Basic RNASeq DGE analysis using R
HSPH-IID Virtual Workshop

Yaoyu E. Wang, Ph.D.
Quantitative Biomedical Research Center (QBRC)
email: qbrc@hsph.harvard.edu

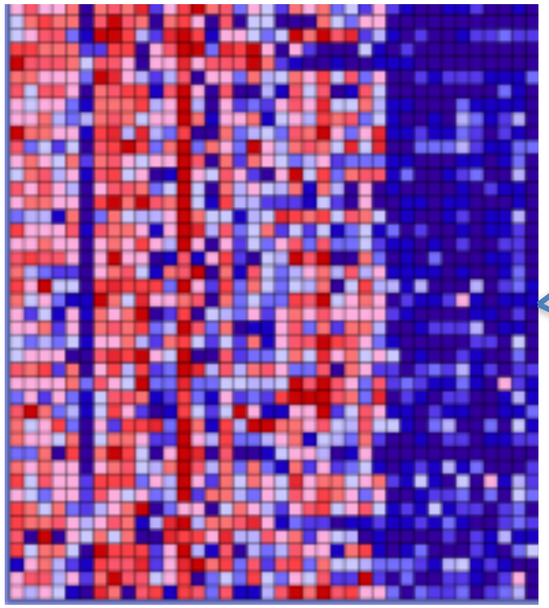
Differential gene expression analysis

Identify genes with statistically significant expression differences between samples of different conditions

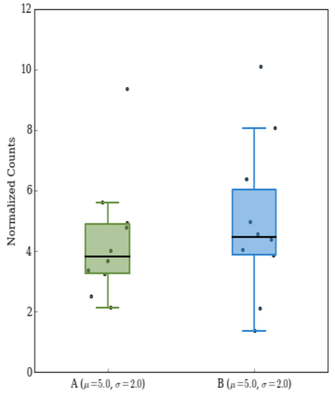
Recap from previous section

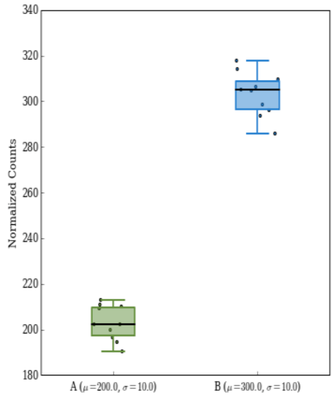
- RNASeq experiment results in short reads (75-150bp) data
- RNASeq data can be quantified by either alignment-based or pseudo-alignment based methods
- Gene counts need to be normalized to remove experimental variation
- Selection of normalization method can affect down stream results
- Unsupervised analysis such as PCA and hierarchical clustering are used for first-pass data exploration

Modeling for Differential Gene Expression



1 test per gene!!!

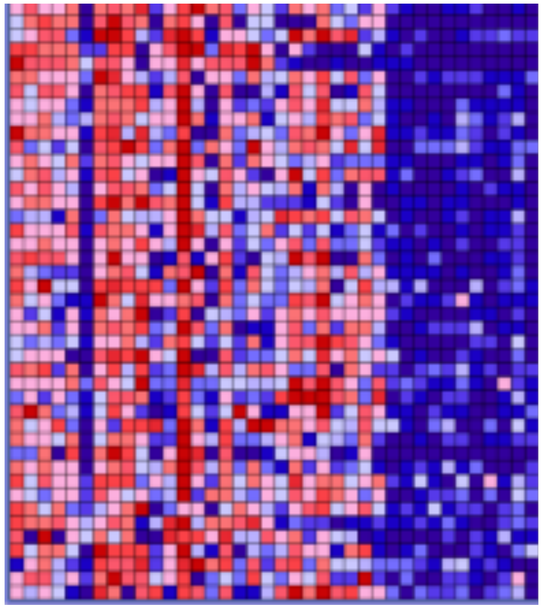




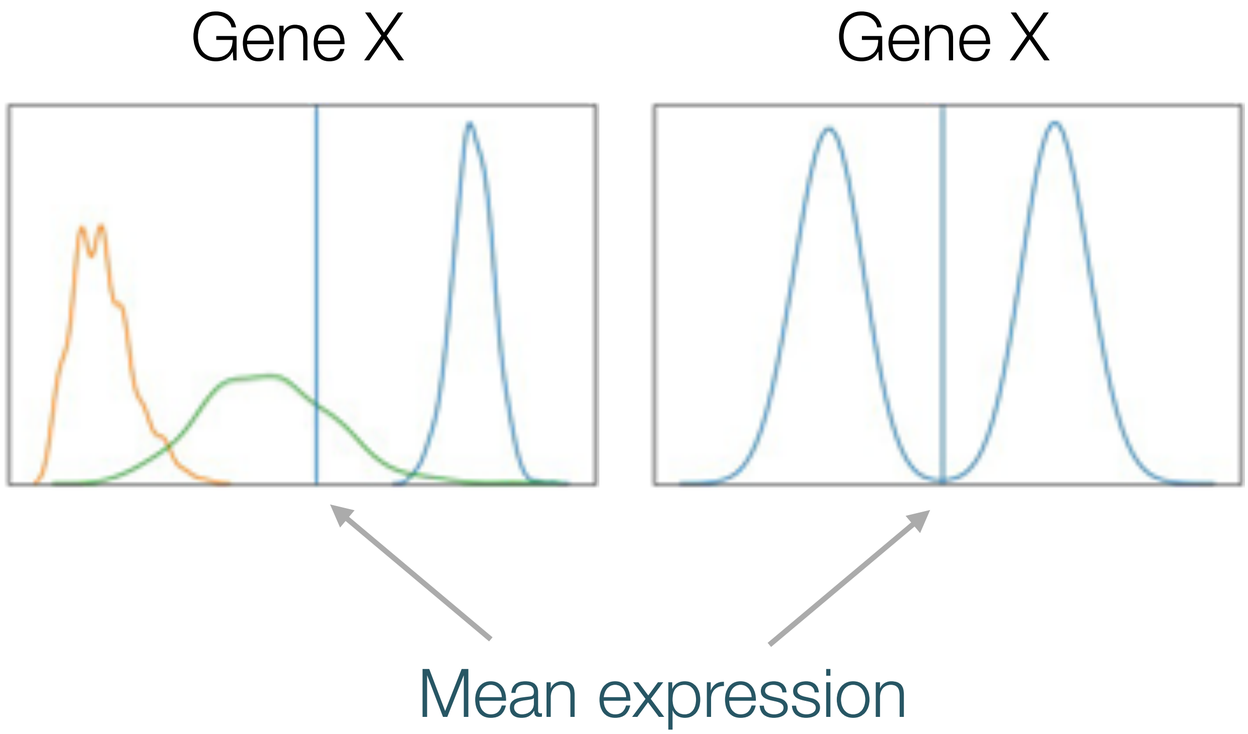
1. Estimate **magnitude** of DGE
   - Report as LogFC (log fold change)
2. Estimate the **significance** of
   - (adjusted) p-values that account for performing thousands of tests

H0: no difference in the read distribution between conditions

Modeling for Differential Gene Expression

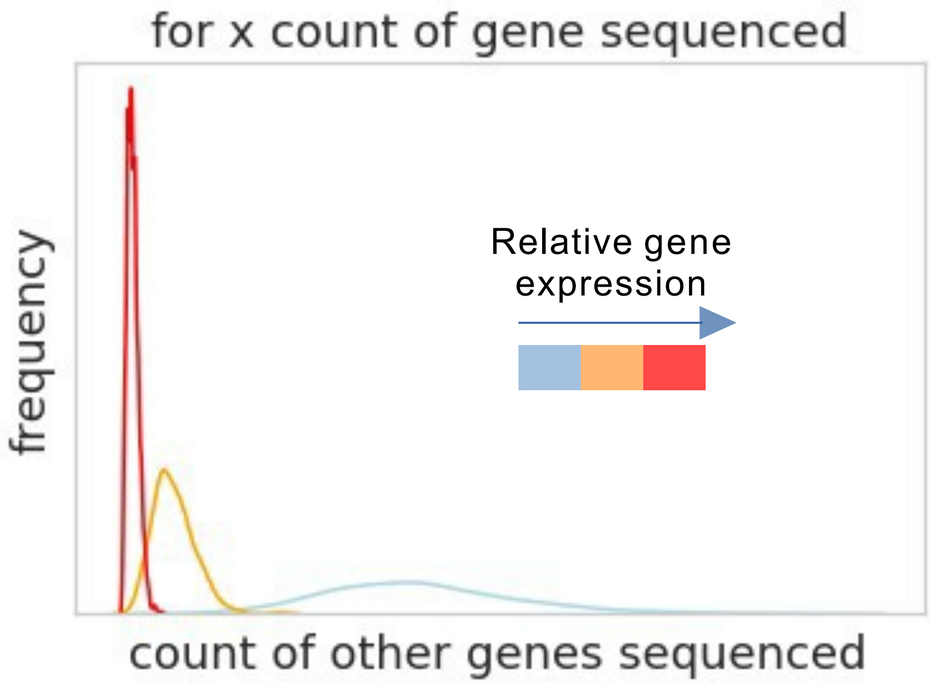


1 Single Expression Value



Modeling for Gene Expression using negative binomial distribution

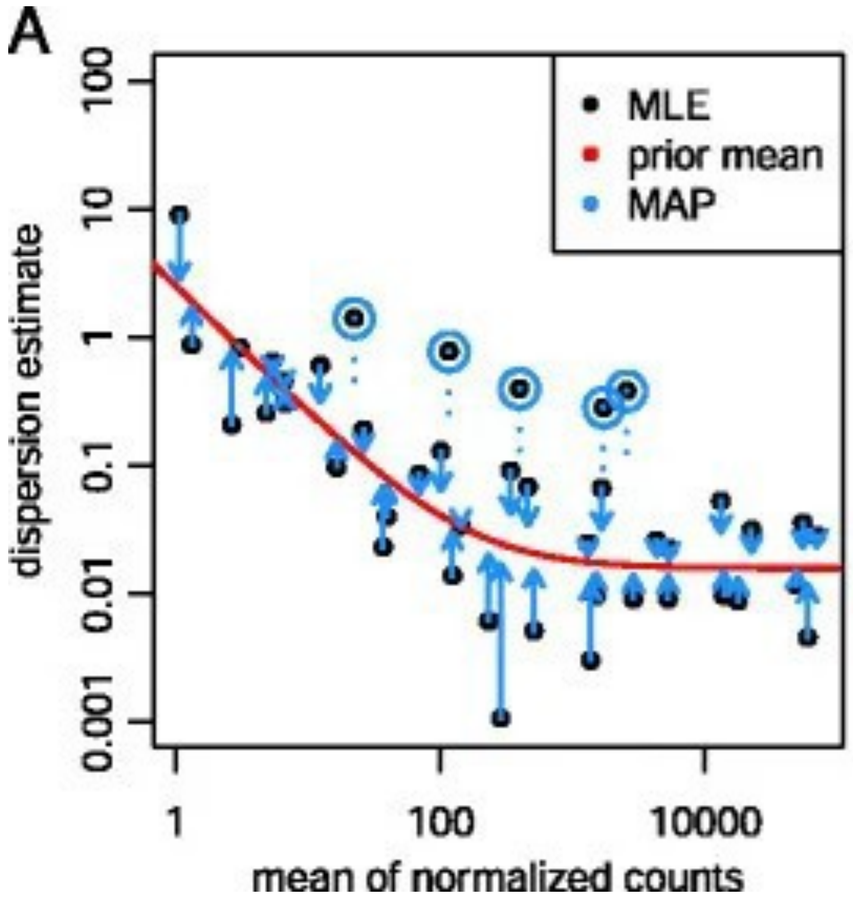
1. Fit a statistical model



Empirically fit a distribution to estimate read count properties by **negative binomial distribution**

Modeling for Gene Expression using negative binomial distribution

1. Shrinkage (of variance)



When Individual gene count is small

- Gene expression has high variance
- High variance = poor statistical power

Reduce the calculated variance (black dots)

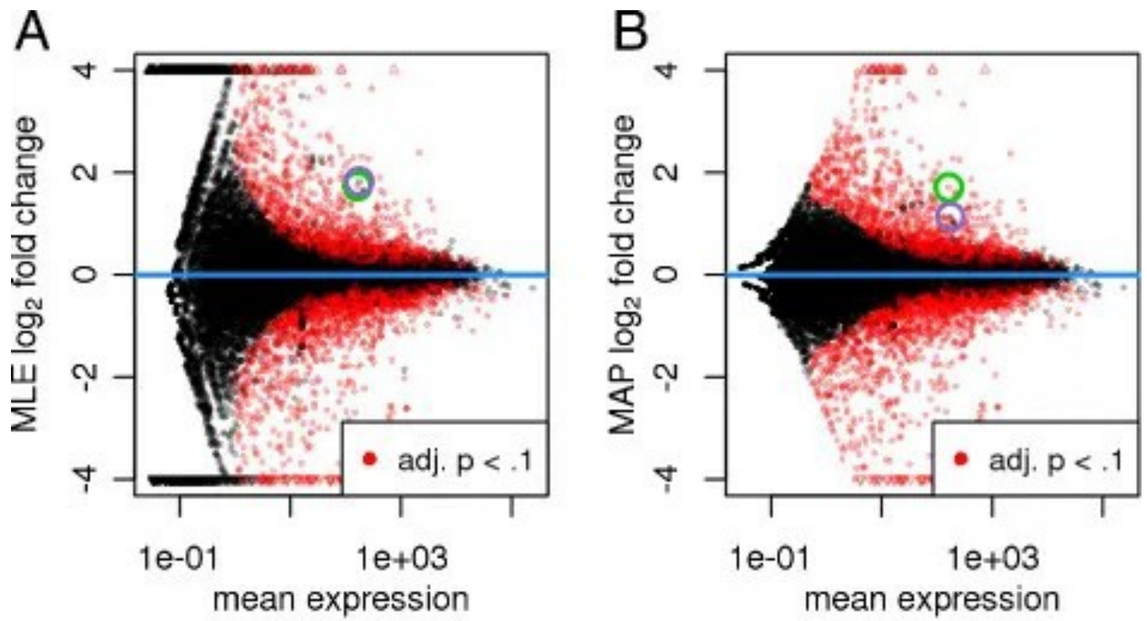
- Use information from other genes to
  - fit a mean dispersion curve (red)
  - Adjust (shrink) variance with this new piece of information (blue arrows)

How shrinkage is done is major differentiator between DGE algorithms (DESEQ2, edgeR, Voom, etc).

Love MI, Genome Biology, 2014

Modeling for Gene Expression using negative binomial distribution

Weighted shrinkage for low counts



No Shrinkage

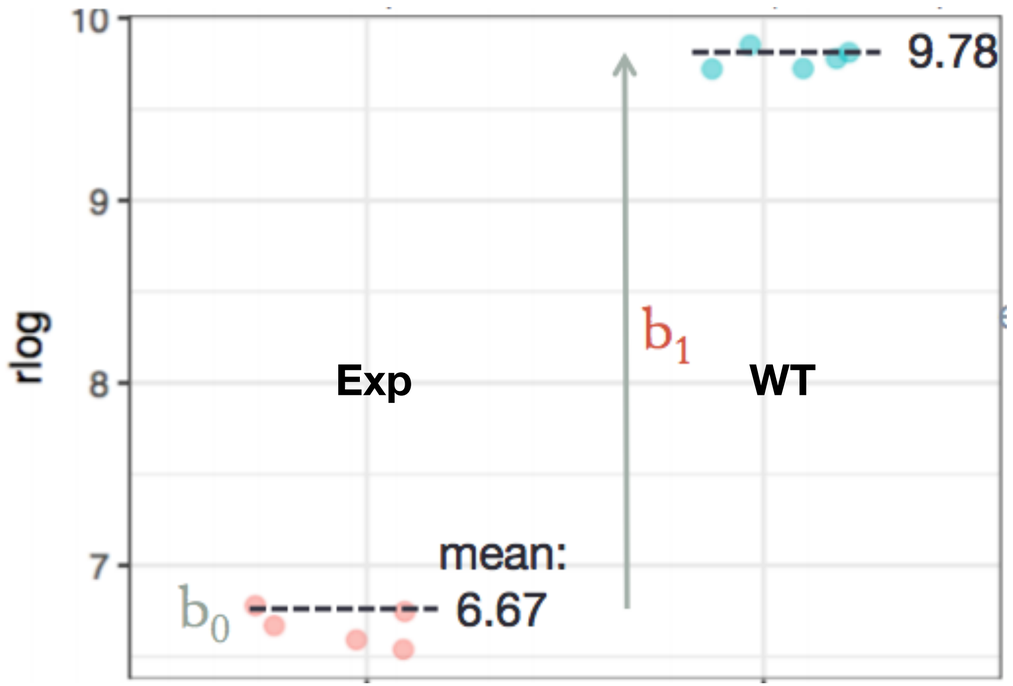
Shrinkage

- Lower counts have intrinsically higher variance
- Weight shrinkage **more** for low count genes

Love MI, Genome Biology, 2014

Modeling for Differential Gene Expression with Linear Model

2. Estimate difference



Estimate the difference between groups using a linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Y : Observed values (expression)
x : Covariates (experiment groups)
β : regression coefficients

Simple Linear Model

KO vs WT: 2 samples, 1 each condition, 1 gene

$$\beta_0 + \beta_1 x_1$$

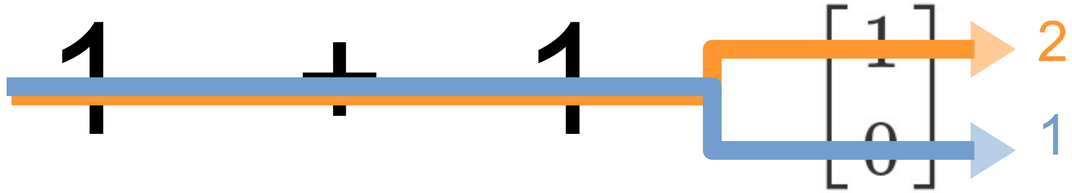
Gene expression at baseline (i.e. WT)

Magnitude of KO effect

KO = 1
WT= 0

Simple Linear Model

KO vs WT: 2 samples, 1 each condition, 1 gene



Two fold change in expression

$$\beta_0 + \beta_1 x_1$$

Gene expression at baseline (i.e. WT)

Magnitude of KO effect

KO = 1

WT = 0

Model more effects in experiment

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Experimental condition

Batch effect

In R: ~ condition + batch + … + time

Hypothesis testing

- Test whether gene expression differences between conditions controlling for covariates are significant. There are two general methods used:

Wald Test:

- Default hypothesis testing method
- Use the estimated standard error of the log2 fold change to test against null hypothesis
- Suitable for contrast in simple linear model: ~condition

Likelihood Ratio Test

- Useful in comparing complex models such as drug-drug interaction model
- Compare a full model against a reduced model to test for reduced term:
  - Full model: ~group+condition
  - Reduce model: ~group

DGE Results

Gene	baseMean	baseMeanA	baseMeanB	foldChange	log2FC	pval	padj
FTL2	94.324	2.319	186.329	80.318	6.327	7.97E-44	2.89E-40
REC8	120.143	229.661	10.626	0.0462	-4.433	4.05E-38	9.32E-35
DLK2	626.928	1026.15	227.706	0.221	-2.171	1.18E-18	1.87E-15
…	…	…	…	…	…	…	…
PDE6b	430.808	301.37	560.239	1.858	0.894	0.328	0.765
LEPREL4	495.854	532.61	459.092	0.862	-0.214	0.328	0.765
NLRP12	4.009	5.466	2.535	0.463	-1.108	0.329	0.766

Commonly used methods DESeq2, edgeR, limma all produce results in similar format

Controlling false-positives by multiple comparisons

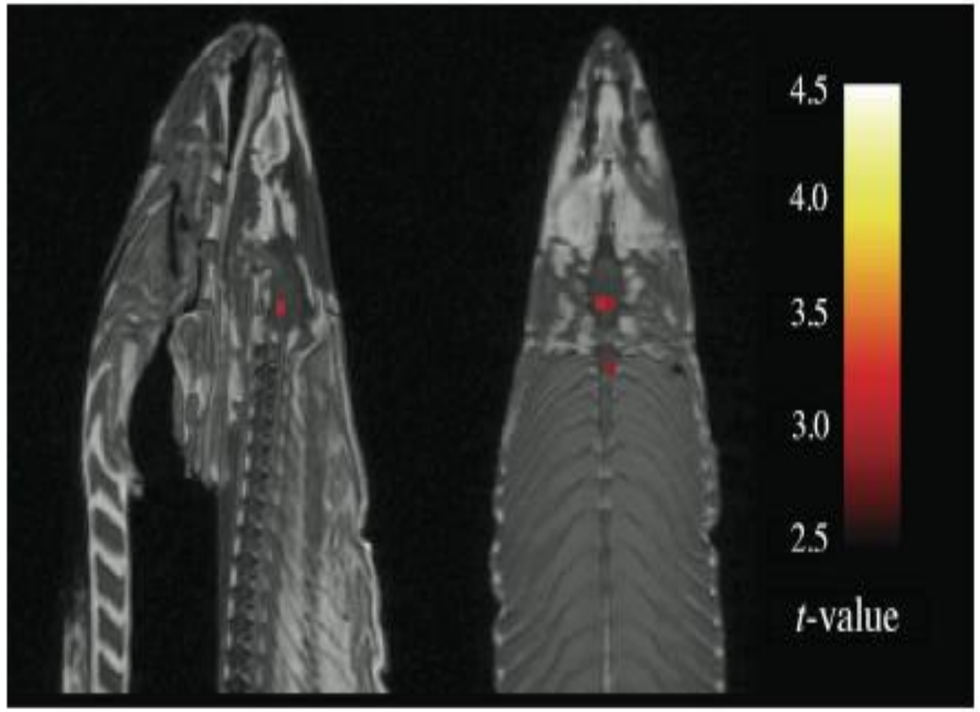
- When the same question is asked thousands of times, some will show up as significant by random

- Most commonly used method for RNASeq is False Discovery Rate (FDR) by Benjamini-Hochberg

$$FDR = Q_e = E[V/(V+R)]$$

V = False Positives
R = True Positives + False Positives

Ask a dead salmon a series of questions...



fMRI with many statistical tests performed (just like testing differential expression on many genes!)

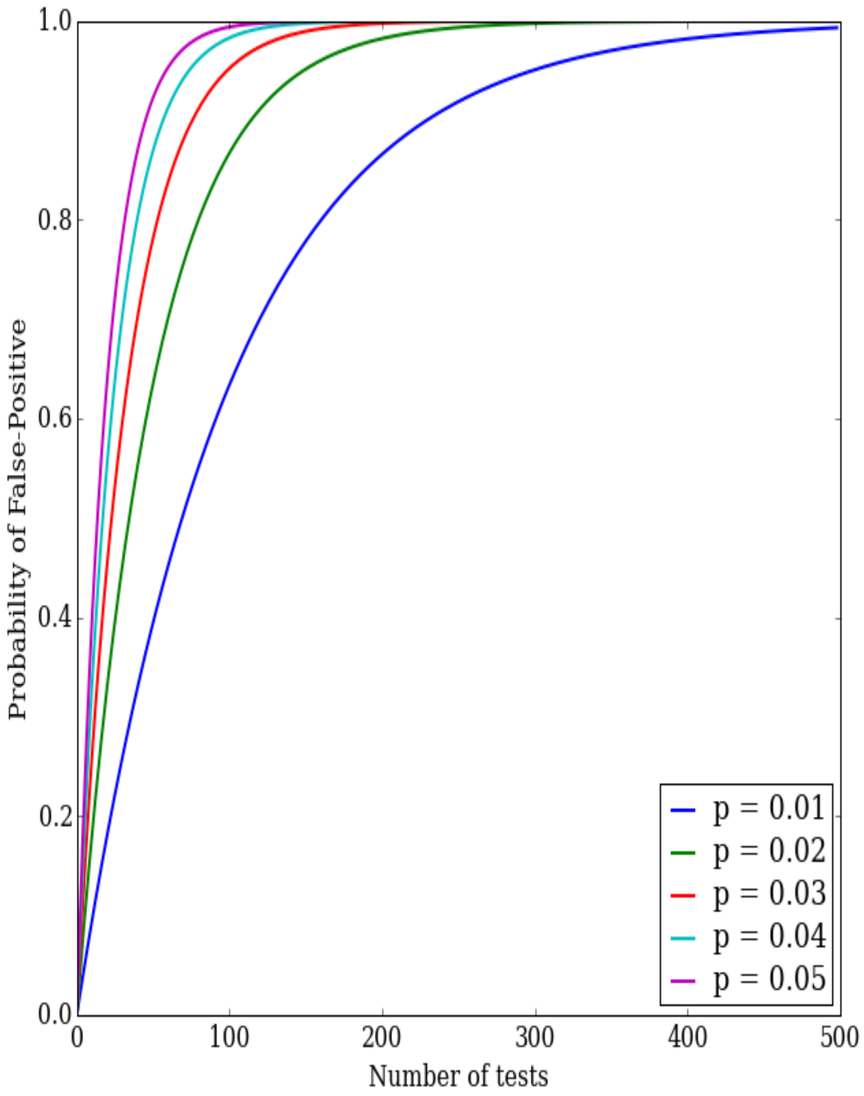
Bennett (2009)

Controlling false-positives

$$f = 1 - (1 - p)^n$$

(Probability of *at least one* false-positive, called FWER)

Aim to control the False-Discovery Rate (FDR), or the proportion of false-discoveries in “all discoveries”

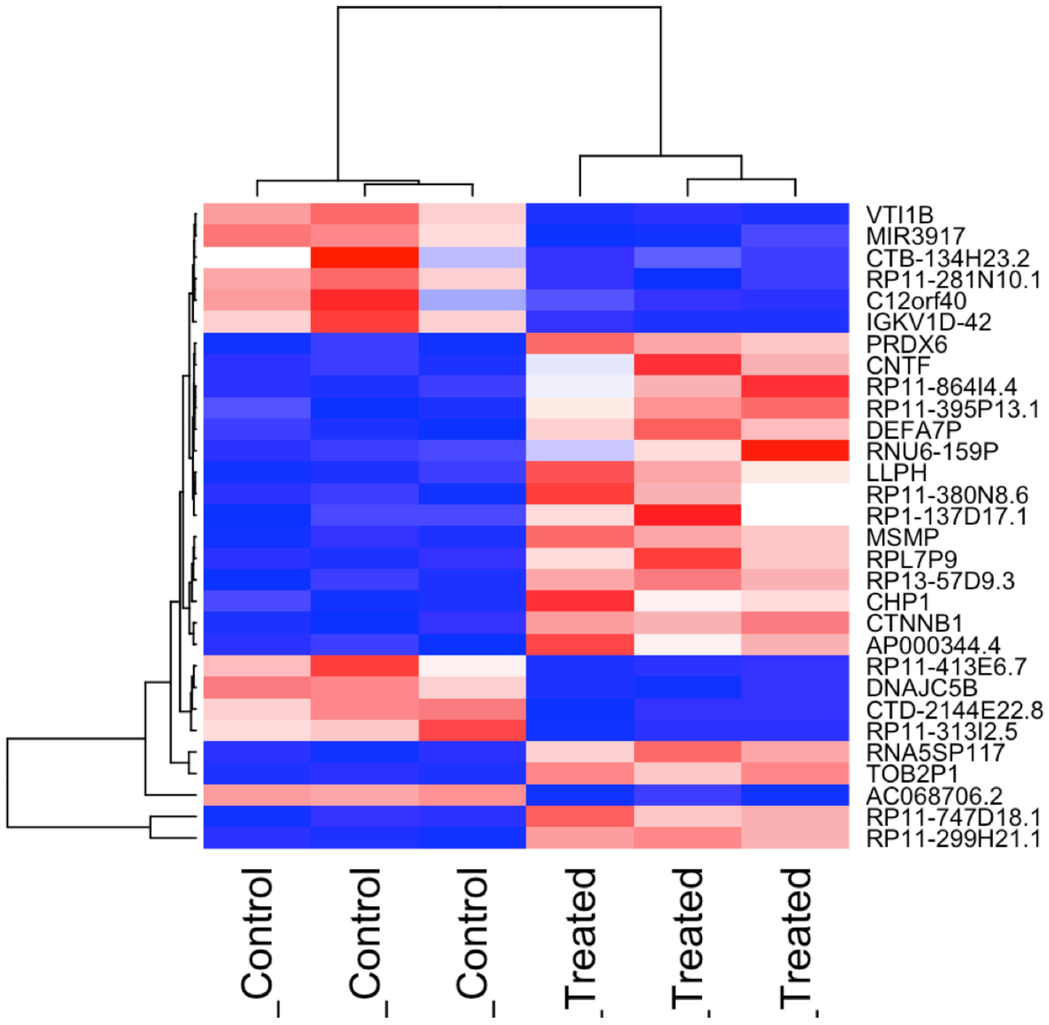


DGE Results Examination

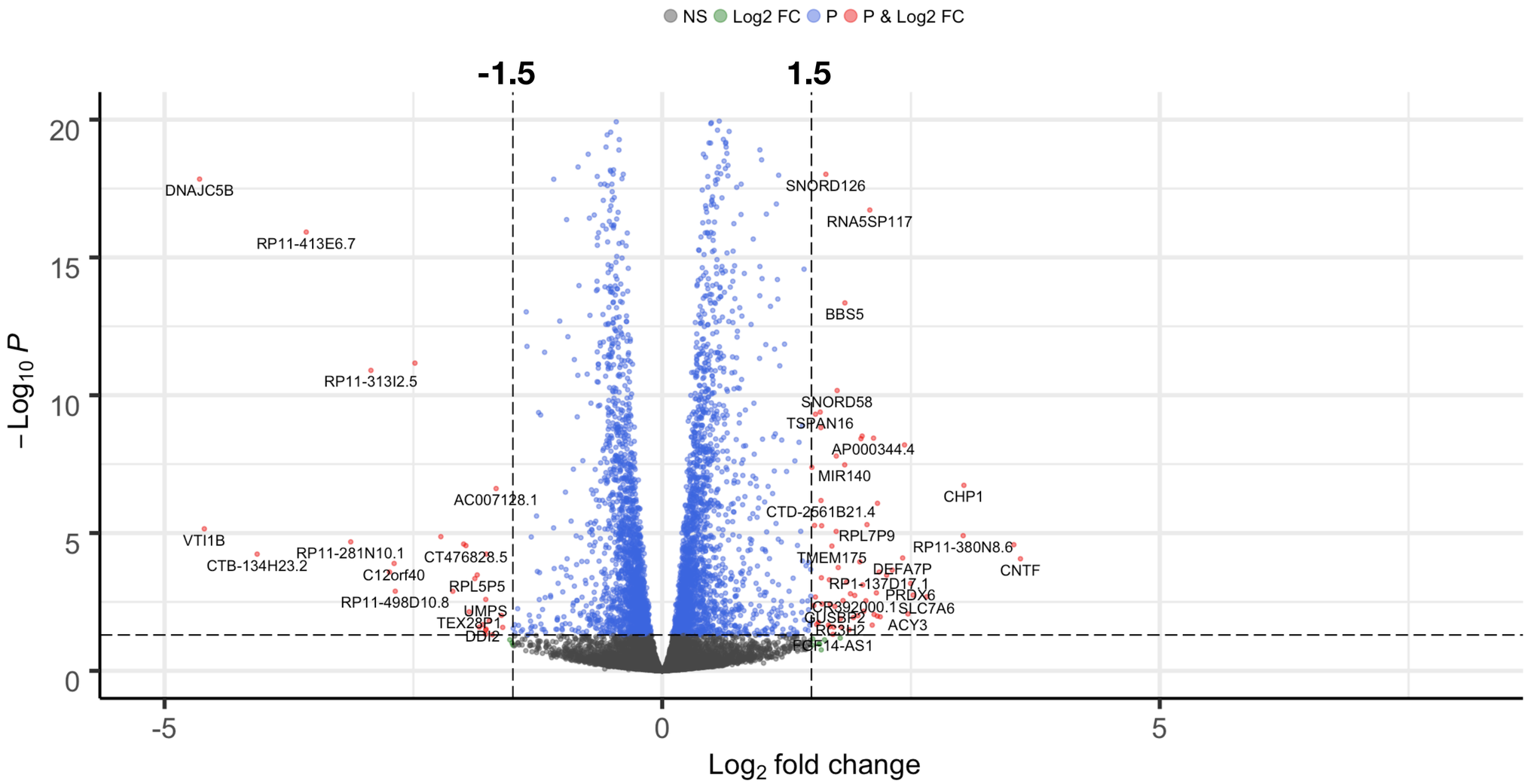
Gene	baseMean	baseMeanA	baseMeanB	foldChange	log2FC	pval	padj
FTL2	94.324	2.319	186.329	80.318	6.327	7.97E-44	2.89E-40
REC8	120.143	229.661	10.626	0.0462	-4.433	4.05E-38	9.32E-35
DLK2	626.928	1026.15	227.706	0.221	-2.171	1.18E-18	1.87E-15
…	…	…	…	…	…	…	…
PDE6b	430.808	301.37	560.239	1.858	0.894	0.328	0.765
LEPREL4	495.854	532.61	459.092	0.862	-0.214	0.328	0.765
NLRP12	4.009	5.466	2.535	0.463	-1.108	0.329	0.766

Filter DGE result table by Log2FC (usually > 1.2) or adjusted P-value

DGE Results: Heatmap



DGE Results - Volcano Plots



Functional Enrichment Analysis

Putting differential expressed genes into biological context using gene annotation databases

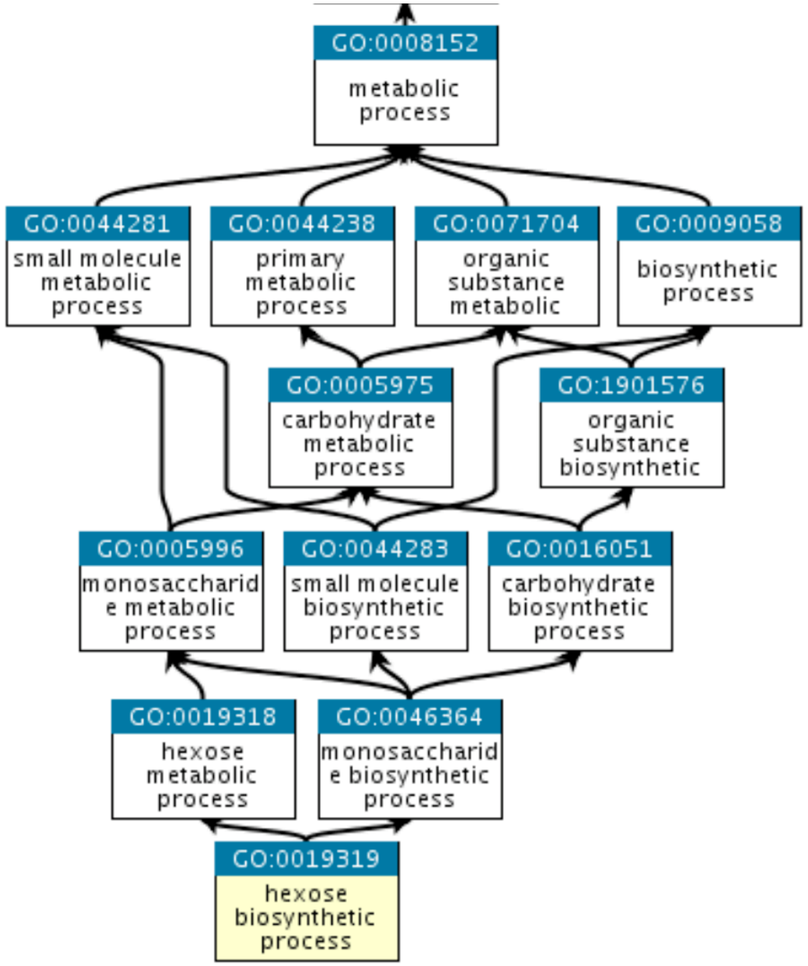
GENEONTOLOGY
Unifying Biology

Gene Ontology Database (GO) - terms that group genes into sets of classes by their annotations

1. **Molecular Function:** Molecular-level activities performed by gene products, such as “catalysis” or “transport”

2. **Cellular Component:** The locations relative to cellular structures in which a gene product performs a function (e.g. “mitochondrion”, “ribosome”)

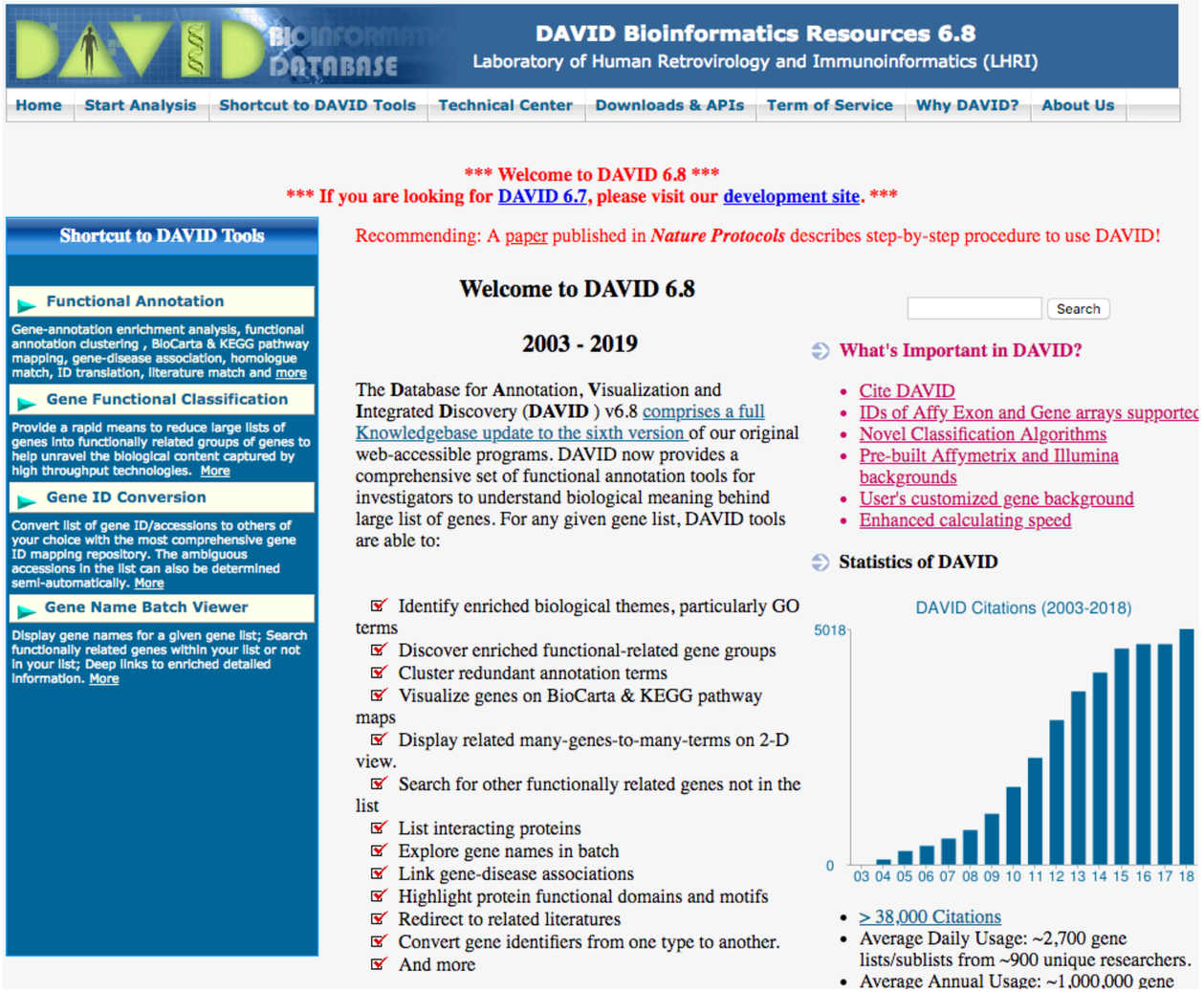
3. **Biological Process:** The larger processes, or ‘biological programs’ accomplished by multiple molecular activities (e.g. “DNA repair”, “signal transduction”)



Loosely hierarchical GO Term structure

Ashburner et al. Nat. Genet. (2000)

Functional Enrichment Analysis with DAVID



Use a modified **Fisher Exact Test** to determine if there is enrichment

Confusion Matrix	Number of genes is DGE	Number of genes is not DGE
Number of genes in pathway *y*	76	20
Number of genes not in pathway *y*	2	29920

p<0.00001!!!

Conclusion: Pathway *y* is differentially regulated

https://david.ncifcrf.gov/

Sherman, et. al, Nat. Prot.(2008)

Functional Enrichment Analysis with DAVID

Use a modified **Fisher Exact Test** to determine if there is enrichment

Confusion Matrix	Number of genes is DGE	Number of genes is not DGE
Number of genes in pathway *y*	76	20
Number of genes not in pathway *y*	2	29920

p<0.00001!!!

Conclusion: Pathway *y* is differentially regulated

MSigDB
Molecular Signatures Database

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets

- Curated Gene Sets from literatures, such as functional pathway (KEGG), gene functional groups. **Most commonly used gene set class**
- Contain domain specific gene sets (H, C6, C7)
- Human genome location (C1) Predicted gene sets (C2, C4)
- GO term (C5)

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 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

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

C5 **GO gene sets** consist of genes annotated by the same GO terms.

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

C7 **immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.

Detecting modest but coordinate changes

The goal of GSEA is to detect modest but coordinated changes in pre-specified sets of related genes by using all genes and their statistical variation values

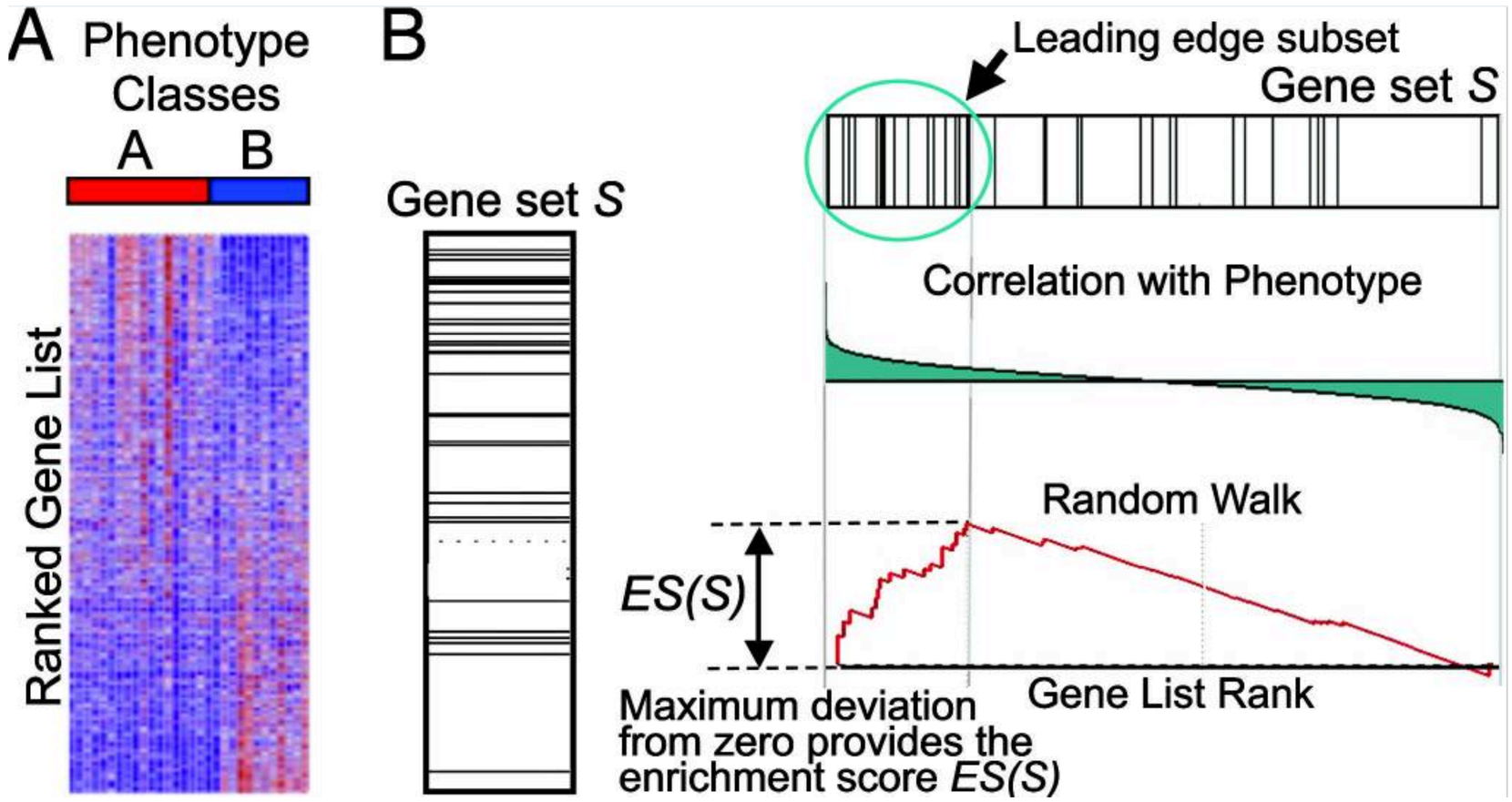
Step 0: Sort Genes into a ranked gene list

Step 1: Calculate Enrichment Score: Compute cumulative sum over ranked genes by summing statistics of gene in a set, and subtracting statistics of genes outside of the set

Step 2: Assess significance using Permutation Test: permute sample phenotype labels

Step 3: Adjust for multiple hypothesis testing: using FDR correction

Detecting modest but coordinate changes



ES: reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes by running-sum statistic

Interpreting GSEA Results

Leading edge analysis

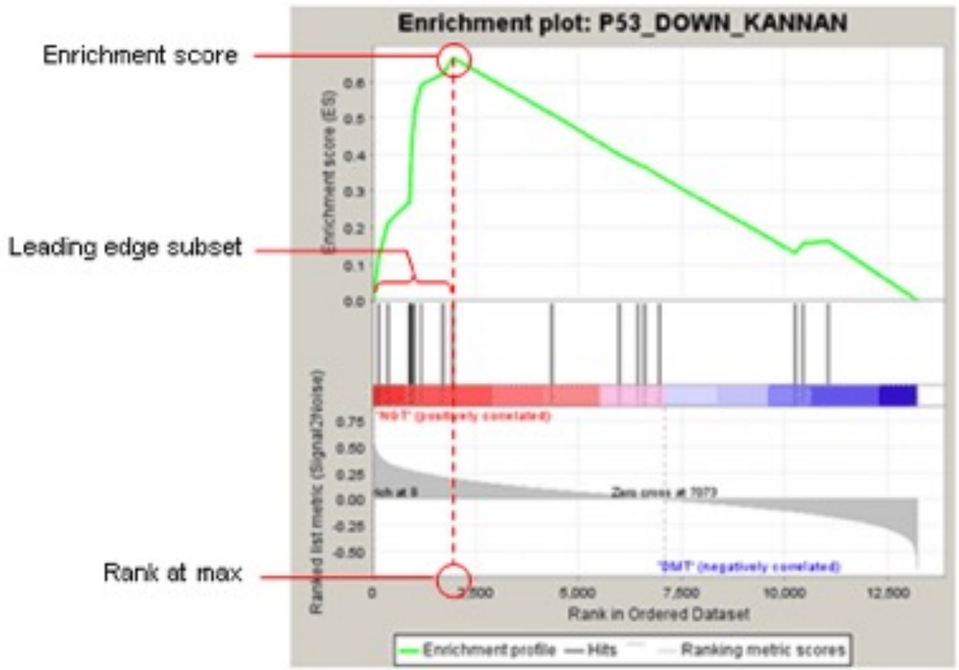


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

Table: Gene sets enriched in phenotype NGT (17 samples) [plain text format]

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	P53_DOWN_KANNAN	Details ...	17	0.67	1.95	0.000	0.017	1.000	1953	tags=53%, list=15%, signal=62%
2	ELECTRON_TRANSPORT_CHAIN	Details ...	81	0.61	1.76	0.027	0.059	1.000	3047	tags=59%, list=23%, signal=77%

Nominal ES (NES): the enrichment score for the gene set after it has been normalized across analyzed gene sets:

$$\text{NES} = \frac{\textit{Actual ES}}{\textit{mean(ESs against all permutations of the dataset)}}$$

Example Analysis:

Input Data

- 3 WT vs 3 Treated Cell line RNA-Seq data
- Single End 75bp RNA-Seq, STAR aligned, HTSeq quantified raw count

We will perform:

1. Install and load libraries from CRAN and Bioconductor
2. Load Data
3. PCA on raw count data
4. Hierarchical Clustering Tree
5. DESeq2 to perform differential gene expression analysis
6. Heatmap
7. Volcano plot
8. GSEA