Transcriptomics Practical Functional Enrichment

HackBio

Melyssa Minto

West Lab, Duke Neurobiology

Computational Biology and Bioinformatics

Transcriptomics pipeline/workflow

Preprocessing

Sample fastqs

transcriptome alignment

Sample bams

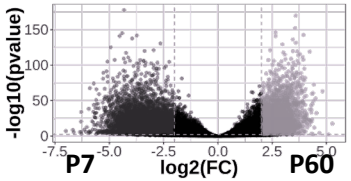
Generate gene counts

Gene Count matrix

Analyses

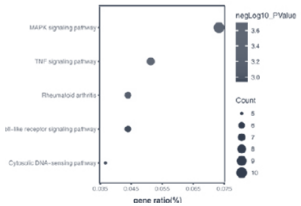
Clustering

Differential Expression



DESeq2 limma Glimma edgeR

Functional Enrichment



GSEA Gene Set Enrichment Analysis

MetaCore

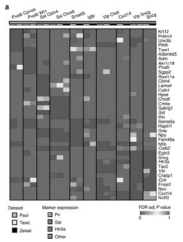
GREAT

KEGG Kyoto Encyclopedia of Genes and Genomes

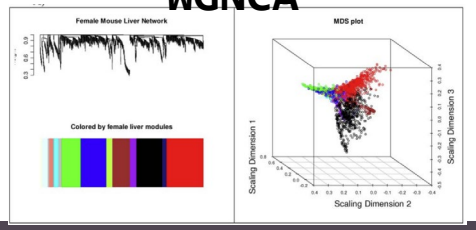
reactome

DAVID BIOINFORMATICS DATABASE

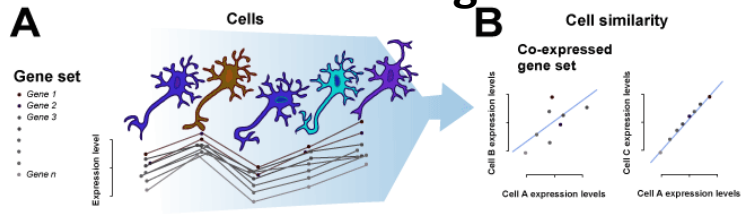
Coregulated Gene Expression



WGNCA



MetaNeighbor



What is enrichment?

- Most enrichment programs use the same statistical test to test for enrichment – Fishers Exact Test

	Differential Expression	NO Differential Expression	Total
IN Transcription Elongation	x	m - x	m
NOT IN Transcription Elongation	k - x	n - (k - x)	n
Total	k	(m + n - k)	m + n

- Test of proportions given a category
- You will need
  1. Annotation
  2. Gene set
  3. Background set

To reduce false positives – resampling based methods should be used

What is enrichment?

- Most enrichment programs use the same statistical test to test for enrichment – Fishers Exact Test

	Gene set	background	
In _____ Category	a	b	a+b
Not In _____ Category	c	d	c+d
	a+c	b+d	N=a+b+c+d

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{N}{a+c}}$$

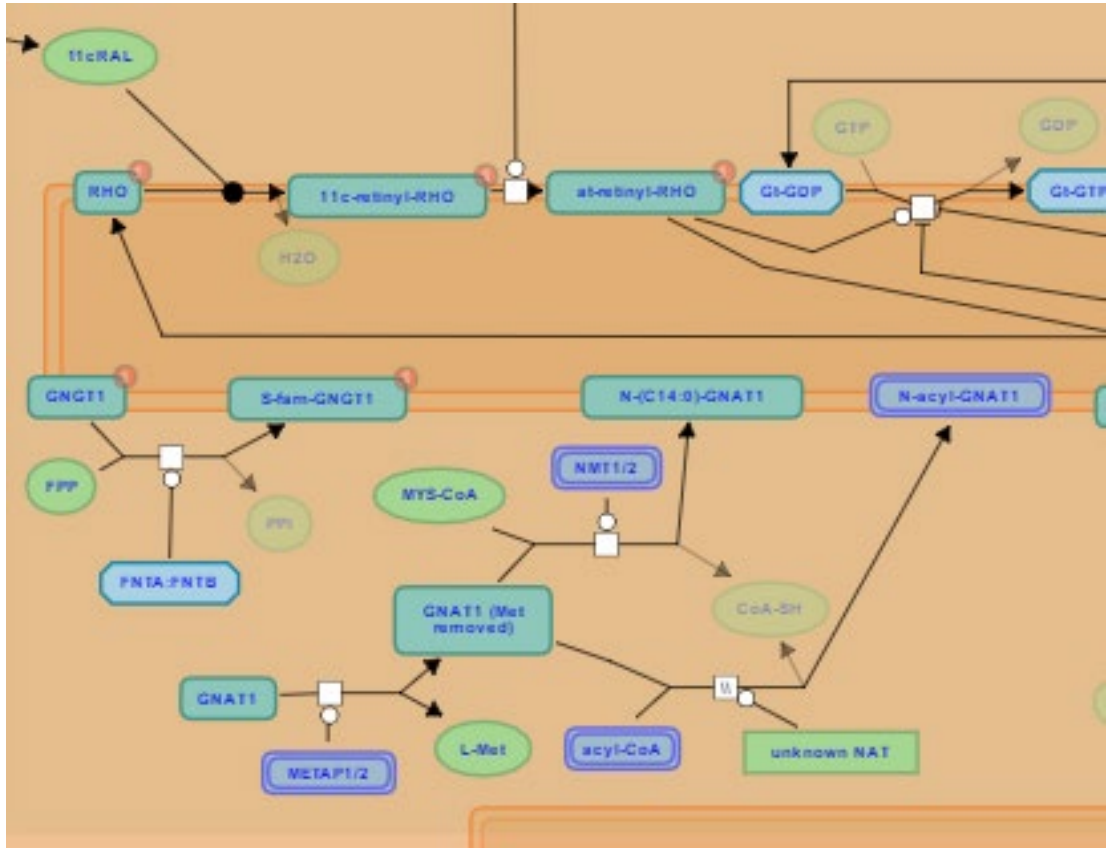
https://towardsdatascience.com/fishers-exact-test-from-scratch-with-python-2b907f29e593

Different ways to categorize data

Different ways to categorize data

Molecular Function	Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as "catalysis" or "transport". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products (*i.e.* a protein or RNA), but some activities are performed by molecular complexes composed of multiple gene products. Examples of broad functional terms are *catalytic activity* and *transporter activity*; examples of narrower functional terms are *adenylate cyclase activity* or *Toll-like receptor binding*. To avoid confusion between gene product names and their molecular functions, GO molecular functions are often appended with the word "activity" (a *protein kinase* would have the GO molecular function *protein kinase activity*).
Cellular Component	The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (*e.g.*, *mitochondrion*), or stable macromolecular complexes of which they are parts (*e.g.*, the *ribosome*). Unlike the other aspects of GO, cellular component classes refer not to processes but rather a cellular anatomy.
Biological Process	The larger processes, or 'biological programs' accomplished by multiple molecular activities. Examples of broad biological process terms are *DNA repair* or *signal transduction*. Examples of more specific terms are *pyrimidine nucleobase biosynthetic process* or *glucose transmembrane transport*. Note that a biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

Different ways to categorize data



Different ways to categorize data

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

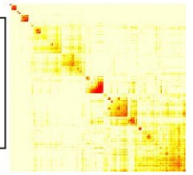
C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

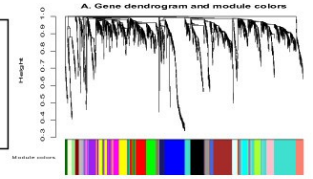
C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

Weighted Gene Correlation Network Analysis (WGCNA)

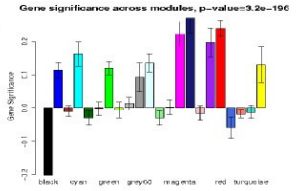
Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression



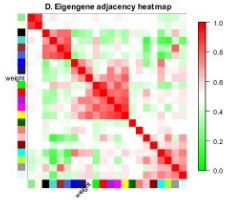
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut



Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules



Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks



Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing