GenePattern

ssGSEAProjection Documentation

Description: Project each sample within a data set onto a space of gene set enrichment scores using the ssGSEA projection methodology described in Barbie et al., 2009.

Author: Pablo Tamayo (Broad Institute) Tamayo@broadinstitute.org, gp-help@broadinstitute.org

Summary

Single-sample GSEA (ssGSEA), an extension of Gene Set Enrichment Analysis (GSEA), calculates separate enrichment scores for each pairing of a sample and gene set. Each ssGSEA enrichment score represents the degree to which the genes in a particular gene set are coordinately up- or down-regulated within a sample.

Discussion

When analyzing genome-wide transcription profiles from microarray data, a typical goal is to find genes significantly differentially correlated with distinct sample classes defined by a particular phenotype (e.g., tumor vs. normal). These findings can be used to provide insights into the underlying biological mechanisms or to classify (predict the phenotype of) a new sample. Gene Set Enrichment Analysis (GSEA) addressed this problem by evaluating whether a priori defined sets of genes, associated with particular biological processes (such as pathways), chromosomal locations, or experimental results are enriched at either the top or bottom of a list of differentially expressed genes ranked by some measure of differences in a gene's expression across sample classes. Examples of ranking metrics are fold change for categorical phenotypes (e.g., tumor vs. normal) and Pearson correlation for continuous phenotypes (e.g., age). Enrichment provides evidence for the coordinate up- or down-regulation of a gene set's members and the activation or repression of some corresponding biological process.

Where GSEA generates a gene set's enrichment score with respect to phenotypic differences across a collection of samples within a dataset, ssGSEA calculates a separate enrichment score for each pairing of sample and gene set, independent of phenotype labeling. In this manner, ssGSEA transforms a single sample's gene expression profile to a *gene set* enrichment profile. A gene set's enrichment score represents the activity level of the biological process in which the gene set's members are coordinately up- or down-regulated. This transformation allows researchers to characterize cell state in terms of the activity levels of biological processes and pathways rather than through the expression levels of individual genes.

In working with the transformed data, the goal is to find biological processes that are differentially active across the phenotype of interest and to use these measures of process activity to characterize the phenotype. Thus, the benefit here is that the ssGSEA projection transforms the data to a higher-level (pathways instead of genes) space representing a more biologically interpretable set of features on which analytic methods can be applied.

This module implements the single-sample GSEA projection methodology described in Barbie et al, 2009.

References

1. 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. http://www.pnas.org/content/102/43/15545.abstract
2. Barbie DA, Tamayo P, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009;462:108-112.

Parameters:

Name	**Description**
input gct file (required)	GCT file containing input dataset's gene expression data.
output file prefix (required)	The prefix used for the name of the output GCT file. If unspecified, output prefix will be set to <prefix of input GCT file>.PROJ. The output GCT file will contain the projection of input dataset onto a space of gene set enrichments scores.
gene sets database	Gene sets database from the GSEA website. Note: *Gene sets database file* and *gene sets database list file* override this parameter.
gene sets database file	Gene sets database files in either GMT or GMX format. Upload a gene set if your gene set is not listed as a choice for the gene sets database parameter. Note: An optional parameter, which when set overrides the *gene sets database* parameter.

gene sets database list file	.txt file containing a list of GMT and GMX gene set description files (one gene set description filename per line). This optional parameter should be used if projecting expression data across gene sets spanning multiple gene sets database files. The listed gene sets database files must be uploaded to GenePattern server. This list file is typically generated using the GenePattern ListFiles module. Note: An optional parameter, which when set overrides the *gene sets database* and *gene sets database file* parameters.
gene symbol column (required)	Input GCT file column containing gene symbol names. In most cases, this will be column 1. (default: *Column 1*)
gene set selection (required)	Comma-separated list of gene set names on which to project the input expression data. Alternatively, this field may be set to *ALL*, indicating that the input expression dataset is to be projected to all gene sets defined in the specified gene set database(s). (default: *ALL*)
sample normalization method (required)	Normalization method applied to expression data. Supported methods are *rank*, *log.rank*, and *log*. (Default: *rank*)
weighting exponent (required)	Exponential weight employed in calculation of enrichment scores. The default value of 0.75 was selected after extensive testing. The module authors strongly recommend against changing from default. (Default: *0.75*)
min gene set size (required)	Exclude from the projection gene sets whose overlap with the genes listed in the input GCT file are less than this value. (Default: 10)

Input Files

1. input expression dataset

   The GCT file containing the input dataset's gene expression data (see the GCT file format information at http://www.broadinstitute.org/cancer/software/genepattern/gp_guides/file-formats/sections/gct). Gene symbols are typically listed in the column with header *Name*; however, GCT files containing RNAi data may list the gene symbol name in alternative columns. The "gene symbol column name" parameter specifies which of the input GCT file's columns contains the gene symbols.

The input GCT file’s row identifiers must draw from the same family of gene identifiers (the same ontology or name space, such as HUGO Gene Nomenclature) as those used to identify genes in the gene sets database file (see next item below). Typically these are human gene symbols.

If a GCT file’s row identifiers are probe IDs, and gene sets are defined through a list of human gene symbols, it will be necessary to collapse all probe set expression values for a given gene into a single expression value and use a human gene symbol to represent that gene. The CollapseDataset GenePattern module can make this transformation.

2. gene sets database file

   An optional GMT or GMX file containing a collection of gene set definitions (see the GMT file format and the GMX file format in the GenePattern file formats documentation.

   In the case of experimentally derived gene sets with two versions (UP/DN defined by the top up-/down-regulated genes in the data) a combined score will be computed, $ES(G_{UP},S)-ES(G_{DN},S)$, and added to the projection. Pairings of up- and down-regulated gene sets have matching gene set names distinguished solely by the “_UP” and “_DN” suffixes.

3. gene sets database list file

   An optional text file containing a list of gene set definition files (each in the GMT or GMX format). Each line in the text file contains a single filename. Typically this file is generated by the GenePattern ListFiles module and is used when projecting expression data onto gene sets defined across multiple gene sets database files. No duplicate gene set names are allowed across the listed gene sets database files.

Output Files

1. output enrichment score dataset

   A GCT file containing the input dataset’s projection onto a space of specified gene sets. This GCT file may serve as input into GenePattern’s many clustering and classification algorithms.

Platform Dependencies

Module type:	Projection
CPU type:	any
OS:	any
Language:	R

GenePattern Module Version Notes

Date	Version	Description
06/11/13	4	Updated list of gene sets to include v4.0 MSigDB collections
02/15/12	3	Updated list of gene sets databases to include v3.1 MSigDB collections; updated FTP download code; made documentation more biologist-friendly.
08/31/12	2	Add support for GMX-formatted gene sets description files
07/03/12	1	Initial version