Analysis Pipeline

Code written by Julián Candia and Mayank Tandon to ensure full reproducibility of the results reported in "The genomic landscape of Mongolian hepatocellular carcinoma" by J. Candia et al. For questions and/or comments, please contact julian.candia@nih.gov

1. Transcriptomics-based analysis

1.1 Unsupervised approach (clustering):

- gene_expr_voom.R: generates voom-normalized (log2-transformed) gene expression matrix.
- gene_filter.R: calculates tumor-vs-nontumor statistics, filters genes based on MAD thresholds applied to tumor samples.
- cluster.R: performs consensus clustering on a range of values for number of clusters and MAD thresholds.
- cluster_stats.R: compares clustering solutions.
- cluster_surv.R: performs survival analysis (Kaplan-Meier, log-rank test) for each clustering solution.
- cluster_relabel.R: best clustering solution (4 clusters) is identified, clusters are relabeled.
- cluster_surv_replot.R: performs pairwise log-rank tests across clusters for the best clustering solution (4 clusters), replots Kaplan-Meier using final cluster labels and color scheme.
- cluster_relabel_alt.R: clusters relabeled based on an alternative clustering solution (2 clusters).
- cluster_surv_replot_alt.R: pairwise log-rank tests across clusters for the alternative clustering solution (2 clusters), Kaplan-Meier plots.
- cluster_sig.R: finds signatures of differentially expressed genes for each cluster.
- cluster_demo_clin.R: tests for association between clusters and demographic/clinical annotations.

1.2 Supervised approach (Cox regression):

- eNetXplorer_data_prep.R: preprocesses data for eNetXplorer analysis.
- eNetXplorer.R: performs eNetXplorer analysis for feature selection based on regularized Cox regression.
- cox.R: performs Cox regression using top features from eNetXplorer analysis, creates boxplot of integrated unsupervised (clustering) - supervised (Cox regression) analysis.
- cox_dendrogram.R: creates dendrogram of integrated unsupervised (clustering) - supervised (Cox regression) analysis.
- cox_risk.R: creates distribution of risk scores, uses threshold to binarize cohort into low- and high-risk groups.

- cox_risk_surv.R: performs survival analysis (Kaplan-Meier, log-rank test) for low- vs high-risk groups.

1.3 Summary of transcriptomics results:

- cluster_heatmap.R: Expression of cluster-specific differentially expressed genes with integrated demographic, clinical, and risk group annotations.

1.4 Comparison to class signatures from other studies:

- class_sig_comp.R: assigns class labels to the Mongolian cohort based on class signatures from other studies, performs chi-square tests to compare class assignments.
- class_sig_comp_circular.R: generates circular plot to visually compare class assignments across studies.

2. Whole Exome Sequencing-based analysis

2.1 Oncoplot and driver gene analysis:

- filter_maf.R: applies filters for variant selection, generates output mafs, generates list of candidate driver genes using MutSigCV q-values and the fraction of mutated samples as selection criteria. It also generates mafs for subcohorts based on HDV status.
- filter_maf_TCGA_LIHC.R: applies filters for variant selection to TCGA-LIHC, generates output mafs for the full cohort and ethnicity-specific (Asian, Caucasian) subcohorts.
- mut_oncoplot.R: generates oncoplot and summary table comparing Mongolia to TCGA-LIHC.
- mut_driver_gene_comp.R: generates driver gene table.
- mut_pca.R: generates PCA plot based on trinucleotide mutation frequencies.
- assoc_mut_annot.R: performs tests for association between clusters and demographic/clinical annotations.
- cooccur_mut.R: performs test for mutation co-occurrence between driver genes.
- explore_maf_hotspots.R: explores the existence of mutational hotspots in driver genes.
- mut_freq_TCGA_byGene.R: performs pan-cancer analysis of mutation frequencies of Mongolian HCC driver genes.
- mut_pval_TCGA_byGene.R: performs hypergeometric test of significance of driver gene mutation frequencies (Mongolia vs pan-cancer).
- mut_heatmap_TCGA_byGene.R: generates heatmap for the pan-cancer analysis of mutation frequencies of Mongolian HCC driver genes.
- mut_freq_TCGA_byLocus.R: performs pan-cancer analysis of mutation frequencies of Mongolian HCC hotspot loci.
- mut_pval_TCGA_byLocus.R: performs hypergeometric test of significance of hotspot loci mutation frequencies (Mongolia vs pan-cancer).
- mut_heatmap_TCGA_byLocus.R: generates heatmap for the pan-cancer analysis of mutation frequencies of Mongolian HCC hotspot loci.

2.2 Mutational signature analysis:

- mut_sig_trinucl_freq.R: calculates trinucleotide frequency distributions for the full cohort and the difference between HDV+ and HDV- subcohorts.
- mut_sig_deconstructSigs.R: generates matrix of subject vs signature weights using COSMIC (v2 and v3) and Environmental Agents (Kucab et al, Cell 2019) Compendia.
- mut_sig_deconstructSigs_plot.R: generates annotated heatmap of subject vs signature weights.