MSIB32500 Advanced Bioinformatics Fall 2017

RNAseq Data Analysis and Clinical Applications, Part II

Riyue Bao, Ph.D.
Research Assistant Professor (Bioinformatics)
Center for Research Informatics & Department of Pediatrics
The University of Chicago

Outline

- Part I (11/18/2017)
  - Introduction to RNAseq technology and clinical applications
  - Hands on: From raw data to gene expression quantification
- Part II (11/25/2017)
  - Differential gene expression analysis and data visualization
  - Hands on: Identification of genes and pathways significantly changed under condition
  - ***Homework assignment***
  - ***Thanksgiving week – Gleacher center physically closed. Class will be on WebEx.***
- Part III (12/02/2017)
  - How to associate gene expression data with clinical outcome
  - Hands on: Use gene expression data to discover tumor subtypes and survival analysis

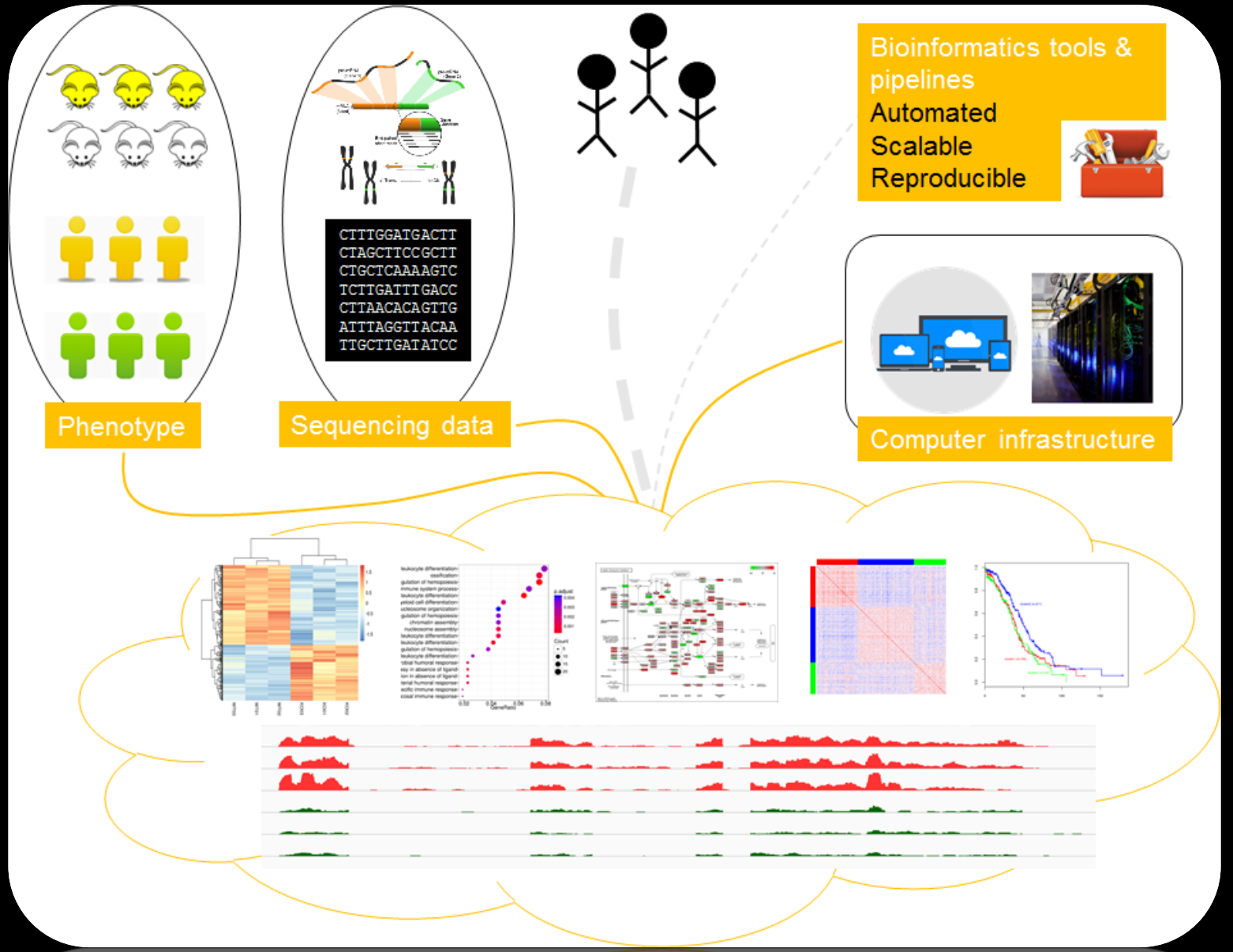
Class materials

- GitHub
  - https://github.com/MScBiomedicalInformatics/MSIB32500
  - This lecture note contains the same contents as the notebook. In addition, the notebook also contains hands-on materials
  - **lecture9.pdf**
  - **Handson9.Rmd**

- Rstudio (or R console) on personal computers (hands on practice)

Objective

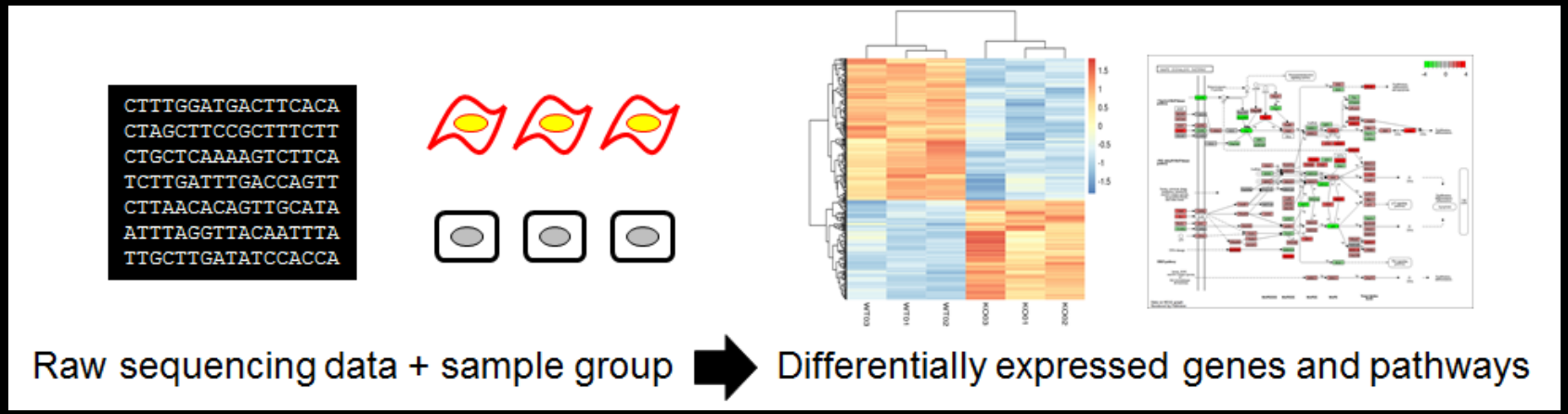
- *(Recap from last class) Rmarkdown HTML fixed; Alignment visualization; Reference databases*
- Detect genes differentially expressed between conditions
- Identify pathways / network enriched in genes of interest
- Generate high-quality figures for publication (PCA, heatmap, sample/gene cluster, GO/pathways, etc.)
- Become familiar with running commands in R / Rstudio


Phenotype
CTTTGGATGACTT
CTAGCTTCCGCTT
CTGCTCAAAAGTC
TCTTGATTTGACC
CTTAACACAGTTG
ATTTAGGTTACAA
TTGCTTGATATCC
Sequencing data
Bioinformatics tools & pipelines
Automated
Scalable
Reproducible
Computer infrastructure

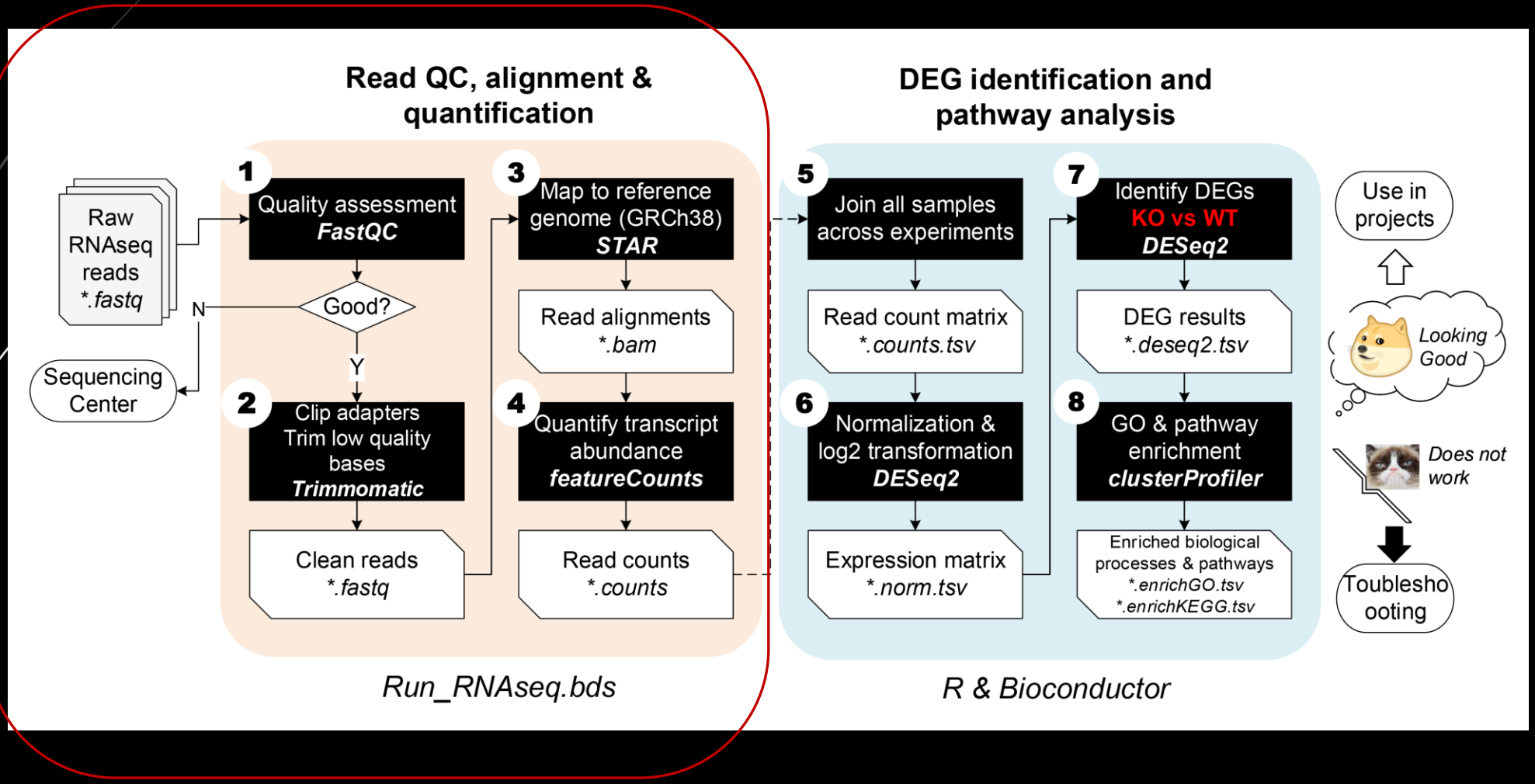
How to perform RNAseq analysis

The good-practice analysis protocol takes 8 major steps.

- **01-04**: From raw sequencing to transcript quantification
- **05-08**: DEG and pathway analysis (11/25, part II)



How to perform RNAseq analysis



IGV (Integrative Genome Viewer)

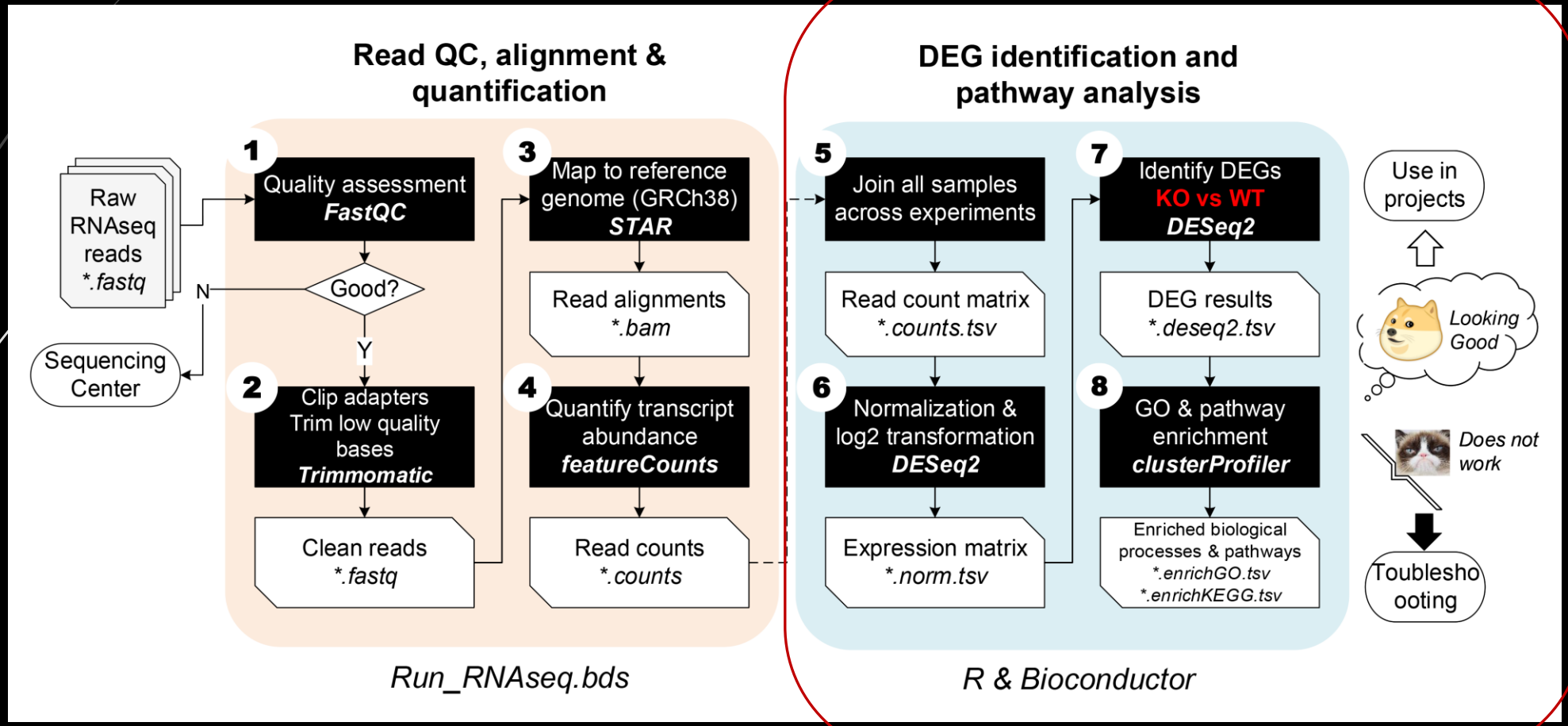
http://software.broadinstitute.org/software/igv/home

- Load existing genomes, or generate custom genomes
- Visualize standard file formats
  - BAM
  - BED
  - GTF
  - … and more!

Reference databases

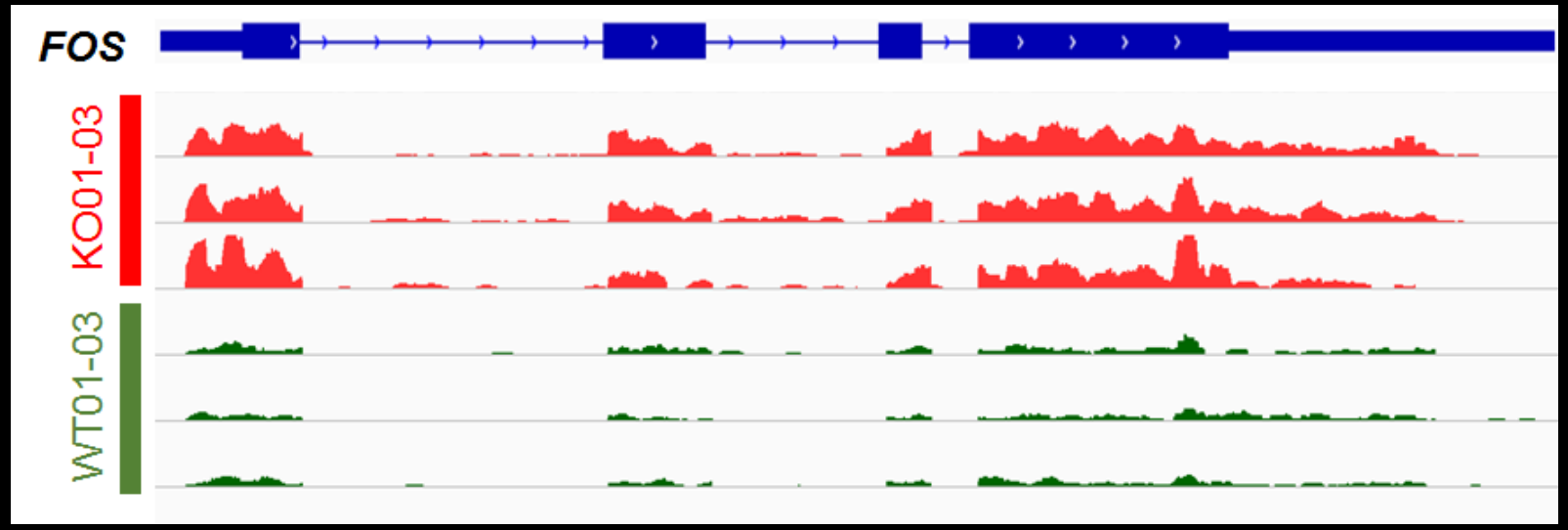
- **Gene annotation database: GENCODE**
  - https://www.gencodegenes.org/
- Ensembl database
  - https://www.ensembl.org/index.html
- UCSC Genome Browser
  - https://genome.ucsc.edu/
- NCBI databases
  - https://www.ncbi.nlm.nih.gov/guide/genomes-maps/

How to perform RNAseq analysis



05-08: Identify differentially expressed genes and pathways: DESeq2, clusterProfiler

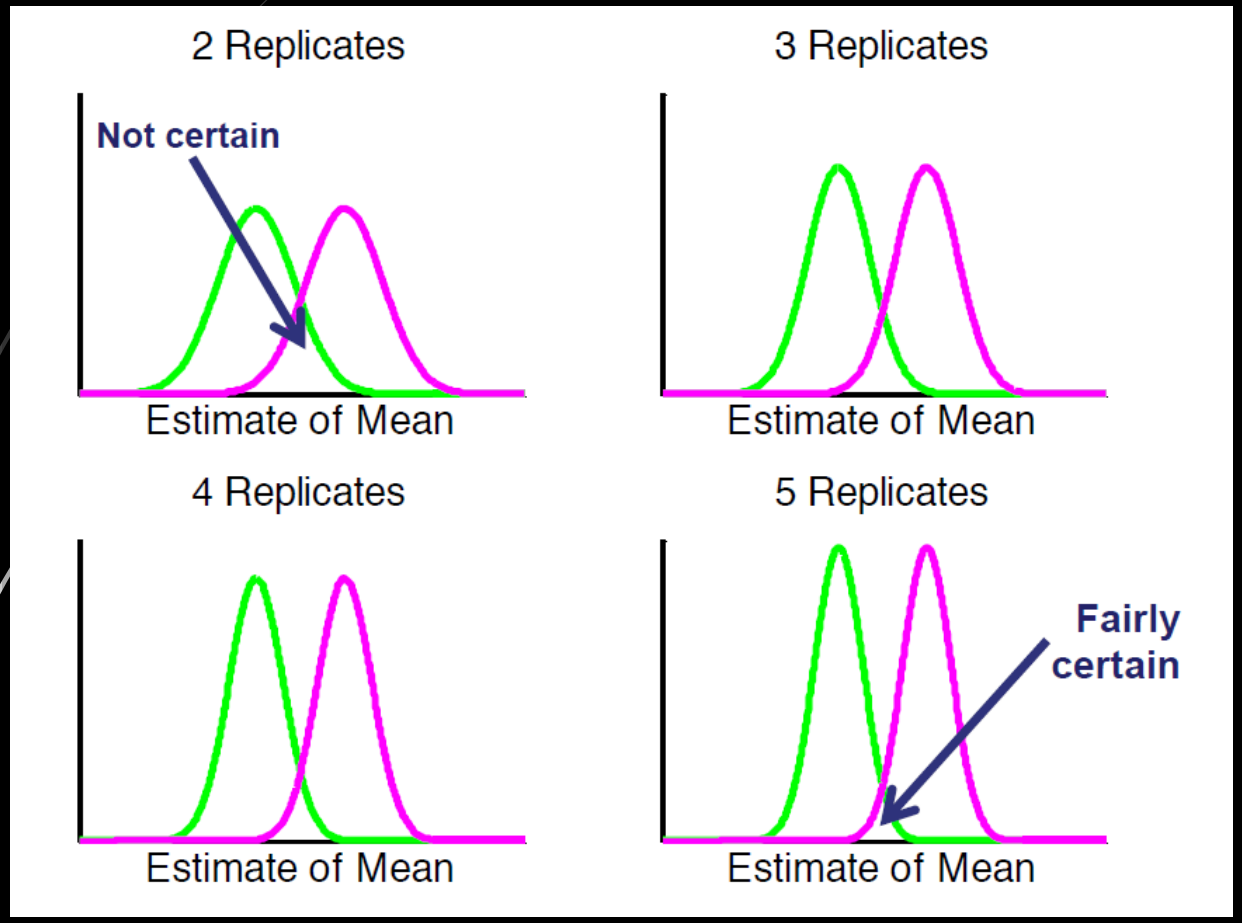
- After steps 01-04, we have generated read alignment and counts for every annotated gene on the genome
- The next step is to utilize the read counts data to detect DEGs
- For example, if we visualize *FOS* gene across 6 samples in genome browser



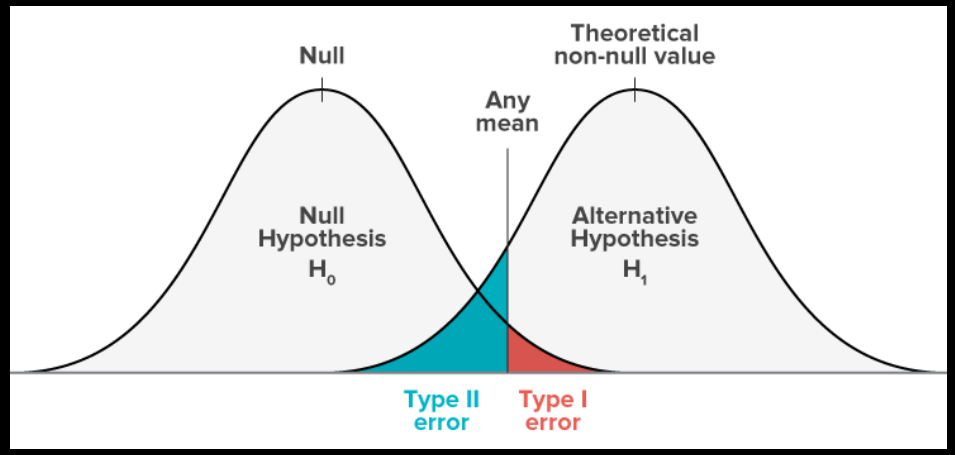
FOS = Fos proto-oncogene, AP-1 transcription factor subunit

DEG detection

Sensitivity: TP/(TP+FN) Specificity: TN/(FP+TN)

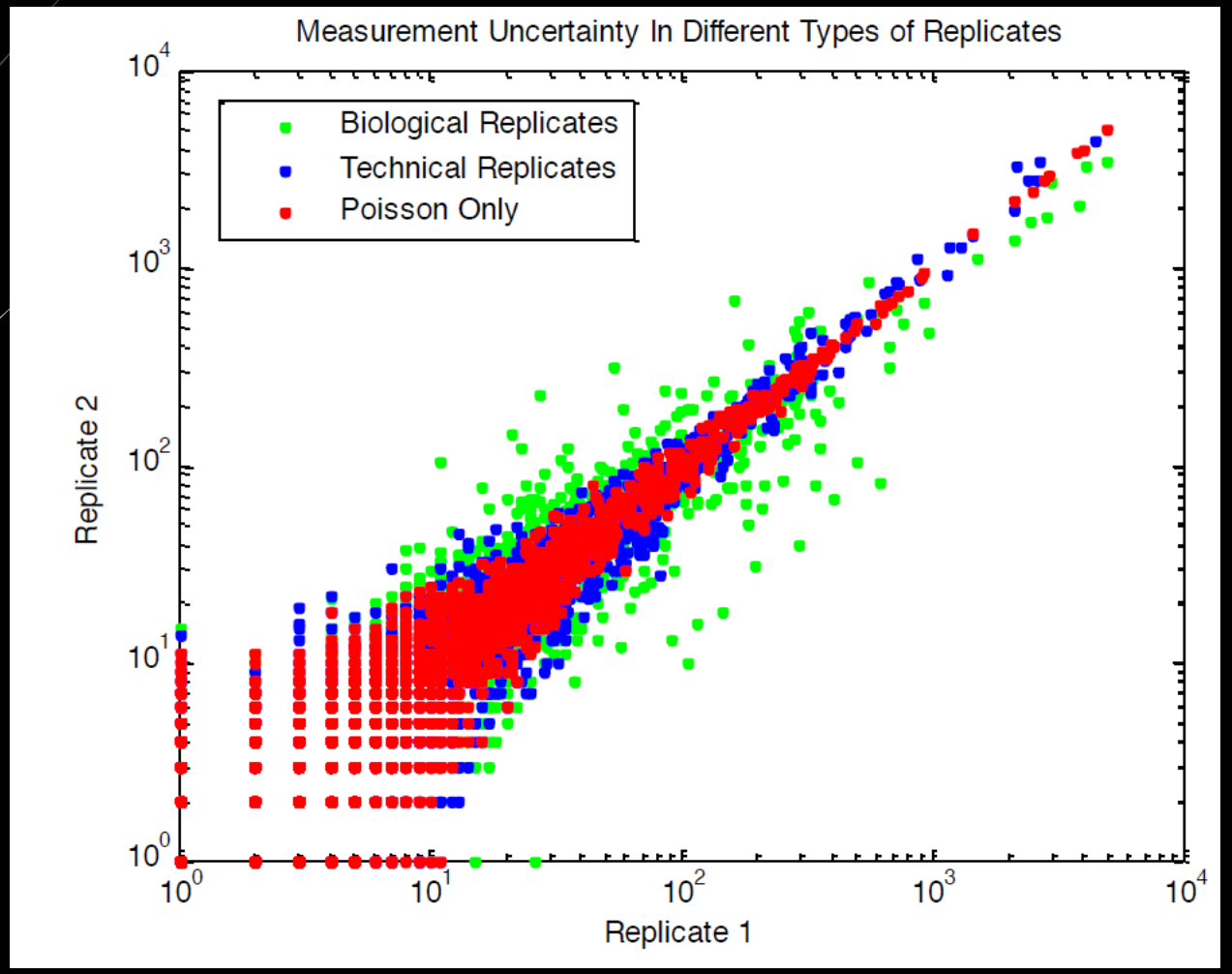


		Reality	
		Positive	Negative
Study Finding	Positive	True Positive (Power) (1-β)	FP Type I Error (α)
	Negative	FN Type II Error (β)	True Negative



More biological replicates per group lead to higher discovery power, sensitivity and specificity.

DEG detection



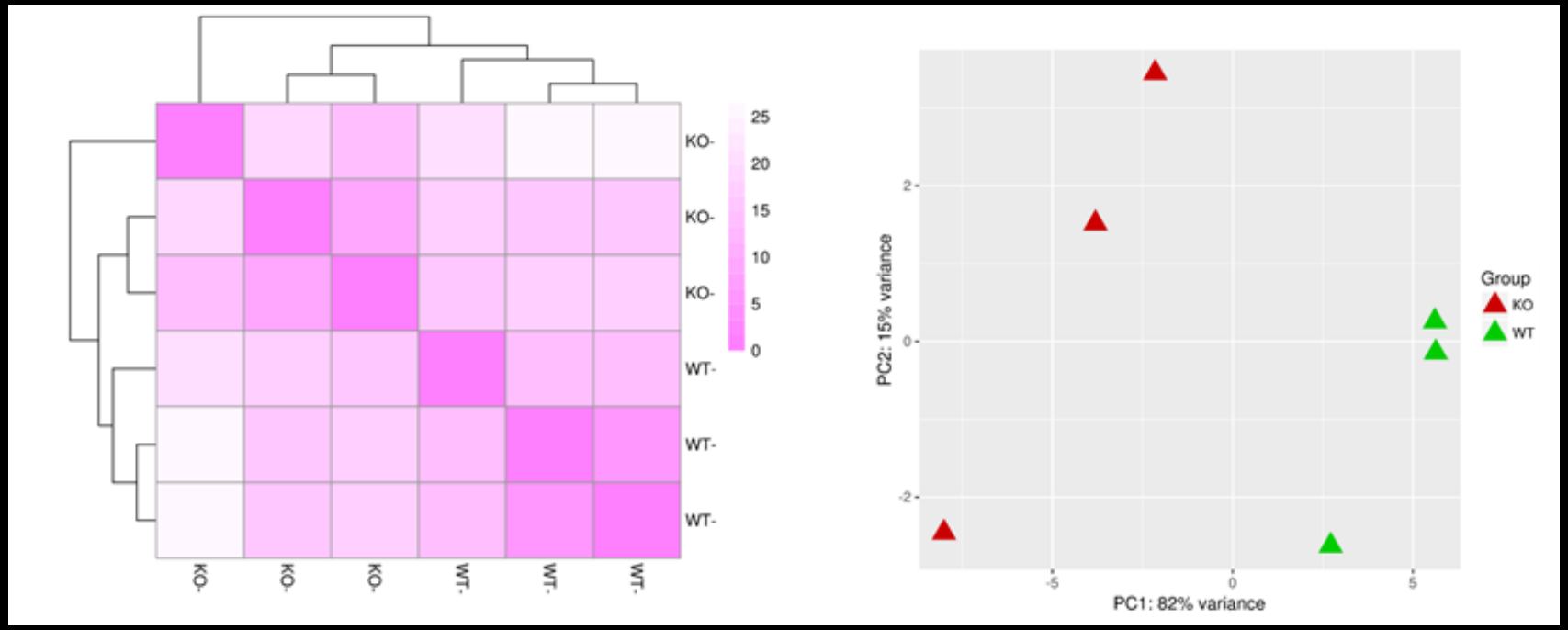
http://www.broadinstitute.org

High correlation is expected between biological replicates.

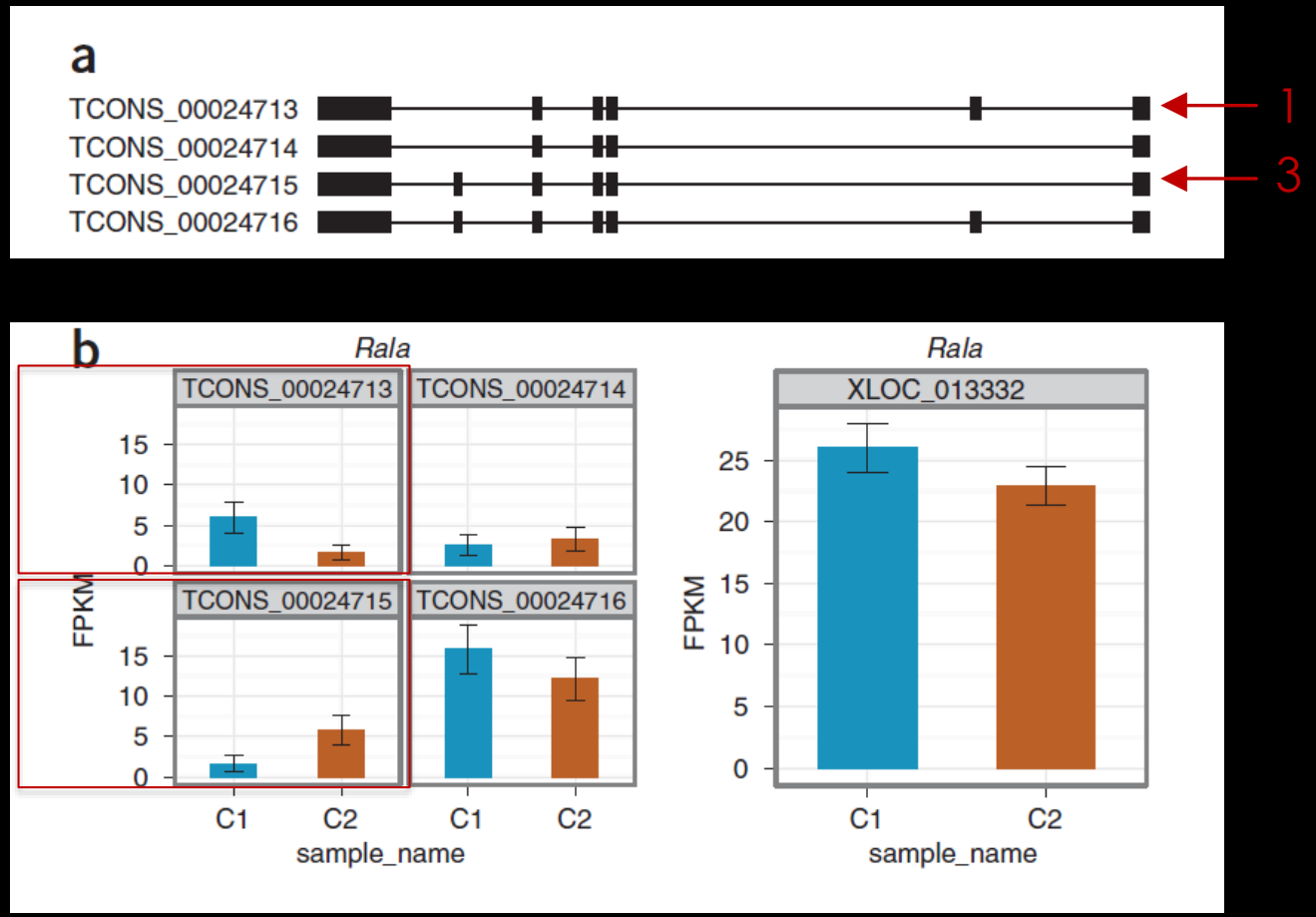
If one sample is an outlier, it can be identified if multiple replicates are included in an experiment.

How to identify an outlier?

- PCA plot (visualization)
- Unsupervised sample clustering based on all genes or *top variable genes (e.g. 1500)*



Transcript vs gene level quantification



Isoform 1 *

Isoform 3 *

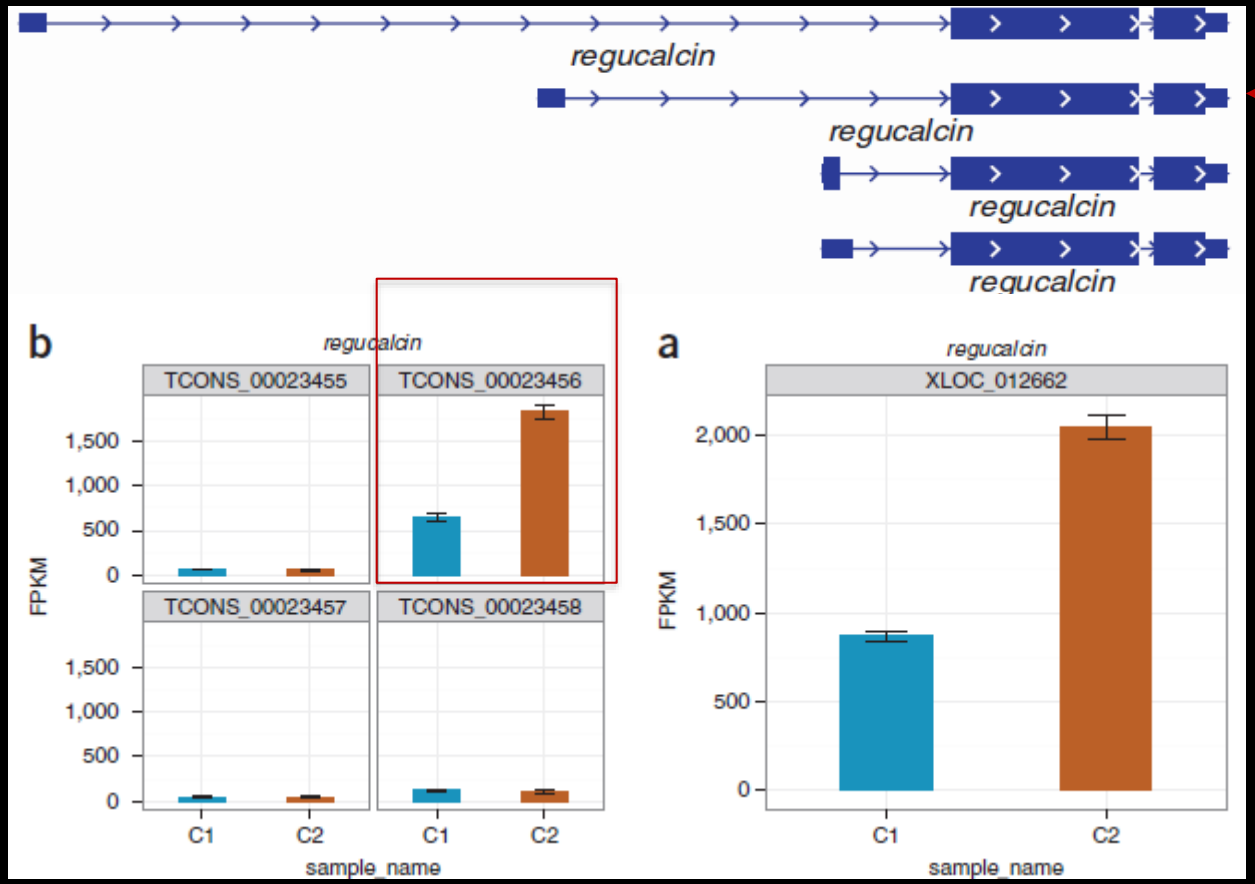
Isoforms 1 – 4

Gene

Difference in gene-level expression is not significant due to variability of isoforms

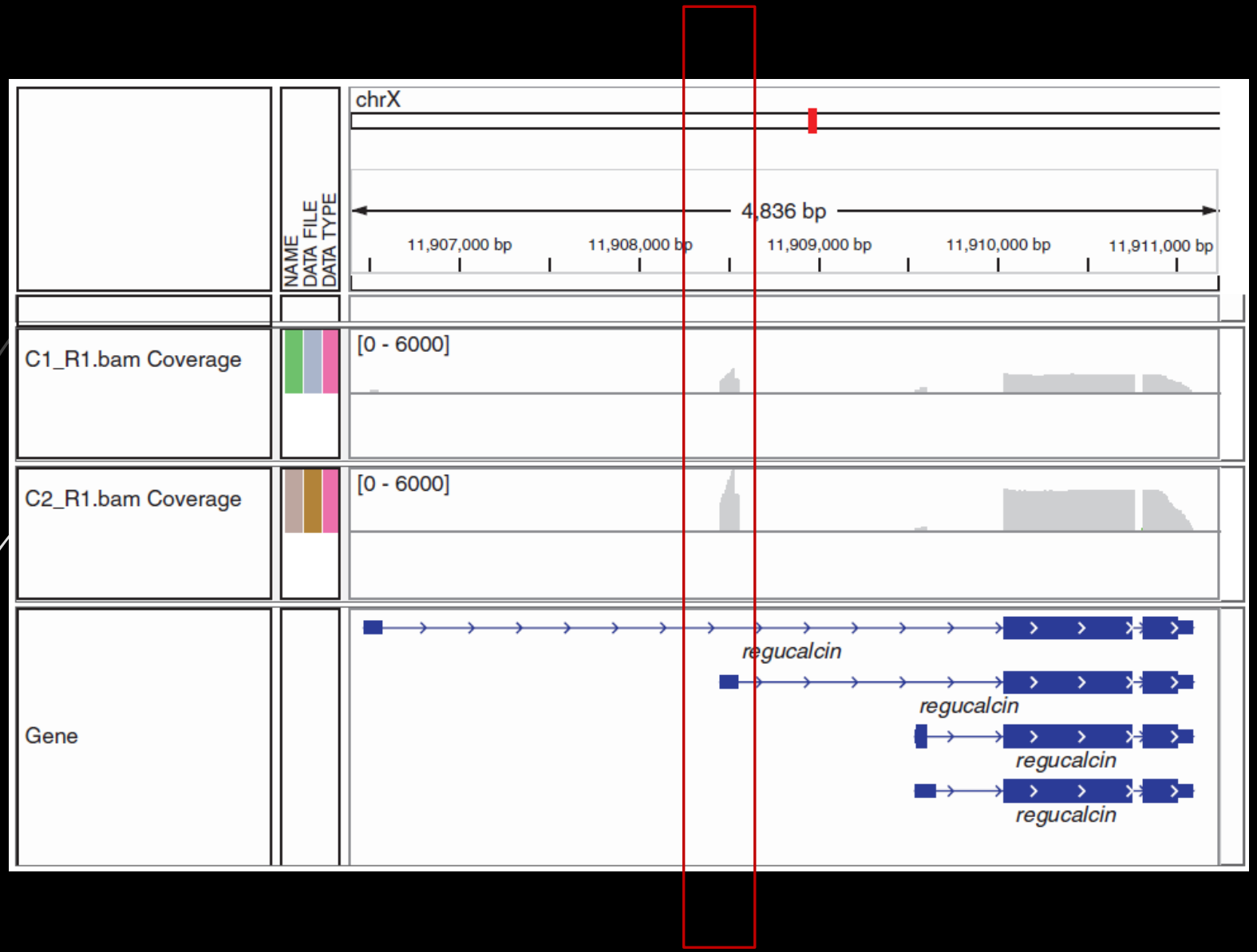
Transcript vs gene level quantification

Isoform 2 *



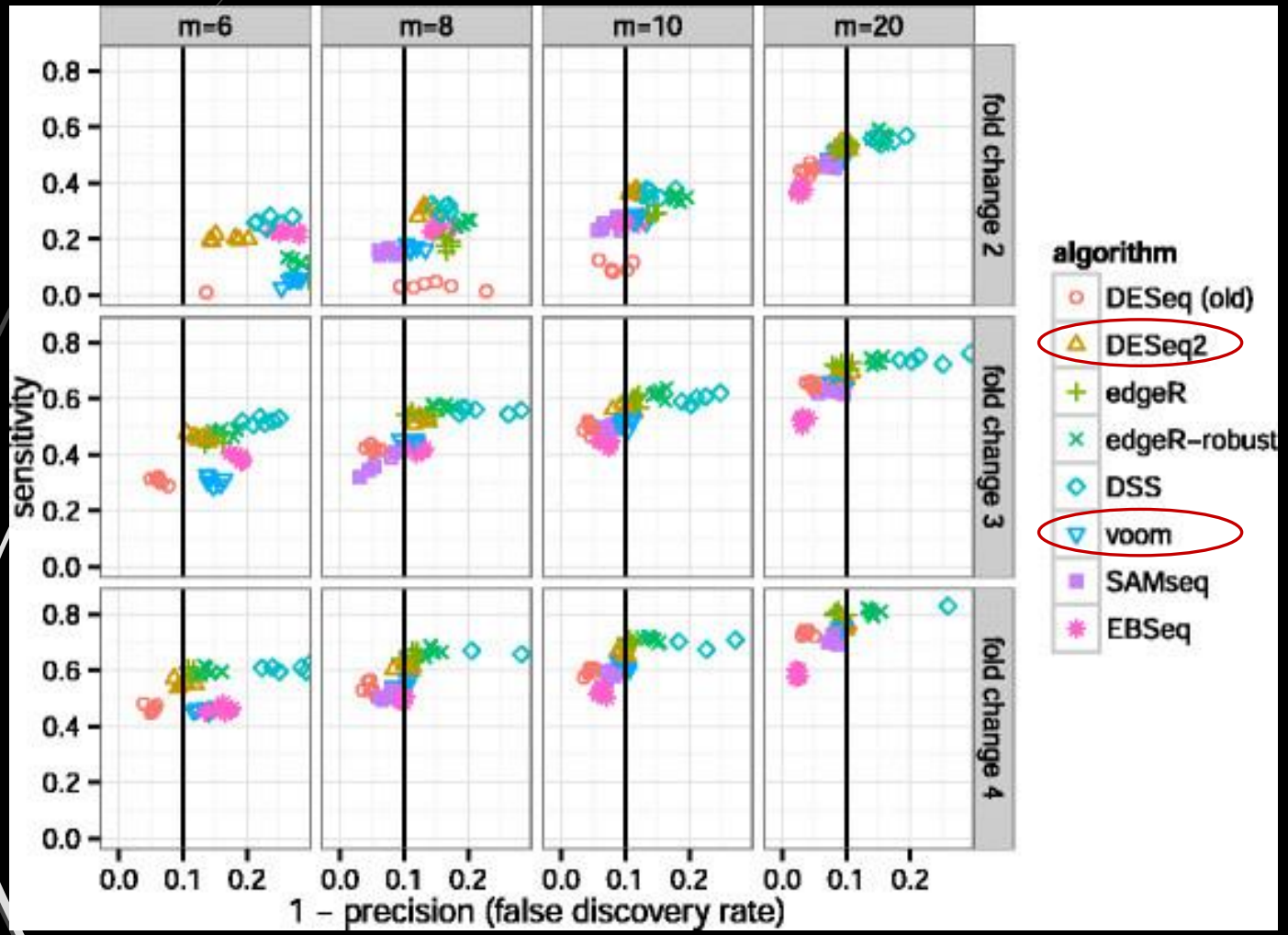
Isoforms 1 – 4 Gene

Difference in gene-level expression is significant, which is largely due to a great increase in the expression of isoform 2



Difference in gene-level expression is significant, which is largely due to a great increase in the expression of isoform 2

Comparison of different DEG identification methods



Sensitivity and precision of algorithms across combinations of sample size and effect size.

DESeq2 and edgeR often had the highest sensitivity of those algorithms that controlled the FDR, i.e., those algorithms which fall on or to the left of the vertical black line.

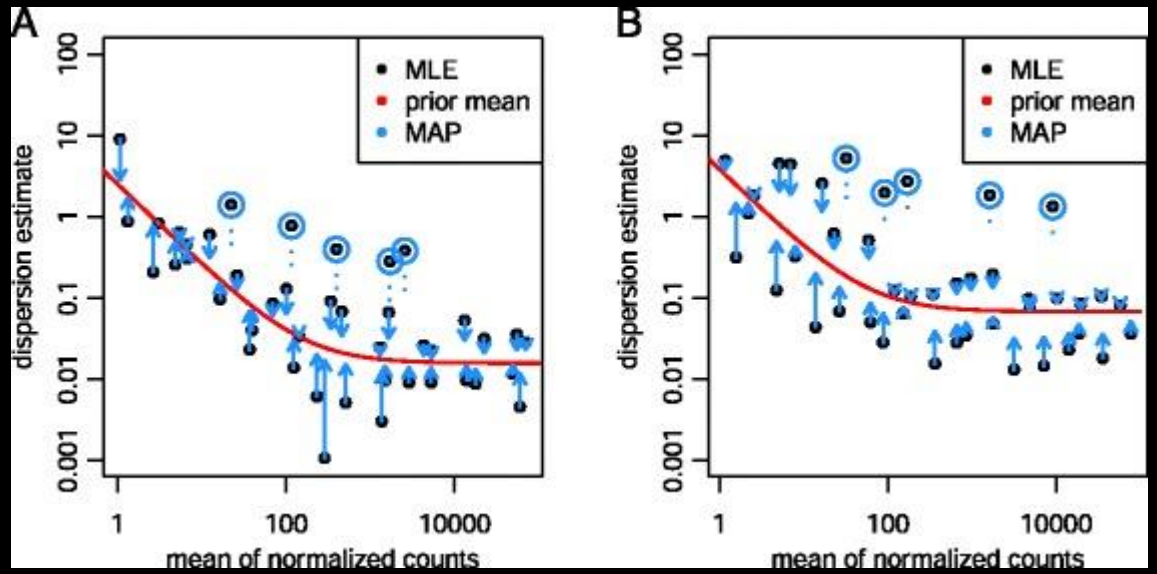
m: total sample size; split into two even-sized groups for comparison

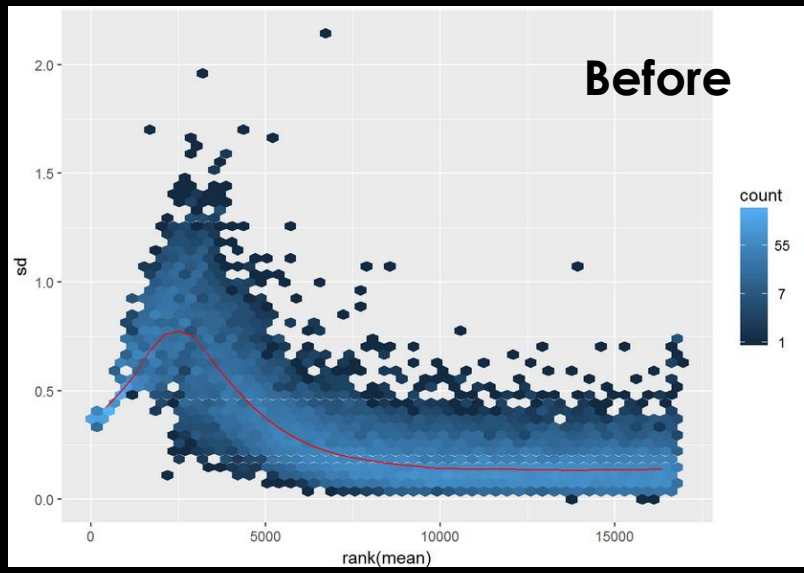
Love et al., Genome Biology 2014

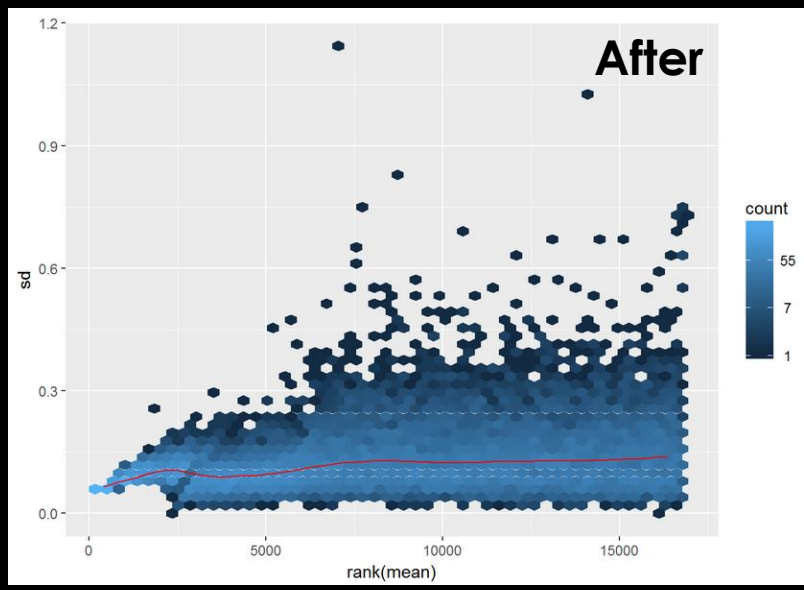
DESeq2

- Count matrix data
- Assume data follow negative binomial distribution (sometimes also called a gamma-Poisson distribution) with mean (μ) and dispersion (α) parameters
- Within-group variability, i.e., the variability between replicates, is modeled by the dispersion parameter alpha, which describes the *variance* of counts
- Empirical Bayes shrinkage for dispersion estimation

MAP, maximum a posteriori; MLE, maximum-likelihood estimate



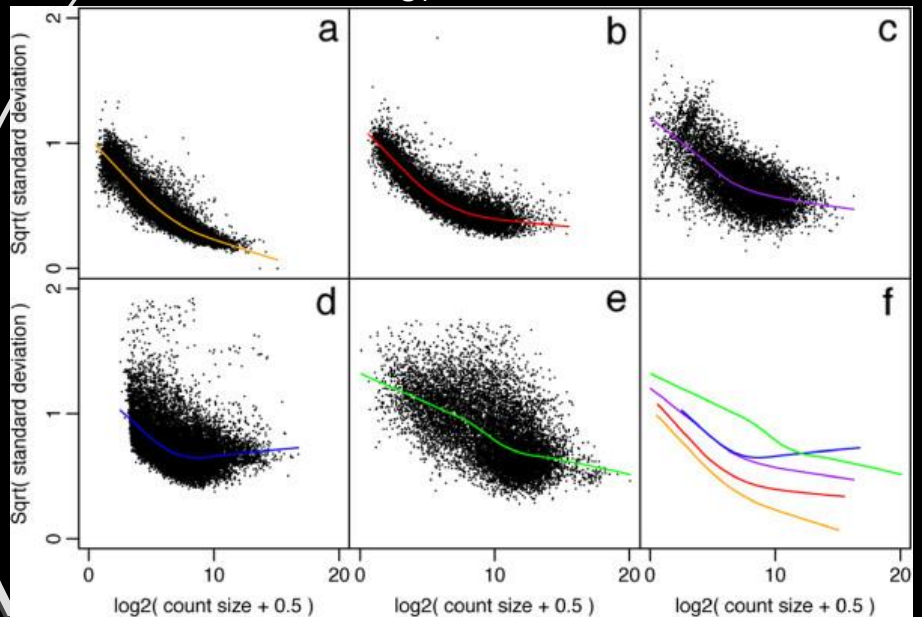




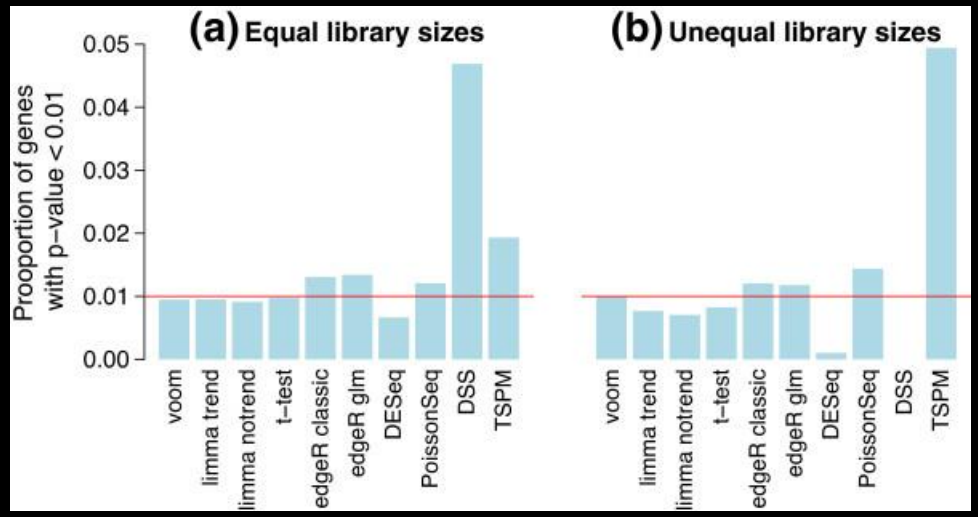
Limma voom (*weighted* algorithm)

- To model the mean-variance relationship than to specify the exact probabilistic distribution of the counts (e.g. NB or Poisson)
- Provide accurate Type I (alpha) and Type II error (beta) control compared to other methods, especially when sample size is small
- *Voom with sample quality weights*

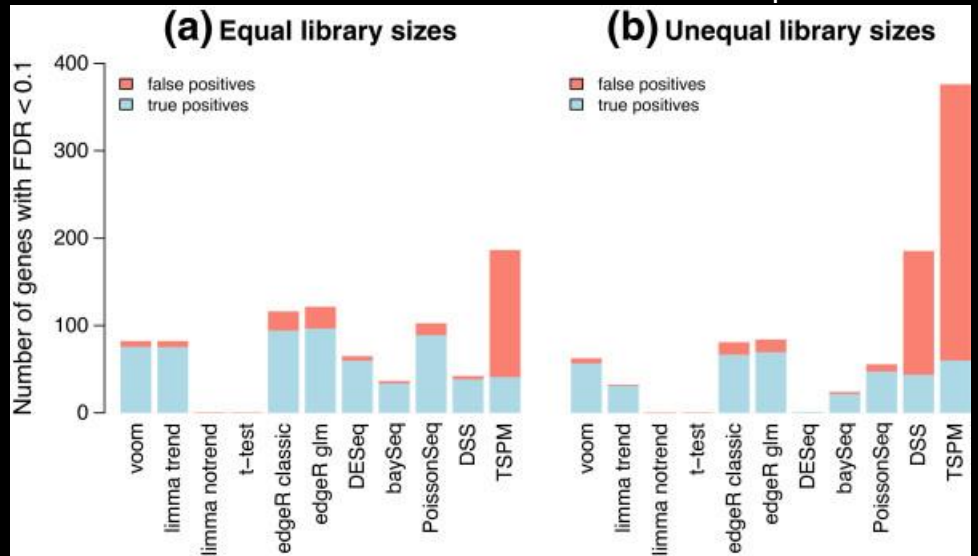
Law et al., Genome Biology 2014



Type I error rates in the absence of true differential expression

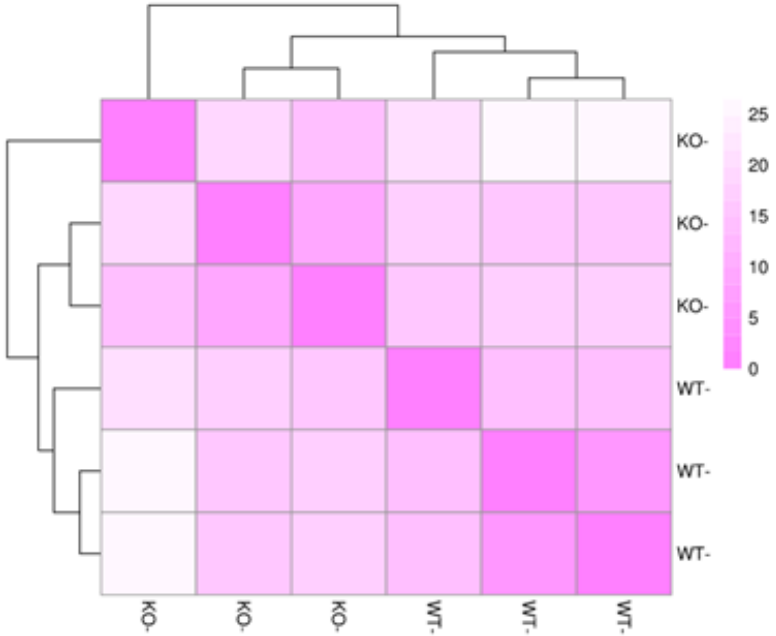


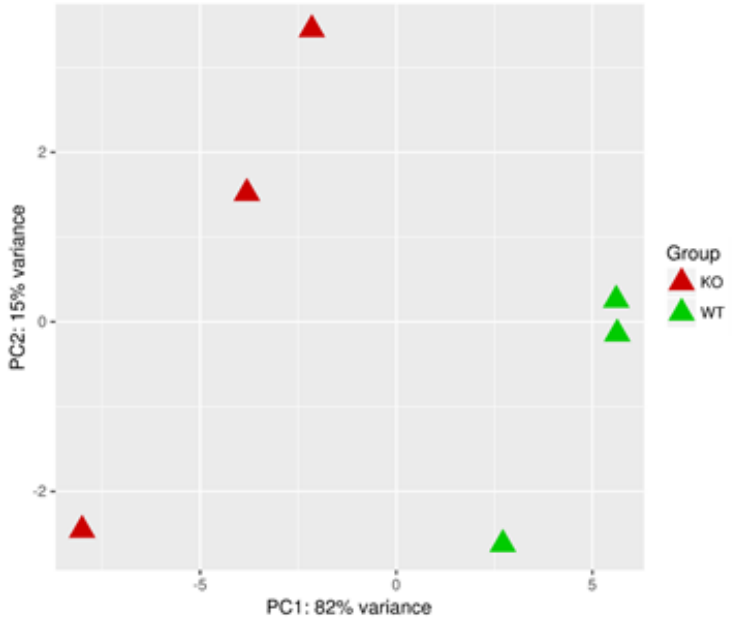
Power to detect true differential expression



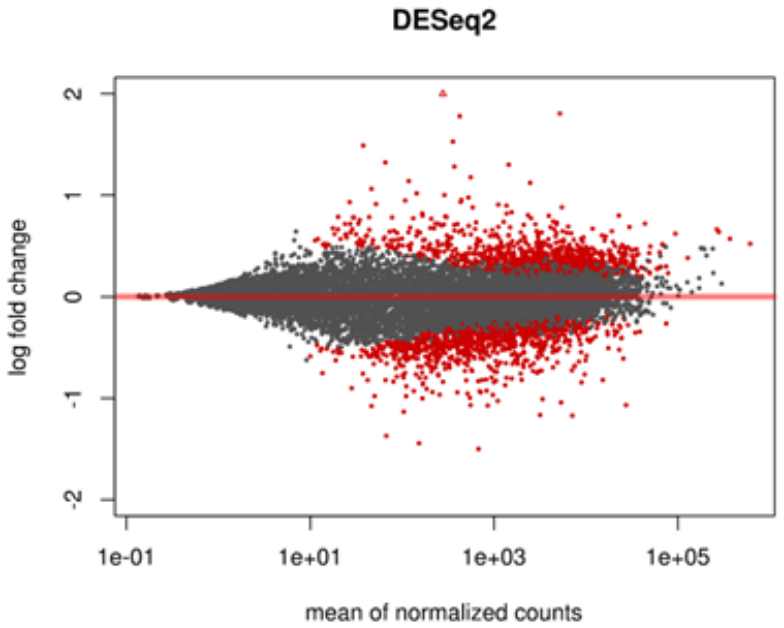
More databases!

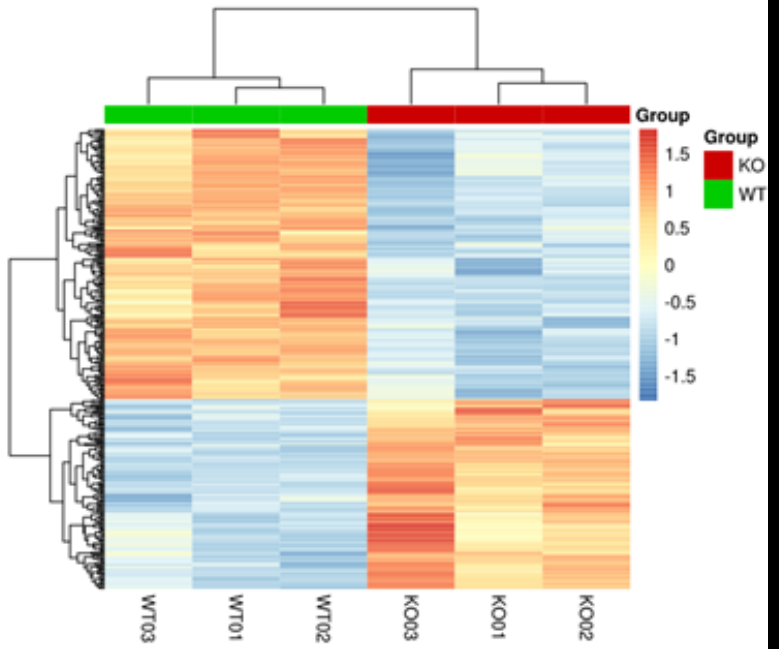
- Gene annotation database: GENCODE
  - https://www.gencodegenes.org/
- Gene Ontology (GO) database: Gene Ontology Consortium
  - http://www.geneontology.org/
- Pathway database: KEGG
  - http://www.genome.jp/kegg/
- Predefined gene sets: MSigDB
  - http://software.broadinstitute.org/gsea/msigdb/

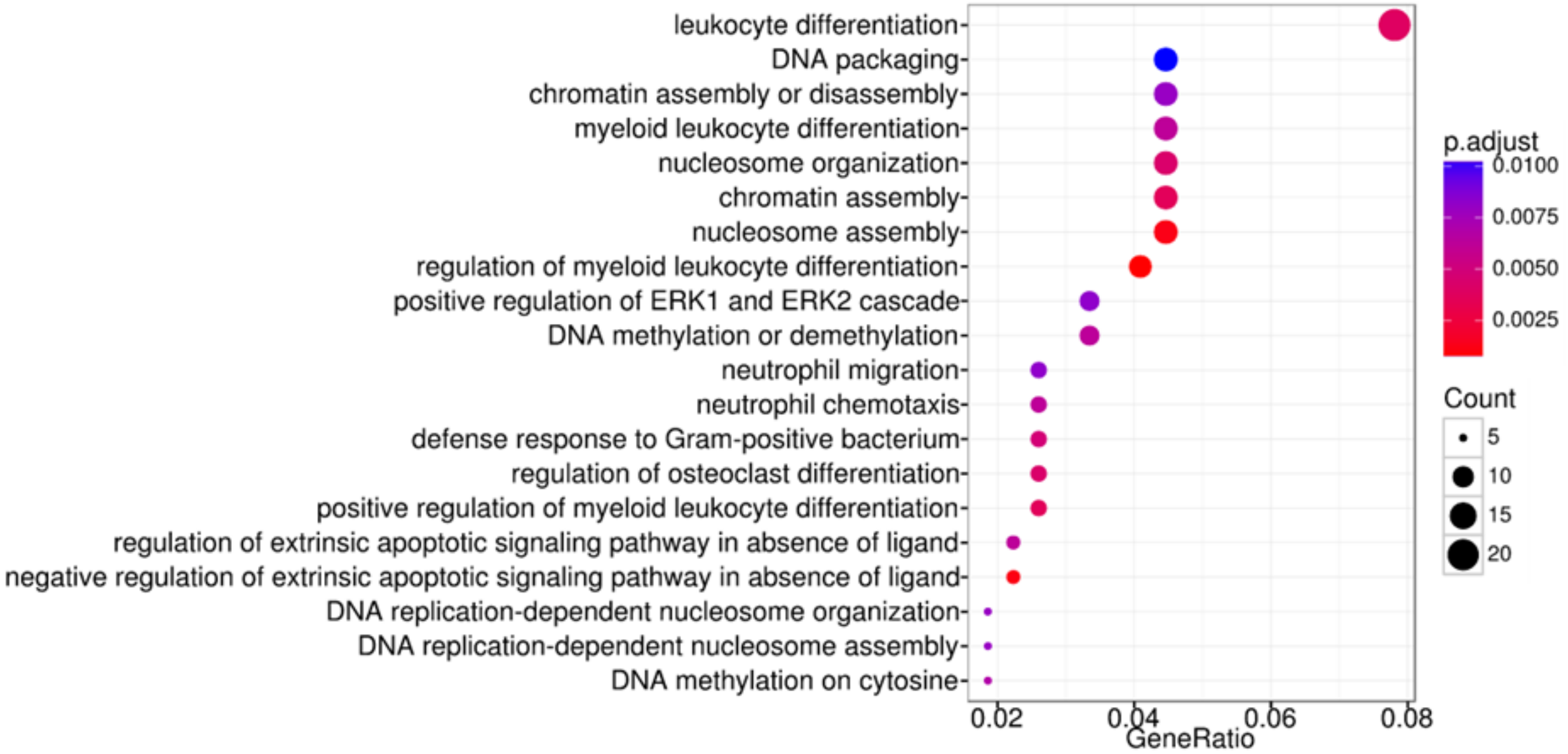


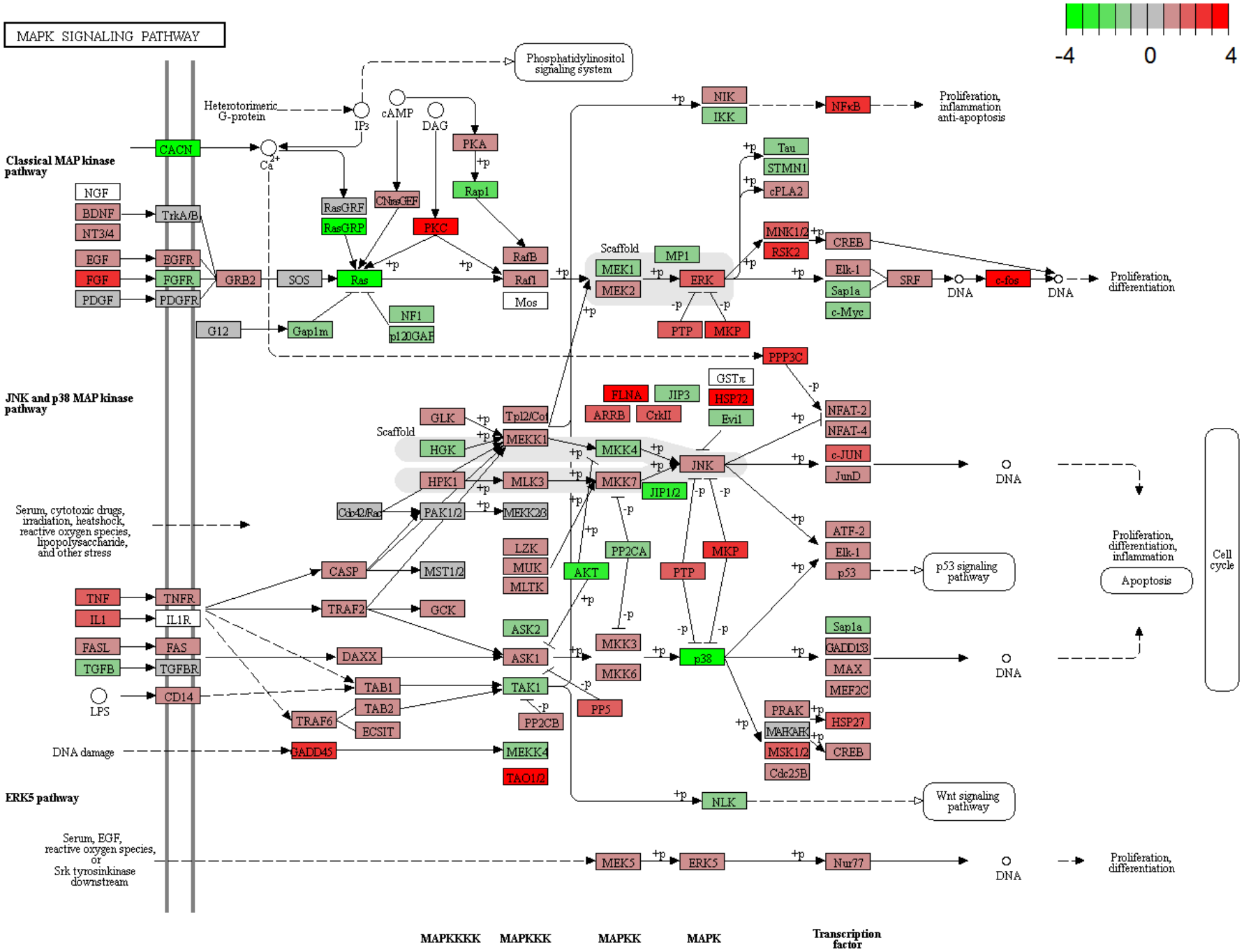



```
[1] "Genes significant = 296 (fc, 1.5, fdr 0.05)"
[1] "Heatmap = 296 genes on the row, 6 samples on the column"
```



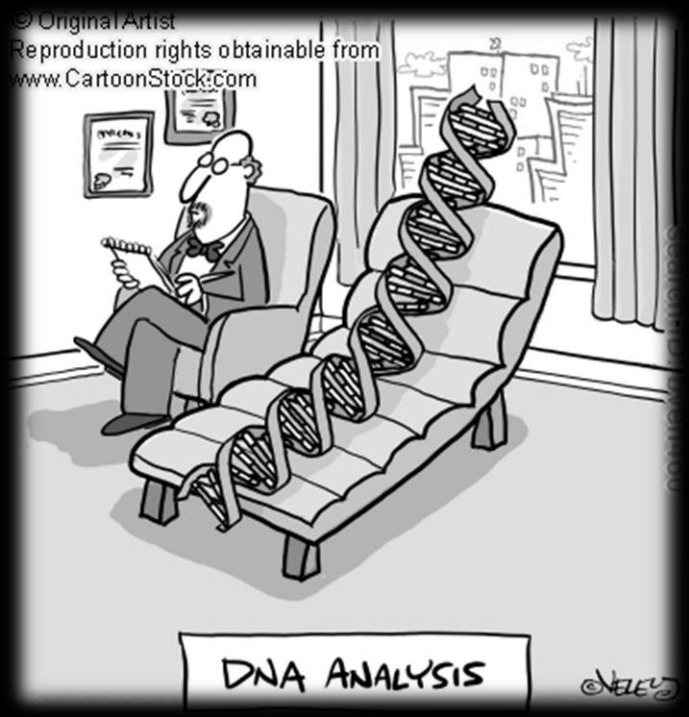



leukocyte differentiation
DNA packaging
chromatin assembly or disassembly
myeloid leukocyte differentiation
nucleosome organization
chromatin assembly
nucleosome assembly
regulation of myeloid leukocyte differentiation
positive regulation of ERK1 and ERK2 cascade
DNA methylation or demethylation
neutrophil migration
neutrophil chemotaxis
defense response to Gram-positive bacterium
regulation of osteoclast differentiation
positive regulation of myeloid leukocyte differentiation
regulation of extrinsic apoptotic signaling pathway in absence of ligand
negative regulation of extrinsic apoptotic signaling pathway in absence of ligand
DNA replication-dependent nucleosome organization
DNA replication-dependent nucleosome assembly
DNA methylation on cytosine
0.02
0.04
0.06
0.08
GeneRatio
p.adjust
0.0100
0.0075
0.0050
0.0025
Count
5
10
15
20



Data on KEGG graph
Rendered by Pathview

Thank you!



Questions

Hands-on practice START

- Open your handson.Rmd on the Github or download to local computer
- https://github.com/MScBiomedicalInformatics/MSIB32500/raw/master/lectures/handson8.html
- *Dataset: two groups (PRDM11 KO vs WT, human U2932 cells), 6 samples*
- *Single-end reads, unstranded libraries*

Sample	Group	Sequencing File	Sequencing Data
KO01	KO	KO01.fastq.gz	74,126,025 reads
KO02	KO	KO02.fastq.gz	64,695,948 reads
KO03	KO	KO03.fastq.gz	52,972,573 reads
WT01	WT	WT01.fastq.gz	55,005,729 reads
WT01	WT	WT02.fastq.gz	61,079,377 reads
WT01	WT	WT03.fastq.gz	66,517,156 reads

Fog. et al. 2015. Loss of *PRDM11* promotes MYC-driven lymphomagenesis. Blood 125(8):1272-81

PRDM11 = PR/SET domain 11


INDEX
EXTRA QUALITY
90