GenePattern

GSEA Documentation

Description: Evaluates a genomewide expression profile and determines whether *a priori* defined sets of genes show statistically significant, concordant differences between two biological states (e.g., phenotypes) or over a continuous phenotype.

Author: Aravind Subramanian, Pablo Tamayo, gp-help@broadinstitute.org

Summary

Gene Set Enrichment Analysis (GSEA) is a powerful analytical method for interpreting gene expression data. It evaluates cumulative changes in the expression of groups of multiple genes defined based on prior biological knowledge. It first ranks all genes in a data set, then calculates an enrichment score for each gene set, which reflects how often members of that gene set occur at the top or bottom of the ranked data set (for example, in expression data, in either the most highly expressed genes or the most underexpressed genes).

Introduction

Microarray experiments profile the expression of tens of thousands of genes over a number of samples that can vary from as few as two to several hundreds. One common approach to analyzing these data is to identify a limited number of the most interesting genes for closer analysis. This usually means identifying genes with the largest changes in their expression values based on a t-test or similar statistic, and then picking a significance cutoff that will trim the list of interesting genes down to a handful of genes for further research.

Gene Set Enrichment Analysis (GSEA) takes an alternative approach to analyzing genomic data: it focuses on cumulative changes in the expression of multiple genes as a group, which shifts the focus from individual genes to groups of genes. By looking at several genes at once, GSEA can identify pathways whose several genes each change a small amount, but in a coordinated way. This approach helps reflect many of the complexities of co-regulation and modular expression.

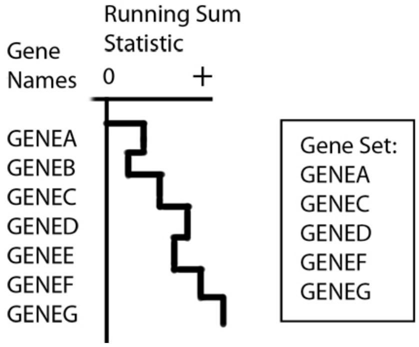
GSEA therefore takes as input two distinct types of data for its analysis:

- the gene expression data set
- *gene sets*, where each set is comprised of a list of genes whose grouping together has some biological meaning; these gene sets can be drawn from the Molecular Signatures Database (MSigDB) or can be from other sources

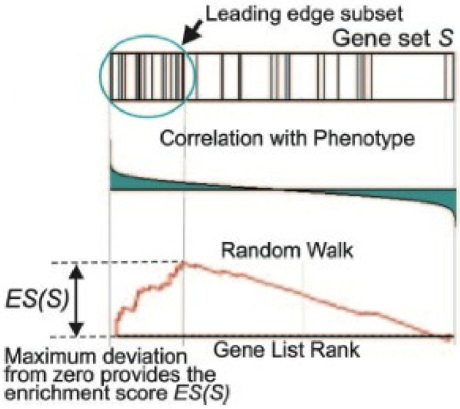
The GSEA GenePattern module uses categorical data for its analysis; that is, datasets that contain two different classes of sample, such as "tumor" and "normal." The GSEA desktop application, available on the GSEA website, has additional functionalities. For instance, the GSEA desktop application can analyze many different types of data, including the ability to assign continuous phenotype labels.

Algorithm

GSEA first ranks the genes based on a measure of each gene's differential expression with respect to the two phenotypes (for example, tumor versus normal). Then the entire ranked list is used to assess how the genes of each gene set are distributed across the ranked list. To do this, GSEA walks down the ranked list of genes, increasing a running-sum statistic when a gene belongs to the set and decreasing it when the gene does not. A simplified example is shown in the following figure.



The enrichment score (ES) is the maximum deviation from zero encountered during that walk. The ES reflects the degree to which the genes in a gene set are overrepresented at the top or bottom of the entire ranked list of genes. A set that is not enriched will have its genes spread more or less uniformly through the ranked list. An enriched set, on the other hand, will have a larger portion of its genes at one or the other end of the ranked list. The extent of enrichment is captured mathematically as the ES statistic.



Next, GSEA estimates the statistical significance of the ES by a permutation test. To do this, GSEA creates a version of the data set with the phenotype labels randomly scrambled, produces the corresponding ranked list, and recomputes the ES of the gene set for this permuted data set. GSEA repeats this many times (1000 is the default) and produces an empirical null distribution of ES scores.

The nominal p-value estimates the statistical significance of a single gene set's enrichment score, based on the permutation-generated null distribution. The nominal p-value is the probability under the null distribution of obtaining an ES value that is as strong or stronger than that observed for your experiment under the permutation-generated null distribution.

Typically, GSEA is run with a large number of gene sets. For example, the MSigDB collection and subcollections each contain hundreds to thousands of gene sets. This has implications when comparing enrichment results for the many sets:

The ES must be adjusted to account differences in the gene set sizes and in correlations between gene sets and the expression data set. The resulting normalized enrichment scores (NES) allow you to compare the analysis results across gene sets.

The nominal p-values need to be corrected to adjust for multiple hypothesis testing. For a large number of sets (rule of thumb: more than 30), we recommend paying attention to the False Discovery Rate (FDR) q-values: consider a set significantly enriched if its NES has an FDR q-value below 0.25.

For more information, see http://www.broadinstitute.org/gsea.

Known Issues

File names

Input expression datasets with the character '-' or spaces in their file names causes GSEA to error.

CLS files

The GSEA GenePattern module interprets the sample labels in CLS files by their order of appearance, rather than via their numerical value, unlike some other GenePattern modules. For example, in the CLS file below:

```
13 2 1
# resistant sensitive
1 1 1 1 1 1 1 1 0 0 0 0 0
```

Most other GenePattern modules would interpret the first 8 samples to be sensitive and the remaining 5 to be resistant. However, GSEA assigns resistant to the first 8 samples and sensitive to the rest. This is because GSEA assigns the first name in the second line to the first symbol found on the third line.

If the sample labels are in numerical order, as below, no difference in behavior will be noted.

```
13 2 1
# resistant sensitive
0 0 0 0 0 1 1 1 1 1 1 1 1
```

Reference

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*. 2005;102(43);15545-15550. (Link)

Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesivor JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1-α responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34:267-273. (link)

GSEA User Guide: http://www.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html

Parameters

NOTE: Certain parameters are considered to be "advanced"; that is, they control details of the GSEA algorithm that are typically not changed. You should not override the default values unless you are conversant with the algorithm. These parameters are marked "Advanced" in the parameter descriptions.

Name	**Description**
expression dataset (required)	This is a file in either GCT or RES format that contains the expression dataset.
gene sets database	This drop-down allows you to select gene sets from the Molecular Signatures Database (MSigDB) on the GSEA website. This provides access to only the most current version of MSigDB. If you want to use files from an earlier version of MSigDB, you will need to download that file from the archived releases on the website and specify it in the *gene sets database file* parameter*.* If you do not select an option here, you MUST upload a file in the *gene sets database file* parameter.
gene sets database file	Allows you to upload a gene set file not available in the current version of MSigDB (and thus not listed in the *gene sets database* parameter drop-down). This file must be in GMT, GMX, or GRP format.
number of permutations (required)	Specifies the number of permutations to perform in assessing the statistical significance of the enrichment score. It is best to start with a small number, such as 10, in order to check that your analysis will complete successfully (e.g., ensuring you have gene sets that satisfy the minimum and maximum size requirements and that the collapsing genes to symbols works correctly). After the analysis completes successfully, run it again with a full set of permutations. The recommended number of permutations is 1000. Default: 1000

phenotype labels (required)	A phenotype label file defines phenotype labels and assigns those labels to the samples in your expression dataset. This is a tab-delimited text file in CLS format. The CLS should contain only two phenotypes, or categorical labels, such as tumor and normal. The GSEA module also supports continuous labels, which are used to analyze a time series experiment or to find gene sets that correlate with a gene of interest (gene neighbors). A continuous label contains a value for each sample, where that series of values defines the phenotype profile. The file name must include the name of the profile (e.g., gseaMat.cls#IncreasingProfile) that you enter in the *target profile* parameter. GSEA interprets CLS files differently than many GenePattern modules. See the Known Issue for more details.
target profile	Name of the target phenotype profile for a CLS file that defines continuous labels. This will be ignored if the CLS file defines a discrete phenotype pair.
collapse dataset (required)	Select *true* to have GSEA collapse each probe set in the expression dataset into a single line of data for the gene, which is identified by its HUGO gene symbol. Be sure that your gene sets and array annotations also use gene symbols as the gene identifier format. Select *false* to use your expression dataset as is, with its native feature identifiers. When you select this option, the chip annotation file (*chip platform* parameter) is optional and you must specify a gene set file (g*ene sets database file* parameter) that identify genes using the same feature (gene or probe) identifiers as is used in your expression dataset. Default: *true*

permutation type (required)	Type of permutations to perform in assessing the statistical significance of the enrichment score. Options are: • phenotype (default): Random phenotypes are created by shuffling the phenotype labels on the samples. For each random phenotype, GSEA ranks the genes and calculates the enrichment score for all gene sets. These enrichment scores are used to create a distribution from which the significance of the actual enrichment score (for the actual expression data and gene set) is calculated. This is the recommended method when there are at least 7 samples in each phenotype. • gene_set: Random gene sets, size matched to the actual gene set, are created and their enrichment scores calculated. These enrichment scores are used to create a null distribution from which the significance of the actual enrichment score (for the actual gene set) is calculated. This method is useful when you have too few samples to do phenotype permutations (that is, when you have fewer than 7 samples in any phenotype). Phenotype permutation is recommended whenever possible. The phenotype permutation shuffles the phenotype labels on the samples in the dataset; it does not modify gene sets. Therefore, the correlations between the genes in the dataset and the genes in a gene set are preserved across phenotype permutations. The gene_set permutation creates random gene sets; therefore, the correlations between the genes in the dataset and the genes in the gene set are not preserved across gene_set permutations. Preserving the gene-to-gene correlation across permutations provides a more biologically reasonable (more stringent) assessment of significance.
chip platform	This drop-down allows you to specify the chip annotation file, which lists each probe on a chip and its matching HUGO gene symbol, used for the expression array. The chip files listed here are from the GSEA website: http://www.broadinstitute.org/gsea/downloads.jsp If you used a chip file not listed here, you will need to upload and specify it in the *chip platform file* parameter. If you do not select an option here, you MUST upload a file in the *chip platform file* parameter.
chip platform file	Allows you to upload a chip annotation file not available in the drop-down list of the *chip platform* parameter. This file must be in CHIP format.

scoring scheme (required)	The enrichment statistic. This parameter affects the running-sum statistic used for the enrichment analysis, controlling the value of p used in the enrichment score calculation. Options are: • classic: p=0 • weighted (default): p=1; a running sum statistic that is incremented by the absolute value of the ranking metric when a gene belongs to the set (see the 2005 PNAS paper for details) • weighted_p2: p=2 • weighted_p1.5: p=1.5
metric for ranking genes (required)	GSEA ranks the genes in the expression dataset and then analyzes that ranked list of genes. Use this parameter to select the metric used to score and rank the genes. The default metric for ranking genes is the *signal-to-noise ratio*. To use this metric, your expression dataset must contain at least three (3) samples for each phenotype. For descriptions of the ranking metrics, see Metrics for Ranking Genes in the GSEA User Guide.
gene list ordering mode (required)	Specifies whether to sort the genes in descending (default) or ascending order.
gene list sorting mode (required)	Specifies whether to sort the genes using the real (default) or absolute value of the gene-ranking metric score.
gene list ordering mode (required)	Specifies the direction in which the gene list should be ordered (ascending or descending).
max gene set size (required)	After filtering from the gene sets any gene not in the expression dataset, gene sets larger than this are excluded from the analysis. Default: 500
min gene set size (required)	After filtering from the gene sets any gene not in the expression dataset, gene sets smaller than this are excluded from the analysis. Default: 15

collapsing mode for probe sets with more than one match (required) (Advanced)	Collapsing mode for sets of multiple probes for a single gene. Used only when the *collapse dataset* parameter is set to *true*. Select the expression values to use for the single probe that will represent all probe sets for the gene. Options are: • Max_probe (default): For each sample, use the maximum expression value for the probe set. That is, if there are three probes that map to a single gene, the expression value that will represent the collapsed probe set will be the maximum expression value from those three probes. • Median_of_probes: For each sample, use the median expression value for the probe set.
normalization mode (required) (Advanced)	Method used to normalize the enrichment scores across analyzed gene sets. Options are: • meandiv (default): GSEA normalizes the enrichment scores as described in Normalized Enrichment Score (NES) in the GSEA User Guide. • None: GSEA does not normalize the enrichment scores.
randomization mode (required) (Advanced)	Method used to randomly assign phenotype labels to samples for phenotype permutations. ONLY used for phenotype permutations. Options are: • no_balance (default): Permutes labels without regard to number of samples per phenotype. For example, if your dataset has 12 samples in phenotype_a and 10 samples in phenotype_b, any permutation of phenotype_a has 12 samples randomly chosen from the dataset. • equalize_and_balance: Permutes labels by equalizing the number of samples per phenotype and then balancing the number of samples contributed by each phenotype. For example, if your dataset has 12 samples in phenotype_a and 10 samples in phenotype_b, any permutation of phenotype_a has 10 samples: 5 randomly chosen from phenotype_a and 5 randomly chosen from phenotype_b.
omit features with no symbol match (required) (Advanced)	Used only when *collapse dataset* is set to *true*. By default (*true*), the new dataset excludes probes/genes that have no gene symbols. Set to *false* to have the new dataset contain all probes/genes that were in the original dataset.

make detailed gene set report (required) (Advanced)	Create detailed gene set report (heat map, mountain plot, etc.) for each enriched gene set. Default: true
median for class metrics (required) (Advanced)	Specifies whether to use the median of each class, instead of the mean, in the *metric for ranking genes*. Default: false
number of markers (required) (Advanced)	Number of features (gene or probes) to include in the butterfly plot in the Gene Markers section of the gene set enrichment report. Default: 100
plot graphs for the top sets of each phenotype (required) (Advanced)	Generates summary plots and detailed analysis results for the top x genes in each phenotype, where x is 20 by default. The top genes are those with the largest normalized enrichment scores. Default: 20
random seed (required) (Advanced)	Seed used to generate a random number for phenotype and gene_set permutations. Timestamp is the default. Using a specific integer valued seed generates consistent results, which is useful when testing software.
save random ranked lists (required) (Advanced)	Specifies whether to save the random ranked lists of genes created by phenotype permutations. When you save random ranked lists, for each permutation, GSEA saves the rank metric score for each gene (the score used to position the gene in the ranked list). Saving random ranked lists is **very memory intensive**; therefore, this parameter is set to false by default.
output file name (required)	Name of the output file. The name cannot include spaces. Default: <expression.dataset_basename>.zip

Input Files

1. GCT or RES file

   This file contains the expression dataset.

2. GMT, GMX, or GRP file (optional, if you do not select a *gene set database* from the drop-down)

A gene set file not available in the current version of MSigDB (and thus not listed in the *gene sets database* parameter drop-down).

3. CLS file

   The GSEA module supports two kinds of class (CLS) files: those with two phenotypes (or categorical labels), such as tumor and normal, and those with continuous labels.
   A continuous label contains a value for each sample, where that series of values defines the phenotype profile. The file name must include the name of the profile (e.g., gseaMat.cls#IncreasingProfile).

   - For a gene of interest, the value for each sample is the expression value of the gene. The phenotype profile is, therefore, the expression profile of the gene of interest.
   - For a time series, the value for each sample is a number chosen to define the desired expression profile. The relative change in the value for each sample defines the relative distance between points in the profile. Assume, for example, that you have five samples taken at 30 minute intervals.

   To define a phenotype profile that shows steadily increasing gene expression, you would choose steadily increasing values for each sample (perhaps the number of minutes elapsed since the initial treatment):

   ```
   #numeric                              #IncreasingProfile 30 60 90 120 150
   ```

   To define a phenotype profile that shows an initial peak and then gradual decrease, you would choose values for each sample that reflect that desired phenotype profile:

   ```
   #numeric                       #PeakProfile 5 20 15 10 5
   ```

4. CHIP file (optional, if you do not select a *chip platform* from the drop-down)

   A chip annotation file not available in the module drop-down list.

Output Files

1. ZIP file containing the result files

   For more information on interpreting these results, see Interpreting GSEA Results in the GSEA User Guide.

Platform Dependencies

Module type:	Gene List Selection
CPU type:	any
OS:	any
Language:	Java 1.6

GenePattern Module Version Notes

Version	Release Date	Description
14	6/11/2013	Update the gene sets database list and the GSEA Java library, added support for continuous phenotypes.
13	9/27/2012	Update and sorted the chip platforms list, changed default value of num permutations to 1000, and updated the GSEA java library.