Supplementary Information to “Assessment of Gene Set Enrichment Analysis using curated RNA-seq-based Benchmarks” by J. Candia and L. Ferrucci

Please notice that several of the algorithms provided below are stochastic, including GSEA’s built-in randomizations to assess pathway significance and our custom scripts for phenotype randomization. Therefore, one would not obtain the exact same results reported in our paper, although you should expect similar results. Contact information: julian.candia@nih.gov

Data

TCGA data download from the NCI GDC portal (https://portal.gdc.cancer.gov):

- From the “Projects” tab, select:

Program = “TCGA”
Data Category = “transcriptome profiling”
Experimental Strategy = “RNA-Seq”

- Select “Open Query in Repository”

- From the Repository tab, select:

Data Type = “Gene Expression Quantification”

This selects 11,274 open-access files. Since downloads are limited to 10,000 files, we split TCGA projects into two manifests as follows:

gdc_manifest.2023-07-19_D1.txt: 9,962 files (TCGA except "breast", "eye and adnexa")
gdc_manifest.2023-07-19_D2.txt: 1,312 files (TCGA "breast", "eye and adnexa")

After downloading the GDC client, you can download all files to your local computer as follows:

./gdc-client download -m <path-to-manifest>

In a similar fashion, you can download the clinical and biospecimen files associated with these samples. We cleaned up and merged the relevant information into “sample_metadata.txt”. This file, together with the manifests described above, are available in the DATA folder.

MSigDB data download (https://www.gsea-msigdb.org/gsea/downloads.jsp#msigdb): download the GMT file msigdb.v2023.1.Hs.symbols.gmt (containing 33,591 human gene sets) into the DATA folder.

R Scripts

1. GSEA (paired-sample analysis)

TCGA_proj_filter.R: preselects TCGA projects with at least 10 subjects with paired "Primary Tumor" / "Solid Tissue Normal" samples, creates directories for each project (15 TCGA projects preselected).

alias_gene_names.R: creates list of protein-coding genes from the downloaded expression matrices, compares against gene names in the MSigDb reference, and determines gene aliases that maximize the gene name overlap between them. Note that expression matrices for all samples contain the same set of genes (harmonized pipeline from GDC).

TCGA_pwy_presel.R: for each TCGA project, creates a list of preselected pathways by matching MSigDb pathway names with manually curated search terms.

TCGA_DEG.R: removes FFPE, generates DEG ranks and saves cleaned datasets

TCGA_DEG_stats.R: calculates DEG statistics (number of DE genes using different thresholds).

TCGA_DEG_volcano.R: generates Volcano plots, additional DEG statistics.

TCGA_GSEA_gs.R: (gene-set randomization) for each TCGA project, it performs Preranked-GSEA. It requires previous install of a local copy of command-line GSEA.

GSEA_merge.R: (gene-set randomization) generates "pval" and "dir" matrices (pathways=rows vs runs=columns). Here, 4 different runs refer to different weight parameters.

TCGA_DEG_rdm.R: (phenotype randomization) for each TCGA project, it imports the expression, gene annotation and sample annotation matrices (cleaned datasets saved by TCGA_DEG.R) and generates n_rdm=1000 realizations of randomized paired data. For each randomization, it generates a "ranks.rnk" file. This is the data input needed for phenotype randomization.

TCGA_GSEA_rdm.R: (phenotype randomization) for each TCGA project, it performs GSEA on the set of random ranks files previously generated by TCGA_DEG_rdm.R.

TCGA_GSEA_ph.R: (phenotype randomization) calculates empirical p-values.

Venn_type.R: generates Venn diagrams based on permutation type ("gs", "ph") by placing user-selected cutoffs on the p-values.

Venn_weights.R: generates Venn diagrams based on weight parameter ("cl", "p1", "p1.5", "p2") by placing user-selected cutoffs on the p-values.

TCGA_pwy_stats.R: calculates the number of significant pathways (at a given significance threshold) across TCGA projects.

TCGA_proj_list_filt.R: based on the tables generated by TCGA_pwy_stats.R, we remove some projects that have too few preselected pathways found significant (TCGA_KICH, TCGA_KIRP, TCGA_UCEC). The remainder of the analyses are made on this filtered list of 12 TCGA projects.

TCGA_pwy_pos.ctrl.R: filters pwy.presel pathways by applying a soft filter requiring p-value<0.05 at least once among the eight modalities (gs and ph permutations and four weight parameter choices).

TCGA_pwy_ref_stats.R: table showing the number of preselected pathways through subsequent selection criteria (size, being detected using at least one of the GSEA modalities).

TCGA_pwy_neg.ctrl.unif.R: for each TCGA project, it creates a list of randomized (negative control) pathways of sizes (number of genes per pathway) drawn from a uniform probability distribution from size_min=15 to size_max=500

NOTE: at this point, in order to generate results for negative-control pathways, we re-run TCGA_GSEA_gs.R, GSEA_merge.R, TCGA_GSEA_rdm.R, and TCGA_GSEA_ph.R choosing collection = "neg.ctrl.unif" and uploading “TCGA_proj_list_filt.txt” as project list.

TCGA_ROC.R: calculates ROC plots and AUC comparing two collections (used as positive and negative controls).

TCGA_AUC.R: generates AUC plot across TCGA projects.

2. GSEA (unpaired-sample analysis)

TCGA_DEG_np.R: modified version of TCGA_DEG.R for unpaired differential gene expression.

TCGA_DEG_rdm_np.R: modified version of TCGA_DEG_rdm.R for unpaired sample randomization.

NOTE: the remaining pieces of the analysis are identical to the paired-sample case.

3. ORA

ORA.R: Over-Representation Analysis using signed / unsigned approaches; gene lists selected based on two different selection criteria: Bonferroni-adjusted p-value<0.05 and Benjamini-Hochberch adjusted p-value<0.05.

ORA_ROC: ROC plots for each TCGA project.

ORA_AUC: AUC and 95% CI across TCGA projects.

4. Enrichment Evidence Scores (EES)

REBC_THCA_comp.R: calculates the EES contingency matrix from the comparison between studies (TCGA-THCA and REBC-THYR), generates balloonplot, runs Fisher's exact test. The process is repeated after grouping EES scores into tumor, undetermined and non-tumor categories.