Pipeline Scripts —Quick Guide Gastric Cancer

00_Covfilecreator.R

- Builds a standardized covariate table (CombinedCovariates.csv): robust text normalization, TNM parsing, and rule-based cleaning.
- Detects Platform from RNA_seq.xlsx / Micro.xlsx headers; leaves missing values as blank (no 'NA').
- Stage is blank for Normal tissue; ensures strict Sample order alignment with expression columns.

01_Covfilecreator.R

- Updated pass of 00 with the same schema and QA: stronger row/column alignment and clearer final logging.
- Same output contract (Sample, Platform, Tissue, Sex, Stage, LocationCode, LaurenCode, Age_Code).
- Designed to be a drop-in replacement with improved resilience to messy inputs.

02_Pvalue.R

- Pairwise dependence testing across covariates: Fisher exact (2×2) or Chi-square (larger tables), with Monte Carlo when needed.
- Treats Missing as an explicit level; adds Spearman trend p-values for ordinal pairs (Stage, LocationCode, LaurenCode, Age_Code).
- Outputs: level_counts.csv, pairwise_pvalues_*.csv (All/TumorOnly/NormalOnly), MANIFEST.csv.

03_geo_combat_umap_O.R

- Microarray-only ComBat after removing Unknown tissue to isolate Normal vs Tumor signal.
- Adaptive batch key: GSE series > Organization×Country > Platform; exports GEO_ComBat_matrix.csv.
- UMAP before/after with top-variance genes post-ComBat; writes GEO_batch×tissue table for QC.

04_normalize_and_merge.R

- Runs ComBat only on multi-sample batches; singletons are kept untouched and recombined afterward.
- Preserves Unknown tissue level; uses tissue in the ComBat design only if both classes are present.
- Saves GEO_ComBat_matrix.csv and GEO_UMAP_4panel.png (and PCA if available).

05_all_combat.R

- Global ComBat with multi vs singleton separation and tissue stratification (Normal/Tumor/Unknown).
- Flexible model adds Sex/Stage/Age when consistent; safeguards against collinearity and sparse strata.
- Outputs All_ComBat_matrix.csv, All_Covariates_augmented.csv, ALL_UMAP_4panel.png.

06_qc_outliers_scaling.R

- Post-ComBat QC: density/box plots, PCA/UMAP colored by Tissue; optional pre vs post if Full_matrix.csv exists.
- Outlier detection via robust PCA (PcaHubert) + IQR rules; merges flags into Any_Outlier.
- Exports Outliers_report.csv, retained_samples.txt, removed_samples.txt.

07_edge.R

- Differential expression (limma): Tumor vs Normal (global) and per-demographic layers when n≥6 and both classes exist.
- Design: Tissue2 (Normal/Tumor/Unknown) plus informative covariates (Study/Platform/Sex/Stage/AgeZ).
- Outputs DGE_global_Tumor_vs_Normal.csv, volcano, and a top-50 heatmap when feasible.

08_gsva.R

- Pathway scoring (GSVA) with ID mapping to SYMBOL; collections: H, KEGG, REACTOME, GOBP (via msigdbr).
- limma on pathway scores (Tumor−Normal) globally and per-demographic; volcano/heatmap plots.
- Network-prep exports: high-variance pathway matrices and cleaned covariates; GSVA_scores_list.rds if present.

09_wgcna_modules.R

- Weighted gene co-expression network (WGCNA) on high-variance genes; pickSoftThreshold with fallback to 6.
- blockwiseModules(networkType='signed'); module eigengenes and module–trait correlations with p-values.
- Optional univariate Cox per module if survival data are available.

10_network_variants.R

- Multiple network flavors in Tumor/Normal: ARACNE, CLR, MRNET (MI-based) and PCIT (partial correlation).
- Top 2,000 variable genes; require n≥10 per group; edge weights: MI (minet) or |ρ| (PCIT).
- Computes method overlap (Jaccard, inter/union) → edge_overlap_jaccard.csv.

11_networks.R

- Per-subgroup ARACNE with survival-informed residualization (limma::removeBatchEffect on significant Cox covariates).
- Discretization (nbins=3) → MI → ARACNE; writes edges, node degrees, preview graph + GraphML.
- Avoids over-adjustment by excluding the grouping covariate from the residualization model.

12_0network.R

- Lightweight MI/ARACNE baseline without residualization; quick recoding of Location/Lauren/Stage/Age groups.
- Builds networks for groups with n≥10; exports edges and compact preview graphs.
- Intended for sanity-checks and baseline comparisons vs 11_networks.R.

13_diffcoexp.R

- DGCA-based differential co-expression between Tumor and Normal using Spearman correlations.
- FDR control (BH); significant pairs (q<0.05) form an undirected graph; components labeled as DiffCoExp modules (DCM).
- Exports DGCA_all_pairs.csv, DGCA_sig_pairs_q05.csv, DiffCoEx_modules.csv.

14_integration_network_dge_gsva.R

- Integration: network hubs (95th percentile degree) ∩ DE genes → HubDrivers_Tumor_ARACNE.csv.
- ME×GSVA correlations linking WGCNA modules to pathway activities; exports matrices and a compact heatmap.
- Centrality table includes degree, betweenness, closeness for Tumor_ARACNE.

15_survival.R

- Auto-maps OS_time/OS_event from Demo.xlsx; Kaplan–Meier by demographics (Sex/Stage) with log-rank p-values.
- Adjusted Cox model: Sex + Stage + Age + Study + Platform + Tissue; exports coefficients and a forest plot.
- Optional KM by GSVA hallmark scores (e.g., EMT/Angiogenesis) via median split.

16_survival_advanced.R

- Cox LASSO on module eigengenes (and optional hub genes) to derive a continuous risk score (Xβ).
- Time-dependent ROC at 12/36/60 (months or days, scale-aware) + KM by risk tertiles.
- Note: fixed HUB_* feature naming; writes selected coefficients and risk scores.

17_external_validation.R

- External validation: recompute module eigengenes in the validation cohort using training module colors.
- Apply CoxLASSO coefficients from training to compute risk; evaluate time-dependent AUC.
- Requires GC_VALIDATION_DIR; writes Risk_scores_VALID.csv and timeROC_AUC_VALID.csv.