Sample Size and Reproducibility of Gene Set Analysis

1st Farhad Maleki
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
farhad.maleki@usask.ca

2nd Katie Ovens
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
katie.ovens@usask.ca

3rd Ian McQuillan
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
mcquillan@cs.usask.ca

4th Anthony J Kusalik
Department of Computer Science
University of Saskatchewan
Saskatoon, Canada
kusalik@cs.usask.ca

***Abstract*—Gene set analysis is widely used to gain insight from gene expression data. Achieving reproducible results is a fundamental part of any expression analysis. In this paper, we propose a systematic approach to study the effect of sample sizes on the reproducibility of the results of 10 gene set analysis methods. To do so, we quantify the concept of reproducibility and use real expression datasets of different sizes. Our findings suggest that, as a general pattern, the results of gene set analysis are more reproducible as sample size increases. However, the smallest sample size for achieving reproducible results are variable across gene set analysis methods. Moreover, for some methods, increasing sample size leads to an increase in the number of false positives.**

***Index Terms*—gene expression, gene set analysis, enrichment analysis, sample size**

I. Introduction

The choice of sample size is an essential factor in experiment design. Although determining sample size for achieving a predetermined power is possible for simple statistical processes, for a sophisticated and complex experiment such as gene set analysis, there is no methodological approach to determine the optimal number of samples for reaching a given statistical power. Consequently, researchers either choose the largest possible sample size considering funding and availability of samples and technicians for conducting the experiments, or they use an arbitrary sample size—as small as two or three samples per treatment. An unnecessarily large sample size results in financial loss and a waste of resources. In many cases, there are also ethical concerns. On the other hand, a small sample size leads to results that are not reliable and reproducible.

Paper published in Proceedings of BIBM 2018, https://doi.org/10.1109/BIBM.2018.8621462.

The effect of sample size on differential gene expression has been studied. Assuming equality of standardized effect size and gene-gene correlation among the differentially expressed genes, Tsi et al. [1] proposed an approach for estimating sample size using the beta-binomial distribution for the two-sample z-test. They reported that under a more general gene-gene correlation structure the proposed method might underestimate sample size. Stretch et al. [2] reported that small sample sizes result in unstable prediction of differentially expressed genes. Schurch et al. [3]—using a case-control study with 48 replicates per class—evaluated 11 tools for detecting differentially expressed genes in RNA-Seq experiments. They reported that, using 3 replicates, differentially expressed genes found by 8 methods cover only 20% to 40% of those genes predicted as being differentially expressed when using all 48 replicates. They also suggested that to predict over 85% of the differentially expressed genes found when considering the entire dataset, i.e. 48 replicates per group, at least 20 replicates are required.

Analysis of data from a typical gene expression experiment usually leads to the prediction of several hundred genes as being differentially expressed. Gaining insight from such a large list of genes is cumbersome and prone to investigator bias(es) towards a hypothesis of interest. To deal with such large lists of genes, gene set analysis—also known as enrichment analysis—is commonly performed. There are various gene set analysis methods available. These methods usually follow relatively complex procedures; therefore, unlike simple statistical tests, it is difficult to find an estimate for the smallest sample size that leads to a desired statistical power and reproducible results. Despite the existence of different studies evaluating the sensitivity and specificity of gene set analysis methods [4]–[6], to the best of our knowledge, there is no systematic analysis of the effect of sample size on the results of these methods. While it is the common belief that reproducibility increases as sample size increases, the extent

of the increase (if any) for different gene set analysis methods is not known. In this research, we address this need with the study of a comprehensive list of gene set analysis methods. We compare these methods on the basis of the reproducibility of their results when applied to expression datasets with sample sizes commonly used in gene expression studies.

II. Data and methodology

A. Data

The availability of public repositories of gene expression datasets such as Gene Expression Omnibus (GEO) [7] makes it possible to obtain original datasets with large samples sizes. To achieve results that are unbiased to the choice of original gene expression dataset, three large, diverse datasets were selected and downloaded from GEO. In this study, non-related case-control experiments in humans from the *Affymetrix GeneChip Human Genome U133 Plus 2.0* microarray platform were used. The three datasets are from the study of 1) renal cell carcinoma tissue (77 controls and 77 cases, GSE53757) [8] 2) skin tissue in psoriasis patients (64 controls and 58 cases, GSE13355) [9] , and 3) gingival tissues (64 controls and 183 cases, GSE10334) [10]. The raw data were analyzed by first reading the CEL files into R using the *GEOquery v2.46.15* R package, and generating the normalized expression table using the *affy v1.56.0* package and *justRMA* normalization.

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.

B. Methodology

To investigate the effect of sample size on the result of gene set analysis, we need to conduct experiments with different sample sizes while keeping all remaining factors—such as phenotype under study, gene expression measurement platforms, the protocol for experimenting, laboratory technician skill level, and environmental condition—constant. Since finding multiple expression datasets for which these factors are constant across experiments is almost impossible, we utilize the following procedure to develop replicate datasets for which these confounding factors are invariable as much as possible.

Given an original case-control dataset with n_C control samples and n_T case samples and a given integer n, where n_C and n_T are relatively large numbers (> 50), $n < n_C$, and $n < n_T$, we develop a balanced case-control dataset from this original dataset by random sampling (without replacement) of n control samples (out of all n_C controls) and n case samples (out of the n_T samples). Hereafter, we refer to this process as the dataset generation procedure. Also, to avoid confusion, we refer to a dataset downloaded from GEO as the original dataset, and a balanced case-control dataset constructed by assembling n case and n control samples simply as a replicate dataset of size $2 \times n$.

To achieve results that are independent of a specific composition of samples, the dataset generation procedure can be repeated to create more replicate datasets each with n controls and n cases. In this research, for each sample size, we repeat the dataset generation process 10 times. The assembled datasets differ due to the nature of random sampling.

Since all the generated datasets are assembled from an original dataset, the confounding factors remain invariable as much as possible. For example, all these datasets have the same platform and protocol, and they have been generated by the same technician(s). This lets us study the effect of sample size on the result of different gene set analysis methods while keeping the confounding factors constant as much as possible.

To investigate the effect of sample size on the result of various gene set analysis methods, the above procedure is used to generate balanced case-control datasets with 3 to 20 samples per group ($3 \leq n \leq 20$). In each generated dataset, we have two groups, one for controls and one for cases, both of equal sample size. Different gene set analysis methods are then applied to these datasets to find the list of differentially enriched gene sets. Next, the results of each method across samples sizes are evaluated using statistical data analysis.

In this research, 10 commonly used gene set analysis methods are tested—PAGE [11], GAGE [12], Camera [13], ROAST [14], FRY (from limma R package) [15], GSEA [16], ssGSEA [17], GSVA [18], PLAGE [19], and over-representation analysis (ORA) [20]. For each method, all replicate datasets of size $2 \times n$ ($n \in \{3, \ldots, 20\}$) are used to conduct gene set analysis. The default parameters, as suggested by each method's authors, are used. The methods are obtained through the following R packages: GSVA, PLAGE, and ssGSEA are run from *GSVA* package version *1.18.0*; ORA is implemented using the *phyper* method from the *stats* package version *3.4.4*; GSEA is run using the *GSEA.1.0.R* script downloaded from the Broad Institute software page for GSEA (http://software.broadinstitute.org/gsea/downloads.jsp); Camera, ROAST, and FRY are obtained from the *limma* package version *3.34.9*; PAGE and GAGE are used from the *gage* package version *2.20.1*. To be consistent across gene set analysis methods, a Benjamini-Hochberg correction [21] for multiple comparisons with a false discovery rate of 0.05 is applied for all gene set analysis experiments. Also, a gene set analysis method uses a gene set database as an input. In this study, the GO gene sets—a subset of *MSigDB* version *6.1* [16]—is used. Hereafter, we refer to this database as $\mathbb{G}$.

For a given original dataset D, first, the dataset generation procedure is used to assemble m replicate datasets $D_1^{(2\times n)}$, ..., $D_m^{(2\times n)}$, each of size $2\times n$. For all experiments, we use 10 replicate datasets ($m = 10$). Next, a gene set analysis method ψ is applied to each $D_i^{(2\times n)}$ $(1 \leq i \leq m)$ and the result $R^{\psi}_{D_i^{(2\times n)}}$ is stored. $R^{\psi}_{D_i^{(2\times n)}}$ is a vector of adjusted p-values where the k^{th} element of this vector represents the

adjusted p-value for testing differential enrichment of the k^{th} gene set of $\mathbb{G}$. $R^{\psi}_{D_i^{(2\times n)}}$ is a vector with a length equal to the number of gene sets in $\mathbb{G}$.

After conducting gene set analysis and generating adjusted p-values for all gene sets in $\mathbb{G}$, a significance level $\alpha = 0.05$ is used to determine the differential enrichment status of all gene sets in $\mathbb{G}$. This is achieved by comparing each element of $R^{\psi}_{D_i^{(2\times n)}}$ against α: if the k^{th} element of $R^{\psi}_{D_i^{(2\times n)}}$ is less than α, the k^{th} gene set of $\mathbb{G}$ is considered as being differentially enriched, and non-differentially enriched otherwise. We define $S^{\psi}_{D_i^{(2\times n)}}$ to be the set of all gene sets predicted as being differentially enriched.

We use the Jaccard index [22] to quantify the reproducibility of the results of a gene set analysis method ψ when applied to two datasets $D_i^{(2\times n)}$ and $D_j^{(2\times n)}$. The value is referred to as an overlap score and defined as follows:

$$J(S_{D_i^{(2\times n)}}, S_{D_j^{(2\times n)}}) = \frac{S_{D_i^{(2\times n)}} \cap S_{D_j^{(2\times n)}}}{S_{D_i^{(2\times n)}} \cup S_{D_j^{(2\times n)}}} \quad (1)$$

A Jaccard index of 0 means no overlap, i.e. no agreement, between the results of gene set analysis and a value of 1 means complete overlap between $S^{\psi}_{D_i^{(2\times n)}}$ and $S^{\psi}_{D_j^{(2\times n)}}$. Hereafter, we refer to the Jaccard index between two sets as the overlap of those two sets.

To asses the reproducibility of a gene set analysis method when using replicate datasets of size $2 \times n$, for each pair of datasets $D_i^{(2\times n)}$ and $D_j^{(2\times n)}$ ($1 \leq i, j \leq m$ and $i \neq j$), we compute the overlap between $S^{\psi}_{D_i^{(2\times n)}}$ and $S^{\psi}_{D_j^{(2\times n)}}$ and place the resulting score in position (i, j) of an upper triangular matrix, called an overlap matrix, which is visualized in Section III. Since the overlap is a symmetric function, i.e. $J(S_{D_i^{(2\times n)}}, S_{D_j^{(2\times n)}}) = J(S_{D_j^{(2\times n)}}, S_{D_i^{(2\times n)}})$, for each sample size $2 \times n$ we have $\frac{m\times(m-1)}{2}$ overlap values. The distribution of these values tells us the extent to which an expression study using a sample size of $2 \times n$ is reproducible. If using a dataset with sample size $2 \times n$ leads to reproducible results, there should be a high overlap between each pair from $S^{\psi}_{D_1^{(2\times n)}}, \ldots, S^{\psi}_{D_m^{(2\times n)}}$. For each method ψ, we construct a multiset $P^{\psi}_{(2\times n)}$—which is a set but with repetition allowed—as follows:

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

After that, for each method ψ, we use the Kruskal-Wallis test to investigate if there is a statistically significant difference between these multisets of overlap scores ($P^{\psi}_{(2\times n)}$) across the different sample sizes ($3 \leq n \leq 20$).

When conducting a gene set analysis on a replicate dataset $D_i^{(2\times n)}$, differentially enriched gene sets can be sorted based on their adjusted p-value. Most researchers select the top gene sets (those with smallest adjusted p-values) for further study and interpretation. Therefore, not only is the consistency between differentially enriched gene sets important but also the order in which these gene sets are reported. We use Kendall's coefficient of concordance [22] to assess the agreement in the order of differentially enriched gene sets among the results of analyzing replicates of the same sample size. A gene set, in order to be considered in this calculation, needs to be predicted as differentially enriched for at least one replicate dataset. The Kendall coefficient of concordance ranges between 0 and 1, where 0 represents no agreement and 1 represents complete agreement.

Furthermore, it is important to determine if—for a method ψ—the results of gene set analysis of a replicate dataset $D_{2\times n}$ is consistent with the results when using a larger sample size, for example the whole dataset D. To do so, we conduct gene set analysis on the original dataset D using each method ψ and calculate S^{ψ}_{D}. Then we construct a multiset $W^{\psi}_{(2\times n)}$ of m overlap scores, as follows:

$$W^{\psi}_{(2\times n)} = \{J(S^{\psi}_{D_i^{(2\times n)}}, S^{\psi}_{D}) \mid 1 \leq i \leq m\} \quad (3)$$

High overlap scores between the results of gene set analysis of replicates of size $2 \times n$ and the whole dataset indicates that a sample size of $2 \times n$ might be sufficient for achieving the same result as a larger dataset.

III. Experimental Results

To visualize the overlap between the results of the gene set analysis of replicate datasets of the same size, we use a collection of modified heat maps to construct a plot, hereafter referred to as a pine plot. A pine plot is a stack of pyramids, where each pyramid—hereafter referred to as a layer—is a triangular heat map of values above the diagonal in a overlap matrix (as described in Section II). This visualizes the overlap between the results of the gene set analysis of replicate datasets of a specific size. More specifically, the colour intensity of cell (i, j) in each layer represents $J(S_{D_i^{(2\times n)}}, S_{D_j^{(2\times n)}})$, i.e. the overlap between $S_{D_i^{(2\times n)}}$ and $S_{D_j^{(2\times n)}}$. When $i = j$ the overlap score is 1. Although uninformative, we keep these cells as a visual reference point always in red in the baseline of each layer. The pine plot in Fig. 1 depicts the overlap score for replicate datasets of size 2×3, 2×5, 2×10, 2×15, and 2×20 analyzed by ORA.

A common pattern across all methods under study is that the overlap score increases as sample size increases. This pattern is also consistent across all three original datasets; therefore, all visualizations and plots in this paper are for one dataset (GSE53757). For instance in Fig. 1, moving from the base layer (sample size of 2×3) to the top layer (sample size of 2×20) shows a transition from blue colour gradients (low overlap scores) to red colour gradients (high overlap scores). However, the increments in overlap score are not the same across all methods. For example, as observed from Fig. 4, 6, and 5, ROAST shows a small amount of overlap between replicates at lower sample sizes while GAGE shows high overlap scores. Also, the overlap scores of replicate datasets for the same sample size are more variable when

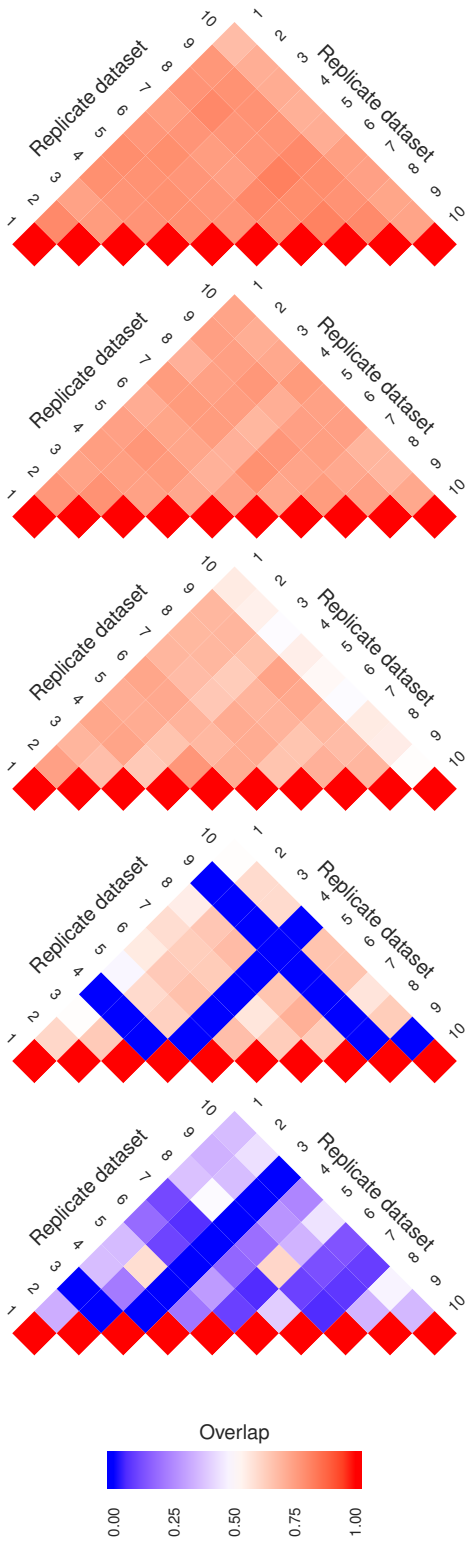

Fig. 1. A pine plot depicting the change in the reproducibility of the results from over-representation analysis (ORA) as sample sizes increases. The concept of reproducibility is quantified by overlap score (Equation 1). Each layer of the pine plot illustrates the overlap score of the results of ORA for pairs of 10 replicate datasets with the same sample size. The layers in the plot, from bottom to top, represent replicates with sample size 2×3, 2×5, 2×10, 2×15, and 2×20. The overlap score ranging from 0 to 1 is represented by gradients from blue to red, separated by white in the middle (overlap of 0.5). The pine plot suggests that the overlap between replicates is very small (low overlap is shown in blue) for a sample size of 3. This gradually improves with more overlap present in replicates with a higher number of samples.

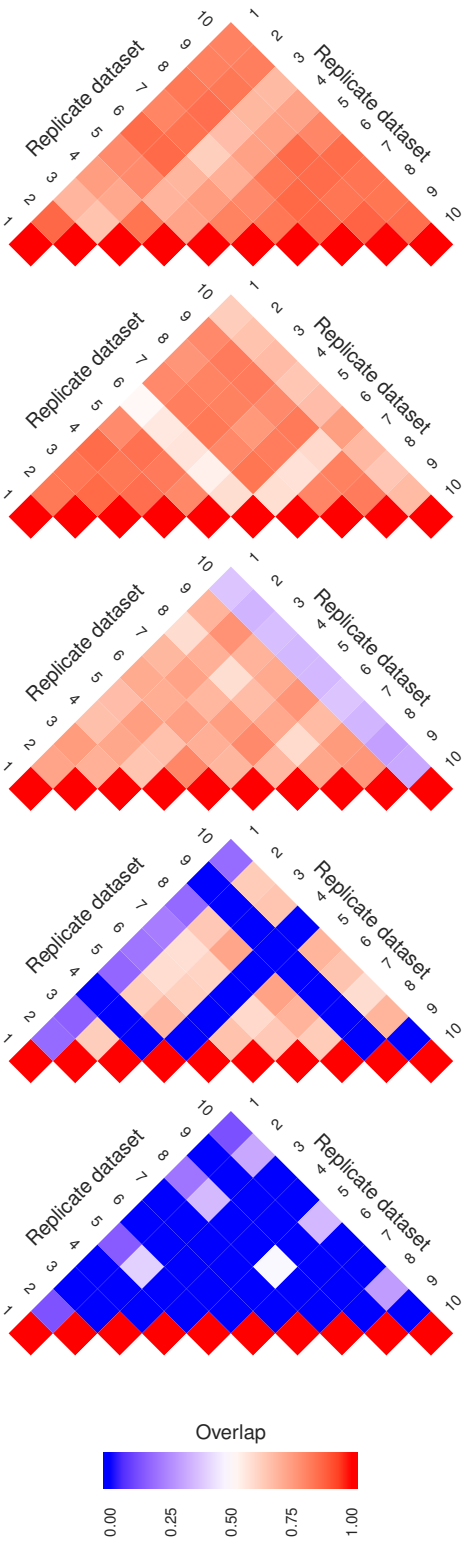

Fig. 2. A pine plot depicting the change in the reproducibility of the results from ROAST as sample size increases (see Fig. 1 caption for more information). The pine plot suggests that the overlap between replicates is very small (low overlap is shown in blue) for a sample size of 3. This gradually improves with much more overlap present in replicates with a higher number of samples.

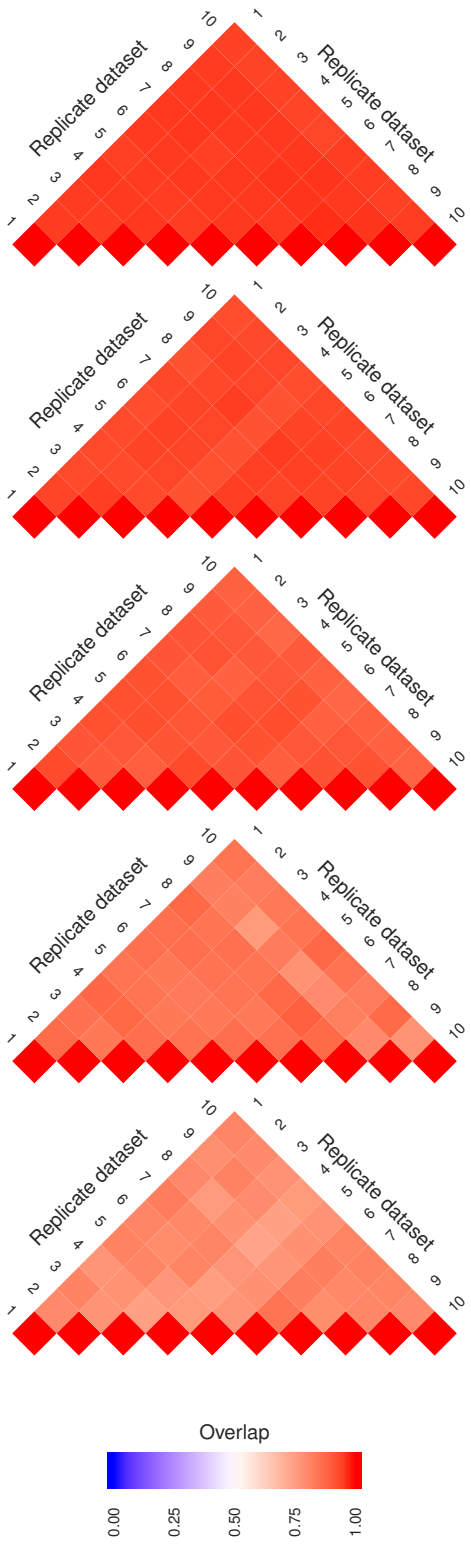

Fig. 3. A pine plot depicting the change in the reproducibility of the results from GAGE as sample size increases (see Fig. 1 caption for more information). The pine plot suggests that the overlap between replicates is larger in comparison to that of ORA and ROAST (see Fig. 1 and Fig. 2). GAGE has much more agreement between replicates using lower sample sizes such as 3, and the overlap scores continue to improve for higher numbers of samples.

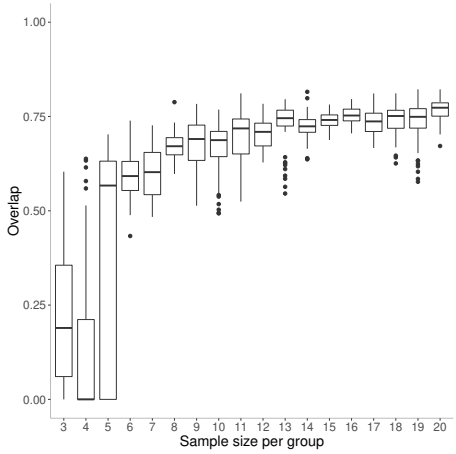

Fig. 4. A box plot showing the distribution of overlap scores resulting from the gene set analysis of replicate datasets with sample size $2 \times n$ ($3 \leq n \leq 20$)—where n is the sample size per group, i.e. n control and n case samples—using ORA. Each box shows the overlap scores resulting from gene set analysis of all pairs from 10 replicate datasets (all of the same sample size). Overlap score agreement is intermediate between ROAST and GAGE.

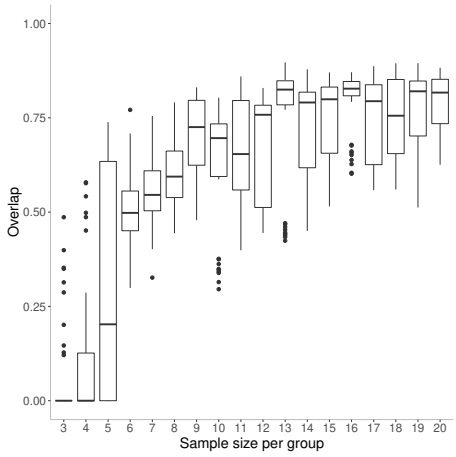

Fig. 5. A box plot showing the distribution of overlap scores resulting from the gene set analysis of replicate datasets with sample size $2 \times n$ ($3 \leq n \leq 20$) using ROAST (see Fig. 4 caption for more information). ROAST produces small overlap scores for small sample sizes. Also, the variation of overlap scores for each sample size is higher than that of GAGE and ORA.

comparing pine plots from ORA or ROAST to the pine plot for GAGE. To investigate if there is a statistically significant difference between the overlap scores across sample sizes, i.e. $P^{\psi}_{(2\times3)}, \ldots, P^{\psi}_{(2\times20)}$, we conduct a Kruskal-Wallis test. Table I shows the p-values resulting from the tests for all the gene set analysis methods under study. The results suggest that there is a significant difference between the overlap scores across sample sizes.

As illustrated by the pine plots in Fig. 1, 2, and 3 and also the box plots in Fig. 4, 5, and 6, there is a substantial increase in the overlap scores as sample size increases. Also, all methods lead to small overlap scores when using a small sample size such as 2×3 . A similar pattern is observed when comparing the overlap between replicates and the whole

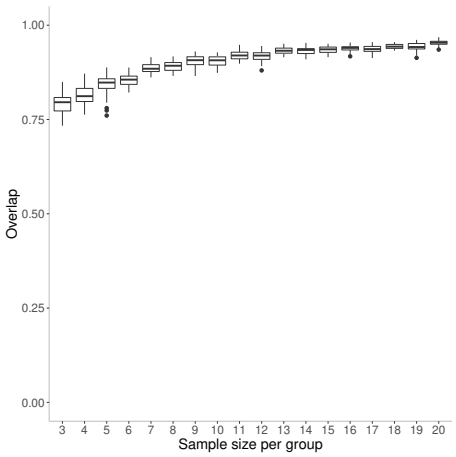

Fig. 6. A box plot showing the distribution of overlap scores resulting from the gene set analysis of replicate datasets with sample size $2\times n$ ($3 \leq n \leq 20$) using GAGE (see Fig. 4 caption for more information). There is an increase in the overlap as sample size increases. Also variation of overlap scores is smaller than that of ORA and ROAST.

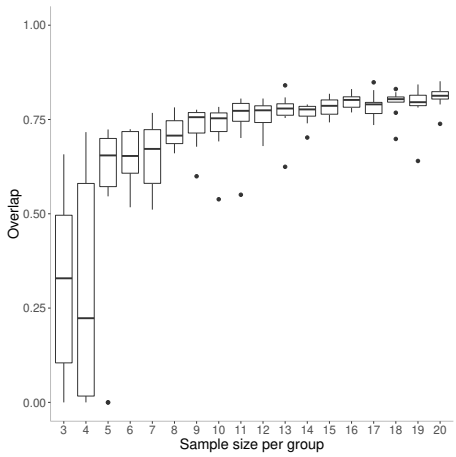

Fig. 7. A box plot showing the distribution of overlap scores resulting from the gene set analysis of each replicate dataset with that of the whole dataset using ORA. A box with the x-coordinate of n shows the overlap scores resulting from the gene set analysis of each of the 10 replicate datasets (all with sample size $2 \times n$) and the result of gene set analysis of the whole dataset. ORA achieves low overlap scores for small samples sizes. Also the variability of overlap scores is higher in comparison to GAGE.

datasets (see Fig. 7, 8, and 9).

Fig. 10 illustrates Kendall's concordance coefficients for replicate datasets across sample sizes. As a common trend, the concordance coefficient increases as sample size increases.

Fig. 11 depicts the average number of gene sets predicted as being differentially enriched in replicate datasets of a given sample size. The number of gene sets predicted as being differentially enriched for GAGE, GSVA, ROAST, and FRY increases as sample size increases, while the rest of the methods show an almost constant number of gene sets predicted as being differentially enriched.

IV. Discussion

In this paper, we proposed a quantitative approach to systematically assess the reproducibility of gene set analysis methods using real expression datasets. Furthermore, we suggested an overlap score to quantify the concept of reproducibility in the context of gene set analysis. Also, we described and used pine plots to visualize the overlap between the results of replicate datasets of the same size. However and more generally, pine plots can be used for visualizing the interaction of several variables while controlling for one or more confounding factors. The only limitation is that the interaction between variables must be definable using a symmetric function. Although this might sound like a limitation on the usability of pine plots as a general-purpose data visualization tool, in practice most scores for measuring the interaction of different variables are symmetric—for example, Pearson correlation and Spearman's rank correlation coefficients. Also, any well-defined metric or distance function [23] can be used with pine plots too. The pine plots in this paper clarify how reproducibility increases when sample size increases. Unlike the box plots, the pine plots clearly illustrate the extent of the overlap between each pair of replicate datasets.

It should be mentioned that although the reproducibility of gene set analysis across replicate datasets is a necessary

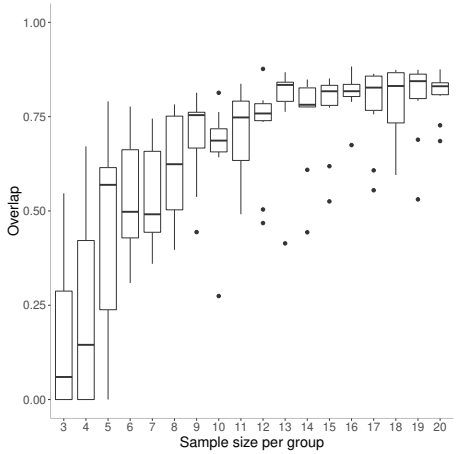

Fig. 8. A box plot showing the distribution of overlap scores resulting from the gene set analysis of each replicate dataset with that of the whole dataset using ROAST (see Fig. 7 caption for more information). ROAST achieves very low overlap scores for small samples sizes with high variability. Also the variability across replicates is higher when compared to GAGE and ORA.

TABLE I
The result of Kruskal-Wallis tests

	Datasets from GEO		
Method	GSE53757	GSE13355	GSE10334
FRY	6.28e-13	2.99e-26	2.60e-13
ORA	5.03e-18	2.39e-19	1.67e-14
PAGE	1.81e-16	5.26e-20	1.50e-10
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
GSEA	1.07e-11	1.45e-15	4.72e-05
ssGSEA	4.71e-25	8.70e-26	1.87e-26
Camera	1.81e-19	2.11e-20	9.74e-05
GSVA	5.30e-20	7.17e-27	4.21e-07

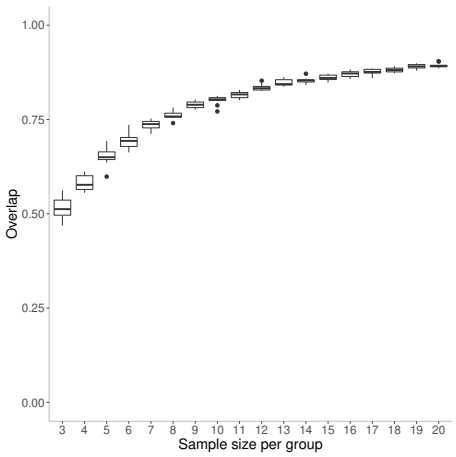

Fig. 9. A box plot showing the distribution of overlap scores resulting from the gene set analysis of each replicate dataset with that of the whole dataset using GAGE (see Fig. 7 caption for more information). GAGE achieves higher overlap scores for small samples sizes compared to ORA and ROAST. Also the variability of overlap scores is lower for GAGE.

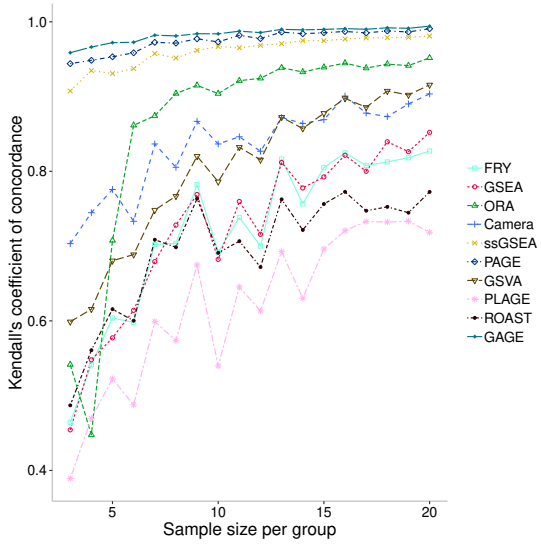

Fig. 10. Kendall's coefficient of concordance for each method under study. The x-axis shows the sample size. The y-axis shows concordance coefficients of the results of gene set analysis of 10 replicate datasets of the same size.

condition for achieving biologically valid results, it is not sufficient. As a hypothetical example, assume a method always reports all gene sets as differentially enriched. Such a method is of no value due to its large number of false positives, but it achieves a maximum overlap score of 1. Therefore, we suggest the study of the sensitivity and specificity together with the study of overlap between gene set analysis methods for future research. This would help alleviate the challenges regarding the lack of gold standard datasets for evaluating gene set analysis methods as it would provide a means to compare methods regardless of the dataset(s) being used for evaluation.

In the rest of this section, we discuss the observations about the number of differentially enriched gene sets reported by the methods under study (Fig. 11), and the Kendall concordance

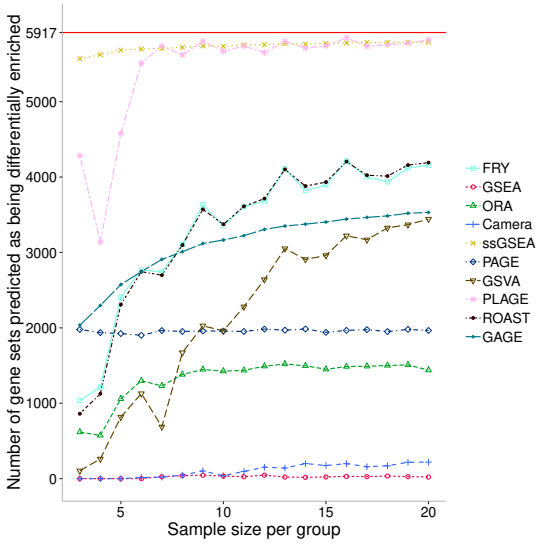

Fig. 11. The number of gene sets predicted as differentially enriched for each method under study. The x-axis shows the sample size. The y-axis shows the average number of gene sets predicted as being differentially enriched across 10 replicate datasets of the same size. The red line parallel to the 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.

coefficients for each method (Fig. 10).

PLAGE reports almost all of the gene sets as differentially enriched regardless of the sample size used. This explains why PLAGE achieves high overlap scores. Since it is unlikely that such a large number of gene sets are all differentially enriched in a living organism, we assume that this method leads to a large number of false positives, i.e. gene sets incorrectly predicted as being differentially enriched. Also, PLAGE's low Kendall concordance coefficients depicted in Fig. 10 shows that the order in which it reports the differentially enriched gene sets is not conserved across replicate datasets. This could be explained by the fact that PLAGE, for each gene set, defines "the activity level in terms of the first eigenvector, 'metagene', in the singular value decomposition" [19]. By ignoring other eigenvectors of an expression profile for a gene set, it cannot entirely capture the variability of expression of genes within a gene set. This causes the gene sets predicted as being differentially enriched by PLAGE to show variation in statistical significance across replicate datasets, and therefore, be ranked differently for each replicate dataset.

ssGSEA also reports almost all gene sets as being differentially enriched for each replicate dataset, regardless of the number of samples being used for gene set analysis. This means that relying on gene sets predicted as being differentially enriched by 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 the most statistically significant gene sets predicted by these methods.

Camera reported a small number of differentially enriched gene sets (average of 108 in Fig. 11) regardless of the sample size used for gene set analysis. Therefore, a small difference in the set of gene sets predicted as being differentially enriched strongly affects the overlap, leading to small values.

GSEA, using sample permutation, produces no enriched gene sets using lower sample sizes. This behaviour was expected since the permutation method for significance assessment requires large sample sizes. For example, in a replicate dataset with 3 controls and 3 cases, there are 20 distinct permutations of controls and cases—the combination of 3 out of 6. In this case, the smallest non-zero p-value is 0.05, which is not considered significant. However, for a sample size of 2×10 or greater, the number of differentially enriched gene sets (average of 15 in Fig. 11) remains steady while Kendall's concordance increases. This suggests that 10 samples per group might be a reasonable lower bound for using GSEA.

For sample sizes larger than 6, ORA remains quite consistent in the number of enriched gene sets reported, although the method appears to be less conservative compared to GSEA.

Since all replicates of size $2 \times n$ are generated from the same original dataset, for each method we expect the number of gene sets predicted as being differentially enriched to remain approximately the same across sample sizes. However, this is not the case with GAGE, GSVA, FRY, and ROAST. For these methods, the number of gene sets predicted as being differentially enriched dramatically increases with the increase in sample size. This increase may partially be responsible for the increase in the overlap scores for these methods as the number of samples increases. Also, FRY and ROAST closely mirror each other in the number of gene sets predicted as being differentially enriched. This is expected as FRY has been proposed to be a fast approximation of ROAST. Also, since it is unlikely to have such a large number of gene sets as being differentially enriched, we assume that these methods may lead to more false positives as sample size increases.

V. Conclusion

This research lays out a systematic methodology for evaluating the reproducibility of gene set analysis methods using quantitative measures. The proposed methodology not only allows for evaluation of the reproducibility of a gene set analysis method across sample sizes but also can be extended to compare the result of different gene set analysis methods for a given dataset. We used this methodology to evaluate the reproducibility of 10 gene set analysis methods across real gene expression datasets. As a general pattern, we observed that overlap score increases with increase in sample size. However, the rate of increase in overlap score is not the same across all methods. We also conjectured that for methods such as GAGE, GSVA, ROAST, and FRY an increase in sample size may lead to an increase in the number of false positives. Also, our findings suggest that for all methods under study achieving reproducible results using small sample sizes—such as 3, 4, or 5 samples per group—is unlikely.

References

[1] C. A. Tsai, S. J. Wang, D. T. Chen, and J. J. Chen, "Sample size for gene expression microarray experiments," *Bioinformatics*, vol. 21, no. 8, pp. 1502–1508, 2004.

[2] C. Stretch, S. Khan, N. Asgarian, R. Eisner, S. Vaisipour, S. Damaraju, K. Graham, O. F. Bathe, H. Steed, R. Greiner *et al.*, "Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature," *PloS One*, vol. 8, no. 6, p. e65380, 2013.

[3] N. J. Schurch, P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes *et al.*, "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" *RNA*, vol. 22, no. 6, pp. 839–851, 2016.

[4] A. L. Tarca, G. Bhatti, and R. Romero, "A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity," *PloS One*, vol. 8, no. 11, p. e79217, 2013.

[5] H. Alavi-Majd, S. Khodakarim, F. Zayeri, M. Rezaei-Tavirani, S. M. Tabatabaei, and M. Heydarpour-Meymeh, "Assessment of gene set analysis methods based on microarray data," *Gene*, vol. 534, no. 2, pp. 383–389, 2014.

[6] R. Mathur, D. Rotroff, J. Ma, A. Shojaie, and A. Motsinger-Reif, "Gene set analysis methods: a systematic comparison," *BioData Mining*, vol. 11, no. 1, p. 8, 2018.

[7] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[8] C. A. Von Roemeling, D. C. Radisky, L. A. Marlow, S. J. Cooper, S. K. Grebe, P. Z. Anastasiadis, H. W. Tun, and J. A. Copland, "Neuronal pentraxin 2 supports clear cell renal cell carcinoma by activating the ampa-selective glutamate receptor-4," *Cancer Research*, vol. 74, no. 17, pp. 4796–4810, 2014.

[9] W. R. Swindell, A. Johnston, S. Carbajal, G. Han, C. Wohn, J. Lu, X. Xing, R. P. Nair, J. J. Voorhees, J. T. Elder *et al.*, "Genome-wide expression profiling of five mouse models identifies similarities and differences with human psoriasis," *PloS One*, vol. 6, no. 4, p. e18266, 2011.

[10] R. T. Demmer, J. H. Behle, D. L. Wolf, M. Handfield, M. Kebschull, R. Celenti, P. Pavlidis, and P. N. Papapanou, "Transcriptomes in healthy and diseased gingival tissues," *Journal of Periodontology*, vol. 79, no. 11, pp. 2112–2124, 2008.

[11] S.-Y. Kim and D. J. Volsky, "PAGE: parametric analysis of gene set enrichment," *BMC Bioinformatics*, vol. 6, no. 1, p. 144, 2005.

[12] W. Luo, M. S. Friedman, K. Shedden, K. D. Hankenson, and P. J. Woolf, "GAGE: generally applicable gene set enrichment for pathway analysis," *BMC Bioinformatics*, vol. 10, no. 1, p. 161, 2009.

[13] D. Wu and G. K. Smyth, "Camera: a competitive gene set test accounting for inter-gene correlation," *Nucleic Acids Research*, vol. 40, no. 17, pp. e133–e133, 2012.

[14] D. Wu, E. Lim, F. Vaillant, M.-L. Asselin-Labat, J. E. Visvader, and G. K. Smyth, "ROAST: rotation gene set tests for complex microarray experiments," *Bioinformatics*, vol. 26, no. 17, pp. 2176–2182, 2010.

[15] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, pp. e47–e47, 2015.

[16] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *PNAS*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[17] D. A. Barbie, P. Tamayo, J. S. Boehm, S. Y. Kim, S. E. Moody, I. F. Dunn, A. C. Schinzel, P. Sandy, E. Meylan, C. Scholl *et al.*, "Systematic RNA interference reveals that oncogenic kras-driven cancers require tbk1," *Nature*, vol. 462, no. 7269, p. 108, 2009.

[18] S. Hänzelmann, R. Castelo, and J. Guinney, "GSVA: gene set variation analysis for microarray and RNA-seq data," *BMC Bioinformatics*, vol. 14, no. 1, p. 7, 2013.

[19] J. Tomfohr, J. Lu, and T. B. Kepler, "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics*, vol. 6, no. 1, p. 225, 2005.

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

[21] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 289–300, 1995.
[22] G. J. Bakus, *Quantitative analysis of marine biological communities: field biology and environment*. John Wiley & Sons, 2007.
[23] N. Loehr, *Advanced linear algebra*. Chapman and Hall/CRC, 2014.