_i, i.e. a gene set that has highest overlap with G_i. The second component performs the same task for each gene set G'_j in $\mathbb{G}'$. $\mathbb{S}(\mathbb{G}, \mathbb{G}')$ is a value between 0 and 1, with 0 showing no similarity between $\mathbb{G}$ and $\mathbb{G}'$ and 1 showing highest similarity, i.e. equality. Each of the two fractions is at most 0.5 due to the division by two, and therefore the similarity score is between 0 and 1.

7.2.3 Presence score

Genes vary in the degree to which they appear in different gene sets of a database. In an extreme case, a gene might not appear in any gene set within a database. We define presence score ($P(g, \mathbb{G})$) to quantify the presence of a gene g within a gene set database $\mathbb{G}$.

$$P(g, \mathbb{G}) = \sum_{G_i \in \mathbb{G}} I(g, G_i) \tag{7.3}$$

where $I(g, G_i)$ is defined as follows:

$$I(g, G_i) = \begin{cases} 1 & g \in G_i \\ 0 & otherwise \end{cases}$$

For example, $P(g, \mathbb{G}) = 0$ means that g is not in any gene sets of $\mathbb{G}$. To further explore the presence of a gene in more specific gene sets, we also calculate the presence score for subsets of $\mathbb{G}$ that contain gene sets of a size less than or equal to a given threshold t. We use $\mathbb{G}|_t$ to represent a gene set database $\mathbb{G}$ that is restricted to gene sets of size less than or equal to t, and we consider $P(g, \mathbb{G}|_t)$. In this paper, we use threshold values of 500, 250, 100, and 50.

To statistically assess if genes are presented equally across the databases, we use a Friedman test, which is an analogous nonparametric version of ANOVA for repeated measures. Presence scores are calculated for a total of 25,052 genes acquired from the Ensembl human genome version 86 as Entrez gene identifiers. Hereafter, we refer to this list of genes as E. To make presence scores comparable, they are normalized by dividing by their corresponding database size.

7.2.4 Permeability score

While the presence score provides insight about the degree to which a gene is represented in a gene set database, it does not measure the co-occurrence of genes from a given list. It is also affected by large gene sets that often are general and less informative; such large gene sets inflate the presence score. The coverage score of L by G_i is $C(G_i, L) = \frac{G_i \cap L}{G_i}$, that represents the proportion of genes in G_i that are present in L. The maximum achievable coverage score of L by $\mathbb{G}$ is defined as follows:

$$T(\mathbb{G}, L) = \max_{G_i \in \mathbb{G}} C(G_i, L) \tag{7.4}$$

To quantify co-occurrences of a list of genes of interest L for a given gene set database $\mathbb{G}$ for a given threshold τ, we define the permeability score of L in $\mathbb{G}$ as follows:

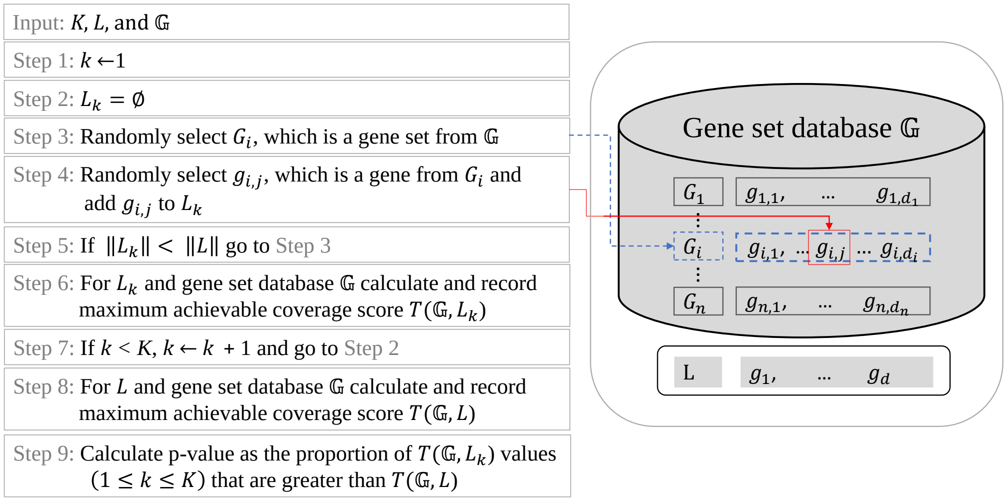
$$\Delta(\mathbb{G}, L, \tau) = \sum_{G_i \in \mathbb{G}} H\left(G_i, L, \tau\right) \tag{7.5}$$

where $H(G_i, L, \tau)$ is defined as follows:

$$H(G_i, L, \tau) = \begin{cases} 1 & C(G_i, L) \geq \tau \\ 0 & otherwise \end{cases}$$

where τ is a given threshold for the proportion of genes in L that are also in G_i. $\Delta(\mathbb{G}, L, \tau)$ represents the number of gene sets in $\mathbb{G}$ where at least $\tau \times 100\%$ of their genes are from L. Notice that the maximum achievable coverage score is the maximum value of τ such that $\Delta(\mathbb{G}, L, \tau) > 0$ or zero if there is none. Thus, for any τ greater than the maximum achievable coverage score, $D(\mathbb{G}, L, \tau) = 0$.

In order to statistically assess if the maximum achievable coverage score for a given list of genes L by a gene set database $\mathbb{G}$ is not solely due to chance, we use a bootstrapping hypothesis test. Fig. 7.1 provides a step-by-step description of the bootstrapping procedure.



Input: K, L, and $\mathbb{G}$

Step 1: $k \leftarrow 1$

Step 2: $L_k = \emptyset$

Step 3: Randomly select G_i, which is a gene set from $\mathbb{G}$

Step 4: Randomly select $g_{i,j}$, which is a gene from G_i and add $g_{i,j}$ to L_k

Step 5: If $\|L_k\| < \|L\|$ go to Step 3

Step 6: For L_k and gene set database $\mathbb{G}$ calculate and record maximum achievable coverage score $T(\mathbb{G}, L_k)$

Step 7: If $k < K$, $k \leftarrow k + 1$ and go to Step 2

Step 8: For L and gene set database $\mathbb{G}$ calculate and record maximum achievable coverage score $T(\mathbb{G}, L)$

Step 9: Calculate p-value as the proportion of $T(\mathbb{G}, L_k)$ values $(1 \leq k \leq K)$ that are greater than $T(\mathbb{G}, L)$

Gene set database $\mathbb{G}$
G_1
$g_{1,1}$, ... g_{1,d_1}
G_i
$g_{i,1}$, ... $g_{i,j}$ ... g_{i,d_i}
G_n
$g_{n,1}$, ... g_{n,d_n}
L
g_1, ... g_d

Figure 7.1: The methodology for significance assessment of the maximum achievable coverage scores. This procedure assesses whether or not a maximum achieved coverage score of a list of genes in a gene set database is due to chance. $L = \{g_1, \ldots, g_d\}$ is a list of genes; $\mathbb{G} = \{G_1, \ldots, G_n\}$ is a gene set database. L, $\mathbb{G}$, and K are inputs, with K representing the number of samplings for the significance assessment. We used $K = 1000$ in this paper. G_i $(1 \leq i \leq n)$ is a set of genes $\{g_{i,1}, \ldots, g_{i,j}, \ldots, g_{i,d_i}\}$ from $\mathbb{G}$. Also, $\emptyset$ represents the empty set.

The three measures presented in this section, as well as the proposed bootstrapping method, provide guidance for selecting an appropriate gene set database for an experiment, rather than an ad-hoc approach which is currently the common practice. Given two databases $\mathbb{G}$ and $\mathbb{G}'$ with a low similarity score, i.e. the existence of substantial differences between these two databases, the natural question is which gene set database should be used for a given experiment? Often a list L of genes that are known or hypothesized to be associated with the phenotype of interest is available. The presence score and permeability score help to choose the gene set database that is more representative of the phenotype of interest. If the majority of genes in L achieve a zero or low presence score for a database, that database is inappropriate for conducting the gene set analysis. The gene set database that achieves a larger permeability score for higher values of τ better represents the phenotype of interest. Finally, the bootstrapping approach presented in this section assesses if the maximum achievable coverage score is due to chance or not.

7.3 Results

Fig. 7.2 shows the number of gene sets available in major subcategories of MSigDB, including Biocarta, KEGG, Reactome, and GO. As the plot shows, there was a major update for gene sets derived from GO in version 5.2. Further, GO gene sets were the same for versions 3.1, 4.0, 5.0, and 5.1, with no update from 2012 to 2016. Also, BioCarta, KEGG, and Reactome were not updated between versions 3.1 and 6.2, while these datasets have been actively updated outside of MSigDB. We obtained KEGG* and GO*, updated versions of KEGG and GO, for further comparisons. Our observations showed that KEGG* and GO* substantially differ from their counterparts from MSigDB version 6.2. For example, KEGG* contains 321 pathways (gene sets), which is almost twice the number of gene sets in the KEGG category of MSigDB version 6.2 (186 gene sets).

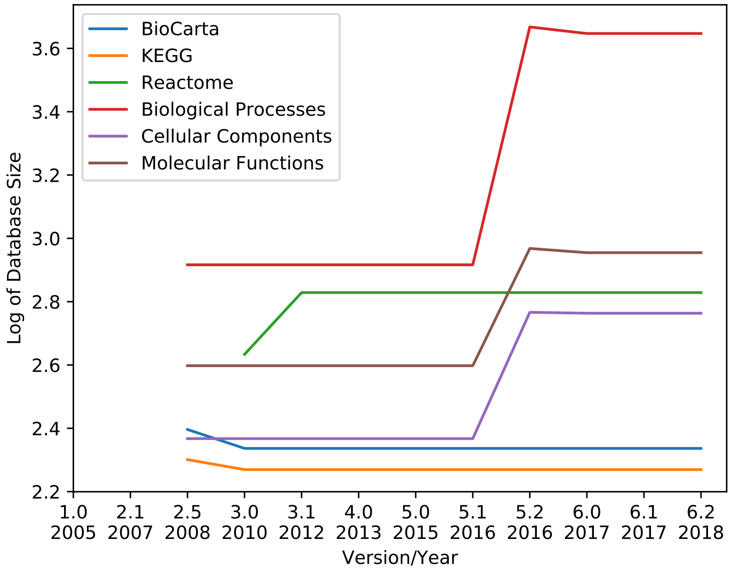


Figure 7.2: The change in the number of gene sets within the major subcategories of MSigDB across different versions. All of these databases have been added since version 2.5 of MSigDB.

Table 7.1 shows the similarity scores (Equation 7.2) between the gene set databases under study. The highest achieved similarity scores were between KEGG and KEGG* with a similarity score of 0.68, followed by GO and GO* with a similarity score of 0.36. These results are not surprising considering the differences in the construction of GO and GO* and also KEGG and KEGG* (see Section 7.2). All other pairwise comparisons of these gene set databases resulted in similarity scores below 0.25.

For a given gene set database $\mathbb{G}$ and the list of genes E, we calculated $P(g, \mathbb{G})$ for each gene in E. A Friedman test for assessing equality of presence score across gene set databases resulted in a chi-square statistic of 32443.38 and a p-value $<$ 1.00E-8 for the normalized presence scores. This suggests that there is a significant difference between the normalized presence scores across gene set databases.

Also, we used Human Phenotype Ontology (HPO) to extract a gene list associated with "Arthritis"

Table 7.1: Similarity scores for pairwise comparison of gene set databases GO, GO*, KEGG, KEGG*, BioCarta, and Reactome.

Method	GO	GO*	KEGG	KEGG*	BioCarta	Reactome
GO	1.00	0.36	0.19	0.18	0.11	0.25
GO*	0.36	1.00	0.17	0.15	0.13	0.24
KEGG	0.19	0.17	1.00	0.68	0.11	0.23
KEGG*	0.18	0.15	0.68	1.00	0.11	0.22
BioCarta	0.11	0.13	0.11	0.11	1.00	0.16
Reactome	0.25	0.24	0.23	0.22	0.16	1.00

(HP:0001369). Since the majority of gene sets contain less than or equal to 250 genes, as shown in Fig. 7.3, we restricted the gene set databases to gene sets of size 250 or less to calculate presence scores. Fig. 7.4 shows plots of the presence scores for "Arthritis" gene list across the gene set databases. The genes in this list are most represented within the GO database. GO* also has each gene from this list in at least one gene set. However, gene set databases like BioCarta and Reactome have far less representation across the genes of this list. In BioCarta in particular, more than half of the genes are not represented at all within the database.

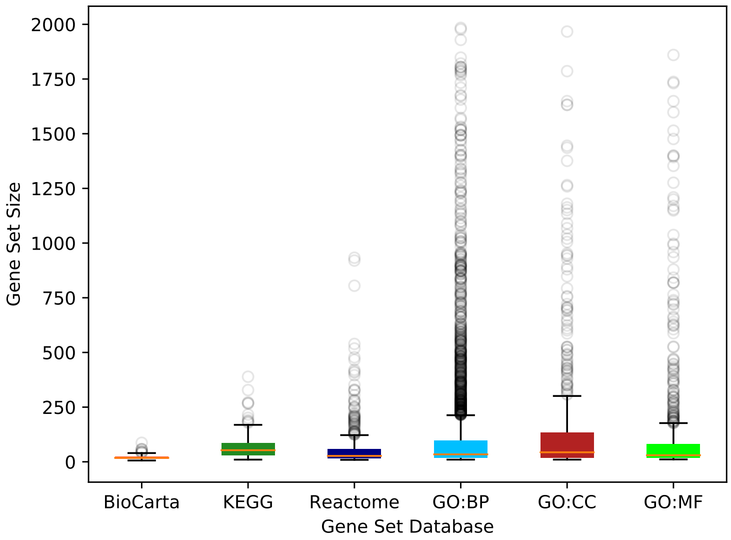


Figure 7.3: Distribution of gene set size for BioCarta, KEGG, Reactome, GO:BP, GO:CC, and GO:MF—each extracted from MSigDB V6.2—where GO:BP, GO:CC, and GO:MF represent GO gene sets from Biological Processes, Cellular Components, and Molecular Functions. GO contains large gene sets of up to 1984 genes. More specifically, the biological processes, a subcategory of GO, contains more gene sets of large sizes (outliers). Outliers are represented by circles.

Fig. 7.5 depicts the number of genes present in GO, GO*, KEGG, KEGG*, BioCarta, and Reactome, where these databases have been restricted to gene sets of size t or less where t equals 500, 250, 100, or 50. Fig. 7.5 also shows the number of genes present in each database with no restriction applied. In general databases such as BioCarta, KEGG, and Reactome do not cover the majority of the known genes in humans. Not all known genes for human are represented within some of these databases, with even fewer genes represented in gene sets with sizes less than 50 genes ($\mathbb{G}|_{50}$). Also, while the Ensembl human genome version

86 includes 25,052 Entrez gene identifiers, MSigDB contains over 30,000 Entrez gene identifiers.

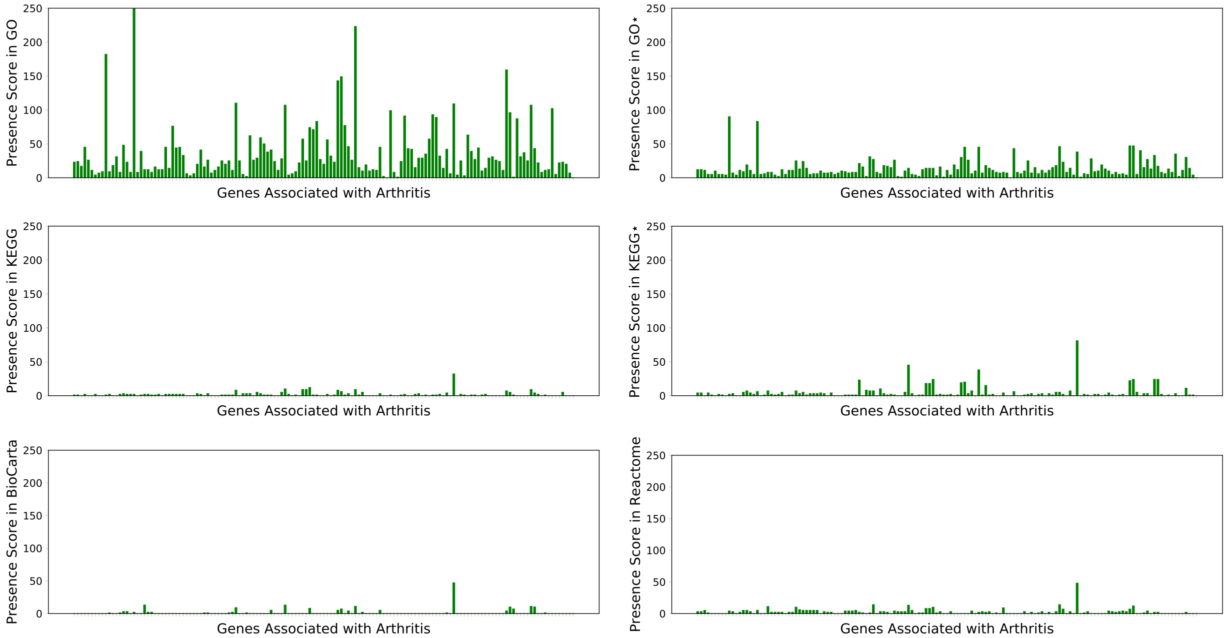


Figure 7.4: The presence scores for genes in the "Arthritis" gene set (HP:0001369) from HPO. All databases have been filtered to only include gene sets of size 250 or less.

The plots in Fig. 7.6 show the permeability scores for genes associated with four phenotypes extracted from HPO across thresholds of values (τ). Some databases show higher permeability scores than others depending on the phenotype of interest. Table 7.2 illustrates the results of the significance assessment of the maximum achievable coverage scores for the gene set databases under study and the phenotypes from HPO as well as a list of genes associated with JIA extracted from literature.

Significant maximum achievable coverage values were reported for different gene set databases depending on the phenotype of interest. The gene set for arthritis (HP:0001369) had a significant maximum achievable coverage score of 0.58 in GO. None of the other databases, including GO* with a maximum achievable coverage score of 0.50, had a statistically significant maximum achievable coverage score. The leukemia gene set (HP:0001909) resulted in a significant maximum coverage values in GO*, followed by GO. Reactome, BioCarta, and both versions of KEGG did not show significant scores for the genes associated with this phenotype. Abnormal leukocyte count (HP:0011893) had significant maximum achievable coverage values in KEGG and both versions of GO, while abnormal ciliary motility (HP:0012262) resulted in significant maximum achievable coverage values in Reactome and both versions of GO. The list of genes associated with JIA only had significant maximum coverage values reported for GO and GO*.

Table 7.2: The results of significance assessment of maximum achievable coverage score for GO, GO*, KEGG, KEGG*, Reactome, BioCarta, and MSigDB. Five gene lists (L) were extracted from HPO as well as a list of genes reported to be associated with JIA extracted from the literature.

Phenotype (L)	Database ($\mathbb{G}$)	Max achievable coverage	p-value
Arthritis (HP:0001369)	GO	0.583	0.000
	GO*	0.500	0.254
	KEGG	0.171	1.000
	KEGG*	0.185	1.000
	BioCarta	0.250	1.000
	Reactome	0.250	0.998
	MSigDB	0.583	0.005
Leukemia (HP:0001909)	GO	0.364	0.014
	GO*	0.600	0.005
	KEGG	0.193	0.931
	KEGG*	0.174	0.997
	BioCarta	0.381	1.000
	Reactome	0.280	0.897
	MSigDB	0.625	0.000
Abnormal leukocyte count (HP:0011893)	GO	0.615	0.001
	GO*	0.667	0.048
	KEGG	0.743	0.008
	KEGG*	0.703	0.636
	BioCarta	0.545	1.000
	Reactome	0.560	0.411
	MSigDB	1.000	0.000
Abnormal cilliary motility (HP:0012262)	GO	1.000	0.000
	GO*	1.000	0.000
	KEGG	0.016	1.000
	KEGG*	0.026	1.000
	BioCarta	0.000	1.000
	Reactome	0.100	0.982
	MSigDB	1.000	0.000
JIA (Appendix E)	GO	0.273	0.001
	GO*	0.429	0.001
	KEGG	0.086	0.694
	KEGG*	0.086	0.773
	BioCarta	0.176	0.916
	Reactome	0.100	0.825
	MSigDB	0.286	0.011

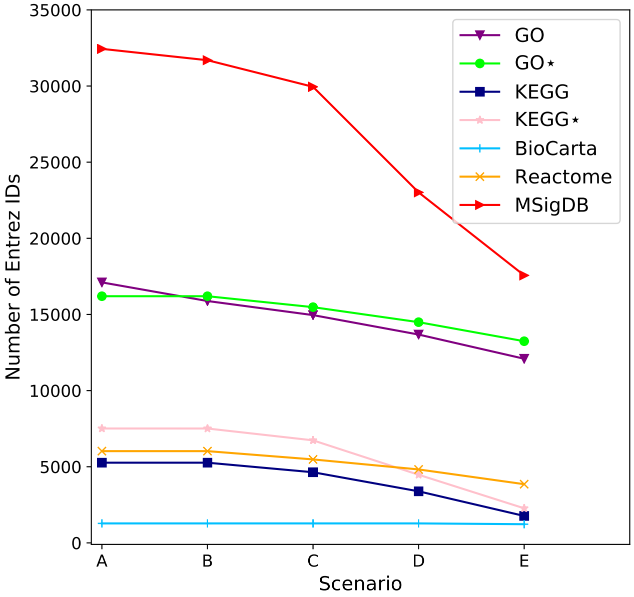


Figure 7.5: The number of genes present in each gene set database (y-axis) across four scenarios A, B, C, and D (x-axis). In scenario A, all gene sets are considered; scenarios B, C, D, and E only consider gene sets with size less than or equal to 500, 250, 100, and 50, respectively. KEGG⋆ and GO⋆ represent the gene sets that were extracted from updated KEGG and GO, respectively.

7.3.1 Biomedical Case Study

Juvenile Idiopathic Arthritis is an autoinflammatory disorder in children that is currently divided into six different subtypes [21]. The exact causes of JIA are unknown, but gene expression and genome-wide genotyping have identified genes that are associated with the development of this disease and the different subtypes. Variants in these genes can impair the normal regulation of inflammatory processes, which involve the activity of many innate regulators including interleukins, chemokines, and matrix metalloproteinases (MMPs) resulting in joint inflammation [20, 23, 11, 22]. Many of the genes associated with JIA belong to a family of genes that provide instructions for making a group of related proteins called the human leukocyte antigen (HLA) complex, which helps the immune system identify proteins made by foreign invaders.

In this experiment, GSEA (gene permutation version) [25] and over-representation analysis (ORA) [9], which are two widely used gene set analysis methods, were utilized with the gene set databases. To take into account gene sets that are potentially relevent but contain less than 15 genes, we considered 5 as the lower bound for the size of gene sets. Also, to avoid large and general gene sets, we used 500 as the upper bound for gene set sizes. A Benjamini-Hochberg correction [6] for multiple comparisons with a false discovery rate of 0.05 was applied to the results of all gene set analysis methods.

The results of these gene set analysis methods using three JIA datasets were used to conduct a biomedical investigation of the outcome of gene set analysis when using different databases. The relationship between maximum achievable coverage values in each database for genes associated with JIA and the enrichment results obtained using each database were investigated.

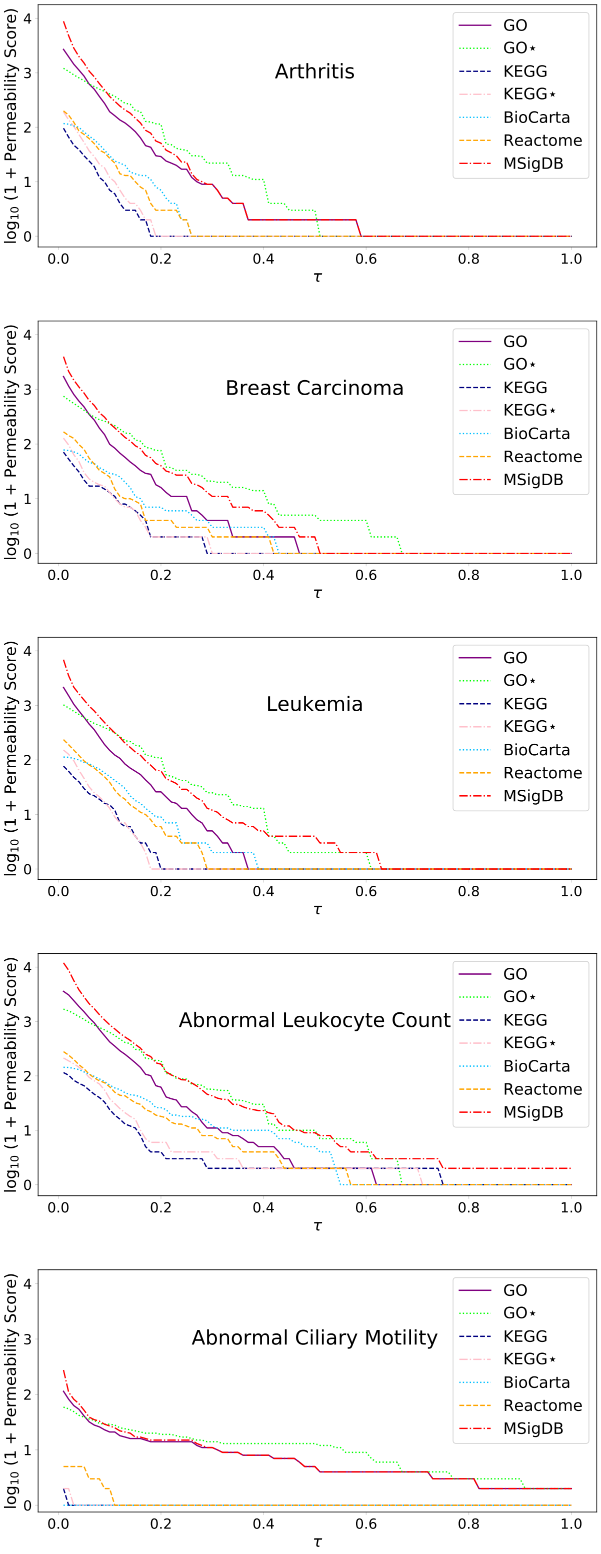


Figure 7.6: The permeability scores for five gene lists. Each gene list L has been extracted from HPO gene-phenotype associations. Each plot corresponds to one list and each trace corresponds to a gene set database. The x-axis shows τ values, i.e. different thresholds for coverage score. The y-axis shows the logarithm of permeability scores. To visualize the permeability scores, a log transformation is applied. Also, to to prevent taking the logarithm of 0, we add a pseudocount of 1.

Table 7.3 provides the number of differentially enriched gene sets predicted for each method across gene set databases and expression datasets. GSE7753, GSE13501, and GSE26544 had 245, 279, and 262 differentially expressed genes, respectively, when performing single gene analysis [9] with a fold change of 2 and an adjusted p-value less than 0.05.

Table 7.3: The number of differentially enriched gene sets reported by each method using each GO, GO*, KEGG, KEGG*, BioCarta, Reactome, and MSigDB when analyzing the JIA datasets.

Method	GSE7753		GSE13501		GSE26554	
	ORA	GSEA-G	ORA	GSEA-G	ORA	GSEA-G
GO	42	184	50	248	55	436
GO*	23	96	26	90	5	86
KEGG	0	10	1	13	0	36
KEGG*	1	34	4	34	1	59
BioCarta	1	4	1	10	0	3
Reactome	0	57	1	59	0	145
MSigDB	588	2399	858	2301	627	3645

The maximum achievable coverage score for JIA was significant only using the GO and GO* databases, with GO* achieving the highest coverage value of 0.429. When using GO as the gene set database, the 20 most significant gene sets reported by ORA included "Humoral immune response" (GO:0006959), "Leukocyte migration" (GO:0050900), "Immune response regulating cell surface receptor signaling pathway" (GO:0002768), "Activation of immune response" (GO:0002253), and "Cytokine mediated signaling pathway" (GO:0019221). Results using GO and GO* were consistent when considering the most differentially enriched gene sets. When using GO*, ORA was able to identify several different gene sets with potential implications for JIA. Some of these potentially related gene sets, such as "Neutrophil degranulation" (GO:0043312), "Haptoglobin binding" (GO:0031720), and "Haptoglobin-hemoglobin complex" (GO:0031838), were not available from the GO database from MSigDB.

Using MSigDB as a whole resulted in substantially more gene sets predicted as being differentially enriched compared to using any of the other databases. The majority of these gene sets predicted as enriched were from "C7: immunologic signatures" in MSigDB. The immunologic signatures have been created using differentially expressed genes reported in immunology literature. These gene sets also have names that are more difficult to interpret without further investigation. Examples include gene sets such as "MODULE 114", "MODULE 114", and "MODULE 151".

Using KEGG, ORA predicted "Systemic lupus erythematosus" (hsa05322), while using KEGG*, ORA predicted "Malaria" (hsa05144) and "Systemic lupus erythematosus". Systemic lupus erythematosus (SLE) and JIA are distinctly different both clinically and pathogenically. SLE is a prototypic autoimmune disease with very specific serologic hallmarks (antibodies to double-stranded DNA and other distinctive autoantibodies). However, both JIA and SLE are associated with inflammation so there are likely to be some downstream immune and inflammatory responses common to many inflammation-mediated diseases [7]. This is also likely the reason "Malaria" is a predicted gene set when using GO*. It is common to treat JIA with antimalarial drugs to try and control the inflammation, though to variable degrees of success [3].

Using KEGG, GSEA resulted in “Hematopoietic cell lineage” (hsa04640), “Antigen processing and presentation” (hsa04612), “Graft versus host disease” (hsa05332), and “Allograft rejection” (hsa05330) predicted as being differentially enriched. GSEA also predicted “Antigen processing and presentation”, “Graft versus host disease”, and “Allograft rejection” when using KEGG* (“Hematopoietic cell lineage” was still significant, but not one of the 20 most significant gene sets). Using KEGG* also resulted in GSEA predicting gene sets including “Complement and coagulation cascade” (hsa04610), and “IL-17 signaling pathway” (hsa04657), “Malaria”, “Systemic lupus erythematosus”, and “Staphylococcus aureus infection” (hsa05150). A complete list of gene sets predicted as most significantly differentially enriched for each method and each database is available upon request.

7.4 Discussion

In this paper, we proposed similarity, presence, and permeability scores for the quantitative study of gene set databases. We also proposed a statistical procedure for significance assessment of maximum achievable coverage score.

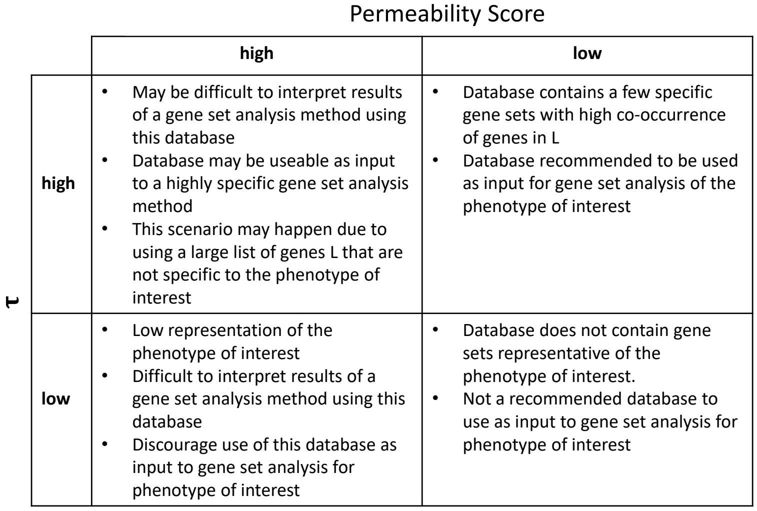
The similarity scores between each pair of GO, GO*, KEGG, KEGG*, BioCarta, and Reactome in Table 7.1 showed that these databases are substantially different from each other. If the databases all had high similarity, any of these gene set databases would be expected to produce comparable results. However, the low similarity scores observed suggest that many of the gene sets within a database are not comparable with gene sets in other databases. Therefore, these databases cannot be used interchangeably in gene set analysis.

We used presence score as an indication of the degree to which a gene is represented in a gene set database. Our study of commonly used gene set databases showed that many genes from the HPO gene lists and the JIA gene list have a presence score of 0, i.e. they are not represented in any gene set of some of the databases. These genes might be differentially expressed in an experiment, but gene set analysis methods will not be able to associate the differential expression of these genes to any gene sets of the database. For example, no gene currently associated with “Abnormal ciliary motility” (HP:0012262) is present in BioCarta. Therefore, performing gene set analysis for an experiment concerning “Abnormal ciliary motility” might lead to no differential enrichment; however, this lack of differential enrichment must not be interpreted as there being no association. As another example, Fig. 7.4 illustrates the presence score for genes associated with “Arthritis” (HP:0001369), which involves many genes linked with the immune system. Many of these genes are absent or rarely present in gene sets of size 250 or less for some databases. Therefore, such gene set databases must not be used for gene set analysis of “Arthritis” related experiments.

We suggest using the presence score as an initial indicator of the utility of a gene set database to be used for a given phenotype. When the majority of genes of interest are not represented in a database, i.e. presence score of zero, such a database must not be used for gene set analysis. To investigate the presence of genes in more specific gene sets that are usually of smaller sizes, we suggest restricting a gene set database to only include gene sets of the size smaller than, or equal to a given threshold. This threshold might vary according to the phenotype of interest and the purpose of the expression study. As a rule of thumb, we suggest considering threshold values of 250, 100, and 50, with the threshold of 50 restricting the database to the most specific gene sets.

Although presence score is an informative indicator of the presence of genes in a given database, it does not account for the co-occurrence of genes within gene sets. Given L—a set of genes of interest—and $\mathbb{G}$—a gene set database—the permeability score measures the co-occurrence of genes from L within gene sets of $\mathbb{G}$. Further, the maximum achievable coverage score shows the highest degree of co-occurrence of genes in L within gene sets of $\mathbb{G}$. To assess if the maximum achievable coverage score for a gene set database is solely due to chance or a consequence of the size of L, we used a bootstrapping hypothesis test and associated a maximum coverage achieved for a database to a p-value. Significant maximum coverage values can be used to select the gene set database most appropriate for conducting gene set analysis in a given experiment.

Figure 7.7: Summary of recommendations for using permeability score and coverage score to select a gene set database.



Permeability Score

		high	low
L	high	• May be difficult to interpret results of a gene set analysis method using this database • Database may be useable as input to a highly specific gene set analysis method • This scenario may happen due to using a large list of genes L that are not specific to the phenotype of interest	• Database contains a few specific gene sets with high co-occurrence of genes in L • Database recommended to be used as input for gene set analysis of the phenotype of interest
	low	• Low representation of the phenotype of interest • Difficult to interpret results of a gene set analysis method using this database • Discourage use of this database as input to gene set analysis for phenotype of interest	• Database does not contain gene sets representative of the phenotype of interest. • Not a recommended database to use as input to gene set analysis for phenotype of interest

For a given L and $\mathbb{G}$, a small value of maximum coverage score indicates that a large proportion of L or its subsets does not co-occur in any gene set of $\mathbb{G}$. Therefore, $\mathbb{G}$ might not be an appropriate choice for studying the phenotype of interest represented by L. It should be noted that genes in L can be chosen by a domain expert or based on the literature consensus. Also, the choice of L can be made prior to conducting an expression study. Human Phenotype Ontology can be used as a reliable source for the choice of L. HPO has been actively developing phenotype-gene associations by curating literature to annotate abnormalities in phenotypes that have been observed in human disease. We discourage using a list of genes L with more than 250 genes (see Figure 7.3) since any gene list larger than this is likely too general and could result in high coverage scores with a wide array of gene sets due to factors such as gene set overlap. In addition, we proposed the bootstrapping approach for significance assessment of the maximum coverage score. Such a bootstrapping approach can be adjusted for any given value of τ with a non-zero permeability score. When the bootstrapping approach leads to a significant p-value for the selected value of τ, Fig. 7.7 shows a summary of recommendations when selecting a database using permeability and coverage score using a list of genes of interest L. What is considered a high permeability and coverage score is at the discretion of the researcher, and based on the phenotype of interest, but it is recommended that the scores are reported as evidence to support the choice of a particular gene set database, just as the number of gene sets predicted as significantly differentially enriched by a gene set analysis method should be reported to assess the usefulness of the results and any downstream interpretation of these results.

An arbitrary choice of a gene set database, which includes gene sets extracted from a single source, decreases the chance of finding biologically meaningful associations. On the other hand, using meta-databases—which include gene sets from several sources— often results in false positives [17, 16]. For example, using MSigDB as the gene set database for analyzing JIA datasets (see Table 7.3) led to a large number of gene sets predicted as being differentially enriched (over several thousand gene sets). Therefore, aggregating multiple gene set databases into one meta-database without considering their biological relevance—with the sole purpose of attaining a high maximum achievable coverage—is not a viable alternative. Rather, we suggest the smallest but the most appropriate gene set database, as measured by the proposed quantitative approach.

The results of gene set analysis using the JIA datasets agree with the results achieved using the maximum achieveable coverage score. We only observed significant maximum achieveable coverage scores for GO and GO* using a list of genes suggested in the literature to be associated with JIA. The gene set in GO* with the highest number of co-occurring genes known to be involved in the development of this disease contained roughly 43% of the genes from our list. GO had a much smaller maximum achieveable coverage score of 0.273, though it was still considered significant. However, a maximum achieveable coverage of 0.273 is low, so the relevance of the gene sets predicted using this database may not be as informative. Using GO* with all three datasets resulted in a partial overlap with the results obtained when using GO. However, GO* achieved the highest maximum achieveable coverage score, which was reflected in predicted gene sets such as “Neutrophil degranulation” in all three datasets. Platelets and neutrophils interact during inflammation, and excessive neutrophil degranulation is a common feature of many inflammatory disorders [15]. Also, two gene sets predicted using GO* were “Haptoglobin binding” and “Haptoglobin-hemoglobin complex”. Haptoglobin can become elevated during infection and inflammation and has been reported as a potential biomarker of JIA [10, 24]. These gene sets are also very specific with 6 and 7 genes, respectively. As “Haptoglobin binding” and “Haptoglobin-hemoglobin complex” are relatively small gene sets (less than 10), MSigDB removes any gene sets that are smaller than 10, which means that these enriched terms would be missed by the gene set analysis of the JIA datasets even though they are highly informative of this disease. Another gene set only reported using KEGG* was “IL-17 signalling pathway”. Interleukin-17 is known to perpetuate inflammatory pathways by inducing fibroblasts to make cytokines and matrix metalloproteinases [2]. This pathway is not a part of the MSigDB database, and as such, it was never predicted as being differentially enriched in any scenario.

Mooney et al. [18] recommended selecting a subset of biologically relevant gene sets in order to perform gene set analysis. However, this might be a biased approach. If a gene set database used for analysis has been filtered to only contain gene sets relevant to a phenotype of interest, then any false positive is also going to be biologically relevant to the phenotype. Considering the lack of specificity of some gene set analysis methods [17], such an approach might lead to incorrect interpretations. Instead, in this work we suggested selecting and using a gene set database in a systematic way that is not prone to an investigator bias(es) toward a hypothesis of interest.

7.5 Conclusions

In this paper, we studied characteristics of gene set databases as an essential component of gene set analysis. We introduced a measure to quantify the similarity between gene set databases. We proposed presence and permeability scores as quantitative measures of occurrence and co-occurrence of genes within gene sets. Further, we proposed a bootstrapping procedure for the significance assessment of permeability score.

The quantitative study of commonly used gene set databases showed that the current gene set databases fall short in representing some phenotypes. Also, some of these databases may not be well maintained and focused on increasing the number of gene sets in a given database. We suggest a shift towards developing and curating gene set databases that are well maintained and cover a wider range of phenotypes.

We also observed that genes known to be associated with some phenotypes are rarely present or completely absent in some widely used gene set databases. The differential expression of these genes cannot be associated with any gene set; therefore, lack of differential enrichment might be interpreted as no association. The use of the proposed scores helps with avoiding such erroneous conclusions. These scores also showed that some databases represent particular phenotypes better than others, and indicate the databases that are most appropriate for performing gene set analysis for a phenotype of interest.

We suggest using permeability score and coverage score as a means to select an appropriate gene set database for a given experiment and avoiding assembling a gene set database by cherry-picking gene sets of interest as it might lead to a biased conclusion.

References

[1] Michiel E Adriaens, Magali Jaillard, Andra Waagmeester, Susan LM Coort, Alex R Pico, and Chris TA Evelo. The public road to high-quality curated biological pathways. *Drug Discovery Today*, 13(19-20):856–862, 2008.

[2] Shilpi Agarwal, Ramnath Misra, and Amita Aggarwal. Interleukin 17 levels are increased in juvenile idiopathic arthritis synovial fluid and induce synovial fibroblasts to produce proinflammatory cytokines and matrix metalloproteinases. *The Journal of Rheumatology*, 35(3):515–519, 2008.

[3] Jie An, Mark Minie, Tomikazu Sasaki, Joshua J Woodward, and Keith B Elkon. Antimalarial drugs as immune modulators: new mechanisms for old drugs. *Annual Review of Medicine*, 68:317–330, 2017.

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[5] Gerald J Bakus. *Quantitative Analysis of Marine Biological Communities: Field Biology and Environment.* John Wiley & Sons, 2007.

[6] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995.

[7] James Bentham, David L Morris, Deborah S Cunninghame Graham, Christopher L Pinder, Philip Tombleson, Timothy W Behrens, Javier Martín, Benjamin P Fairfax, Julian C Knight, Lingyan Chen, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics*, 47(12):1457, 2015.

[8] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, 2013.

[9] Sorin Drăghici. *Statistics and Data Analysis for Microarrays using R and Bioconductor.* CRC Press, Boca Raton, FL, 2016.

[10] Ndate Fall, Michael Barnes, Sherry Thornton, Lorie Luyrink, Judyann Olson, Norman T Ilowite, Beth S Gottlieb, Thomas Griffin, David D Sherry, Susan Thompson, et al. Gene expression profiling of peripheral blood from patients with untreated new-onset systemic juvenile idiopathic arthritis reveals molecular heterogeneity that may predict macrophage activation syndrome. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 56(11):3793–3804, 2007.

[11] LB Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome Research*, 10(10):1435–1444, 2000.

[12] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[13] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, et al. Expansion of the human phenotype ontology (hpo) knowledge base and resources. *Nucleic Acids Research*, 47(D1):D1018–D1027, 2018.

[14] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[15] Ton Lisman. Platelet–neutrophil interactions as drivers of inflammatory and thrombotic disease. *Cell and Tissue Research*, 371(3):567–576, 2018.

[16] Farhad Maleki and Anthony J. Kusalik. Gene set overlap: An impediment to achieving high specificity in over-representation analysis. In *12th International joint conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*, pages 43–54, Prague, Czech Republic, February 2019.

[17] Farhad Maleki, Katie L. Ovens, Elham Rezaei, Alan M. Rosenberg, and Anthony J. Kusalik. Method choice in gene set analysis has important consequences for analysis outcome. In *12th International joint conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*, pages 182–193, Prague, Czech Republic, February 2019.

[18] Michael A Mooney and Beth Wilmot. Gene set analysis: A step-by-step guide. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(7):517–527, 2015.

[19] Darryl Nishimura. Biocarta. *Biotech Software & Internet Report*, 2(3):117–120, 2001.

[20] Ross E Petty, Ronald M Laxer, Carol B Lindsley, and Lucy Wedderburn. *Textbook of Pediatric Rheumatology*. Elsevier Health Sciences, 2015.

[21] Ross E Petty, Taunton R Southwood, Prudence Manners, John Baum, David N Glass, Jose Goldenberg, Xiaohu He, Jose Maldonado-Cocco, Javier Orozco-Alcala, Anne-Marie Prieur, et al. International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *The Journal of Rheumatology*, 31(2):390, 2004.

[22] Sampath Prahalad. Genetic analysis of juvenile rheumatoid arthritis: approaches to complex traits. *Current problems in pediatric and adolescent health care*, 36(3):83, 2006.

[23] Sampath Prahalad, Mary H Ryan, Edith S Shear, Susan D Thompson, Edward H Giannini, and David N Glass. Juvenile rheumatoid arthritis: linkage to hla demonstrated by allele sharing in affected sibpairs. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 43(10):2335–2338, 2000.

[24] Margalit E Rosenkranz, David C Wilson, Anthony D Marinov, Alisha Decewicz, Patrick Grof-Tisza, David Kirchner, Brendan Giles, Paul R Reynolds, Michael N Liebman, VS Kumar Kolli, et al. Synovial fluid proteins differentiate between the subtypes of juvenile idiopathic arthritis. *Arthritis & Rheumatism*, 62(6):1813–1823, 2010.

[25] Aravind Subramanian et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[26] Vinicius Tragante, Johannes MIH Gho, Janine F Felix, Ramachandran S Vasan, Nicholas L Smith, Benjamin F Voight, Colin Palmer, Pim van der Harst, Jason H Moore, and Folkert W Asselbergs. Gene set enrichment analyses: lessons learned from the heart failure phenotype. *BioData Mining*, 10(1):18, 2017.

[27] Lina Wadi, Mona Meyer, Joel Weiser, Lincoln D Stein, and Jüri Reimand. Impact of outdated gene annotations on pathway enrichment analysis. *Nature Methods*, 13(9):705, 2016.

8 Summary and Future Research

In this chapter, we provide a short summary of the previous chapters of the thesis, highlighting its contributions, and discuss some avenues for future research.

8.1 Summary and contributions

In Chapter 2, we reviewed gene set analysis methods, classified them based on their components, discussed the shortcomings and strengths of each class, and established the direction of the research in this thesis. Considering the gained insight from the literature review about the lack of consensus about the best practices in gene set analysis methods, we designed and conducted two studies to investigate the reproducibility of the results of gene set analysis methods.

In Chapter 3, we proposed a methodology to evaluate the reproducibility of a gene set analysis method across sample sizes when analyzing a real expression dataset. Our study of thirteen commonly used gene set analysis methods showed that the results of most methods are not reproducible for small sample sizes. Therefore, one should be skeptical of the results from gene set analyses conducted with small sample sizes.

In Chapter 4, we investigated the consistency of the results of different gene set analysis methods when analyzing an expression dataset. Our data analysis revealed that the outcome of gene set analysis significantly differs across methods. Also, we observed that many gene set analysis suffer from a lack of specificity.

To further investigate the specificity of gene set analysis methods, in Chapter 5, we studied the impact of gene set overlap—which has been suggested as a potential reason for lack of specificity [6]—on the results over-representation analysis. We proposed a measure to quantify gene set overlap and visualized gene set overlap for several gene set databases. We observed that gene set overlap is a ubiquitous phenomenon across gene set databases. Also, our statistical analysis showed that there is a significant negative correlation between gene set overlap and the specificity of over-representation analysis.

Since gene set overlap was shown to affect the results of a gene set analysis method, it should be considered in the design and evaluation of these methods. However, this has not been the case in the evaluation of gene set analysis methods. Due to the lack of gold standard datasets, where the enrichment status of gene sets of a database are known *a priori*, artificial expression datasets and simulated gene set databases with oversimplified assumptions have been used. Expression profiles have often been simulated with constant or zero gene-gene correlation and with normally distributed expression values. Gene set databases also have been simulated to be a collection of non-overlapping gene sets of equal sizes. Simulating expression datasets and gene set databases that preserve the characteristics of real data is not a straightforward task. Considering the insight gained from Chapters 2, 3, 4, and 5 on reproducibility, sensitivity, and specificity of gene set analysis methods, we developed Silver (see Chapter 6). Silver is a framework for generating expression datasets that preserves the characteristics of real data. It also offers a means for quantitative evaluation of the results of different gene set analysis methods using the generated data. We showed the utility of Silver by developing expression data sets and evaluating the results of ten commonly used gene set analysis methods. The results confirmed the above key challenges facing gene set analysis and showed the inability of some of the most widely used gene set analysis methods to detect differential enrichment of gene sets with a mixture of up- and down-regulated genes. These shortcomings have been previously overlooked. Therefore, we suggest Silver as a means for evaluating new gene set analysis methods, rather than using unrealistic and oversimplified datasets.

Since a gene set database serves as an input to gene set analysis, in Chapter 7 we quantitatively study gene set databases. There has been limited research on the quality of gene set databases and the common approach has been to extend the size of databases by adding more and more gene sets. We defined measures to quantify the similarity of gene set databases, the degree to which a given gene is represented in gene sets of a database, and the proportion of genes of interest that are present in a given gene set (coverage score).

Further, we proposed a methodology for the significance assessment of a maximum achievable coverage score given a list of genes and a gene set database. This methodology allows an unbiased selection of appropriate gene set databases to be used for a given experiment; such a systematic methodology has not been previously available, and researchers have have relied on their intuition or selected a gene set database in an ad-hoc manner. The study of widely used gene set databases such as GO [2], KEGG [4], Reactome [3], and BioCarta [5] using the proposed measures showed that genes associated with different phenotypes are not equally represented in gene set databases. In addition, our study showed that merging different databases without systematic curation to generate massive meta-databases leads to a lack of specificity and difficulty in interpreting the results of gene set analysis.

Collectively, methodologies presented in Chapters 3, 4, 5, 6, and 7 of this thesis not only make it possible to evaluate the current gene set analysis methods and gene set databases in a systematic and quantitative manner, but also they can be used to guide the design and development of future methods and databases for gene set analysis. As such, they should improve the sensitivity, specificity, and reproducibility of gene set analysis, contributing to gaining better insight from the data resulting from high-throughput experiments.

8.2 Future work

In this section, we extend the ideas for future research presented in each chapter of the thesis and offer possible avenues for addressing challenges in gene set analysis.

8.2.1 Reproducibility of gene set analysis methods

In this research, we proposed a quantitative approach for assessing the reproducibility of gene set analysis methods across sample sizes. The proposed method naturally extends to evaluating the reproducibility of other methods such as gene regulatory network or coexpression network prediction, where there is no gold standard (ground truth) dataset available. The proposed methodology can be seamlessly applied in such contexts.

For a given sample size, reproducibility, though a necessary condition, is not a sufficient condition; a given gene set analysis method should achieve high specificity and sensitivity scores to achieve biologically valid results. We suggest using Silver, the proposed framework for evaluating gene set analysis methods, to synthesize datasets where the enrichment status of gene sets are known *a priori*. The simulated datasets then can be used for the study of specificity, sensitivity, and reproducibility of results of gene set analysis methods across sample sizes.

8.2.2 Evaluation of gene set analysis methods

Using a quantitative approach, we studied the results of ten gene set analysis methods. Statistical analysis showed that there is a significant difference between the results of most gene set analysis methods. This conclusion was further confirmed by a biological evaluation of the gene set analysis methods when analyzing a juvenile idiopathic arthritis dataset. We suggest using more datasets across different phenotypes for conducting the biomedical evaluation. Further, we suggest the study of gene set analysis in the absence of differential expression using datasets that preserve the true characteristics of real data. This can be accomplished using only data from control samples, or alternatively using Silver without differential enrichment of gene sets, in contrast to the current approach using phenotype permutation, which does not preserve the true characteristics of the expression data (See Chapter 6).

8.2.3 Specificity and gene set overlap

In this thesis, we proposed a methodology to study the relationship between gene set overlap and specificity of ORA. We quantified gene set overlap and visualized it for several gene set databases. The statistical data analysis suggested that there is a significant correlation between gene set overlap and specificity of ORA. We found that gene set overlap is a ubiquitous phenomenon across gene set databases, partially due to the existence of multifunctional genes.

Since gene set overlap can contribute to the lack of specificity of a gene set analysis method, it must be considered in developing and evaluating gene set analysis methods and also gene set databases. Current methods that consider gene set overlap sacrifice sensitivity in favour of specificity. We suggest developing gene set analysis methods that achieve high specificity without significant loss in sensitivity to be the purpose of developing new gene set analysis methods. More specifically, we highly discourage the evaluation of gene set analysis methods using artificial gene set databases of non-overlapping gene sets. Such contexts are substantially different from the real gene set database and often lead to misleadingly high specificity values.

We studied the effect of gene set overlap on the specificity of ORA. Conducting similar studies for other well-established gene set analysis methods is suggested as future research.

8.2.4 Benchmark datasets

Our study suggested that the lack of consensus about a gene set analysis method to be used for a given experiment can be mainly attributed to the absence of gold standard datasets, where the enrichment status of gene sets is known. In this thesis, we proposed Silver for evaluating gene set analysis methods. Silver synthesizes datasets without making oversimplifying assumptions that have been commonly made in the evaluation of gene set analysis methods and have led to inconsistent and contradictory guidelines.

We suggest developing a publicly available repository of benchmark datasets generated using Silver. This repository can help a researcher to choose the best method for a given experimental design by providing measures of performance for different methods across different experimental designs. Also, by automating the evaluation of gene set analysis methods, this repository can set the direction for developing new gene set analysis methods. To be biologically on par with the known biology of a given phenotype, the list of differentially expressed genes in each benchmark dataset must be chosen by experts in that specific phenotype. The high level and intuitive design of Silver make it easy to use for a wide range of users including experts in domains with little exposure to software development.

The current version of Silver uses a *t-test* statistic to define the differential expression of genes. Silver was designed to be easily extendable. Incorporating other measures of differential expression is suggested as future research. One such measure can be the Wilcoxon rank-sum test. Having benchmark datasets synthesized using various measures of differential expression reveal and prevent developing benchmark datasets that are biased in favour or against a given gene set analysis method.

8.2.5 Gene set databases

Online knowledge-bases such as GO [2], KEGG [4], Reactome [3], and BioCarta [5] have been used to develop gene set databases. Also, computational methods have been utilized to create gene sets from the expression datasets that are publically available through GEO and ArrayExpress. In addition, gene sets have been extracted from gene expression literature.

There are two common approaches for choosing a gene set database when conducting gene set analysis: using gene sets obtained from a single source such as KEGG or GO and using gene sets obtained from several sources including online knowledge-bases and/or computational methods. Each of these approaches come with their shortcomings. Single source databases often include a small number of gene sets. Informative gene sets (relatively small gene sets) in single source databases often represent a fraction of known genes for a species of interest and differential expression of the genes that are absent in these informative gene sets will be ignored in gene set analysis. On the other hand, merging different gene set databases with the goal of generating a massive meta-database without biological justification and systematic curation is not recommended. Such an approach could introduce more gene set overlap and consequently more false positives. Another shortcoming of such multi-source databases is that often they are either not updated or partially updated with gene sets from some sources not being updated for several years. Since for a typical user evaluating these databases is not practical, often outdated gene set databases of low quality have been used for gene set analysis.

For meta-databases that extract or curate gene sets from several sources, we suggest providing timestamps for gene sets to track the last update for each source. Updating gene sets from one source and providing a version and time-stamp for the whole meta-database makes it difficult to monitor which part of the meta-database is updated and which part is not.

In gene set analysis, a gene set is represented as a collection of genes that are involved in a biological component, process, or function. Information such as up-regulated and down-regulated genes in a pathway is currently not represented in databases used for gene set analysis. Consequently, detecting such gene expression patterns is difficult, if not impossible. Moreover, having the direction of change in gene expression can facilitate distinguishing gene sets that have a large proportion of their gene in common but with a substantially different direction in their expression patterns. Therefore, we suggest developing such a gene set database as future research to improve the specificity of gene set analysis methods.

In addition, we suggest extracting gene sets from literature through text mining. Currently, such gene sets are available in commercial gene set analysis tools such as Ingenuity Pathways Analysis from QIAGEN (http://www.ingenuity.com) and Bibliosphere by Intrexon (http://www.genomatix.de) [1]. Making such knowledge publicly available would be of great value to the research community. Also, providing a mechanism for researchers to curate the extracted gene sets would be beneficial in increasing the quality of such gene sets.

References

[1] Michiel E Adriaens, Magali Jaillard, Andra Waagmeester, Susan LM Coort, Alex R Pico, and Chris TA Evelo. The public road to high-quality curated biological pathways. *Drug Discovery Today*, 13(19-20):856–862, 2008.

[2] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[3] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477, 2013.

[4] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[5] Darryl Nishimura. Biocarta. *Biotech Software & Internet Report*, 2(3):117–120, 2001.

[6] Cedric Simillion, Robin Liechti, Heidi E.L. Lischer, Vassilios Ioannidis, and Rémy Bruggmann. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics*, 18(1):151, 2017.

Appendix A

Supplementary Material for Chapter 2

In this Appendix, we provide a set of examples of inconsistent and contradictory guidelines in gene set analysis from literature.

- Goeman and Bühlmann [3] argued that gene sampling as a method of significance evaluation for gene set statistics leads to a statistical model that does not agree with the performed biological experiments, and results in misleading interpretations. They discouraged using competitive gene set analysis methods since they rely on gene sampling for significance evaluation.
- According to an empirical study, Liu et al. [6] concluded that Globaltest, ANCOVA and SAM-GS achieve comparable statistical power.
- Based on a simulation study of self-contained gene set analysis methods, Fridley et al. [2] concluded that the Globaltest and the Fisher's method are the most statistically powerful.
- Based on an experimental study of 132 datasets, Hung et al. [4] reported that the Wilcoxon rank sum test statistic and the Weighted Kolmogorov Smirnov score, utilized by GSEA method, better cover predictions made by the mean test, the median test, the χ^2 test, and Hotelling's T^2 test, compared to the converse.
- Ackermann and Strimmer [1] discouraged using GSEA, and argued that the enrichment score, used by GSEA, is not as reliable as simpler methods such as mean or median score.
- Irizarry et al. [5] discouraged using GSEA as it is based on Kolmogorov Smirnov test, which is well-known for its lack of sensitivity. Alternatively, they suggested using the z-score and the χ^2 test as gene set scores. Using an experimental study, they showed that these simple methods outperform GSEA.
- Tamayo et al. [7] strongly disagreed with Irizarry et al. [5] and argued that the method proposed by Irizarry et al. ignores gene-gene correlations that can significantly affect the gene set analysis results. They also showed that the method suggested by Irizarry et al. suffers from a lack of specificity.
- Wu and Lin [8] studied several gene set analysis methods and concluded that the choice of gene set analysis method may affect the outcome of gene set analysis.

References

[1] Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, 2009.

[2] Brooke L Fridley, Gregory D Jenkins, and Joanna M Biernacka. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One*, 5(9):e12693, 2010.

[3] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[4] Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, 13(3):281–291, 2011.

[5] Rafael A Irizarry, Chi Wang, Yun Zhou, and Terence P Speed. Gene set enrichment analysis made simple. *Statistical Methods in Medical Research*, 18(6):565–575, 2009.

[6] Qi Liu, Irina Dinu, Adeniyi J Adewale, John D Potter, and Yutaka Yasui. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, 8(1):431, 2007.

[7] Pablo Tamayo, George Steinhardt, Arthur Liberzon, and Jill P Mesirov. The limitations of simple gene set enrichment analysis assuming gene independence. *Statistical Methods in Medical Research*, 25(1):472–487, 2016.

[8] Michael C Wu and Xihong Lin. Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways. *Statistical Methods in Medical Research*, 18(6):577–593, 2009.

Appendix B

Supplementary Material for Chapter 3

Table B.1: Kruskal-Wallis test results show that there is a statistically significant difference between the reproducibility of gene set analysis methods across sample sizes for all three original datasets.

Method	GSE53757	GSE13355	GSE10334
FRY	6.28e-13	2.99e-26	2.60e-13
GSEA-S	1.07e-11	1.45e-15	4.72e-05
GSEA-G	5.99e-12	6.72e-19	3.58e-13
ORA	5.03e-18	2.39e-19	1.67e-14
Camera	1.81e-19	2.11e-20	9.74e-05
ssGSEA	4.71e-25	8.70e-26	1.87e-26
PAGE	1.81e-16	5.26e-20	1.50e-10
GSVA	5.30e-20	7.17e-27	4.21e-07
PLAGE	2.10e-05	4.88e-06	4.89e-03
ROAST	1.37e-14	1.80e-26	1.10e-13
GAGE	2.34e-28	3.37e-28	4.51e-27
GlobalTest	6.73e-21	7.29e-25	2.58e-16
PADOG	7.10e-06	1.79e-17	7.18e-01

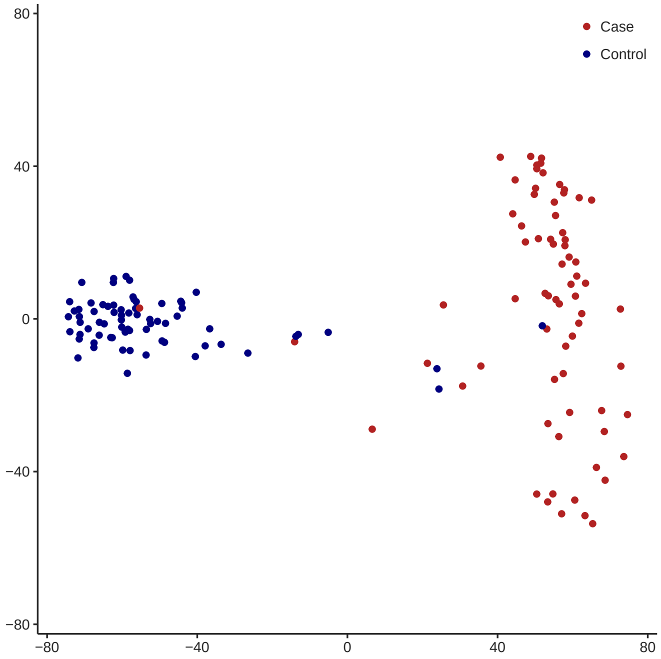


Figure B.1: MDS plot showing the relation between samples in dataset GSE53757. Each sample is represented as a point on the plot. The control samples are coloured in dark red and the case samples are coloured in blue.

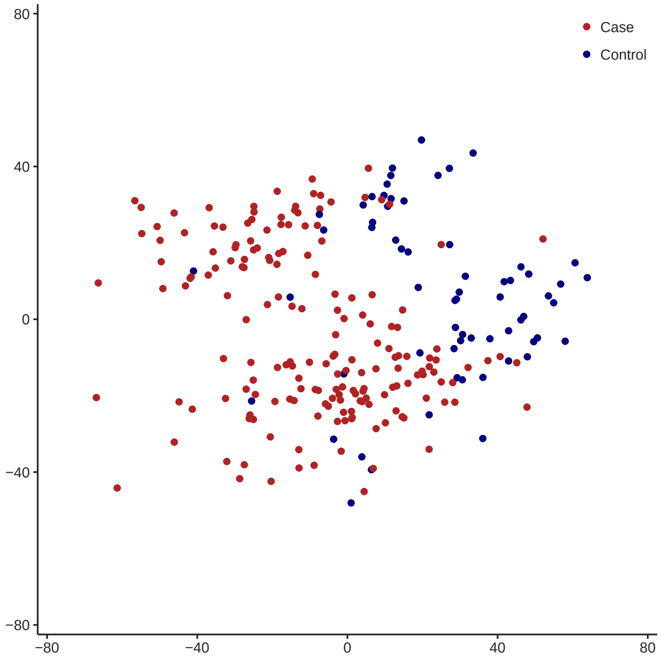


Figure B.2: MDS plot showing the relation between samples in dataset GSE10334. Each sample is represented as a point on the plot. The control samples are coloured in dark red and the case samples are coloured in blue.

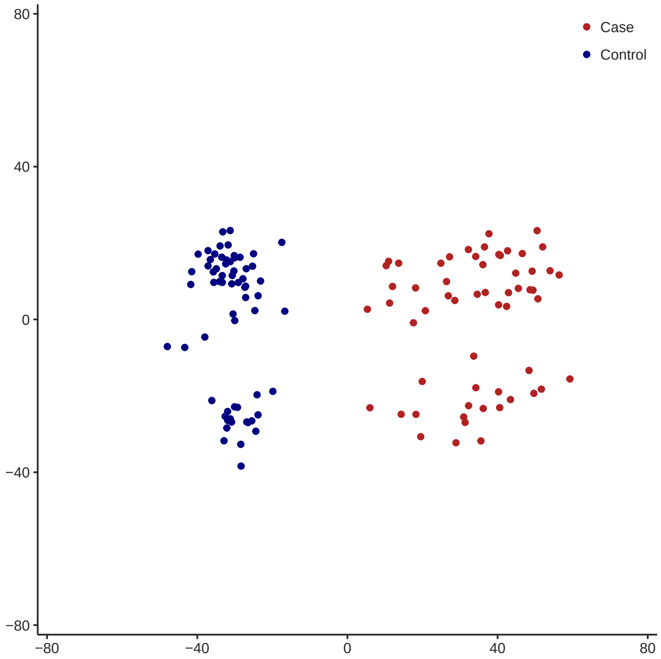


Figure B.3: MDS plot showing the relation between samples in dataset GSE13355. Each sample is represented as a point on the plot. The control samples are coloured in dark red and the case samples are coloured in blue.

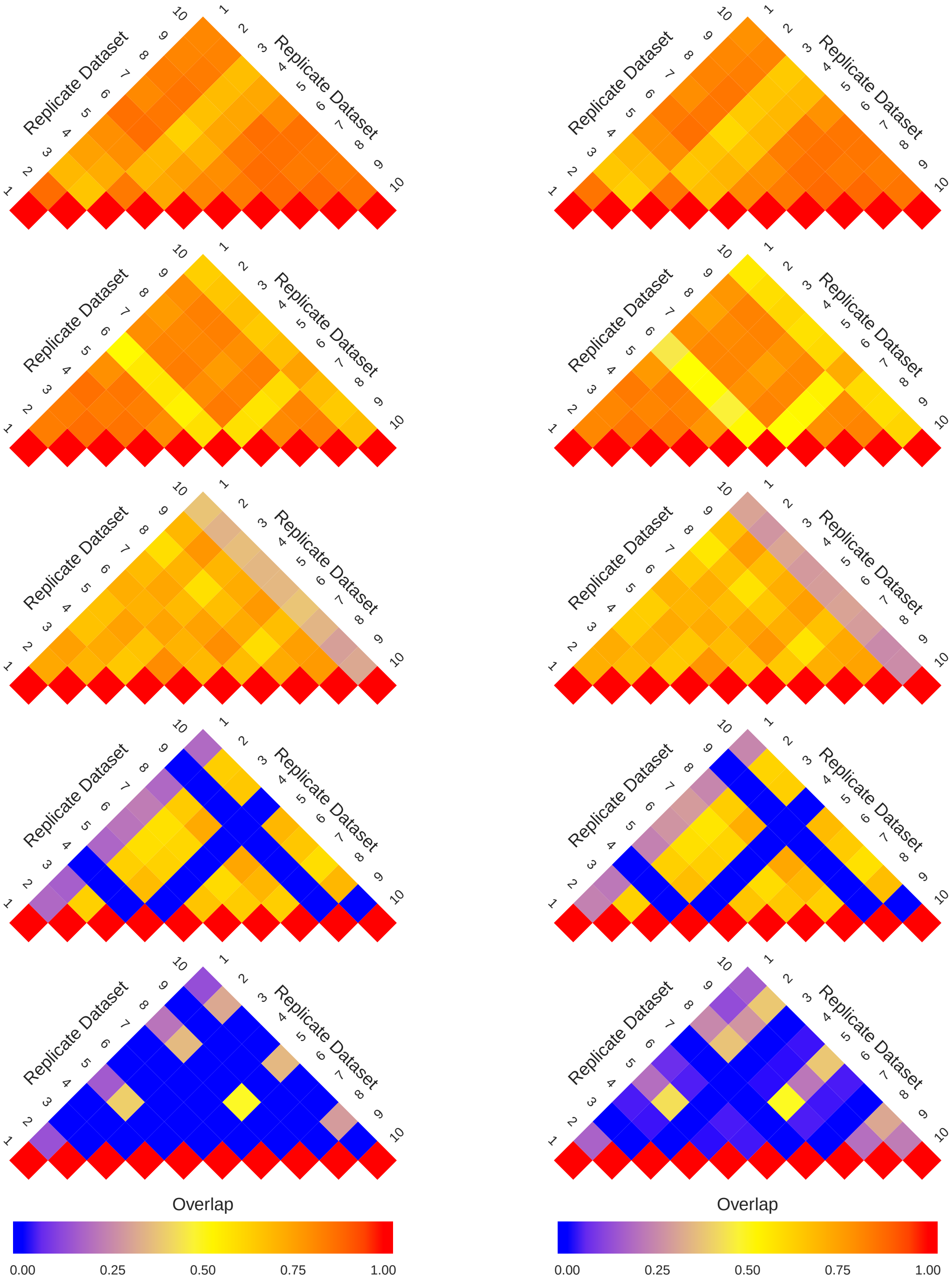


Figure B.4: Pine plots for dataset GSE53757 showing reproducibility of the results from ROAST (left) and FRY (right) across sample sizes. Reproducibility is quantified by overlap score (Equation 4.2). Each layer of the pine plot illustrates the overlap score of the results of a method for 10 replicate datasets with the same sample size. From top to bottom, the pine plot shows replicates with sample size 2 × 20, 2 × 15, 2 × 10, 2 × 5, and 2 × 3. The overlap score ranges from 0 to 1 represented by a gradient from blue to red, respectively, separated by yellow in the middle (overlap of 0.5).

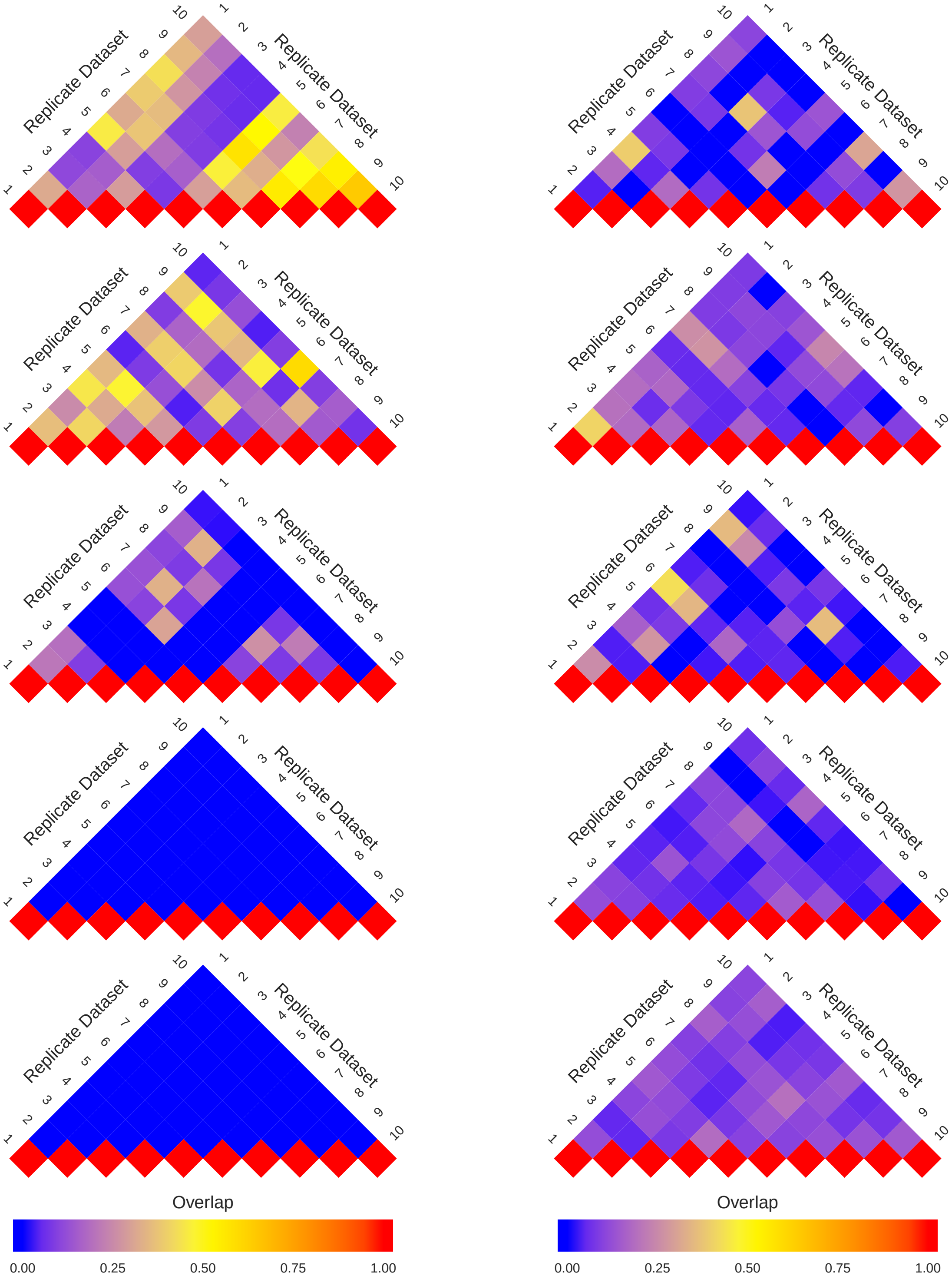


Figure B.5: Pine plots for dataset GSE53757 showing reproducibility of the results from Camera (left) and PADOG (right) across sample sizes. Reproducibility is quantified by overlap score (Equation 4.2). Each layer of the pine plot illustrates the overlap score of the results of a method for 10 replicate datasets with the same sample size. From top to bottom, the pine plot shows replicates with sample size 2 × 20, 2 × 15, 2 × 10, 2 × 5, and 2 × 3. The overlap score ranges from 0 to 1 represented by a gradient from blue to red, respectively, separated by yellow in the middle (overlap of 0.5).

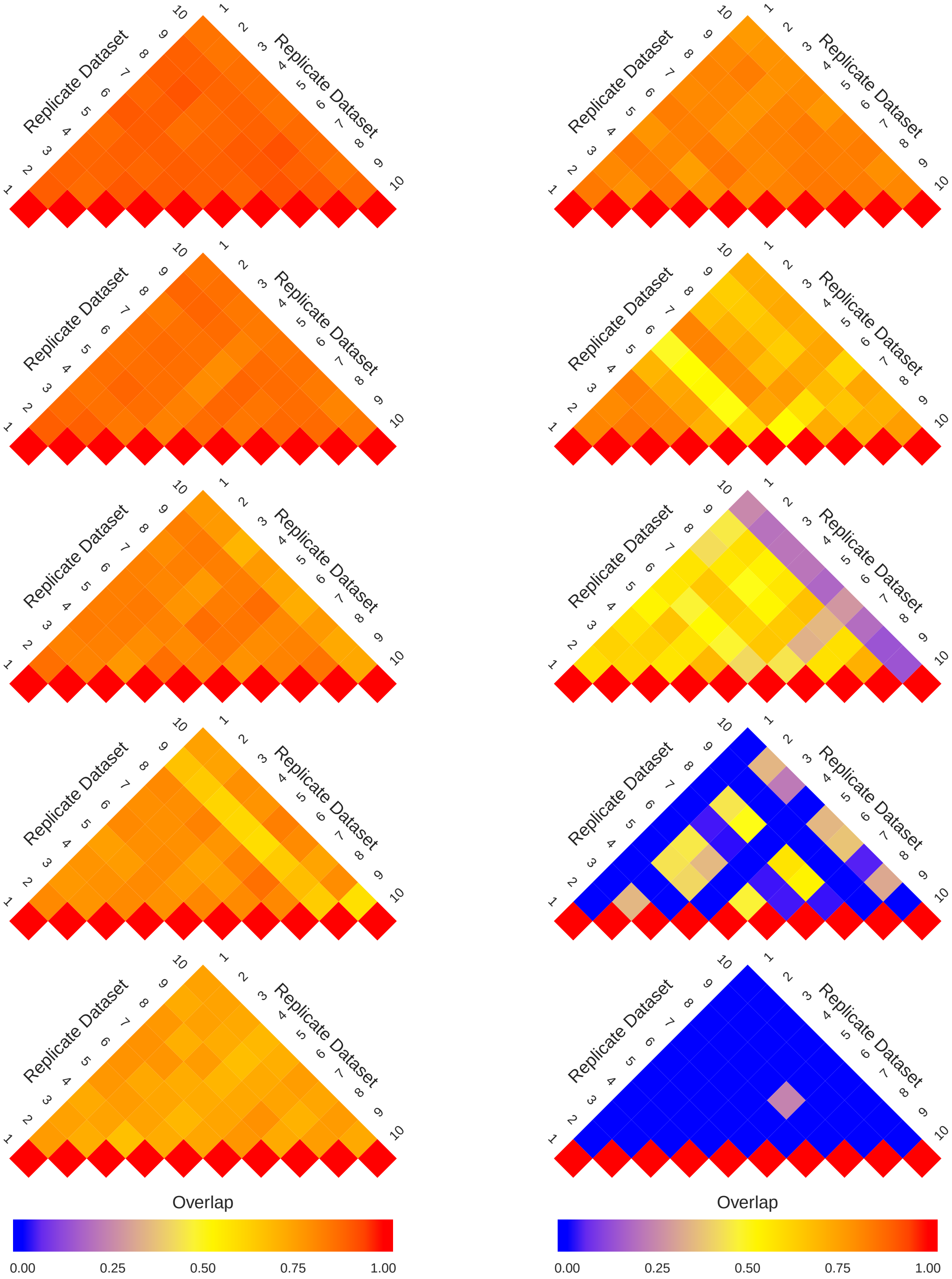


Figure B.6: Pine plots for dataset GSE53757 showing reproducibility of the results from PAGE (left) and GSVA (right) across sample sizes. Reproducibility is quantified by overlap score (Equation 4.2). Each layer of the pine plot illustrates the overlap score of the results of a method for 10 replicate datasets with the same sample size. From top to bottom, the pine plot shows replicates with sample size 2 × 20, 2 × 15, 2 × 10, 2 × 5, and 2 × 3. The overlap score ranges from 0 to 1 represented by a gradient from blue to red, respectively, separated by yellow in the middle (overlap of 0.5).

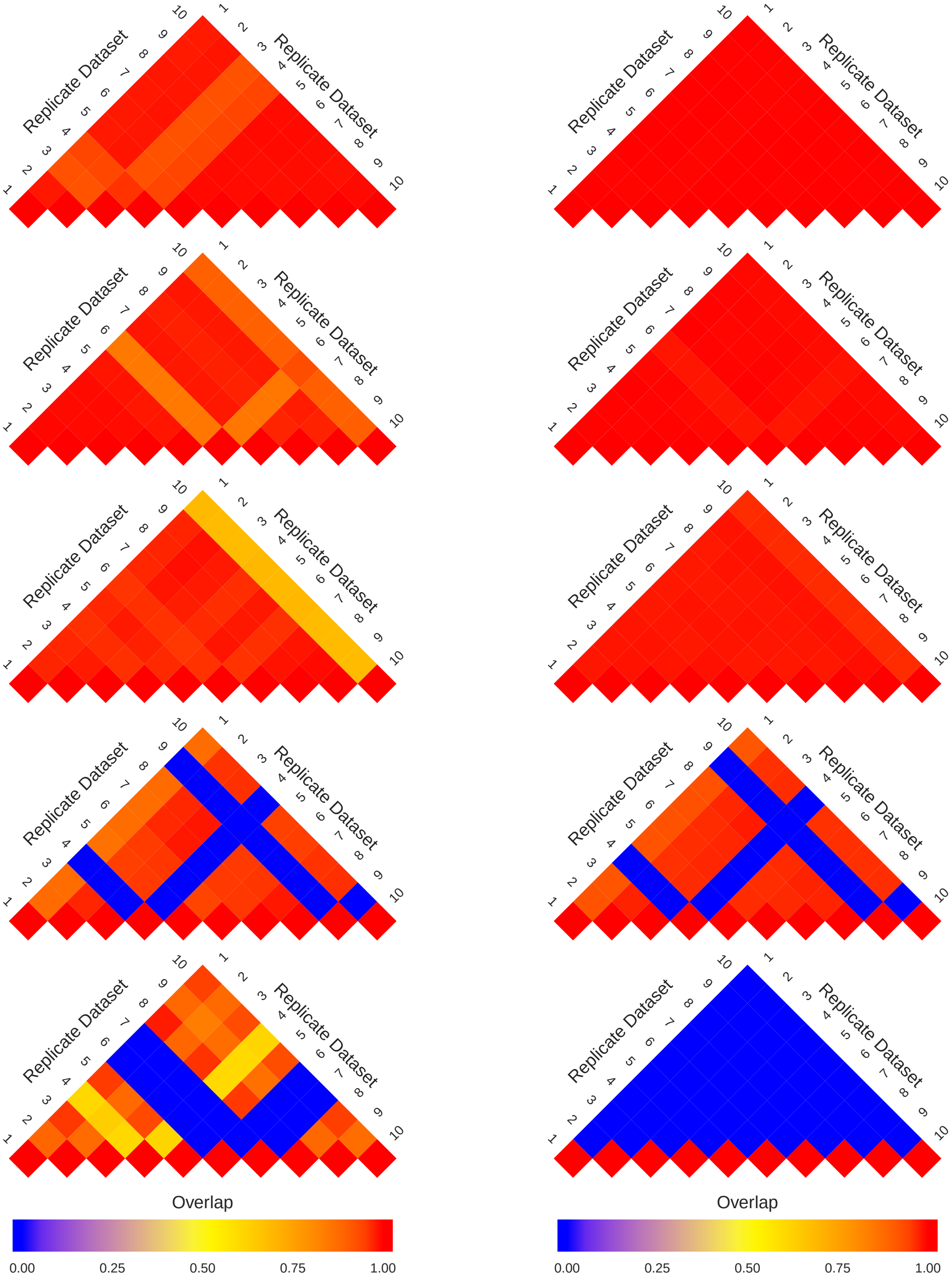


Figure B.7: Pine plots for dataset GSE53757 showing reproducibility of the results from PLAGE (left) and GlobalTest (right) across sample sizes. Reproducibility is quantified by overlap score (Equation 4.2). Each layer of the pine plot illustrates the overlap score of the results of a method for 10 replicate datasets with the same sample size. From top to bottom, the pine plot shows replicates with sample size 2 × 20, 2 × 15, 2 × 10, 2 × 5, and 2 × 3. The overlap score ranges from 0 to 1 represented by a gradient from blue to red, respectively, separated by yellow in the middle (overlap of 0.5).

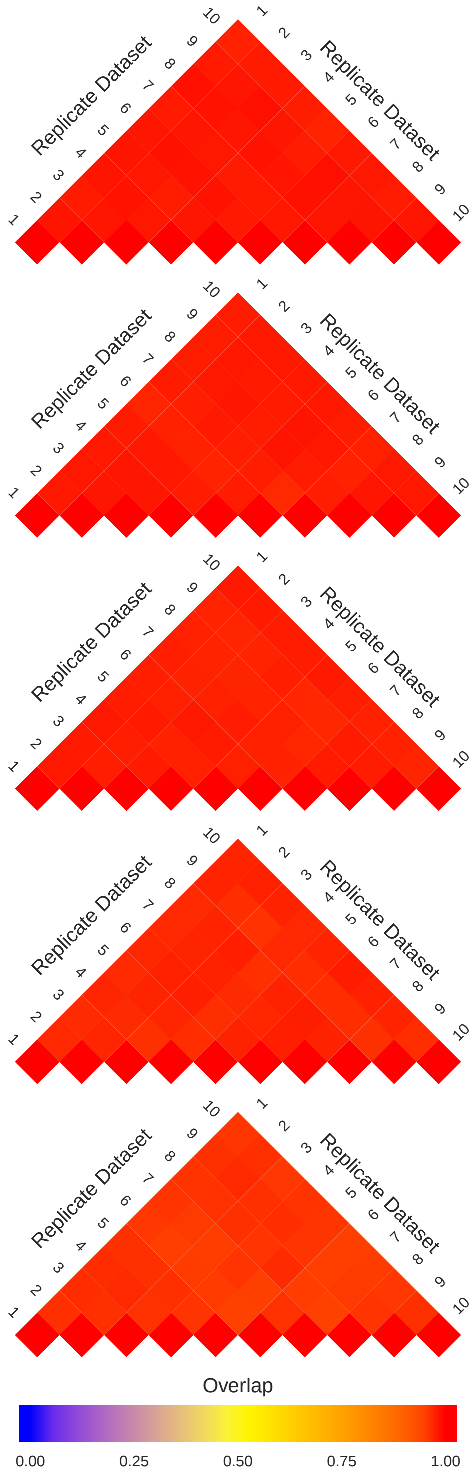


Figure B.8: Pine plots for dataset GSE53757 showing reproducibility of the results from ssGSEA across sample sizes. Reproducibility is quantified by overlap score (Equation 4.2). Each layer of the pine plot illustrates the overlap score of the results of a method for 10 replicate datasets with the same sample size. From top to bottom, the pine plot shows replicates with sample size 2×20, 2×15, 2×10, 2×5, and 2×3. The overlap score ranges from 0 to 1 represented by a gradient from blue to red, respectively, separated by yellow in the middle (overlap of 0.5).

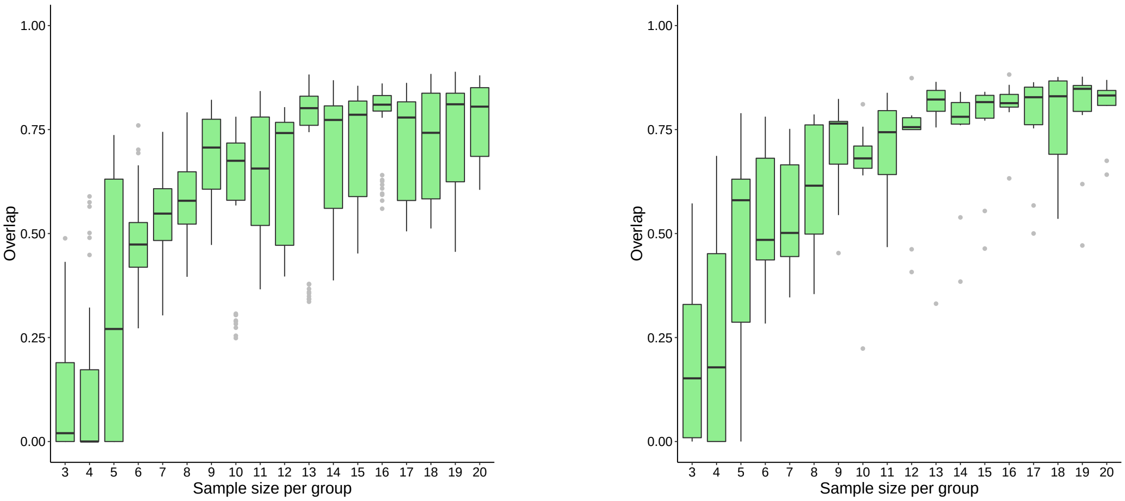


Figure B.9: Box plots showing the distribution of overlap scores resulting from gene set analysis using FRY when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

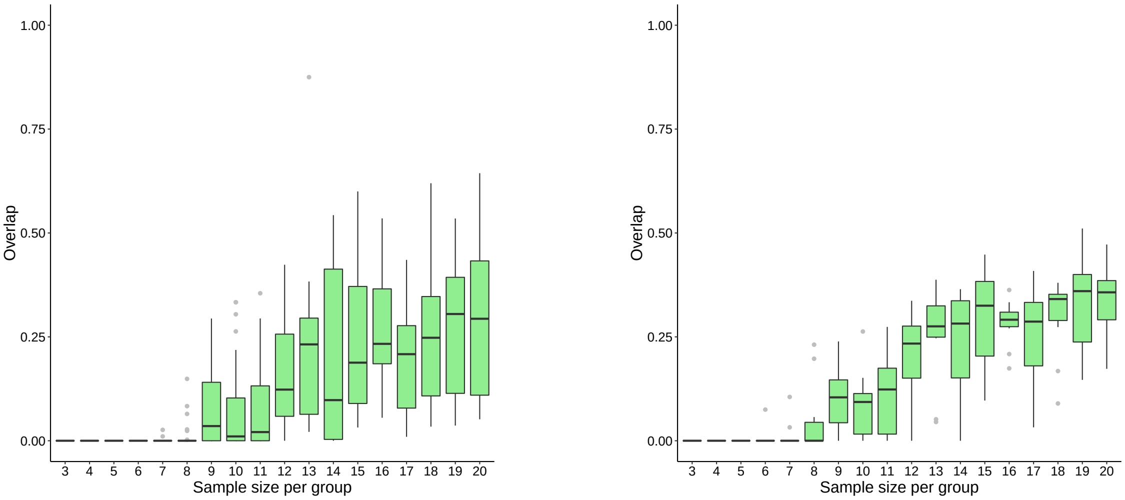


Figure B.10: Box plots showing the distribution of overlap scores resulting from gene set analysis using Camera when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

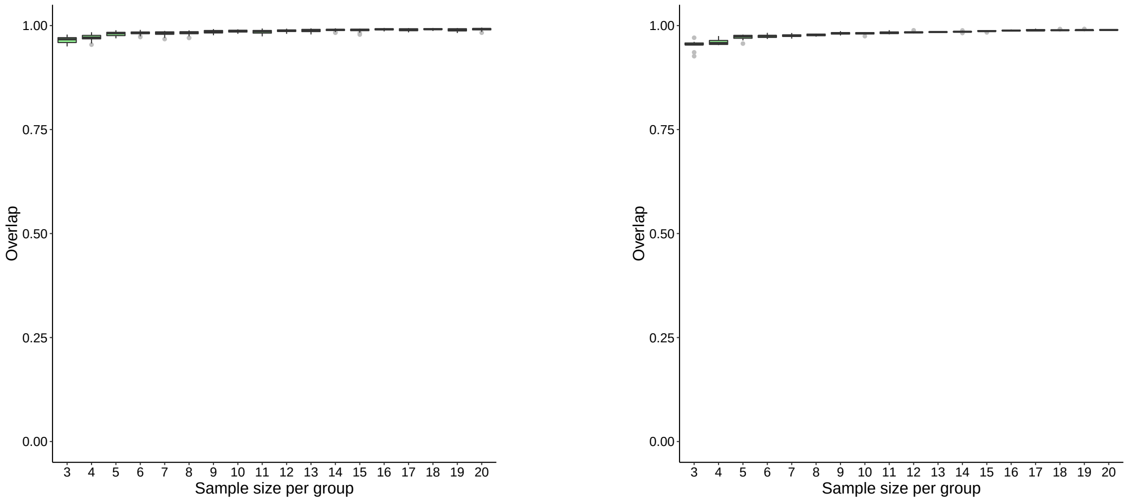


Figure B.11: Box plots showing the distribution of overlap scores resulting from gene set analysis using ssGSEA when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

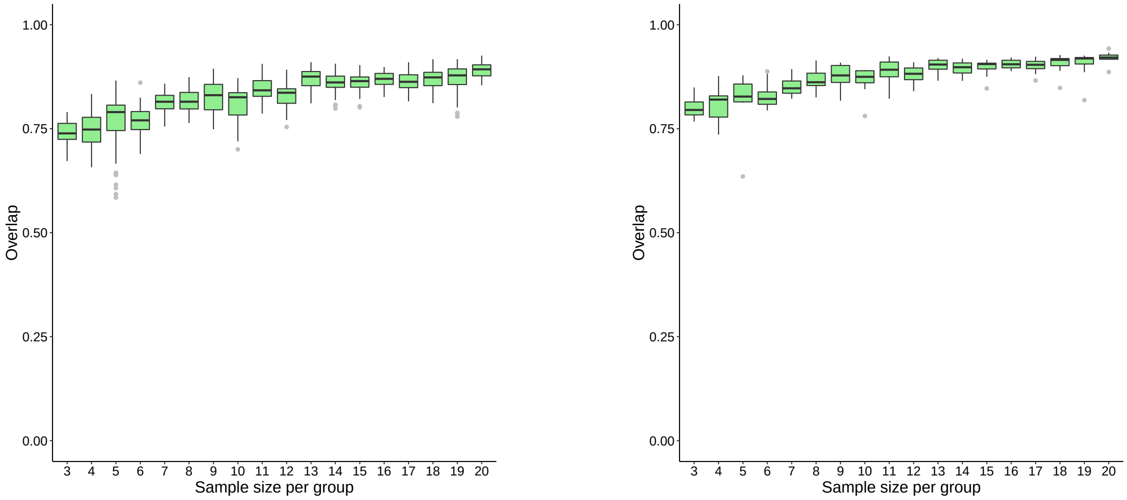


Figure B.12: Box plots showing the distribution of overlap scores resulting from gene set analysis using PAGE when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

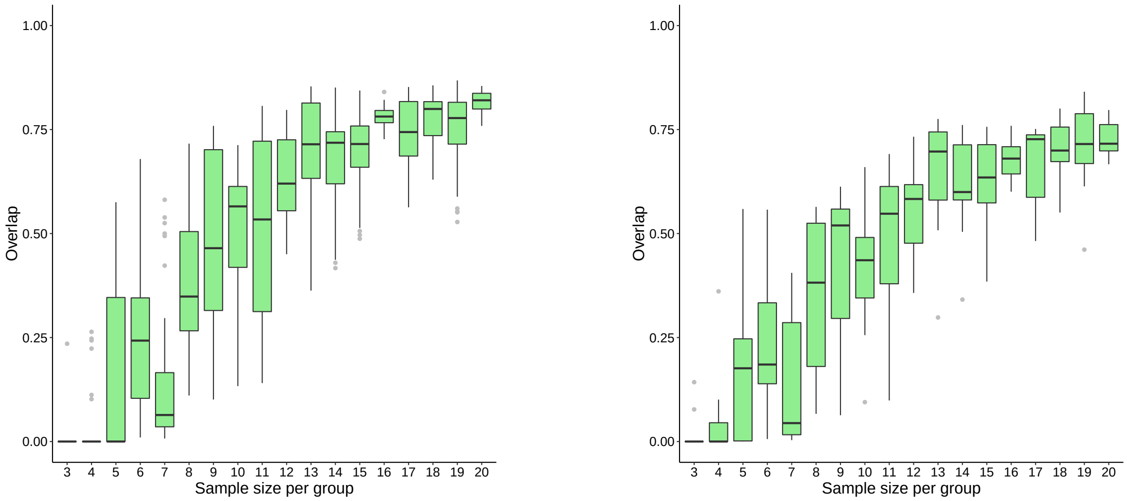


Figure B.13: Box plots showing the distribution of overlap scores resulting from gene set analysis using GSVA when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

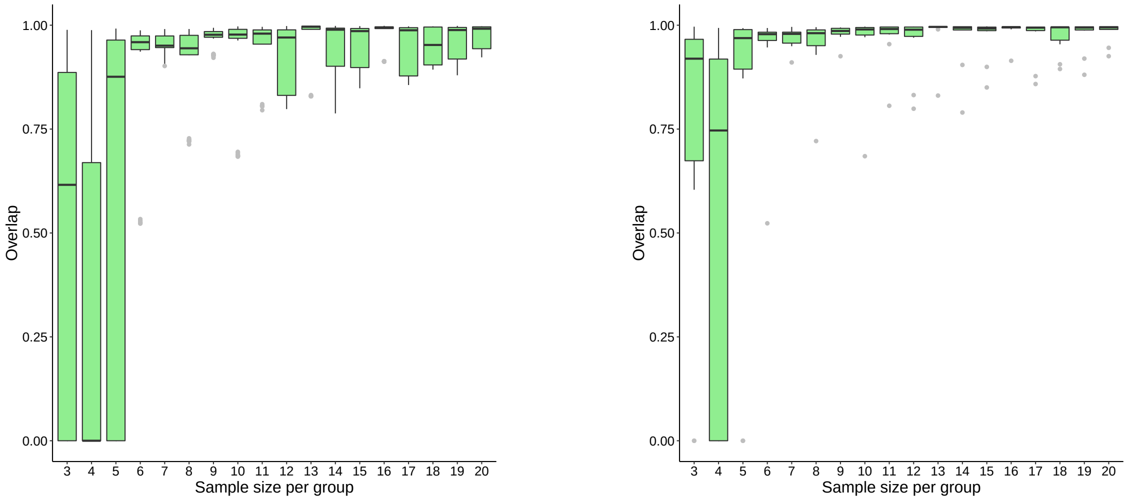


Figure B.14: Box plots showing the distribution of overlap scores resulting from gene set analysis using PLAGE when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

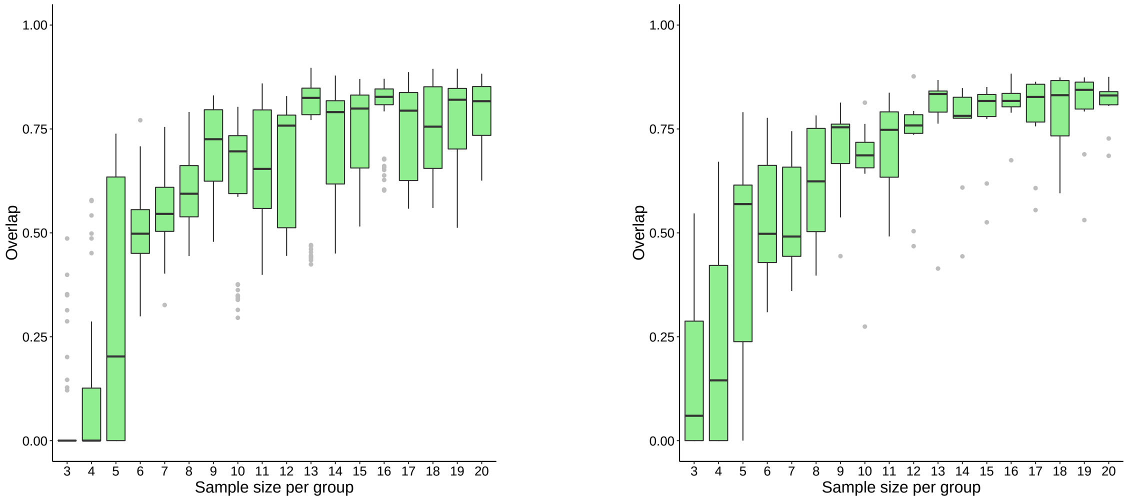


Figure B.15: Box plots showing the distribution of overlap scores resulting from gene set analysis using ROAST when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

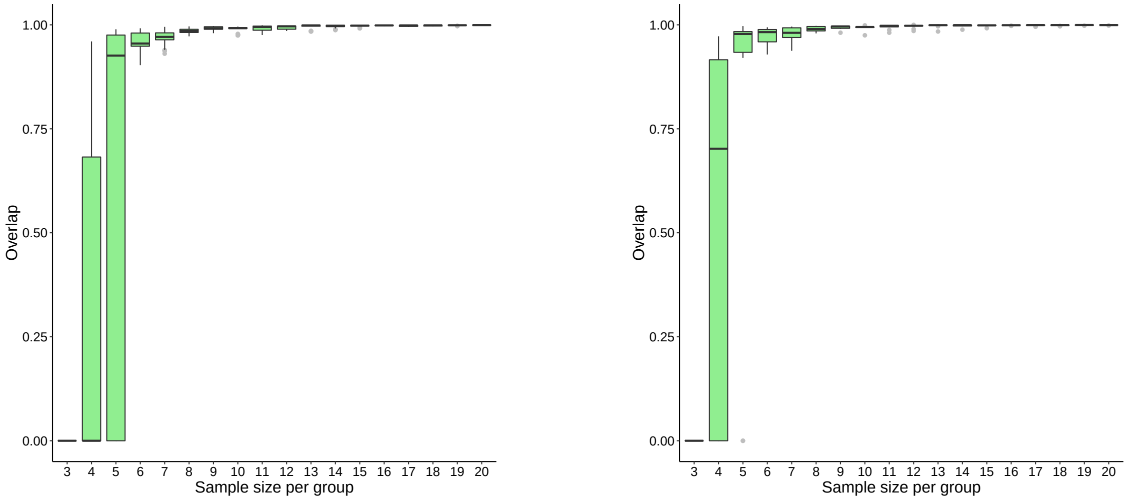


Figure B.16: Box plots showing the distribution of overlap scores resulting from gene set analysis using GlobalTest when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

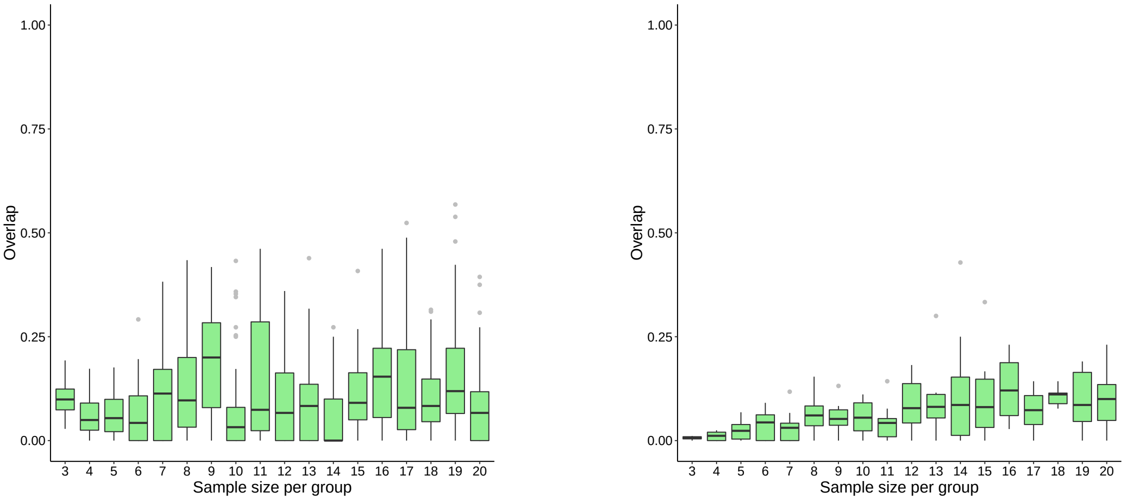


Figure B.17: Box plots showing the distribution of overlap scores resulting from gene set analysis using PADOG when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

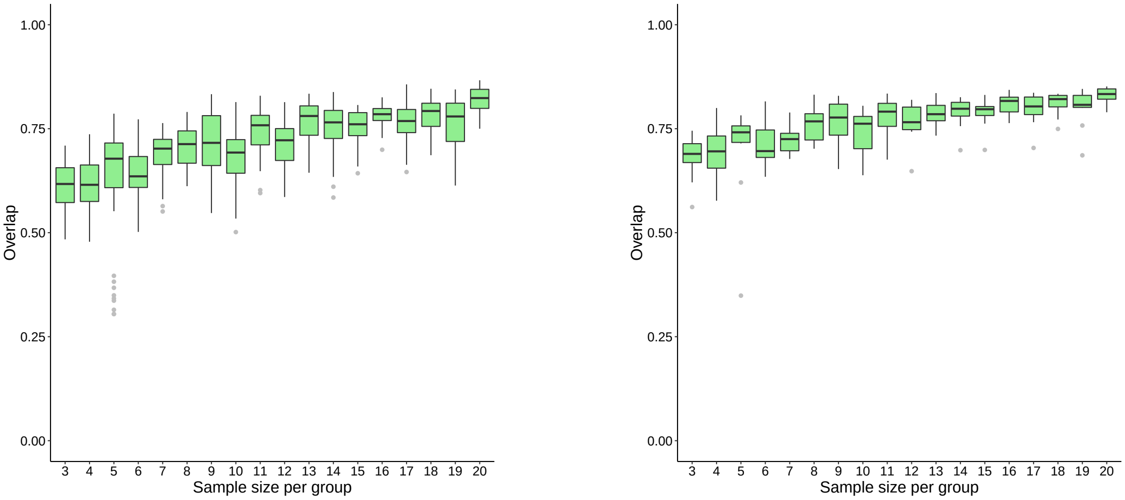


Figure B.18: Box plots showing the distribution of overlap scores resulting from gene set analysis using GSEA-G when using the original dataset GSE53757 for generating replicate datasets. The panel on the left shows the overlap scores from replicate datasets, while that on the right depicts the overlap scores of each replicate dataset and the whole dataset. See Figure 3.3 caption for more information.

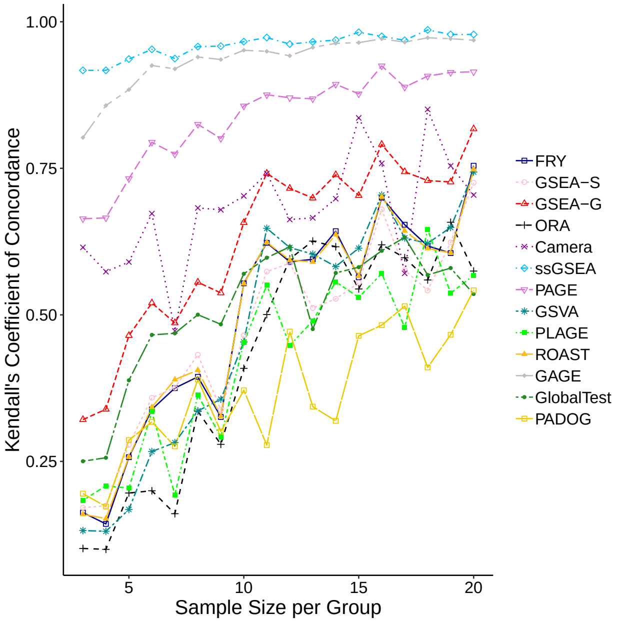


Figure B.19: Kendall's coefficient of concordance for each method under study when using the original dataset GSE10334 for generating replicate datasets. The x-axis shows the sample size. The y-axis shows concordance coefficients of the results of gene set analysis of 10 replicate datasets of the same size.

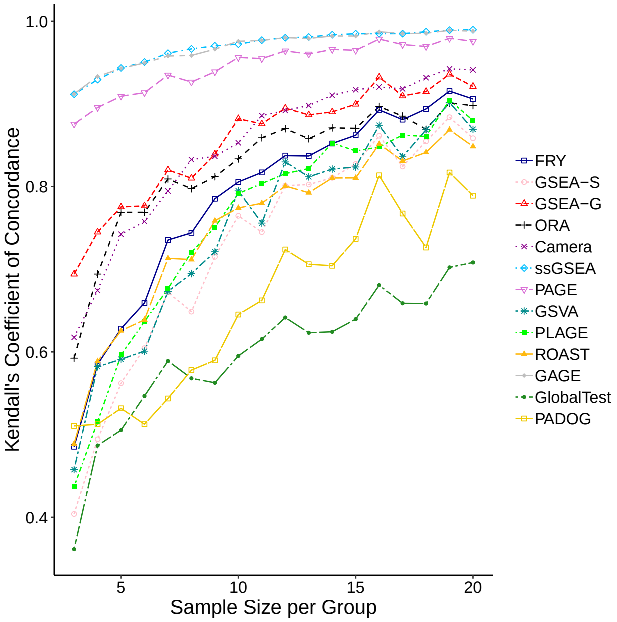


Figure B.20: Kendall's coefficient of concordance for each method under study when using the original dataset GSE13355 for generating replicate datasets. The x-axis shows the sample size. The y-axis shows concordance coefficients of the results of gene set analysis of 10 replicate datasets of the same size.

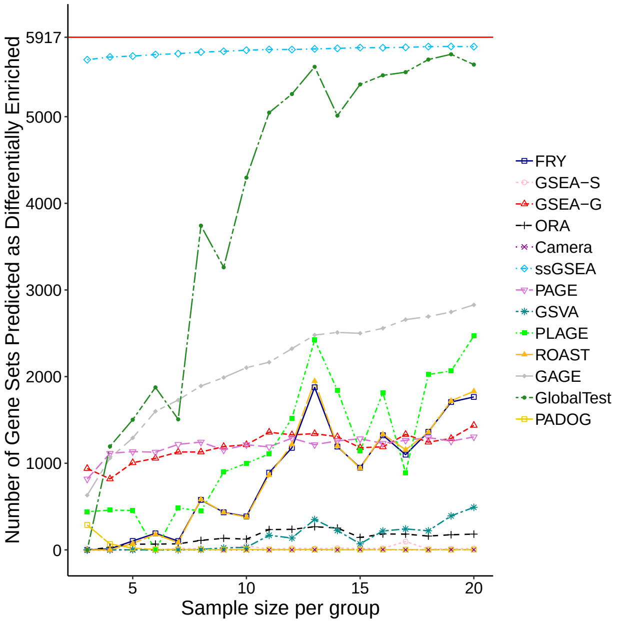


Figure B.21: The number of gene sets predicted as differentially enriched for each method under study when using the original dataset GSE10334 for generating replicate datasets. The x-axis shows the sample size per group. The y-axis shows the average number of gene sets predicted as differentially enriched across 10 replicate datasets of the same size. The red line parallel to the x-axis shows the size of the gene set database being used, i.e. the maximum possible number of gene sets that could be predicted as being differentially enriched.

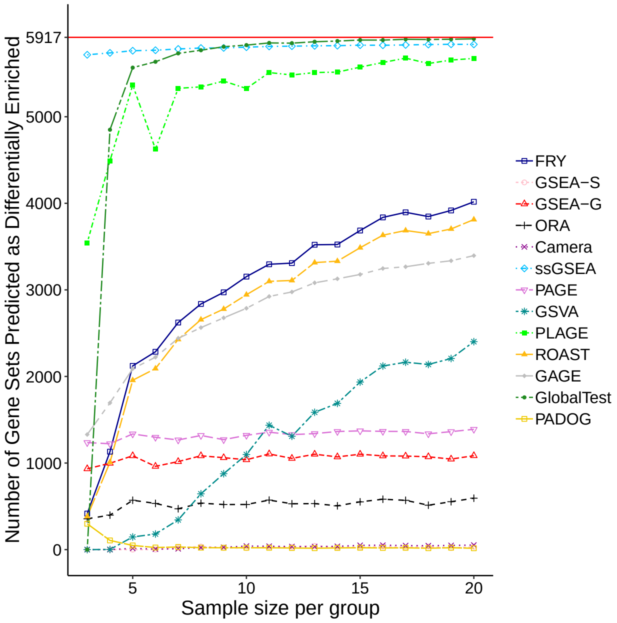


Figure B.22: The number of gene sets predicted as differentially enriched for each method under study when using the original dataset GSE13355 for generating replicate datasets. The x-axis shows the sample size per group. The y-axis shows the average number of gene sets predicted as differentially enriched across 10 replicate datasets of the same size. The red line parallel to the x-axis shows the size of the gene set database being used, i.e. the maximum possible number of gene sets that could be predicted as being differentially enriched.

Table B.2: Average (μ) and standard deviation (σ) of the number of differentially enriched gene sets reported by each method for control-control experiment when using the original dataset GSE53757 for generating replicate datasets. Since both phenotypes have been randomly chosen from control samples of a real dataset (GSE53757), no differentially enriched gene set is expected. The reported gene sets are considered as false positives. Methods with a large number of reported gene sets suffer from a lack of specificity.

Sample size per group		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
FRY	μ	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	σ	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GSEA-S	μ	0.0	0.0	0.0	0.0	17.2	6.5	10.2	13.1	16.1	7.0	14.0	10.4	16.4	11.1	9.5	10.3	11.3	9.9
	σ	0.0	0.0	0.0	0.0	18.5	4.2	8.8	12.8	14.9	4.4	16.8	9.6	15.9	6.3	7.8	7.4	7.4	11.9
GSEA-G	μ	621.8	578.1	596.6	555.8	609.1	559.5	677.8	792.0	884.5	789.0	720.6	511.8	639.9	556.5	565.6	758.0	707.5	425.5
	σ	463.4	580.2	454.5	450.4	441.4	476.2	420.7	472.5	526.7	528.7	442.4	348.0	531.0	471.6	441.3	655.6	326.1	338.3
ORA	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	σ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Camera	μ	0.0	0.0	0.8	0.2	0.0	0.3	38.2	28.3	54.1	18.5	67.0	38.8	102.2	55.5	62.5	72.9	55.3	129.3
	σ	0.0	0.0	2.4	0.4	0.0	0.5	107.6	59.9	83.2	18.5	80.7	29.9	98.2	64.6	61.4	55.1	34.8	78.4
ssGSEA	μ	5157.7	5174.8	5185.0	5220.1	5219.1	5231.9	5233.5	5249.3	5249.2	5256.5	5251.9	5261.5	5264.1	5271.0	5278.2	5276.9	5282.5	5283.0
	σ	1712.0	1717.0	1720.3	1731.6	1731.1	1735.4	1735.9	1741.1	1741.0	1743.4	1741.9	1745.0	1745.9	1748.3	1750.6	1750.2	1752.1	1752.2
PAGE	μ	1189.5	1071.5	1082.9	1095.2	1155.4	1206.6	1246.0	1343.0	1169.7	1225.2	1221.6	1020.3	1228.2	1139.4	1115.6	1128.4	1208.5	1055.5
	σ	426.8	509.0	472.1	424.1	417.4	456.1	473.4	477.4	530.3	516.4	467.7	409.7	538.5	412.9	546.4	472.9	471.7	473.0
GSVA	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	σ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PLAGE	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	σ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ROAST	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	σ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GAGE	μ	941.9	1291.3	1472.2	1718.0	1979.4	2073.3	2214.6	2317.6	2373.8	2452.0	2540.7	2553.8	2632.5	2647.8	2683.5	2741.0	2746.3	2779.1
	σ	364.0	488.9	510.9	592.4	665.2	698.2	740.2	772.1	792.0	824.3	847.1	848.3	874.6	880.6	892.6	910.4	911.9	923.5
GlobalTest	μ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	σ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
PADOG	μ	264.8	78.6	24.7	6.3	7.3	6.2	7.1	5.0	4.1	4.4	6.5	6.3	6.6	7.6	4.2	3.8	6.7	5.7
	σ	106.6	37.5	12.1	4.9	4.5	6.0	7.7	3.3	4.9	4.3	5.6	6.6	5.3	5.7	4.5	5.2	4.6	4.8

Appendix C

Supplementary Material for Chapter 4

Table C.1: The results of Cochran's Q test for all datasets.

Dataset	Q Statistic	Degrees of freedom	p-value
GSE53757	32133.2	9	<2.2e-16⋆
GSE13355	33592.8	9	<2.2e-16⋆
GSE10334	31661.4	9	<2.2e-16⋆
GSE26554	29326.9	9	<2.2e-16⋆

⋆ 2.2e-16 is the smallest p-value reported by *cochran.qtest* method from *RVAideMemoire*

Table C.2: The p-values of Wilcoxon sign tests for pairwise comparisons of the results of the methods using dataset GSE53757.

	Camera	FRY	GAGE	GSEA	GSVA	ORA	PAGE	PLAGE	ROAST
FRY	4.94e-324								
GAGE	0.00e+00	2.51e-101							
GSEA	4.47e-28	0.00e+00	0.00e+00						
GSVA	0.00e+00	2.37e-09	1.46e-57	0.00e+00					
ORA	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00				
PAGE	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.62e-44			
PLAGE	0.00e+00	5.21e-282	0.00e+00	0.00e+00	0.00e+00	4.94e-324	4.94e-324		
ROAST	0.00e+00	4.41e-01	4.69e-103	0.00e+00	2.47e-10	0.00e+00	0.00e+00	9.51e-277	
ssGSEA	4.94e-324	5.41e-207	0.00e+00	0.00e+00	1.15e-266	0.00e+00	0.00e+00	2.04e-09	1.57e-202

Note that a p-value of 0.00e+00 is smaller than the representable floating-point precision.

Table C.3: The p-values of Wilcoxon sign tests for pairwise comparisons of the results of the methods using dataset GSE13355.

	Camera	FRY	GAGE	GSEA	GSVA	ORA	PAGE	PLAGE	ROAST
FRY	0.00e+00								
GAGE	0.00e+00	4.08e-104							
GSEA	1.25e-09	9.88e-324	0.00e+00						
GSVA	0.00e+00	4.66e-159	3.03e-01	0.00e+00					
ORA	2.78e-101	9.88e-324	0.00e+00	1.05e-152	0.00e+00				
PAGE	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	9.55e-152			
PLAGE	9.88e-324	6.49e-233	0.00e+00	9.88e-324	0.00e+00	0.00e+00	0.00e+00		
ROAST	0.00e+00	5.33e-15	6.33e-74	0.00e+00	5.39e-134	0.00e+00	0.00e+00	4.37e-272	
ssGSEA	0.00e+00	7.62e-253	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	1.02e-03	1.81e-295

Note that a p-value of 0.00e+00 is smaller than the representable floating-point precision.

Table C.4: The p-values of Wilcoxon sign tests for pairwise comparisons of the results of the methods using dataset GSE10334.

Camera	Camera	FRY	GAGE	GSEA	GSVA	ORA	PAGE	PLAGE	ROAST
FRY	0.00e+00								
GAGE	0.00e+00	1.55e-11							
GSEA	2.70e-03	4.94e-324	0.00e+00						
GSVA	0.00e+00	2.36e-100	2.04e-01	0.00e+00					
ORA	4.41e-37	4.94e-324	0.00e+00	4.44e-23	4.94e-324				
PAGE	0.00e+00	4.94e-324	0.00e+00	0.00e+00	0.00e+00	0.00e+00			
PLAGE	0.00e+00	9.91e-132	6.79e-212	4.94e-324	3.28e-229	0.00e+00	0.00e+00		
ROAST	0.00e+00	5.00e-02	1.93e-12	0.00e+00	1.70e-104	0.00e+00	4.94e-324	7.23e-128	
ssGSEA	0.00e+00	0.00e+00	0.00e+00	4.94e-324	0.00e+00	4.94e-324	4.94e-324	6.10e-145	0.00e+00

Note that a p-value of 0.00e+00 is smaller than the representable floating-point precision.

Table C.5: The p-values of Wilcoxon sign tests for pairwise comparisons of the results of the methods using dataset GSE26554.

	Camera	FRY	GAGE	GSEA	GSVA	ORA	PAGE	PLAGE	ROAST
FRY	0.00e+00								
GAGE	0.00e+00	1.05e-128							
GSEA	3.920e-67	0.00e+00	0.00e+00						
GSVA	0.00e+00	3.78e-60	3.44e-24	0.00e+00					
ORA	5.00e-05	0.00e+00	0.00e+00	5.33e-39	0.00e+00				
PAGE	9.58e-105	0.00e+00	1.90e-320	9.41e-285	0.00e+00	2.07e-173			
PLAGE	4.94e-324	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00	0.00e+00		
ROAST	0.00e+00	1.76e-12	1.03e-65	0.00e+00	3.86e-65	0.00e+00	0.00e+00	0.00e+00	
ssGSEA	0.00e+00	0.00e+00	0.00e+00	4.94e-324	0.00e+00	4.94e-324	0.00e+00	3.06e-19	0.00e+00

Note that a p-value of 0.00e+00 is smaller than the representable floating-point precision.

Appendix D

Supplementary Material for Chapter 5

Algorithm 1 Calculation of specificity of ORA

```
Input:
    $\mathbb{G} = \{G_j \mid 1 \leq j \leq m\}$: A gene set database
    $\mathbb{L} = \{L_i \mid 1 \leq i \leq l\}$: A set of differentially
                                expressed gene lists
    $U$: A set of genes used as background set
       for ORA
    $\alpha$ : The significance level
    $\gamma$ : The threshold value used for identifying
         true positives
Output:
     Specificity value corresponding to
     each $L_i \in \mathbb{L}$

i = 1
while $i \leq l$ do
  $j = 1$
  while $j \leq m$ do
    $p_j = ORA(G_j, L_i, U)$
    $j = j + 1$
  end while
  Calculate $p_j^{adjusted}$ as the adjusted p-value corresponding to $p_j$, where $(1 \leq j \leq m)$
  $\mathbb{G}_i^+ \leftarrow \{G_k \mid p_k^{adjusted} < \alpha \text{ and } 1 \leq k \leq m\}$
  $\mathbb{G}_i^- = \mathbb{G} - \mathbb{G}_i^+$
  Calculate $T_i^+(\gamma)$ and $T_i^-(\gamma)$ using Equations 5.5 and 5.6
  Calculate $TP_i$, $FP_i$, $TN_i$, and $FN_i$ using Equations 5.7, 5.8, 5.9, and 5.10
  Calculate $SPC_i$ using Equation 5.11
  $i = i + 1$
end while
```

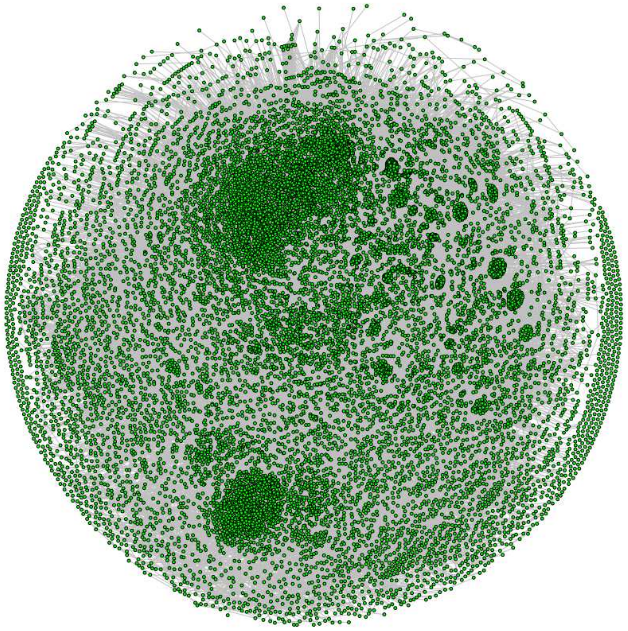


Figure D.1: The graph representing the overlap between gene sets in GeneSetDB. In this graph, each vertex represents a gene set in GeneSetDB, and each edge represents an overlap with Jaccard coefficient greater than or equal to 0.5 between two gene sets. The "hairball" is the result of a large number of gene sets with a substantial overlap (≥ 0.5) with each other.

Table D.1: The result of Shapiro-Wilk tests for different values of γ. All p-values are less than 0.0000001.

γ	W-Statistic	p value
0.10	0.783470	<0.0000001
0.20	0.773921	<0.0000001
0.30	0.771523	<0.0000001
0.40	0.770568	<0.0000001
0.50	0.770193	<0.0000001
0.60	0.769961	<0.0000001
0.70	0.769868	<0.0000001
0.80	0.769840	<0.0000001
0.90	0.769828	<0.0000001
0.99	0.769821	<0.0000001

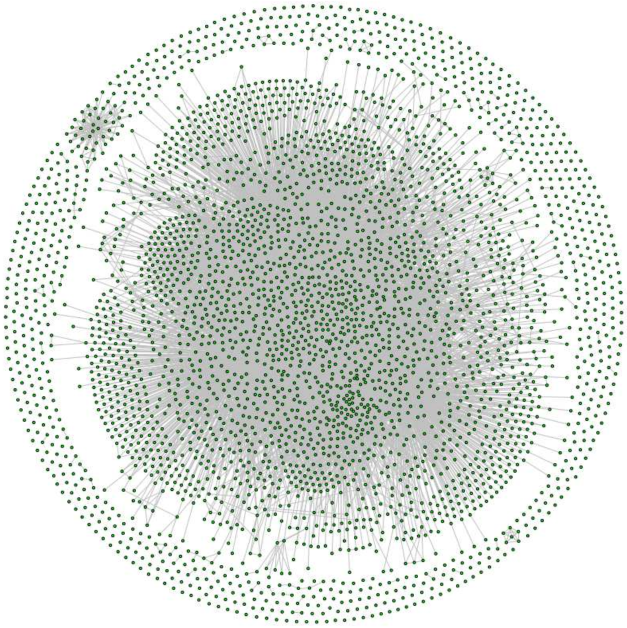


Figure D.2: The graph representing the overlap between gene sets in GeneSigDB. In this graph, each vertex represents a gene set in GeneSigDB, and each edge represents an overlap with Jaccard coefficient greater than or equal to 0.5 between two gene sets. The "hairball" is the result of a large number of gene sets with a substantial overlap (≥ 0.5) with each other.

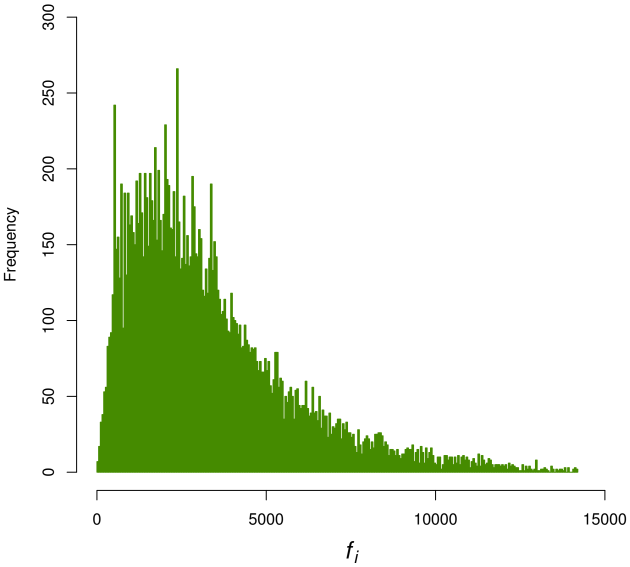


Figure D.3: A frequency plot for f_i values in GeneSetDB illustrates the prevalence of gene set overlap. For each gene set G_i in a gene set database $\mathbb{G}$ (GeneSetDB here), f_i is the number of gene sets G_j ($j \neq i$) in $\mathbb{G}$ with a non-zero overlap with G_i.

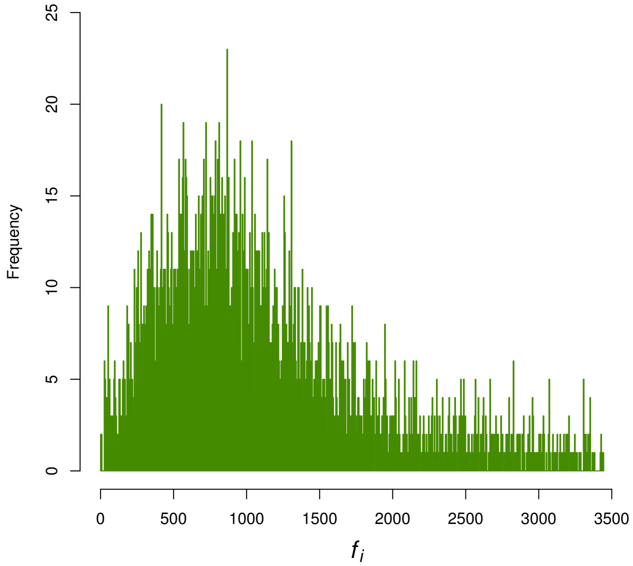


Figure D.4: A frequency plot for f_i values in GeneSigDB illustrates the prevalence of gene set overlap. For each gene set G_i in a gene set database $\mathbb{G}$ (GeneSigDB here), f_i is the number of gene sets G_j $(j \neq i)$ in $\mathbb{G}$ with a non-zero overlap with G_i.

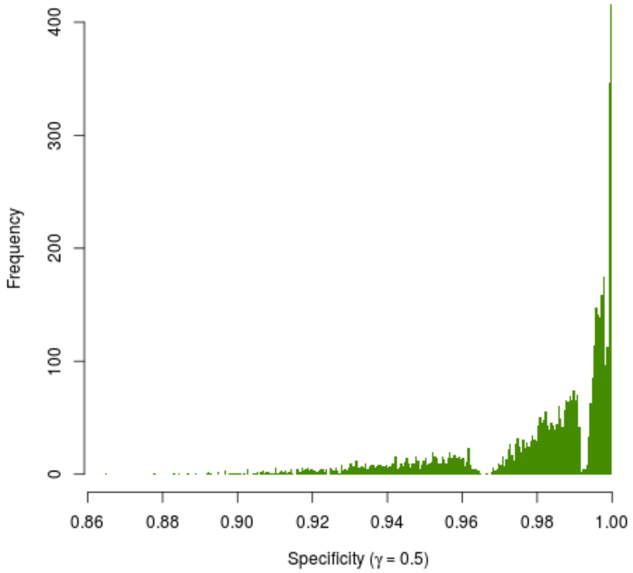


Figure D.5: The histogram of the specificity values ($\gamma = 0.5$). Obvious deviation of the histogram from a bell-shaped curve suggests that the specificity values are not normally distributed. A similar pattern was observed for other values of γ.

Appendix E

Supplementary Material for Chapter 7

Table E.1: List of genes associated with JIA that were extracted from literature.

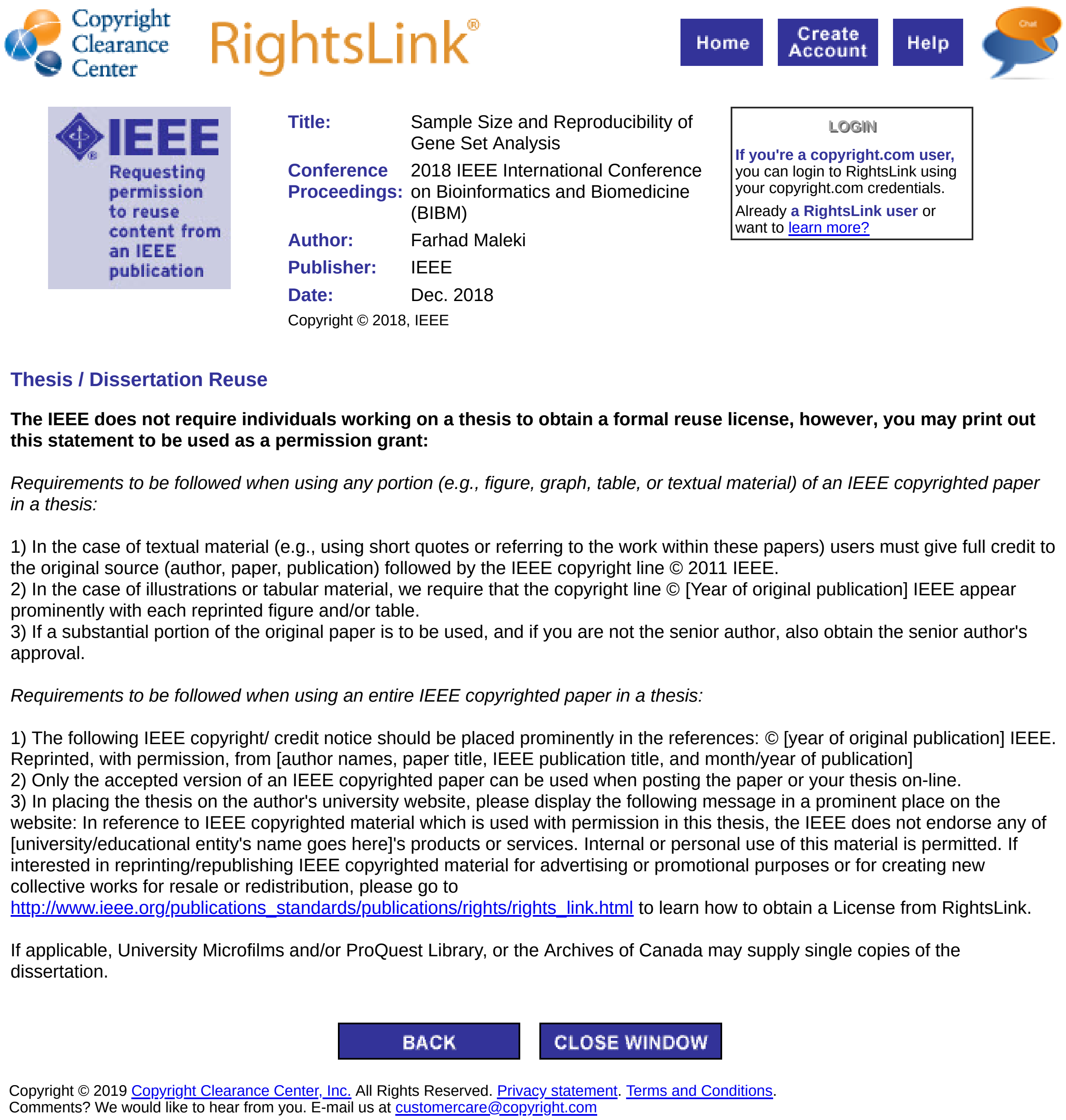
Gene ID	Gene Symbol	Gene ID	Gene Symbol
7124	TNF	6556	SLC11A1
3569	IL6	51752	ERAP1
3586	IL10	51752	ERAP1
64127	NOD2	79679	VTCN1
3557	IL1RN	7332	UBE2L3
4282	MIF	7297	TYK2
929	CD14	7185	TRAF1
3558	IL2	5771	PTPN2
861	RUNX1	111343	Il1
6352	CCL5	3560	IL2RB
114548	NLRP3	4055	LTBR
26191	PTPN22	4012	LNPEP
3596	IL13	8838	CCN6
4210	MEFV	64167	ERAP2
3559	IL2RA	221895	JAZF1
3593	IL12B	5619	PRM1
864	RUNX3	9051	PSTPIP1
7128	TNFAIP3	677	ZFP36L1
3570	IL6R	57511	COG6
59067	IL21	3899	AFF3
149233	IL23R	116028	RMI2
3659	IRF1	79722	ANKRD55
6775	STAT4	57198	ATP8B2
727	C5	79899	PRR5L
3554	IL1R1	164832	LONRF2

Appendix F

Licences to republish

"The IEEE does not require individuals working on a thesis to obtain a formal reuse license" (see the attached form). Also, according to the copyright form for the papers in Chapters 4 and 5, the authors retain the rights to publish the article on their own websites or their employer's websites as long as full credit to the original source is provided. Inquiries regarding the permission to reproduce the papers in Chapters 4 and 5 are in process. That documentation will be added to this appendix as soon as it arrives.

The inquiry about reusing the papers presented at the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC) and the permission is as follows:


Copyright Clearance Center
RightsLink®
Home
Create Account
Help
Chat
IEEE
Requesting permission to reuse content from an IEEE publication

Title: Sample Size and Reproducibility of Gene Set Analysis

Conference Proceedings: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

Author: Farhad Maleki

Publisher: IEEE

Date: Dec. 2018

LOGIN

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials.

Already **a RightsLink user** or want to learn more?

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

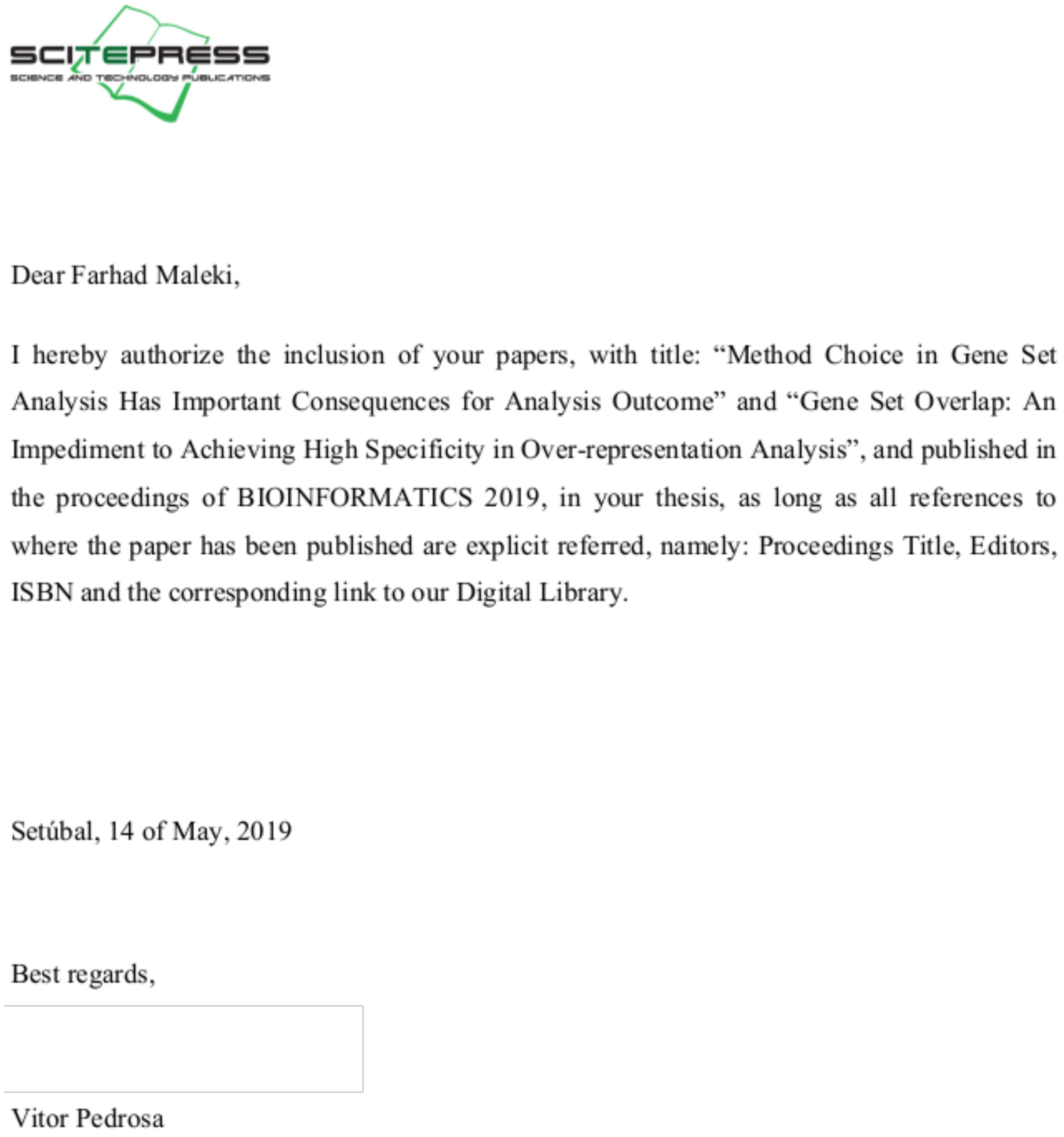
Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.



SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Dear Farhad Maleki,

I hereby authorize the inclusion of your papers, with title: “Method Choice in Gene Set Analysis Has Important Consequences for Analysis Outcome” and “Gene Set Overlap: An Impediment to Achieving High Specificity in Over-representation Analysis”, and published in the proceedings of BIOINFORMATICS 2019, in your thesis, as long as all references to where the paper has been published are explicit referred, namely: Proceedings Title, Editors, ISBN and the corresponding link to our Digital Library.

Setúbal, 14 of May, 2019

Best regards,

Vitor Pedrosa

ACM Publishing License and Audio/Video Release

Title of the Work: Gene Set Databases: A Fountain of Knowledge or a Siren Call?
Submission ID:bcb087
Author/Presenter(s): Farhad Maleki (Univ. of Saskatchewan); Katie Ovens (Univ. of Saskatchewan); Ian McQuillan (Univ. of Saskatchewan); Elham Rezaei (Univ. of Saskatchewan); Alan M. Rosenberg (Univ. of Saskatchewan); Anthony J. Kusalik (Univ. of Saskatchewan)
Type of material: Full Paper

Publication and/or Conference Name: ACM-BCB'19: 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics Proceedings

1. Glossary

2. Grant of Rights

(a) Owner hereby grants to ACM an exclusive, worldwide, royalty-free, perpetual, irrevocable, transferable and sublicenseable license to publish, reproduce and distribute all or any part of the Work in any and all forms of media, now or hereafter known, including in the above publication and in the ACM Digital Library, and to authorize third parties to do the same.

(b) In connection with software and "Artistic Images and "Auxiliary Materials, Owner grants ACM non-exclusive permission to publish, reproduce and distribute in any and all forms of media, now or hereafter known, including in the above publication and in the ACM Digital Library.

(c) In connection with any "Minor Revision", that is, a derivative work containing less than twenty-five percent (25%) of new substantive material, Owner hereby grants to ACM all rights in the Minor Revision that Owner grants to ACM with respect to the Work, and all terms of this Agreement shall apply to the Minor Revision.
(d) If your paper is withdrawn before it is published in the ACM Digital Library, the rights revert back to the author(s).

☑ A. Grant of Rights. I grant the rights and agree to the terms described above.

□ B. Declaration for Government Work. I am an employee of the national government of my country and my Government claims rights to this work, or it is not copyrightable (Government work is classified as Public Domain in U.S. only)

Are any of the co-authors, employees or contractors of a National Government? ○ Yes ● No

---

3. Reserved Rights and Permitted Uses.

(a) All rights and permissions the author has not granted to ACM in Paragraph 2 are reserved to the Owner, including without limitation the ownership of the copyright of the Work and all other proprietary rights such as patent or trademark rights.

(b) Furthermore, notwithstanding the exclusive rights the Owner has granted to ACM in Paragraph 2(a), Owner shall have the right to do the following:

(i) Reuse any portion of the Work, without fee, in any future works written or edited by the Author, including books, lectures and presentations in any and all media.

(ii) Create a "Major Revision" which is wholly owned by the author

(iii) Post the Accepted Version of the Work on (1) the Author's home page, (2) the Owner's institutional repository, (3) any repository legally mandated by an agency funding the research on which the Work is based, and (4) any non-commercial repository or aggregation that does not duplicate ACM tables of contents, i.e., whose patterns of links do not substantially duplicate an ACM-copyrighted volume or issue. Non-commercial repositories are here understood as repositories owned by non-profit organizations that do not charge a fee for accessing deposited articles and that do not sell advertising or otherwise profit from serving articles.

(iv) Post an "Author-Izer" link enabling free downloads of the Version of Record in the ACM Digital Library on (1) the Author's home page or (2) the Owner's institutional repository;

(v) Prior to commencement of the ACM peer review process, post the version of the Work as submitted to ACM ("Submitted Version" or any earlier versions) to non-peer reviewed servers;

(vi) Make free distributions of the final published Version of Record internally to the Owner's employees, if applicable;

(vii) Make free distributions of the published Version of Record for Classroom and Personal Use;

(viii) Bundle the Work in any of Owner's software distributions; and

(ix) Use any Auxiliary Material independent from the Work.

When preparing your paper for submission using the ACM TeX templates, the rights and permissions information and the bibliographic strip must appear on the lower left hand portion of the first page.

The new ACM Consolidated TeX template Version 1.3 and above automatically creates and positions these text blocks for you based on the code snippet which is system-generated based on your rights management choice and this particular conference.

NOTE: For authors using the ACM Microsoft Word Master Article Template and Publication Workflow, The ACM Publishing System (TAPS) will add the rights statement to your papers for you. Please check with your conference contact for information regarding submitting your source file(s) for processing.

Please copy and paste \setcopyright{acmlicensed} before \begin{document} and please copy and paste the following code snippet into your TeX file between \begin{document} and \maketitle, either after or before CCS codes.

```
\copyrightyear{2019}
\acmYear{2019}
\acmConference[ACM-BCB '19]{10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics}{September 7--10, 2019}{Niagara Falls, NY, USA}
```

\acmBooktitle{10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7--10, 2019, Niagara Falls, NY, USA}
\acmPrice{15.00}
\acmDOI{10.1145/3307339.3342146}
\acmISBN{978-1-4503-6666-3/19/09}

If you are using the ACM Microsoft Word template, or still using an older version of the ACM TeX template, or the current versions of the ACM SIGCHI, SIGGRAPH, or SIGPLAN TeX templates, you must copy and paste the following text block into your document as per the instructions provided with the templates you are using:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA
© 2019 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3342146

NOTE: Make sure to include your article's DOI as part of the bibstrip data; DOIs will be registered and become active shortly after publication in the ACM Digital Library. Once you have your camera ready copy ready, please send your source files and PDF to your event contact for processing.

---

4. ACM Citation and Digital Object Identifier.

(a) In connection with any use by the Owner of the Definitive Version, Owner shall include the ACM citation and ACM Digital Object Identifier (DOI).
(b) In connection with any use by the Owner of the Submitted Version (if accepted) or the Accepted Version or a Minor Revision, Owner shall use best efforts to display the ACM citation, along with a statement substantially similar to the following:

> "© [Owner] [Year]. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive version was published in {Source Publication}, https://doi.org/10.1145/{number}."

5. Audio/Video Recording

I hereby grant permission for ACM to include my name, likeness, presentation and comments in any and all forms, for the Conference and/or Publication.

I further grant permission for ACM to record and/or transcribe and reproduce my presentation as part of the ACM Digital Library, and to distribute the same for sale in complete or partial form as part of an ACM product on CD-ROM, DVD, webcast, USB device, streaming video or any other media format now or hereafter known.

I understand that my presentation will not be sold separately as a stand-alone product without my direct consent. Accordingly, I give ACM the right to use my image, voice, pronouncements, likeness, and my name, and any biographical material submitted by me, in connection with the Conference and/or Publication, whether used in excerpts or in full, for distribution described above and for any associated advertising or exhibition.

Do you agree to the above Audio/Video Release? ● Yes ○ No

6. Auxiliary Material

Do you have any Auxiliary Materials? ○ Yes ● No

7. Third Party Materials

In the event that any materials used in my presentation or Auxiliary Materials contain the work of third-party individuals or organizations (including copyrighted music or movie excerpts or anything not owned by me), I understand that it is my responsibility to secure any necessary permissions and/or licenses for print and/or digital publication, and cite or attach them below.

● We/I have not used third-party material.
○ We/I have used third-party materials and have necessary permissions.

8. Artistic Images

If your paper includes images that were created for any purpose other than this paper and to which you or your employer claim copyright, you must complete Part IV and be sure to include a notice of copyright with each such image in the paper.

● We/I do not have any artistic images.
○ We/I have any artistic images.

---

9. Representations, Warranties and Covenants

The undersigned hereby represents, warrants and covenants as follows:

(a) Owner is the sole owner or authorized agent of Owner(s) of the Work;

(b) The undersigned is authorized to enter into this Agreement and grant the rights included in this license to ACM;

(c) The Work is original and does not infringe the rights of any third party; all permissions for use of third-party materials consistent in scope and duration with the rights granted to ACM have been obtained, copies of such permissions have been provided to ACM, and the Work as submitted to ACM clearly and accurately indicates the credit to the proprietors of any such third-party materials (including any applicable copyright notice), or will be revised to indicate such credit;

(d) The Work has not been published except for informal postings on non-peer reviewed servers, and Owner covenants to use best efforts to place ACM DOI pointers on any such prior postings;

(e) The Auxiliary Materials, if any, contain no malicious code, virus, trojan horse or other software routines or hardware components designed to permit unauthorized access or to disable, erase or otherwise harm any computer systems or software; and

(f) The Artistic Images, if any, are clearly and accurately noted as such (including any applicable copyright notice) in the Submitted Version.

☑ I agree to the Representations, Warranties and Covenants.

10. Enforcement.

At ACM's expense, ACM shall have the right (but not the obligation) to defend and enforce the rights granted to ACM hereunder, including in connection with any instances of plagiarism brought to the attention of ACM. Owner shall notify ACM in writing as promptly as practicable upon becoming aware that any third party is infringing upon the rights granted to ACM, and shall reasonably cooperate with ACM in its defense or enforcement.

11. Governing Law

This Agreement shall be governed by, and construed in accordance with, the laws of the state of New York applicable to contracts entered into and to be fully performed therein.

Funding Agents

1. NSERC Discovery Grant award number(s):2016-06172, 05486-2017

DATE: **06/27/2019** sent to farhad.maleki@usask.ca at **23:06:06**