The Effect of Gene Set Overlap on Specificity of Over-representation Analysis

Farhad Maleki and Anthony J. Kusalik
Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

Abstract

Motivation: Gene set analysis methods are widely used to analyse 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 investigate the hypothesis that gene set overlap is an underlying cause of the low specificity in over-representation analysis.

Results: We show that gene set overlap is a ubiquitous phenomenon across gene set databases. Using statistical analysis, we also show that there is 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 establishes the direction for developing new gene set analysis methods.

Contact: farhad.maleki@usask.ca

1 Introduction

High-throughput "omics" technologies have widely been used to investigate biological questions that require screening of a large number of biomolecules. The main challenge facing these technologies is analysis of 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 the data resulting from high-throughput experiments [1]. 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 [2], GeneSigDB [3], GeneSetDB [4], Go-Elite [5], and Enrichr [6] are among the most widely used gene set databases.

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 [7]. 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. [4] used GeneSetDB to analyse a list of 79 differentially expressed Affymetrix probe sets resulting from an experiment where endothelial cells were induced to undergo apoptosis [8]. 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 even more challenging than interpreting the initial list of probes.

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. [7] 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, this 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

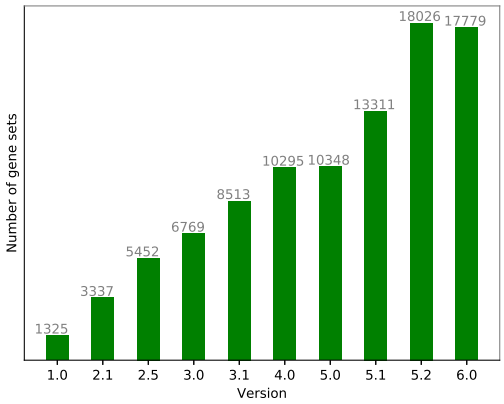

Figure 1: The number of gene sets in different versions of MSigDB

low specificity of gene set analysis methods. To deal with overlap between gene sets, PADOG [9] 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 [10] 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 [11] utilises a Bayesian approach that considers the overlap between GO categories to reduce the number of false positives. SetRank [12] 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. Fig. 1 illustrates the growth of MSigDB across its various versions. This gene set database has undergone a 13-fold increase in the number of gene sets compared to its first version. This steep growth leads to an increase in the number of gene sets overlapping with each other.

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 [13]. Besides its popularity, another reason for choosing ORA for evaluating gene set overlap is the previous work of Tarca et al. [7] that reported ORA as among the methods with low false positive rates in the absence of differentially expressed genes.

The rest of the paper is organised as follows. In Section 2 of the paper, we briefly describe ORA. In Section 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 [4], GeneSigDB [14], and MSigDB [2], which are well-established gene set databases. In Section 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, we describe the experimental results; using the methodology introduced in Section 4, we statistically investigate the effect of gene set overlap on the specificity of ORA. More specifically, we statistically assess the correlation between gene set overlap and specificity of ORA. In Section 6, 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 7 offers a short summary and conclusion.

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 [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. 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".

The ORA method can be outlined as follows. Suppose that data analysis for an experiment using a high-throughput technology predicts a set of differentially expressed genes L. Also assume 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 occurrence in L, 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

Table 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 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

U, i.e. all genes in U but not in G_i. Table 1 depicts the contingency table representation of the ORA problem, where $\| \bullet \|$ is the cardinality operator.

Assuming that genes are selected using a simple random sampling approach, this problem can be modeled using a hypergeometric distribution [26]. Accordingly, the probability of having n'_i genes from G_i among differentially expressed genes, i.e. L, is as follows:

$$f(n'_i; n, \|G_i\|, \|L\|) = \frac{\binom{\|G_i\|}{n'_i} \times \binom{n-\|G_i\|}{\|L\|-n'_i}}{\binom{n}{\|L\|}} \tag{1}$$

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 2.

$$p = \sum_{j=n'_i}^{\|G_i\|} f(j; n, \|G_i\|, \|L\|) = 1 - \sum_{j=0}^{n'_i-1} f(j; n, \|G_i\|, \|L\|) \tag{2}$$

As a toy example, assume that there are 12 genes under study ($n = 12$), G_i is a gene set with 4 genes ($\|G_i\| = 4$) involved in some biological function, 4 genes are predicted as differentially regulated ($\|L\| = 4$), and 3 of the differentially expressed genes belong to G_i ($n'_i = 3$). According to Fisher's exact test, under the null hypothesis that genes in G_i and $\overline{G_i}$ are equally likely to be differentially expressed, the p-value is equal to $f(3; 12, 4, 4) + f(4; 12, 4, 4) = 0.066$. Hence, using a typical significant level of 0.05, we cannot conclude that G_i is differentially enriched.

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 [27], KEGG [28], and OMIM [29] makes it possible to develop gene set databases. There are many publicaly available gene set databases including L2L [30], SignatureDB [31], CCancer [32], GeneSigDB [14], GeneSetDB [4], and MSigDB [2]. The latter three are widely used for gene set analysis.

MSigDB is the gene set database integrated with GSEA [33]. MSigDB acquires gene sets through manual curation and computational methods [2]. As a meta-database, MSigDB extracts gene sets from several sources including GO [27], KEGG [28], Reactome [34], and BioCarta [35].

GeneSigDB is a database of gene sets extracted from published experimental expression studies of genes, proteins, or miRNAs. GeneSigDB relied on PubMed searches to collect the 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 papers or supplementary documents.

GeneSetDB, as a 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 [4].

3.1 Gene set overlap and ORA: A hypothetical example

To show how overlap of gene sets may 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 Fig. 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

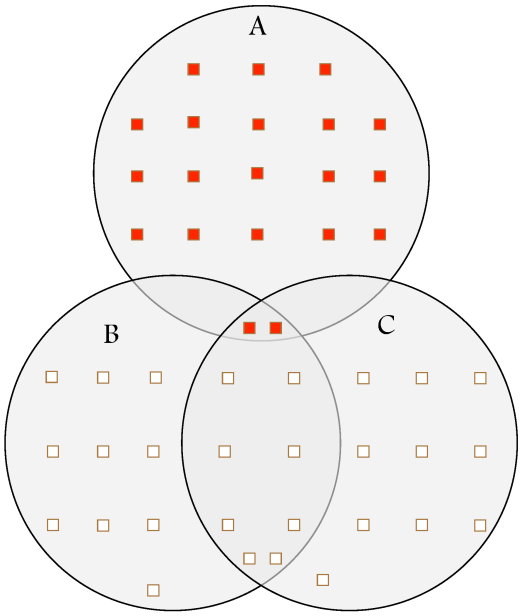

Figure 2: A hypothetical example: gene set overlap leading to lack of specificity of ORA. Each circle represents a gene set. Rectangles coloured in red and white represent differentially expressed and non-differentially expressed genes, respectively. Gene sets B and C are predicted as being differentially enriched by ORA solely due to partial overlap with A, a truly differentially enriched gene set.

the rest of the genes are coloured in white. As shown in Fig. 2, all genes in A have been predicted as being differentially expressed. Table 2 illustrates the contingency table for over-representation of B. The contingency table for over-representation of C is the same.

According to Fisher's exact test, both B and C are predicted as being differentially enriched with a 95% confidence level ($p\text{-}value = 0.0167$ and 1 degree of freedom). This result for gene sets B and C is solely due to their overlap with A. This example shows that gene set overlap may lead to a lack of specificity in gene set analysis methods.

In this paper, we use GeneSigDB version 4, GeneSetDB for Human, and MSigDB version 6.0, unless stated otherwise.

Table 2: The contingency table for the over-representation of B

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

3.2 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:

$$J(A, B) = \frac{\|A \cap B\|}{\|A \cup B\|} \tag{3}$$

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.

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:

$$O(L_i, \mathbb{G}) = \sum_{G_j \in \mathbb{G}} J(L_i, G_j) \tag{4}$$

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 $O(L_i, \mathbb{G})$.

3.3 Visualization of gene set overlap

We visualize a gene set database as a graph, where each gene set is represented as a vertex of the graph. For gene sets G_i and G_j in the gene set database, there is an edge between G_i and G_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. The graph visualisation can be generated using Fruchterman Reingold layout [36] in Gephi (version 0.9.2) [37].

In order to visualise these gene set databases in an informative way, we only show edges with a Jaccard coefficient greater than or equal to 0.5 while retaining all nodes. Even after removing edges with a Jaccard coefficient less than 0.5, one can see the gene set overlap remains rampant. The "hairballs" in Fig. 3 and supplemental Fig. S1 and Fig. S2 are due to the existence of a large number of edges, i.e. pairs of overlapped gene sets. These graphs highlight the existence of gene set overlap as a ubiquitous phenomenon in gene set databases.

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 $f_i = \|\{G_j \mid O(G_i, G_j) > 0 \ (j \neq i) \ and \ G_j \in \mathbb{G}\}\|$, which is the number of gene sets G_j $(j \neq i)$ in $\mathbb{G}$ with a non-zero overlap with G_i. Then we use a frequency plot to show the distribution of f_i values. Fig. 4, Fig. S3, and Fig. S4 illustrate the distribution of f_i values for MSigDB, GeneSetDB, and GeneSigDB, respectively. These figures are in agreement with Fig. 3, Fig. S1, and Fig. S2 and show the prevalence of gene set overlap in the aforementioned gene set databases.

Fig. 4 suggests that gene set overlap in MSigDB has a mixture model. The mixture model 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 Fig. 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 any other gene sets. Also, there are gene sets that overlap with the majority of gene sets in MSigDB. For example, the gene set associated with "cellular response to organic substance" GO term (GO:0071310) has a 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 sets including the gene sets defined using relatively more specific GO terms.

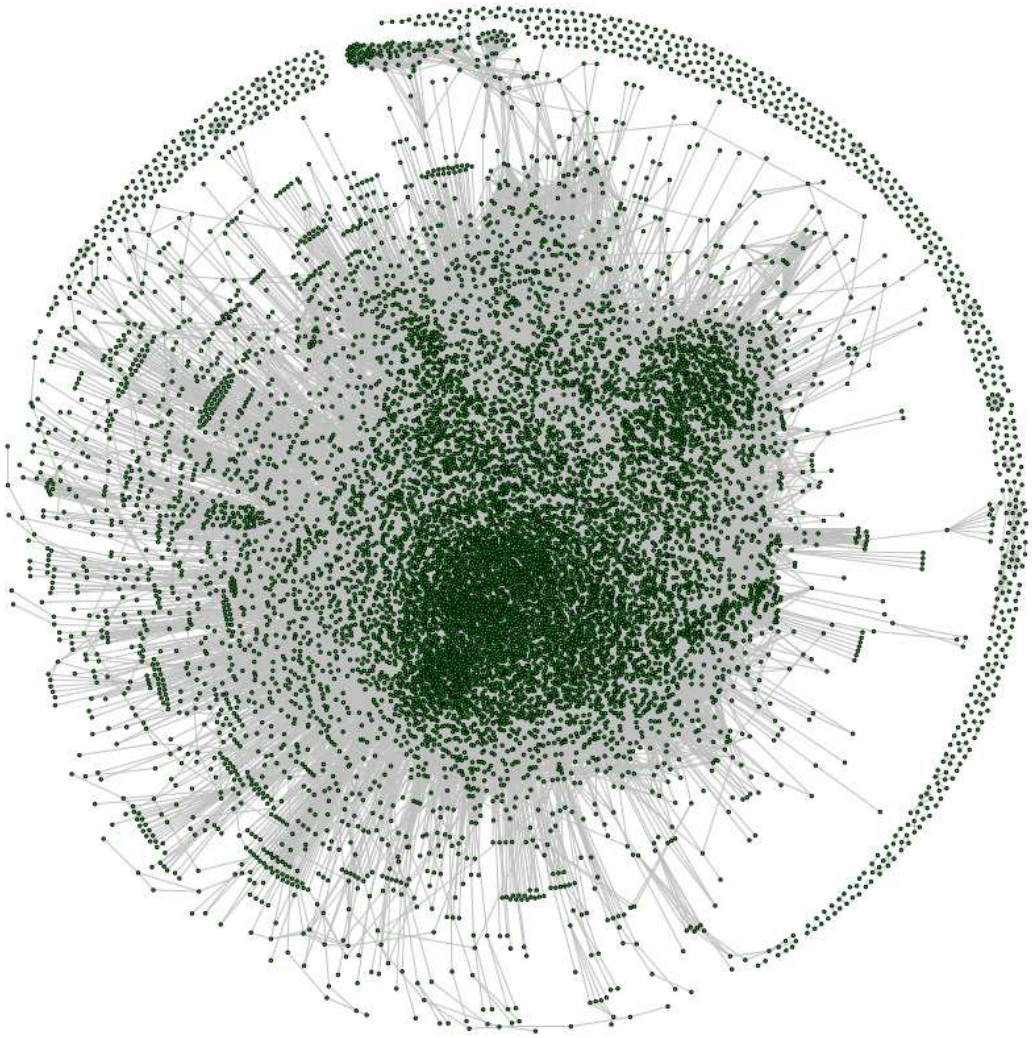

Figure 3: The graph representing the overlap between gene sets in MSigDB. In this graph, each vertex represents a gene set in MSigDB, and each edge represents an overlap with Jaccard coefficient greater than or equal to 0.5 between two gene sets (see Equation 3). The "hairball" is the result of a large number of gene sets that overlap with each other.

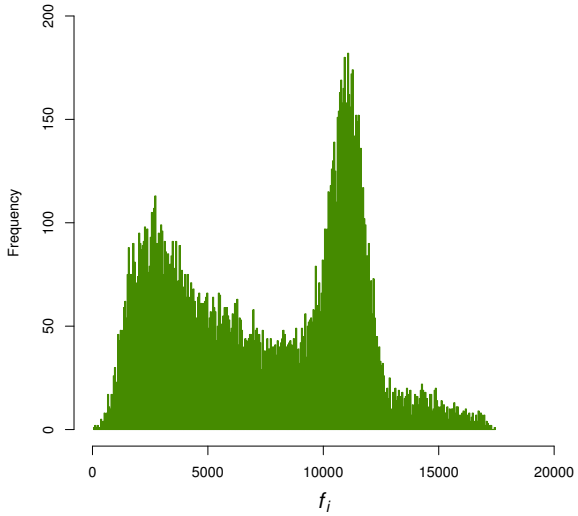

Figure 4: A frequency plot for f_i values in MSigDB illustrates the prevalence of gene set overlap. For each gene set G_i in a gene set database $\mathbb{G}$ (MSigDB here), f_i is the number of gene sets G_j $(j \neq i)$ in $\mathbb{G}$ with a non-zero overlap with G_i.

4 Methodology

In this section, the methodology used for a quantitative evaluation of the effect of gene set overlap on the specificity of ORA is described.

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. 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.

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.

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 of 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. We denote the set of truly differentially enriched and truly nondifferentially enriched gene sets by $T_i^+(\gamma)$ and $T_i^-(\gamma)$, respectively. For a given L_i and $\mathbb{G}$, Equations 5 and 6 define $T_i^+(\gamma)$ and $T_i^-(\gamma)$.

$$T_i^+(\gamma) = \{G_j \in \mathbb{G} \mid \frac{\|G_j \cap L_i\|}{\|G_j\|} \geq \gamma\} \tag{5}$$

$$T_i^-(\gamma) = \{G_j \in \mathbb{G} \mid \frac{\|G_j \cap L_i\|}{\|G_j\|} < \gamma\} \tag{6}$$

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

Since there is no consensus among the gene set analysis community for the value of γ, we consider a wide range of values for γ and investigate the main hypothesis for each value of γ.

Given a significance level α, a value of γ, a set of differentially expressed genes L_i, a background set U, and a gene set database $\mathbb{G} = \{G_j \mid 1 \leq j \leq m\}$, for each gene set G_j in $\mathbb{G}$, ORA is run using L_i, U, and G_j as inputs to calculate a p-value p_j. We calculate $p_1, \ldots, p_m$—the p-values corresponding to the over-representation of gene sets $G_1, \ldots, G_m$ in $\mathbb{G}$ according to Equation 2 (see Section 2). To adjust for multiple comparisons, the Benjamini-Hochberg FDR procedure is used [38]. All gene sets with an adjusted p-value of less than or equal to α are accepted as significant gene sets. $\mathbb{G}_i^+$ 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 7, 8, 9, and 10.

$$TP_i = T_i^+ \cap \mathbb{G}_i^+ \tag{7}$$
$$FP_i = \mathbb{G}_i^+ - T_i^+ \tag{8}$$
$$TN_i = T_i^- \cap \mathbb{G}_i^- \tag{9}$$
$$FN_i = \mathbb{G}_i^- - T_i^- \tag{10}$$

Using these values, specificity (SPC_i) is calculated according to Equation 11.

$$SPC_i = \frac{\|TN_i\|}{\|TN_i\| + \|FP_i\|} \tag{11}$$

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 $\mathbb{L} = \{L_i \mid 1 \leq i \leq n\}$.

Algorithm 1 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 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 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 $\mathbb{G}$. Also, ImmuneSigDB [39] version 6.0 was used as $\mathbb{L}$. Each gene set in ImmuneSigDB has been created by identifying differentially expressed genes in a raw dataset from the Gene Expression Omnibus (GEO), where the dataset has been used in the immunology literature. Therefore, a gene set 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 gene set 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.

Algorithm 1 Calculation of specificity of ORA

Input:

$\mathbb{G} = \{G_j \mid 1 \leq j \leq m\}$: A gene set database

$\mathbb{L} = \{L_i \mid 1 \leq i \leq n\}$: A set of differentially expressed gene list

U: A set of genes used as background set for ORA

α : The significance level

γ : The threshold value used for identifying true positives

Output:

Specificity value corresponding to each $L_i \in \mathbb{L}$

$i = 1$
while $i \leq n$ **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 and 6
  Calculate TP_i, FP_i, TN_i, and FN_i using Equations 7, 8, 9, and 10
  Calculate SPC_i using Equation 11
  $i = i + 1$
end while

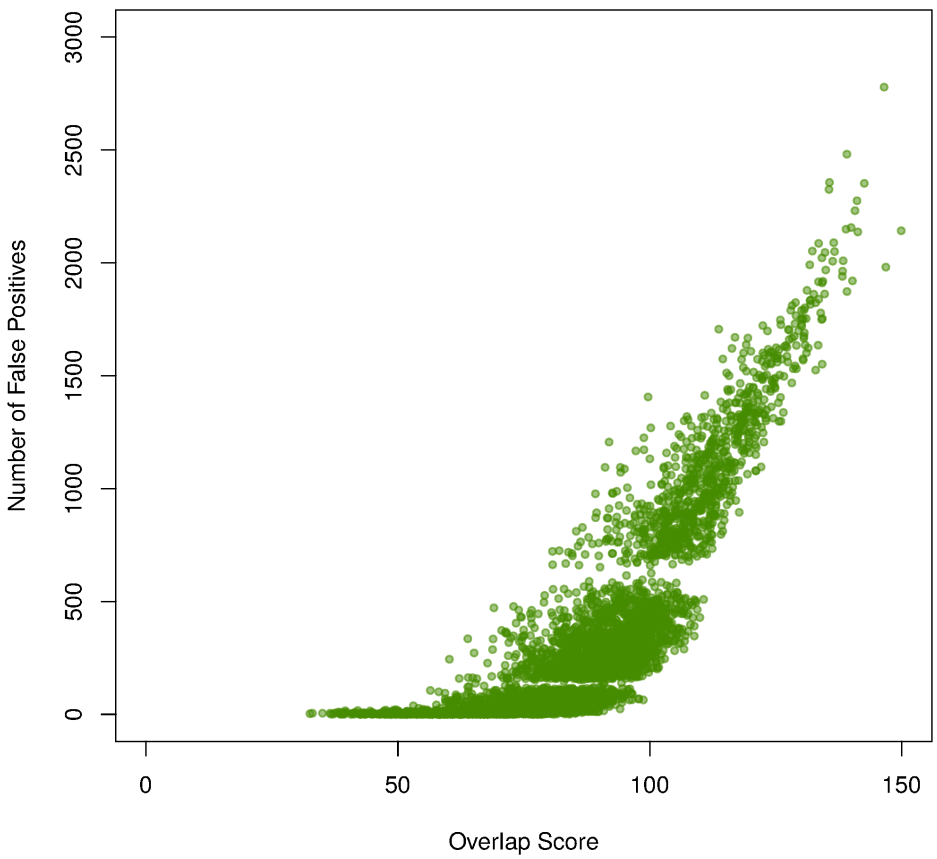

Figure 5: Number of false positives increases as overlap score increases ($\gamma = 0.5$)

For each value of γ, Algorithm 1 was run to calculate SPC_i corresponding to each $L_i \in \mathbb{L}$. Fig. 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 normality assumption. To test the null hypothesis that specificity values are normally distributed, we used the Shapiro-Wilk test [40]. Supplemental Table S1 shows the results for the aforementioned

Table 3: The result of Spearman rank correlation test for different values of γ.

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

values of γ. Considering the result of Shapiro-Wilk test, as confirmed by the histogram in supplemental Fig. S5, we conclude that specificity values are not normally distributed. Therefore, a Spearman's rank correlation coefficient test was conducted for each value of γ to test the null hypothesis that there is no correlation between specificity and overlap scores. Table 3 shows the result of this test for various values of γ. Considering these results, we can conclude that there is a strong negative correlation between gene set overlap scores and specificity of ORA.

6 Discussion

In this paper, we demonstrated that there is a significant negative correlation between the specificity of ORA and gene set overlap. The low specificity of ORA for gene sets with high overlap score decreases the reliability of ORA results, increases error, and reduces the chance of reproducibility of the results. These make the interpretation of the ORA results difficult and prone to investigator biases towards a hypothesis of interest.

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. [41] 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 gene sets based on GO is the child-parent relationship between GO terms. GO terms are organised 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 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 [42, 43, 44]. 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 low specificity and 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 [7]). SetRank is another gene set analysis method that achieves a higher specificity but sacrifices sensitivity. These suggest that more research on developing gene set analysis methods that achieve higher specificity without sacrificing sensitivity is required.

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 higher false positive rate. There is a need to focus on quality rather than sheer quantity in developing 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 if done manually. 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. 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 can be used to study scenarios with several differentially enriched gene sets and to further study the relationship between gene set overlap and specificity of ORA.

Since the input to ORA is a list of differentially expressed genes, we utilized ImmuneSigDB [39] for evaluating ORA using Algorithm 1. However, some gene set analysis methods require an expression matrix that represents expression level of genes under study across control and case samples. Adjusting Algorithm 1 for evaluating such gene set analysis methods requires an extra step in developing expression profiles with the differentially enriched gene set(s) encoded in expression values. We suggest extending Algorithm 1 in this way as future research.

The proposed methodology was used to study the effect of gene set overlap in the context of ORA. It can be adapted to evaluate other gene set analysis methods as well. For example, we suggest the study of the relationship between gene set overlap and the specificity of GSEA [33], which is another well-established gene set analysis method, as future work. Also, Algorithm 1 can be used seamlessly with sensitivity or accuracy instead of specificity.

7 Conclusion

In this paper, we studied the effect of gene set overlap on the result of ORA. We showed that gene set overlap is a ubiquitous phenomenon across gene set databases. We developed a framework for evaluating the effect of gene set overlap on ORA using quantitative measures such as specificity, sensitivity, and accuracy. Using the proposed method and statistical analysis, we showed that there is a significant negative correlation between gene set overlap and specificity of ORA.

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 differen-

tially enriched gene sets. Identifying and characterising the problem, which was accomplished in this paper, is the first step towards this goal.

References

[1] 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.

[2] 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.

[3] 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 gkr901, 2011.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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*, page gkx356, 2017.

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[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):D457–62, Jan 2016.

[29] 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 Res*, 30(1):52–55, Jan 2002.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

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

[37] 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.

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

[39] 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.

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

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

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

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

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