GSEA Pathway Analysis Pipeline

This pathway analysis pipeline script was written by Julián Candia. For questions and/or comments, please contact the author at julian.candia@nih.gov .

0) Set up: install fgsea, dplyr.

1) Input: it is assumed that you are comparing the expression of two groups. Here, we call them "cases" and "controls". A sample input file is provided at DATA/DEG_topTable.txt. Columns are:

ENSG: Ensembl gene ID.

name: primary gene symbol.

logFC: median log-fold change of cases relative to controls. This is only used to assess the direction of change (positive sign when cases are over-expressed relative to controls, negative sign when controls are higher than cases).

P.Value: statistical significance of differential expression obtained from an upstream analysis.

2) Run GSEA.R:

2.1) Set ref_label to choose which gene sets you are interested in. All gene sets from MSigDB are available, including:

Hallmarks: Hallmark gene sets.

Canonical_Pathways: Reactome, KEGG, WikiPathways, PID, and Biocarta gene sets.

Chem_Genet_Perturb: Chemical and genetic perturbations.

miR_Targets: microRNA targets.

TF_Targets: transcription factor targets.

GO_BP: Gene Ontology biological processes.

GO_CC: Gene Ontology cellular components.

GO_MF: Gene Ontology molecular functions.

Immune: Immunologic gene sets.

Cell_Type: Cell type signature gene sets.

Others (not listed here, but available), include cancer-related gene sets. See MSigDB for details. A good place to start is by choosing just "Hallmarks" (50 pathways) and/or "Canonical_Pathways". The currently available reference is MSigDB v7.4 (latest as of Nov 2021). Follow instructions below to update the reference if needed. Custom gene sets may be added to the collection in a similar way.

2.2) Choose n_run. Because GSEA is a stochastic process, here we generate n_run runs and then choose pathways based on consensus. Recommended settings are n_run=10 for a trial run, followed by n_run=1000 for full statistics. Using full statistics, this run may take several hours to complete (especially if some of the larger gene sets are considered).

2.3) Other parameters can be changed (adj_pval_thres, pwy_sel_thres, gene_sel_thres, seed, eps) but sensible defaults are provided, so that tweaking them will generally be unnecessary.

adj_pval_thres: for each GSEA run, this is the adjusted p-value threshold for pathway selection (default=0.05).

pwy_sel_thres: this is the percentage of runs that a pathway is required to have been selected in, in order for the pathway to show up in the final list (default=80). For instance: if a given pathway has been selected in 754 out of 1000 runs (75.4%), the pathway will not be deemed significant and will not show up in the final selection (if the default value pwy_sel_thres=80 was used).

gene_sel_thres: from the total times a given pathway was selected, this is the percentage that a given gene is required to have been chosen as leading edge gene, in order for it to show up in the final list (default=80).

seed: fixed random number seed to ensure reproducibility.

eps: boundary for calculating pathway p-values used by the Monte Carlo procedure implemented in fgseaMultilevel. The default value is eps=1.e-10 but it can be set to zero for better (albeit slower) estimation.

2.4) Output files: For each reference gene set chosen, the output consists of two files deposited by default in the RESULTS folder: "<ref_label>_n<n_run>.txt", "<ref_label>_n<n_run>_pwys_per_gene.txt"

2.5) Additional details: This script checks for gene symbols in the primary input that are not contained in any gene set of interest, and will try to find any matching aliases. This procedure makes use of a mapping file of gene aliases generated by the script genenames.R. If you need to update this mapping file, follow the instructions below.

3) Run gene_pathway_overlaps.R. This script eliminates redundancies by generating a minimal set of significant pathways. When two significant pathways overlap, it will pick the larger one. As a result, if one pathway refers to a more specific process contained in another, more general one, the former will be dropped. One important parameter is frac_thres, the gene fraction of a target pathway that needs to overlap with some other pathway, for the former to be considered a candidate for removal. For instance, frac_thres=0.95 corresponds to 95% overlap. It must be <=1, and filtering is more lenient the closer it is to 1 (that, is more pathways will be kept in the final solution). The process generates cycles of the number of pathways squared, so computing time increases exponentially with the number of pathways considered. The output consists of two files deposited by default in the RESULTS folder that contain the merged, filtered pathways (same formatting as output files from previous step).

INSTRUCTIONS TO UPDATE THE REFERENCE GENE SETS (IF NEEDED):

1) Download the full MSigDB gene set collection from http://www.gsea-msigdb.org/gsea/downloads.jsp

All gmt files should be deposited at REF/MSigDB

2) Download gene names (and aliases) from https://www.genenames.org/download/custom with the custom data selection indicated in the screenshot REF/HGNC/genenames.png

The downloaded file should be deposited at REF/HGNC/genenames.txt

3) Run genenames.R. This script will generate the REF/HGNC/genenames_map.txt file where all gene symbol aliases are mapped to HUGO identifiers and Ensembl gene IDs (ENSG).