SubMap

Description: Maps subclasses between two data sets
Author: Yujin Hoshida (Broad Institute) gp-help@broadinstitute.org

Summary: It is usually difficult to combine multiple independent microarray data sets for the purpose of clustering due to various sources of biases including platform differences [1]. Given a pair of independent microarray data sets with sample subclass information, Subclass Mapping searches for matching pairs of subclasses between two input data sets [2]. Any subclass information, e.g., subclass found by unsupervised clustering, clinical phenotype, etc., can be used as input. Similarity between subclasses is measured using the Gene Set Enrichment Analysis (GSEA) [3]. Mapping result is represented as a subclass association (SA) matrix filled with p-values for each subclass association. By clustering the SA matrix, the global structure and correspondence of subclasses observed in both data sets appears.

The settings used in the original paper will require relatively long computation time. To get a sense of the optimal resolution of subclassification to be assessed (i.e. number of candidate subclasses defined in each input data set), the SubMapBrowser module can be used. To reduce computation time, the SubMapBrowser module (by default) uses a relatively small number of class-label permutations for the computation of p-values. The SubMap module (by default) uses a larger number of permutations to compute more accurate p-values.

Input data sets should have common identifiers. The intersection of these data sets is automatically extracted.

References:

1. Larkin JE, et al. Independence and reproducibility across microarray platforms. Nat Methods, 2005. 2;337-44
2. Hoshida Y, et al. Subclass Mapping: Identifying Common Subtypes in Independent Disease Data sets. PLoS ONE 2(11): e1195, 2007
3. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102;15545-50

Parameters:

Name	Description	Choices
datasetA file	Input dataset A (gct), should have common gene ID with dataset B **Note**: Remove spaces from sample names.	
datasetB file	Input dataset B (gct), should have common gene ID with dataset A **Note**: Remove spaces from sample names.	
classA file	Input class label A (cls), 3rd line should be	

	numeric. **Note**: Class labels are sequential numbers beginning with 1. If the labels in the cls file start at 0, the SubMap module automatically adds 1 to all of the labels.	
classB file	Input class label B (cls), 3rd line should be numeric. **Note**: Class labels are sequential numbers beginning with 1. If the labels in the cls file start at 0, the SubMap module automatically adds 1 to all of the labels.	
num marker genes	Number of marker genes to be mapped. We recommend using the default value.	Default: 100
num perm	Number of random permutations for enrichment score (ES). Using a relatively large number increases the accuracy of the p-value. We recommend using the default value.	Default: 100
num perm fisher	Number of random permutations for Fisher's statistics. We recommend using the default value.	Default: 1000
weighted score type	Weight enrichment by correlation vector (signal-to-noise ratio). We recommend using the default value unless you are familiar with GSEA.	Default: yes
null distribution	Null distribution method. We recommend using the default value.	▪ pool (default): pool permutations for all cells of SA matrix; ▪ each: use permutations for each cell
p value correction	P-value correction method. For small numbers of classes (2~3 classes for each dataset), we recommend using the Bonferroni correction.	▪ Bonferroni (default) ▪ FDR: Benjamini and Hochberg, J Royal Stat Soc B, 1995. 57:289;
cluster rows	Cluster dataset A's subclass in heatmap of SA matrix.	yes (default); no
cluster columns	Cluster dataset B's subclass in heatmap of SA matrix.	yes (default); no
nominal p value matrix	Create heatmap for each nominal-p matrix.	yes (default); no
create legend	Create legend for heatmap.	yes (default); no
random seed	Random seed for permutations.	47365321 (default)
output filename	Name of output files containing the SA matrices, summary of enrichment score (ES) matrix, nominal p-values, and corrected p-values.	

Output Files:

1. <output.filename>_SubMapResult.txt: summary of the results
2. <output.filename>_<Bonferroni, FDR>_SAmatrix.gct: the SA matrix
3. <output.filename>_<Bonferroni, FDR>_SAmatrix.png: heatmap of the SA matrix

If nominal p value matrix is yes:

4. <output.filename>_nominal_p_matrix_<AonB, BonA>.gct: the nominal p value matrix
5. <output.filename>_nominal_p_matrix_<AonB, BonA>.png: heatmap of the nominal p value matrix

If create legend is yes:

6. legend.png

Platform dependencies:

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