Bioinformatics Advance Access published July 20, 2007

GSEA-P: A desktop application for Gene Set Enrichment Analysis

Aravind Subramanian, Heidi Kuehn, Joshua Gould, Pablo Tamayo, Jill P. Mesirov*

Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Associate Editor: Dr. Olga Troyanskaya

ABSTRACT

Gene Set Enrichment Analysis (GSEA) is a computational method that assesses whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states. We report the availability of a new version of the Java based software (*GSEA-P 2.0)* that represents a major improvement on the previous release through the addition of a leading edge analysis component, seamless integration with the Molecular Signature Database (MSigDB) and an embedded browser that allows users to search for gene sets and map them to a variety of microarray platform formats. This functionality makes it possible for users to directly import gene sets from MSigDB for analysis with GSEA. We have also improved the visualizations in *GSEA-P 2.0* and added links to a new form of concise gene set annotations called Gene Set Cards. These additions, as well as other improvements suggested by over 3500 users who have downloaded the software over the past year have been incorporated into this new release of the *GSEA-P* Java desktop program.

Availability: *GSEA-P 2.0* is freely available for academic and commercial users and can be downloaded from http://www.broad.mit.edu/GSEA.

Contact: gsea@broad.mit.edu

1 INTRODUCTION

The selection of differentially expressed genes helps associate biological phenotypes with their underlying molecular mechanisms thereby providing insights into biological function. However, analyzing and interpreting a given list of genes can be challenging due to the difficulty of objectively evaluating members of a given pathway or functional class represented in a gene list. Additionally, single gene marker based approaches can fail to detect transcriptional programs that are distributed across an entire network of genes yet are subtle at the level of individual genes.

To address this problem we previously introduced a statistical methodology called Gene Set Enrichment Analysis (GSEA) for determining whether a given gene set is significantly enriched in a list of gene markers ranked by their correlation with a phenotype of interest (Mootha et al., 2003), (Subramanian & Tamayo et al., 2005). The method has been successfully used to discover metabolic pathways altered in human diabetes (Mootha et al., 2003), compare expression profiles of mouse and humans (Sweet-Cordero et al., 2005), reveal more consistency between independent lung cancer outcome datasets at the gene set level than at the single gene level (Subramanian & Tamayo et al., 2005), and characterize molecular phenotypes in acute megakaryoblastic leukemia (Bourquin et al., 2006), amongst many other applications.

Given a list of genes, ranked by the correlation of their genome-wide expression profiles with one of two phenotypes, GSEA seeks to estimate the significance of the over-representation of an independently defined set of genes, S, in the highly correlated or anti-correlated genes in the list. To evaluate this degree of "enrichment" the GSEA method calculates an Enrichment Score (ES) by walking down the list, increasing a cumulative sum when a gene is in S and decreasing it if a gene is not in S. The size of the increment depends on the gene-phenotype correlation. The ES is the maximum deviation from zero of the cumulative sum and can be interpreted as a weighted Kolmogorov-Smirnov statistic. The genes in the gene set S that appear in the ranked list before the point where the running sum achieves the ES are called the leading-edge subset and are particularly important in evaluating the results of GSEA analysis. The significance of a gene set's ES is estimated by an empirical phenotype-based permutation test procedure. When an entire database of gene sets is scored, an adjustment must be made to the resulting p-values to account for multiple hypotheses testing. GSEA normalizes the ES for each gene set to account for the variation in set sizes, yielding a normalized enrichment score (NES), and calculates a false discovery rate (FDR) corresponding to each NES. The FDR gives an estimate of the probability that a set with a given NES represents a false positive finding; it is computed by comparing the tails of the observed and permutation-computed null distributions for the NES.

The power of GSEA depends on how well the gene sets used to assess enrichment represent meaningful coordinated gene expression behavior that reflects actual biological processes. The more accurately gene sets represent specific transcriptional processes relevant for a particular cellular state the better they will perform as GSEA queries. For this reason, the definition and curation of gene sets is of paramount importance. We have begun a process of systematically collecting gene sets into a Molecular Signature Database (MSigDB).

The MSigDB contains over 3000 gene sets of different types: (i) sets representing genes in the same chromosome or cytogenetic band, (ii) gene sets representing metabolic and signaling pathways from eight publicly available, manually curated pathway databases, (iii) genes reported in the literature as coexpressed in response to genetic or chemical perturbations, (iv) genes sharing conserved upstream regulatory motifs, and (v) sets of genes in expression neighborhoods of cancer-related genes. Users may use this resource or define their own gene sets relevant to the process or phenotype they are investigating.

Version 1.0 of the *GSEA-P* software and MSigDB were originally released in Spring of 2005. There are currently over 3500 registered users. The new version 2.0 of both the software and the database represent a substantial enhancement of the features, interface, and content, which we describe below.

2 FEATURES

Version 2.0 of the *GSEA-P* Java desktop software contains a complete implementation of the GSEA methodology, including

[*] To whom correspondence should be addressed.

leading edge analysis, as well as several usability improvements based on user feedback. New features include a gene set browser to search, download and map gene sets from the MSigDB database. We also have developed a website with comprehensive software documentation and Gene Set Cards with annotations including the source and biological relevance of MSigDB gene sets. A complete list of the new and improved features is in Supplemental Table 1.

Enrichment analysis: In enrichment analysis, a user seeks to determine whether the members of a gene set are over-represented at the top (or bottom) of a ranked list of markers which have been ordered by their correlation with a specified phenotype. This functionality is central to the *GSEA-P 2.0* software and is accessed via the 'Run GSEA page'. Users select a dataset, phenotype, and a gene set collection and set parameters to run an enrichment analysis. We have improved this interface by enabling conversion of the dataset and gene sets to the same identifier format (i.e. gene symbols) before running the analysis (see 'Chip2Chip' description below and Supplementary Figure 1). To address the need for alternative or specialized gene ranking procedures, we now provide a vehicle within the *GSEA-P* software for use with a user-provided ranked gene list. Importantly, in this new release, enrichment results are saved to an XML formatted local database and hence are available for downstream analysis with other GSEA components (see 'Leading edge analysis' below) and integration with other software programs.

Enrichment reports: *GSEA-P 2.0* produces richly annotated HTML reports of enrichment results. In addition to statistical details such as the enrichment score, p-value and FDR, we now provide a link to gene set annotations at the MSigDB website. These annotations allow users to view the full details of the provenance and content of a gene set in a structure similar to that of the GeneCards resource (Rebhan, 1997). The GSEA report also contains improved enrichment plots (Supplementary Figure 2).

Leading edge analysis**:** After an enrichment analysis has been performed, it is often useful to examine and compare the genes in high scoring sets which occur before the maximum of the running enrichment score. These genes can be thought of as the core of a gene set that drives the enrichment signal. By grouping leading edge subsets, high scoring gene sets can often be categorized into similar and distinct biological processes.

To facilitate leading edge analysis, *GSEA-P 2.0* provides an interactive viewer that can be run after a GSEA process completes. The user selects gene sets for leading edge analysis after which the program: (1) computes the core matrix over all selected gene sets, (2) clusters this matrix, and (3) visualizes the result in a heat map (Supplementary Figure 3A). Additionally, similarities between gene sets can be visualized by the Jacquard coefficient (Supplementary Figure 3B).

Batch analysis mode**:** To support the analysis of a large number of datasets or the integration of GSEA into a data analysis pipeline, *GSEA-P 2.0* can run in "headless" mode as part of a shell script or load sharing facility. The analysis performed and the reports produced in this mode are identical to those produced with the graphical user interface.

Mapping identifiers between platforms with Chip2Chip: Microarray platforms come from a number of manufacturers who use a variety of identifiers to represent gene transcripts. Additionally, cross-species comparisons require ortholog mappings. Several tools such as NetAffx (Liu et al., 2003) provide the ability to map a given list of genes between platforms. However, these programs are often restricted to a particular vendor or are cumbersome to use when mapping a large collection of gene sets as they are tailored to map a single input list. To address this need, the *GSEA-P 2.0* software provides a new utility called Chip2Chip that maps identifiers between platforms. Currently, *GSEA-P 2.0* supports mappings between 93 platforms. Chip2Chip can convert between Entrez gene symbols and any of these platforms or between identifiers for any two of these chip types (Supplementary Figure 4).

Integrated gene set browser & query interface: To enable users of *GSEA-P 2.0* to easily access the substantially enlarged MSigDB 2.0 collection we have embedded a gene set browser into the software which enables users to quickly search MSigDB for gene sets using an intuitive graphical user interface (Supplementary Figure 5). By providing an integrated program, we enable the seamless interoperation of gene set analytics with the MSigDB gene sets database.

Documentation: The website accompanying *GSEA-P 2.0* includes extensive documentation: a user guide describing all aspects of the software, an illustrated tutorial, a frequently asked questions section, as well as four examples of GSEA analysis and results. The documentation is packaged into a GSEA Wiki site which will grow over time.

ACKNOWLEDGEMENTS

The authors wish to thank members of the Cancer Program at the Broad Institute for suggestions. They also thank Jide software for a free license to their component suite.

REFERENCES

Bourquin, J. P., Subramanian, A., Langebrake, C., Reinhardt, D., Bernard, O., Ballerini, P., Baruchel, A., Cave, H., Dastugue, N., Hasle, H., *et al.* (2006). Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. Proc Natl Acad Sci U S A *103*, 3339-3344.

Liu, G., Loraine, A. E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D., and Siani-Rose, M. A. (2003). NetAffx: Affymetrix probesets and annotations. Nucleic Acids Res *31*, 82-86.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., *et al.* (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet *34*, 267-273.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D (1997). GeneCards: encyclopedia for genes, proteins and diseases, In Weizmann Institute of Science, Bioinformatics Unit and Genome Center.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). From the Cover: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J. J., Ladd-Acosta, C., Mesirov, J., Golub, T. R., and Jacks, T. (2005). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat Genet *37*, 48-55.