Transcriptional Regulation and Expression Facility

trex_info@cornell.edu

Take our Survey!

Sign up for our List-Serv!

*Send an email message to

TREX-GENEREG-L-request@cornell.edu

with **"join"** as the subject

Upcoming Events

Sign up
at our website
to indicate interest
and receive priority
admission

- **TREx Workshops!**

  RNA Extraction: 1 day workshop – early October

  RNA-seq walkthrough: 4 week workshop – mid October

  Biological Insights: 1 day workshop – early December

- **Tech Talks**: 4th Tuesday of the Month

- **BRC Bioinformatics Facility Workshops**

  Introduction to BioHPC Cloud (September 9th+11th)

  Linux for Biologists (September 16th-October 2nd, M+W)

  RNA-Seq Data Analysis (October 14th-30th, M+W)

Coming Soon to TREx

- **New and Improved Project Submission Form**

  Available on our web site in early September

- **New service: ATACseq**

  Assay for **T**ransposase-**A**ccessible **C**hromatin by sequencing

  Identify promoters, enhancers, motifs enriched in open chromatin
  expressed genes, 'poised' genes (vs RNAseq)

  Researcher provides intact nuclei (preserving native state)

  Goal: launch by the end of 2019

  Contact us if you are interested in early access (beta-testing)

  trex_info@cornell.edu

Transcriptional Regulation and Expression Facility

“RNA-SEQ ANALYSIS: BIOLOGICAL DISCOVERY”

Jen Grenier

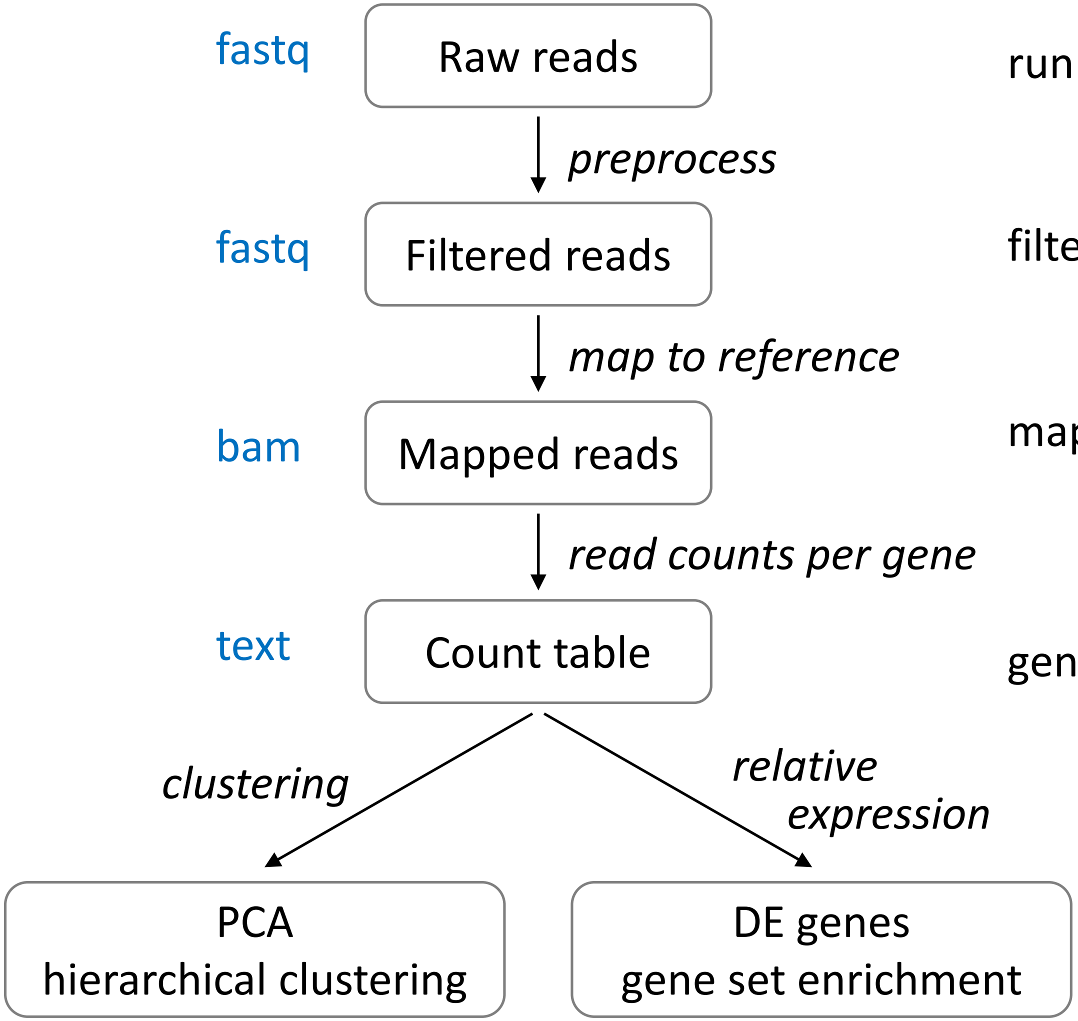
Ann Tate

Christine Butler

Faraz Ahmed

trex_info@cornell.edu

RNAseq Analysis: Reads to Counts



Pipeline

fastq — Raw reads

preprocess

fastq — Filtered reads

map to reference

bam — Mapped reads

read counts per gene

text — Count table

clustering → PCA
hierarchical clustering

relative expression → DE genes
gene set enrichment

Data QC

run stats, fastQC

filtered read count, fastQC

mapping rate: genome and transcriptome

gene body distribution (3’ bias?)

RNAseq Analysis

Unsupervised

Analysis of expressed, variable genes
independent of sample groups

Principal components analysis
Hierarchical clustering

Global signal

Supervised

Analysis of differential expression
between sample groups

Relative expression (A vs B)
log2(fold-change)
DE genes
Gene set enrichment analysis

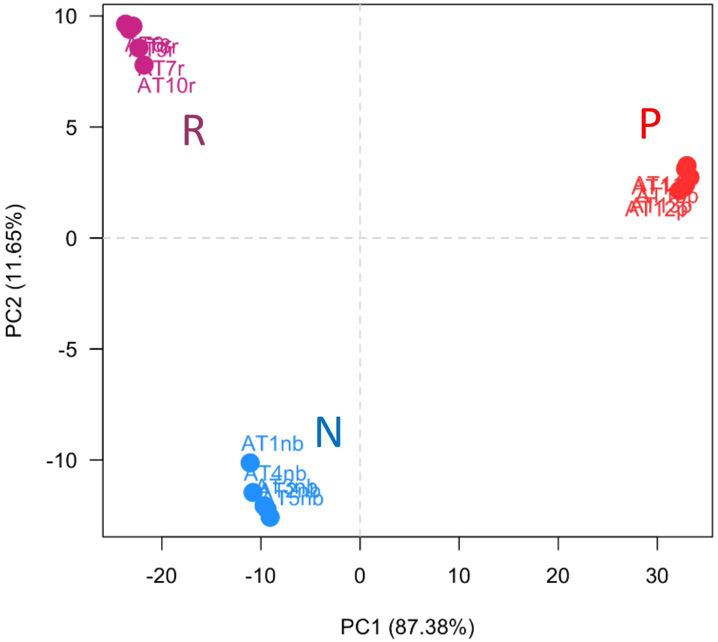
Experimental signal

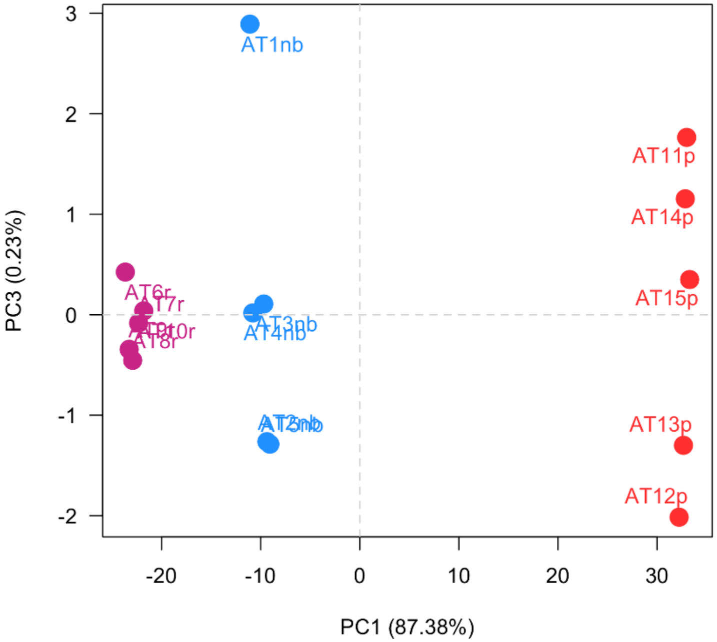
RNAseq Analysis: Clustering

Unsupervised comparison of expression profiles between samples

PCA: Dimensionality reduction

~10,000 expressed genes for 15 samples → 15 principal components
PC1 explains the greatest amount of variation in the dataset, then PC2, …
Samples with similar principal components have more similar profiles



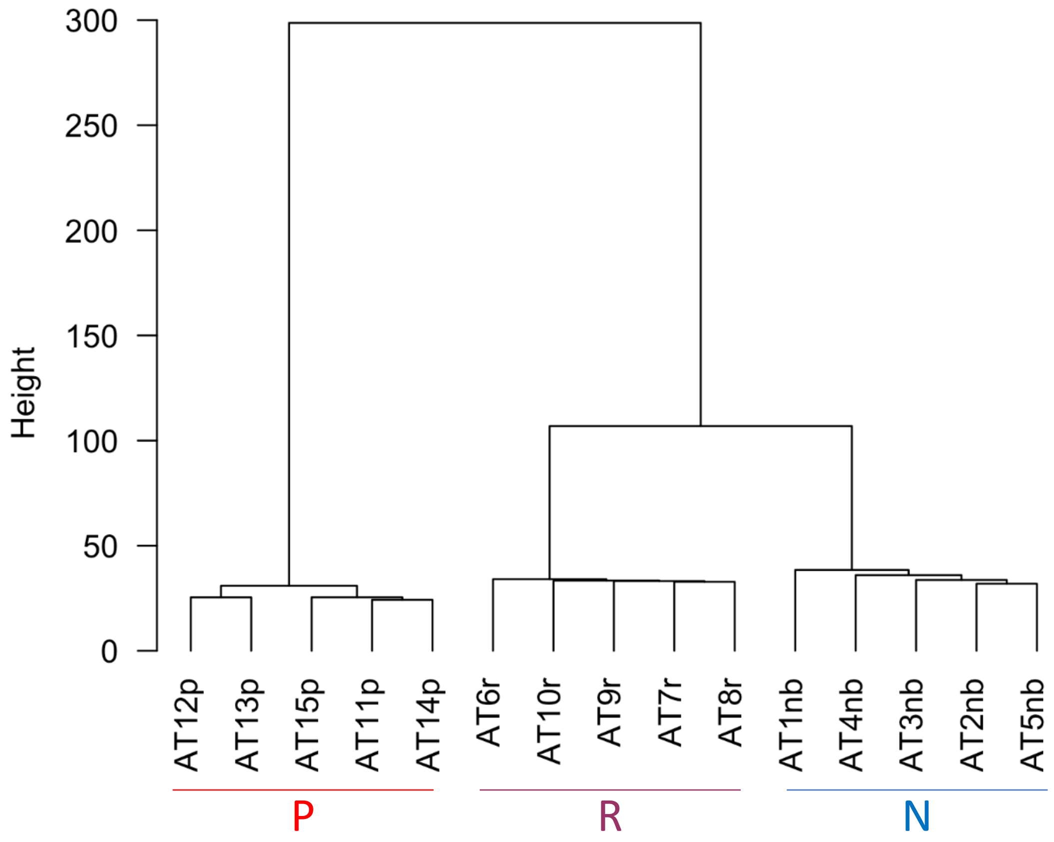


RNAseq Analysis: Clustering

Unsupervised comparison of expression profiles between samples

Hierarchical clustering

Distance matrix → sample ‘tree’

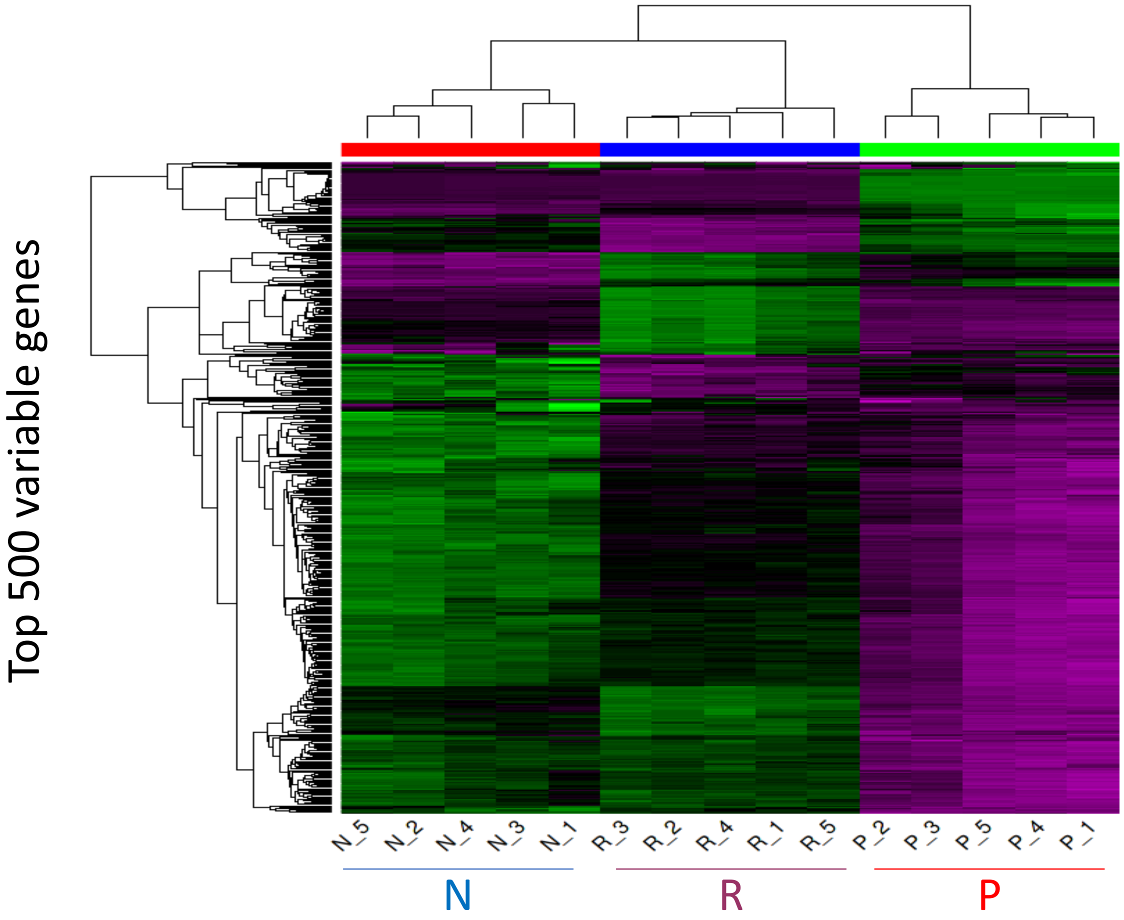


RNAseq Analysis: Clustering

Unsupervised comparison of expression profiles between samples

2D Hierarchical clustering

Distance matrices → sample 'tree' **and** gene 'tree' with heatmap



row-normalized heatmap

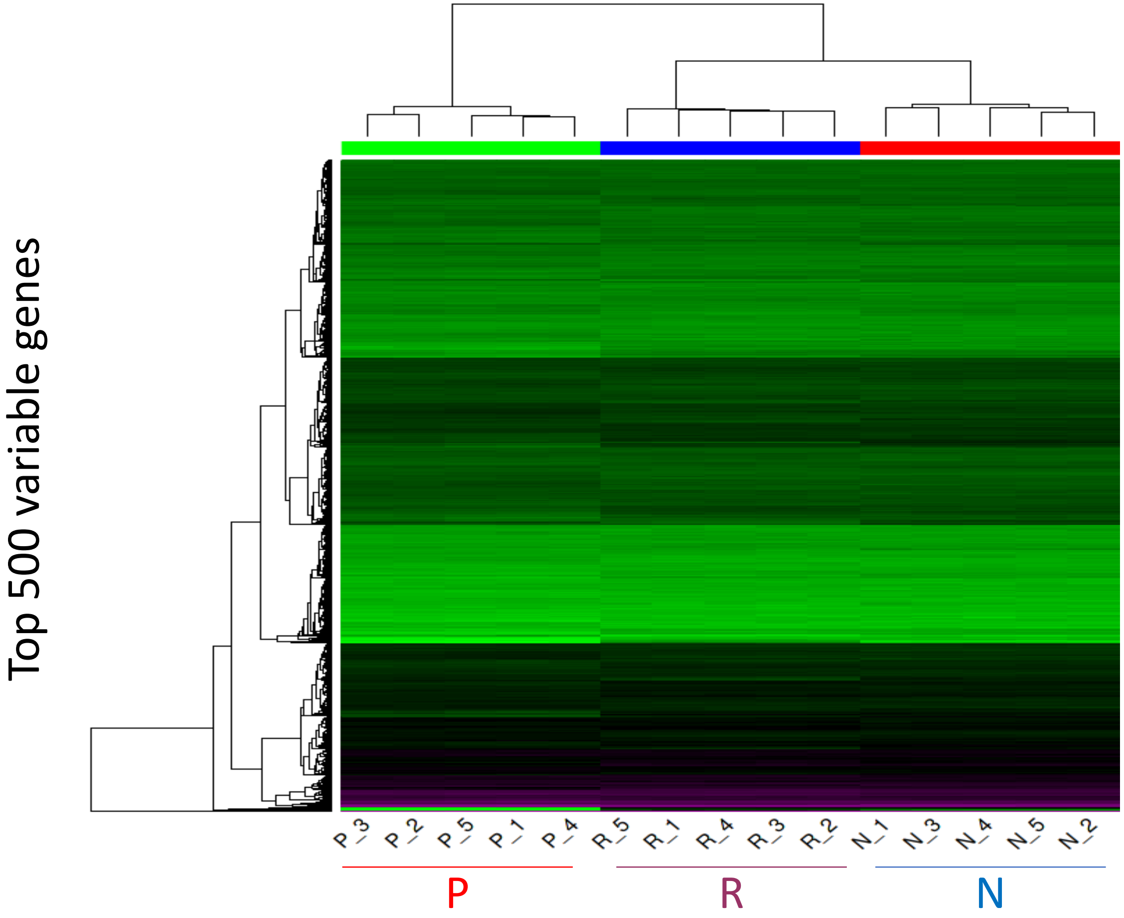
gene clustering: ***differences between samples***

RNAseq Analysis: Clustering

Unsupervised comparison of expression profiles between samples

2D Hierarchical clustering

Distance matrices → sample ‘tree’ **and** gene ‘tree’ with heatmap



CPM heatmap

gene clustering: ***expression level***

RNAseq Analysis: Clustering

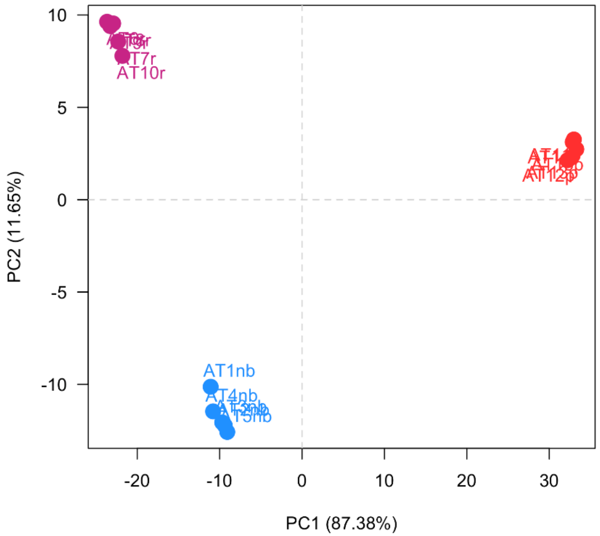
Software tools

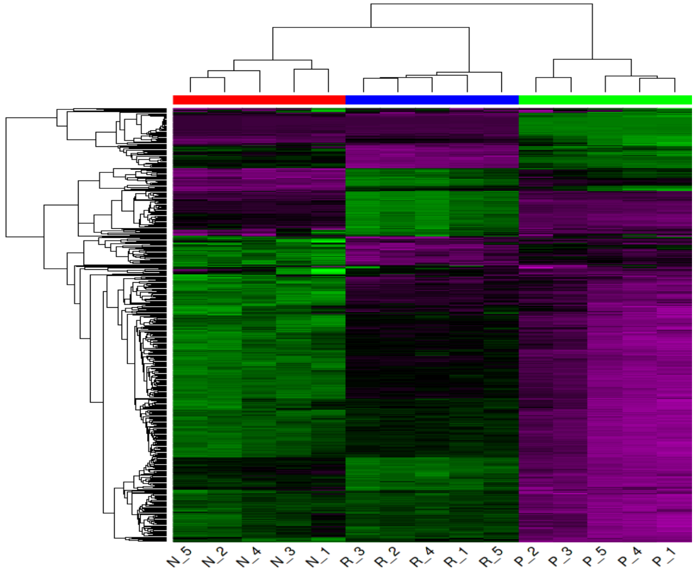
R (RStudio)

★ IDEP

JMP (SAS)

★ Heatmapper.ca





RNAseq Analysis

Unsupervised

Analysis of expressed, variable genes
independent of sample groups

Principal components analysis
Hierarchical clustering

Global signal

Supervised

Analysis of differential expression
between sample groups

Relative expression (A vs B)
log2(fold-change)
DE genes
Gene set enrichment analysis

Experimental signal

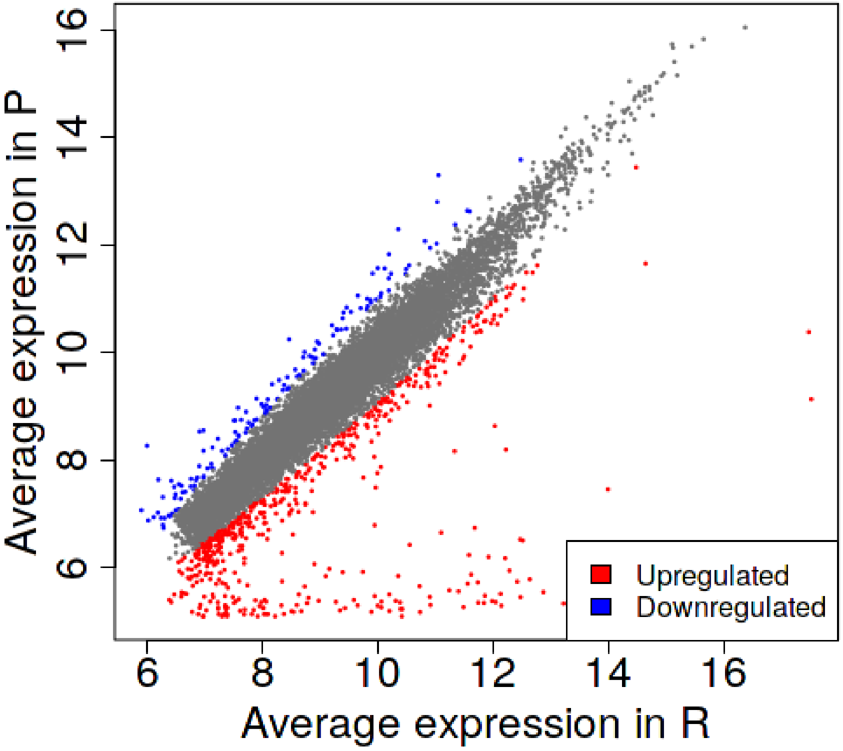
RNAseq: Relative Expression

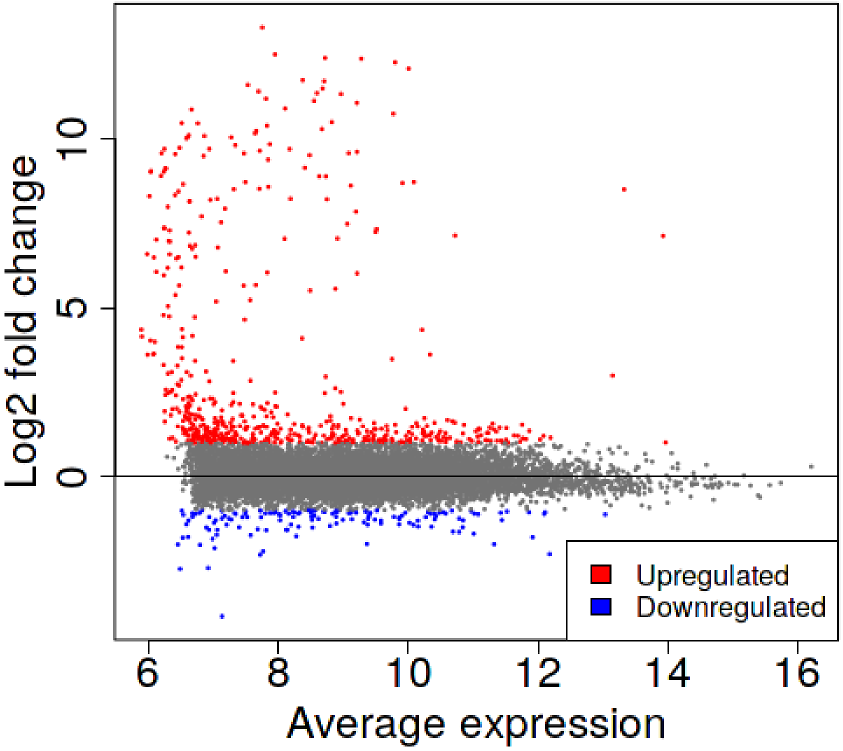
Supervised comparison of expression profiles between samples

Statistical test for differential expression: Appropriate statistical model for RNAseq data

Non-uniform mean-variance relationships → negative binomial distribution

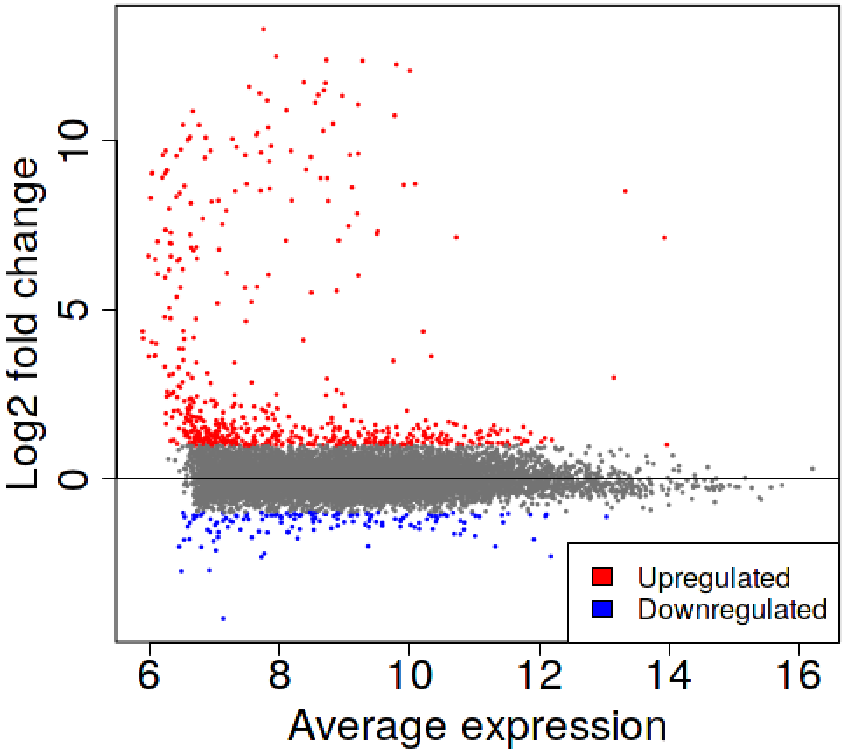
Software: DEseq2, EdgeR, cuffdiff





RNAseq: Biological Discovery

What is interesting / important about differentially expressed genes?



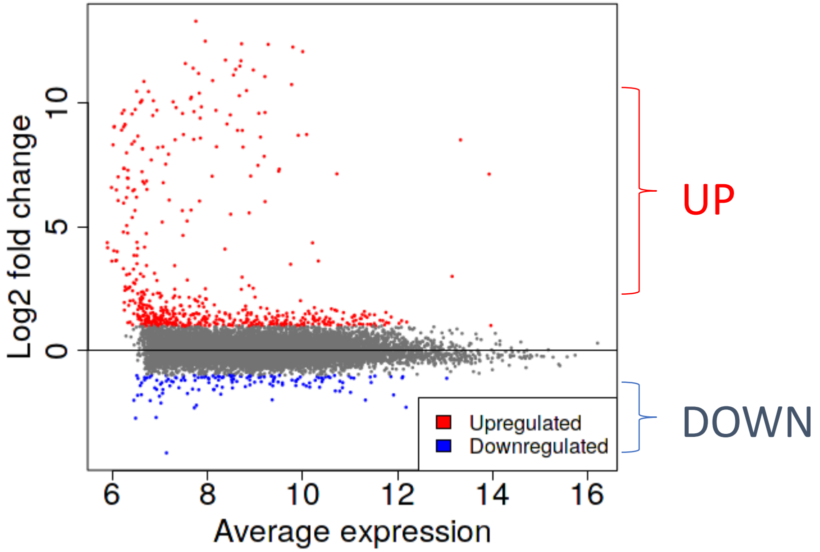
Enrichment
in upregulated
genes

Enrichment
in downregulated
genes

RNAseq: Biological Discovery

DE gene enrichment: Software tools

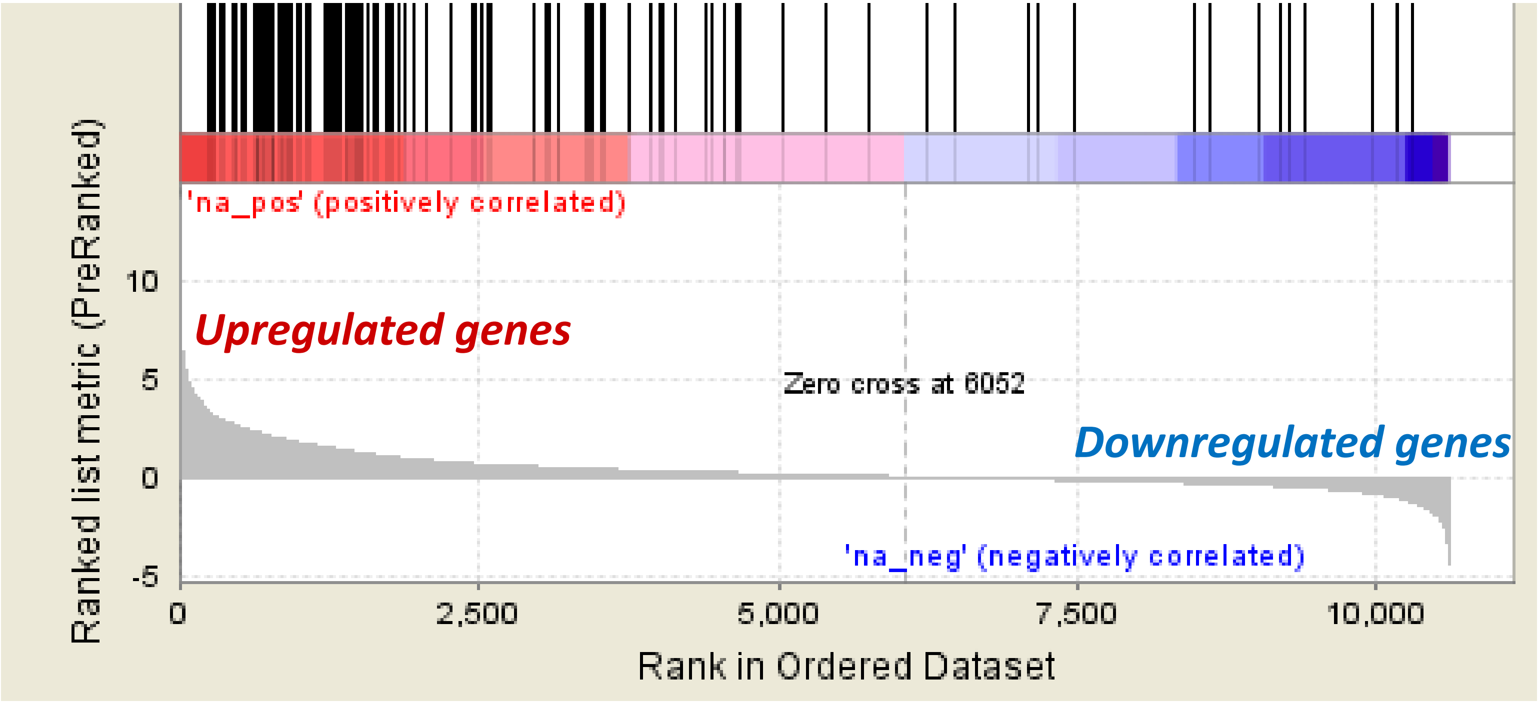
- ★ **Panther**
- **DAVID**
- **Reactome**



RNAseq: Biological Discovery

Gene Set Enrichment Analysis (GSEA)

"A computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states."



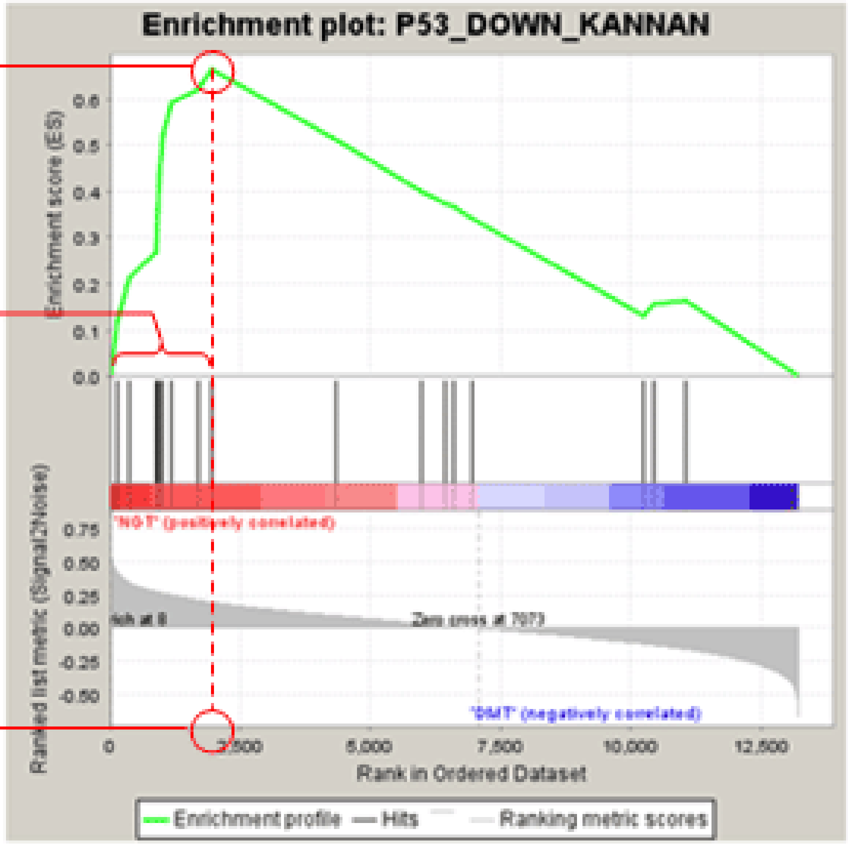
Upregulated genes

Downregulated genes

Genes ranked by log2FC

RNAseq: Biological Discovery

GSEA Enrichment Plot



Enrichment score

Leading edge subset

Rank at max

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

RNAseq: Biological Discovery

Running GSEA for RNAseq

.rnk file col1 = gene names/IDs
col2 = log2FC
use all expressed genes (~10,000 rows)

optional
.gmt file custom gene set
or use built-in Molecular Signatures DB

.rnk file gene identifiers must match gene set!

Use parameters recommended for RNAseq