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. In this manner ssGSEA projects a single sample's gene expression profile from the space of single genes onto the space of gene sets. Any supervised or unsupervised machine learning technique can then be applied to the resulting projected dataset. 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.

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.

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	Name of GCT file containing input dataset's gene expression data.

output file prefix	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 GSEA website.
gene sets database file	Gene sets database - .gmt. Upload a gene set if your gene set is not listed as a choice for the gene sets database parameter.
gene sets database list file	Name of the file containing a list of GMT 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.
gene symbol column name	Name of column in input GCT file containing gene symbol names. In most cases, this will be the *Name* column. (default: *Name*)
gene set selection	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	Normalization method applied to expression data. Supported methods are *rank*, *log.rank*, and *log*. (Default: *rank*)
weighting exponent	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	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 in the GenePattern file formats documentation). Gene symbols are typically contained in the column with header *Name*; however, GCT files containing RNAi data may contain 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.

2. gene set db list file

   An optional text file containing a list of gene set definition files (each in the GMT format – see the GMT file format in the GenePattern file formats documentation). 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