Gene Set Enrichment Analysis: A powerful tool for gene expression interpretation in genome-wide studies

Alnakeb Derar[1], Papadopoulou Theodora[1]

[1]ULB

Abstract

Genome-wide expression analysis is a prevalent method in biomedical research. However, the traditional single-gene analysis methods often miss the broader biological context inside organisms. Here, we present Gene Set Enrichment Analysis (GSEA) allowing to interpret gene expression data based on the biological classification of genes. The benefit of this approach is that it considers all genes in an experiment and focuses on their accumulative changes inside predefined gene sets, rather than focusing on individual genes. We also demonstrate how GSEA extracts meaningful information from complex datasets regarding the biological alterations across different states or experimental conditions.

Introduction

The rapid development of transcriptomics technologies led to an exponential increase in the volume of data produced, which facilitated gene expression profiling under various experimental conditions [1], [2]. However, comparing thousands of differentially expressed genes between those conditions is not biologically interpretable. Diseases, typically, evolve due to perturbations that occur inside biological pathways rather than isolated gene alterations [3]. Thus, focusing on the additive gene expression changes within predefined gene sets clustered together based on biological knowledge can shed light on biological mechanisms underlying phenotypic disparities. A common experiment involves the comparison of mRNA expression levels of several genes among samples belonging to different groups, such as control versus treated cells. These genes are then ranked based on specific metrics, such as statistical significance (p-value) and biological relevance (Fold change) between the sample groups. Focusing on the individual genes in the extremes of the list poses issues during the interpretation of the output, which strongly depends on the biologist's expertise. Usually genes act cooperatively, forming sets of genes involved in cellular processes. Thus, analyzing genes individually relies on predefined thresholds and might overlook crucial effects on biological pathways [4]. Especially in microarray data, modest biological differences between conditions might be eclipsed by the technical noise, challenging the detection of statistically significant changes in individual genes. Gene Set Enrichment Analysis addresses those constraints by focusing on sets of genes, sharing biological function, chromosomal location, regulation, or co-expression, as indicated by biological knowledge. We demonstrate GSEA's ability to analyze gene expression data from male vs female lymphoblastoid cells (unpublished data), showing, as biologically expected, enrichment of pathways related to Y chromosome for this comparison.

Methods

Having a predefined set of genes S representing a pathway, and a ranked list of genes L. GSEA aims to identify whether the genes from the set S are randomly distributed throughout the expression dataset L or if they tend to appear in the extremities of L.

In addition to genes having intense expression changes, GSEA also accounts for tenuous but cumulative expression changes among the genes inside each set, reflecting the coordinated regulation of functionally related genes.

Inputs

GSEA receives as input:

1. An unranked expression dataset D, including the level of expression for all genes across all samples.
2. The gene set S (derived from the Molecular Signature Database, MSigDB, which is freely available)

representing the pathway of interest to be studied.

3. A class of distinction C representing the index of the phenotype of interest in alphabetical order.

4. An exponent p, which is the weight of the step in the running sum. Note that with $p = 0$ we would be implementing a Kolmogorov–Smirnov statistic; we choose $p = 1$ so that every gene is weighted by its fold change value.

Ranking

We start by ranking the expression dataset D based on a statistic, such as fold change or p-value. Given a class of distinction C, we compute the mean level of expression for each gene among samples of one phenotype. We then calculate the fold change of gene expression between the two phenotypes and we, subsequently, rank the genes according to their fold change in decreasing order. This results in the ranked gene list L.

Enrichment Score

Given a ranked expression dataset L, a predefined gene set S, and an exponent p, the algorithm runs through L and computes a score by shifting through the ranked list L=g_1,...,g_N, adding a running-sum statistic every time a gene belongs to set S and decreasing when it does not.

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$
$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}. \qquad [1]$$

This involves computing the fraction of genes present in S weighted by their fold change up to the position i in L (P_{hits}) minus the fraction of the genes not present in S (P_{Misses}). The highest deviation from 0 during the running sum is the Enrichment Score (ES) [1]. This can be visualized with a plot of the running-sum score for each set of genes ($Figure1$).

A set S in which most of the genes appear towards the top or the bottom of the ranked list L will have a relatively high $ES(S)$, whereas a randomly distributed S will have a low $ES(S)$.

Statistical Significance of ES

To assess the significance of the ES, we compute the P-value using the permutation method: we randomly reassign the phenotype tags to samples, generate a new ranked list, and re-calculate the ES of the gene set. This process is repeated 1,000 times to create an empirical null distribution of all the ESs, denoted as ES_{NULL}. It is important to note that the permutations of phenotype tags do not affect the gene-to-gene correlation. Subsequently, the P-value assesses the statistical significance of the set based on the distribution of the ESs generated by the permutations. For an ES higher or equal to zero P-value corresponds to the number of ESs in the null distribution that are greater than the observed ES, divided by the number of ESs in the null distribution being higher or equal to zero. The P-value of an ES lower than zero is computed reversely.

Multiple hypothesis testing

When computing the ES for multiple pathways, the p-value is adjusted using the Benjamini-Hochberg (BH) procedure with a threshold of < 0.2 to reject the null hypothesis. The BH procedure controls the false discovery rate (FDR), which is the expected proportion of false positives among the rejected hypotheses:

1. Sort the p-values in ascending order:

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$$

2. Compute the critical value for each p-value:

$$\frac{i}{m}\alpha$$

where i is the rank of the p-value in the sorted list, m is the total number of tests, and α is the significance level.

3. Reject all p-values that do not meet the criteria, otherwise reject the null hypothesis:

$$p_{(i)} \le \frac{i}{m}\alpha$$

Leading edge subset

To understand which genes play an important role to the ES of each set it is possible to extract those members of the set that are located before and at the position

of the ranked list L corresponding to the highest deviation from zero for the running-sum statistic. In other words, the leading edge is the core of the gene set that represents the enrichment score. Being aware of the leading edge could shed light to the biological alterations inside a pathway that can explain the different phenotypes.

Results

Here, we demonstrate GSEA's ability to extract pertinent biological information from complex datasets. We used the freely available resources from the Broad Institute, and especially the MSigDB database, containing the sub catalogs C1 and C2 [5].

Male vs Female Lymphoblastoid cells

We exploited mRNA expression profiles (12903 genes) of 15 males and 17 females lymphoblastoid-derived cell data (not published) and tested whether the tool was able to distinguish the sex of the individuals based on their gene expression data. We utilized C_1 sub catalog from the $MSigDB$, classifying genes based on their chromosomal topology in cytogenic bands. We first compared males to females and expected to find the $chrYp11$ and $chrYq11$ pathways enriched, since they contain genes located on the Y chromosome. Indeed, those pathways found to be enriched, as demonstrated by the high ES shown in $Figure1$. $Figure2$ lists the $P-values$ of the analyzed pathways. We then compared the female vs male, and reversely, gene expression changes for genes located in the bands of chromosome X, specifically for pathways $chrXp11$, $chrXq12$. We did not find enrichment in those pathways. This is biologically interpretable, since genes located in X chromosomes compensate their expression and thus they are not lower expressed in males [6].

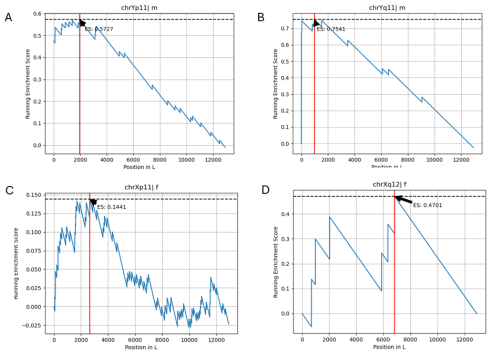

Figure 1: Enrichment Score. The distribution of the running sum statistic (y axis) for the ranked list of genes (x axis) derived from the lymphoblastoid cell line gene expression levels. (A) male vs female comparison of gene set "chrYp11" showed high enrichment, ES=0.5727, (B) male vs female comparison of "chrYq11" gene set, also showed an ES=0.7541, (C) female vs male comparison of chrXp11 had an ES=0.1441 and (D) female vs male comparison of "chrXq12" demonstrated ES=0.412. The genes ranked by their Fold Change of expression. ES corresponds to the maximum deviation from zero of the running sum statistic. The gene sets derived from the C1 cytogenic gene collection. ES=Enrichment Score.

Gene Set	P-value
chrYp11, male vs female	<0.01
chrYq11, male vs female	<0.01
chrXp11, female vs male	>0.1
chrXq12, female vs male	>0.1

Figure 2: P-values for the comparisons stated in Figure 1.

Subsequently, we compared the ESs among different gene sets and visualized them with a plot, as provided in $Figure3$. In this plot gene sets of one phenotype are ranked based on their ES and colored according to their $P-value$. This allows to extract the gene sets that have been affected the most between two phenotypes.

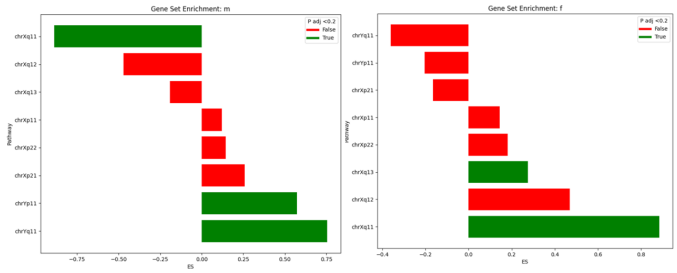

Figure 3: Comparison of ESs among gene sets of each phenotype colored by their $P - values$. The left plot visualizes the ranking of ESs between gene sets after comparing male vs female gene expression changes. Reversely, the right plot draws the comparison of gene set ESs after contrasting female vs male gene expression differences.

Finally, we choose the top-5 leading edge genes for every pathway, those genes represent the most enriched genes in the following pathways in male VS female comparison (over-expressed and uner-expressed):

Pathway	Leading Edge Genes
chrYp11	{AMELY, ASMT, ASMTL, CD99, CRLF2}
chrXp21	{CYBB, DMD, GK, IL1RAPL1, MAGEB1}
chrYq11	{BPY2, CDY1, DAZ2, DDX3Y, EIF1AY}
chrXq13	{ABCB7, ARR3, CDX4, CITED1, CXCR3}
chrXp11	{AKAP4, ALAS2, APEX2, ATP6AP2, BCOR}
chrXp22	{ACE2, AMELX, AP1S2, ARHGAP6, ARSD}
chrXq11	{ARHGEF9, MTMR8}
chrXq12	{AR, EDA2R, HEPH, MSN, OPHN1}

Figure 4: Leading edges for every analyzed pathway (male and female)

Discussion

Gene Set Enrichment Analysis is a useful method for extracting biologically meaningful inferences based on gene expression data. It shifts the focus of research from individual gene changes to coordinated alterations in biologically relevant sets of genes. GSEA facilitates the interpretation of differentially expressed genes by focusing on gene sets, revealing otherwise hidden biological pathways and mechanisms. It also enhances the reproducibility of the analysis, especially when manipulating poorly annotated genes. This approach also enables the detection of subtle and accumulative changes in gene expression, that would otherwise be overlooked, by increasing the signal-to-noise ratio of highly correlated genes. Additionally, the leading-edge aspect can help pinpoint the subset of genes primarily driving the enrichment signal implicated in phenotypic differences.

GSEA differs in two aspects compared to several existing tools based on ontology or pathway information [7]-[8]-[9]. While other tools rely only on overlap statistics, i.e. cumulative hypergeometric distribution, for defining whether a list of differentially expressed genes is over-represented inside a pathway, GSEA takes into consideration the whole list of genes, regardless of their individual significance levels. Moreover, by exploiting permutation testing, GSEA evaluates significance respecting gene to gene correlations, leading to a more precise null distribution. GSEA offers the possibility to be applied on different pathway databases, even manually created. With a vast repository of gene sets across different biological pathways, functionally or topologically related genes, this tool enables the spherical comprehension of different experimental conditions. It also allows us to alter the existing gene sets. Overall, GSEA enables the connection of prior knowledge with new data shedding light on the behavior of genes across different experimental conditions.

References

[1] Lowe, R., Shirley, N., Bleackley, M., Dolan, S. Shafee, T. “Transcriptomics technologies, PLoS Comput Biol 13”. In: (2017).

[2] Lockhart, D. J. et al. “Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14”. In: (1996), pp. 1675–1680.

[3] Coleman, W. B. Tsongalis, G. J. “Molecular Pathology: The Molecular Basis of Human Disease. Molecular Pathology: The Molecular Basis of Human Disease”. In: (2017), pp. 1–802.

[4] J. et al. Rahnenführer. “Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. BMC Medicine”. In: (2023), pp. 1–54.

[5] *GSEA — MSigDB — Human MSigDB Collections.* URL: https://www.gsea-

msigdb . org / gsea / msigdb / human / collections.jsp#C1.

[6] Prothero, K. E., Stahl, J. M. Carrel, L. “Dosage compensation and gene expression on the mammalian X chromosome: one plus one does not always equal two.” In: (2009), p. 637.

[7] Doniger, S. et al. “MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.” In: (2003).

[8] Zhong, S. et al. “GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space.” In: (2004).

[9] Berriz, G. F., King, O. D., Bryant, B., Sander, C. Roth, F. P. “Characterizing gene sets with FuncAssociate.” In: (2003).