BST281: Genomic Data Manipulation, Spring 2017

Monday 10: Gene expression assays

Gene expression or transcriptional activity provides a global snapshot of molecular dynamics.

 Proteins/metabolites hard to measure, but RNA provides a more uniform intermediate.

Transcriptional measurements provide the ability to:

 Associate genes with biological processes / environmental conditions / stimuli / chemistry / regulation / etc.

 Diagnostic / prognostic biomarker for human (or other) sample outcomes.

Microarrays were originally developed for sequencing.

 Array one or more known sequences per gene in preassigned locations.

 Extract RNA from condition(s) of interest, reverse transcribe with fluorescent label.

 Quantify the amount of transcript using fluorescence (i.e. binding) at each location.

Can be either one or two channel depending on control strategy (internal vs. external) and quantification.

QC issues are educational for essentially all 'omics assays, can be at probe, gene (probeset), or experiment level.

 Spots must be located correctly, local background removed, spatial defects handled, batches normalized...

Identify probes, genes, or experiments that should just be tossed.

 Then normalize what's left, and optionally re-impute remaining missing values.

Normalization of expression data often uses MA plots:

 Y axis = M = ratio (red / green for two channel, current / median for single)

 X axis = A = average (red + green, or current + median)

 You want something that's vertically centered and straight (no systematic or intensity-dependent bias).

Median centering, quantile normalization, and LOWESS smoothing adjust plots.

 Median is global, quantile/LOWESS include local elements - ensure similar distributions among samples.

 (GC)RMA extends these to a hierarchical model including per-probe, -gene, -experiment, and -nucleotide.

RNA-seq is a mashup of DNA sequencing methods + QC with single channel microarray analysis.

 Transcript levels measured by read counts rather than fluorescence intensity.

 Tricky part during mapping is dealing with gaps to cross exon junctions; also resolve ambiguous isoforms.

 De novo assembly methods also exist, which extend DNA assembly techniques to allow multiple isoforms.

Single cell RNA-seq (scRNA-seq) replaces "experiments" with single cells, prone to noise / zeros / missing values.

 Differences are mostly in data generation to deal with extremely low input biomass.

 Allows analysis of heterogeneity / cell types / etc. but requires dealing with sparse, noisy, low count data.

Data live in GEO, ArrayExpress, Gene Expression Atlas, GDC Data Portal, BioGPS, and others.

# Textbooks

Expression databases: Pevsner, Chapter 2 p27-28

RNA biology: Pevsner, Chapter 10 p433-470

Expression QC: Pevsner, Chapter 11 p479-498

# Literature

[Widespread aneuploidy revealed by DNA microarray expression profiling. Hughes, Nature Genetics 2000](https://www.ncbi.nlm.nih.gov/pubmed/10888885)

[A survey of best practices for RNA-seq data analysis. Conesa, Genome Biology 2016](https://www.ncbi.nlm.nih.gov/pubmed/26813401)

[Design and computational analysis of single-cell RNA-sequencing experiments. Bacher, Genome Biology 2016](Design%20and%20computational%20analysis%20of%20single-cell%20RNA-sequencing%20experiments.)