BST281: Genomic Data Manipulation, Spring 2017

Wednesday 10: Transcriptomic analyses

Gene expression data are representative of many 'omics analyses over numerical matrices.

Rows = features (genes), columns = samples (experiments/assays/etc.), entries = abundances.

Analyses generally fall into supervised vs. unsupervised categories.

Supervised = inference / prediction based on prior knowledge (labeled examples).

Unsupervised = pattern discovery; generally less well defined, better for visualization.

Differential expression analysis is a supervised class comparison task.

Identify genes with significant expression differences, by some model, among covariates (e.g. case/control).

Fancy t-tests: SAM, limma, DESeq, edgeR, Cuffdiff, etc.

Generates gene lists that can be interpreted by:

Overlap testing (hypergeometric / Fisher's exact): DAVID, GOrilla, etc.

Enrichment analysis: GSEA

Network search: GeneMANIA, IMP, etc.

Expression matrices can also be used to predict labels for features or samples.

Case/control for samples, gene function for features: class prediction.

Can use any machine learner: SVM, kNN, decision trees, etc.

Evaluate by cross-validation / bootstrap to assess test error (in comparison to training); beware overfitting!

Clustering is an unsupervised method good for class discovery.

Takes as input a (dis)similarity measure, finds major patterns in data.

Hierarchical clustering generates a tree by iteratively joining the most similar feature pairs.

Linkage can be single (min), complete (max), average, centroid, etc.

K-means clustering generates k discrete partitions by iteratively grouping the closest feature representatives.

Evaluate cluster quality by stability and/or within-to-between (dis)similarity ratios.

Prediction strength, silhouette width, cophenetic correlation, etc.

Ordination (PCA/PCoA/(N)/MDS/t-SNE/etc.) is also an unsupervised analysis good for overview visualization.

# Textbooks

Transcriptomics: Pevsner, Chapter 8 p279-325

Chapter 9, p331-373

# Literature

[Cluster analysis and display of genome-wide expression patterns. Eisen, PNAS 1998](https://www.ncbi.nlm.nih.gov/pubmed/9843981)

[RNA-Seq analysis in MeV. Howe, Bioinformatics 2011](https://www.ncbi.nlm.nih.gov/pubmed/21976420)

[voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Law, Genome Biology 2014](https://www.ncbi.nlm.nih.gov/pubmed/24485249)

[Machine learning applications in genetics and genomics. Libbrecht, NRG 2015](https://www.ncbi.nlm.nih.gov/pubmed/25948244)