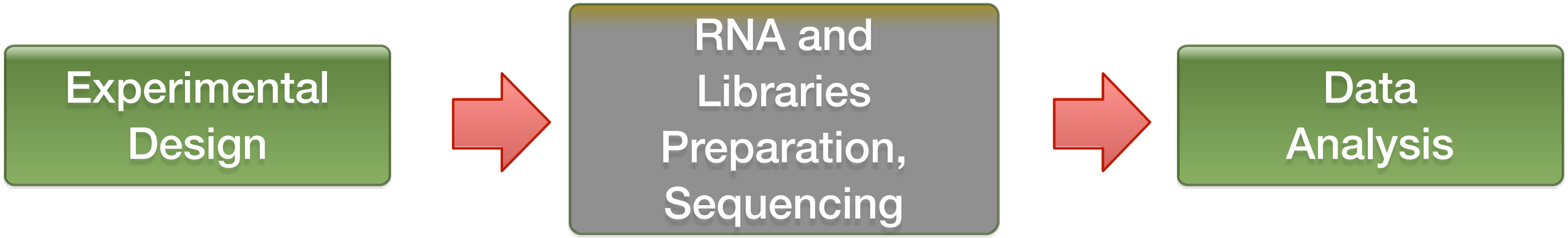
RNA-seq: Experimental Design and Differential Expression Analysis

Jinliang Yang
PostDoc Scholar
Jeffrey R-I lab
Feb. 23, 2015

RNA-seq Outline



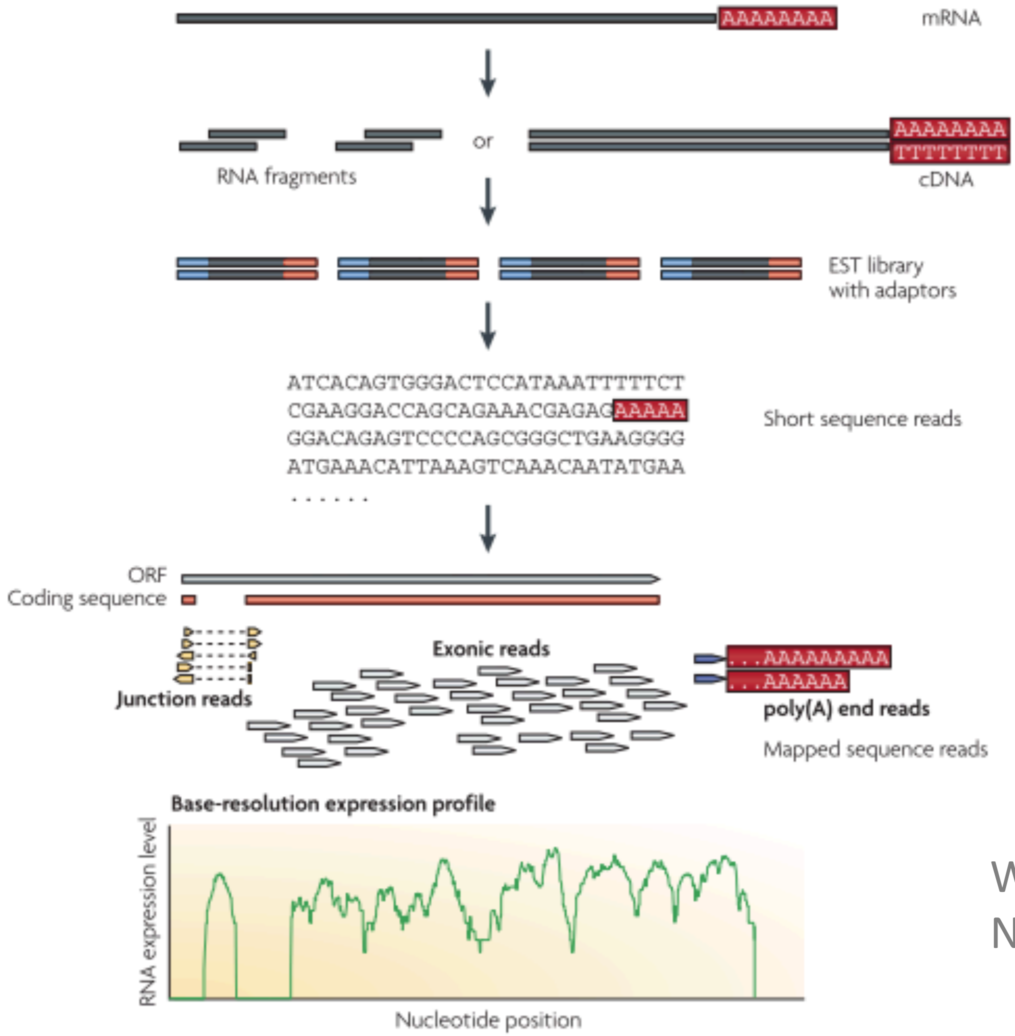
Experimental Design

RNA and Libraries Preparation, Sequencing

Data Analysis

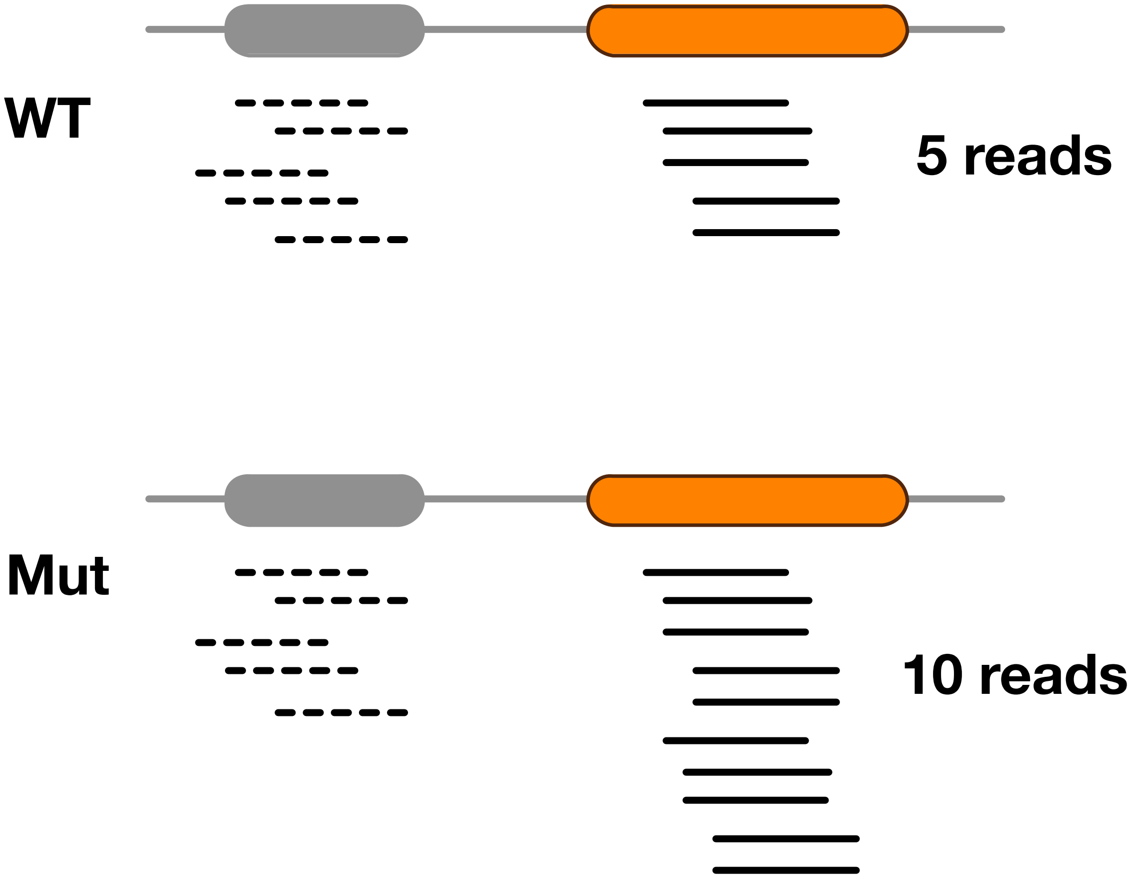
- Overview of RNA-seq Experiment
- Experimental Design
- Sequencing
- Data Analysis (Differential Expression)

Overview of RNA-seq Experiment



Wang *et al.*, 2009,
Nature Review Genetics

RNA-seq: A Toy Example



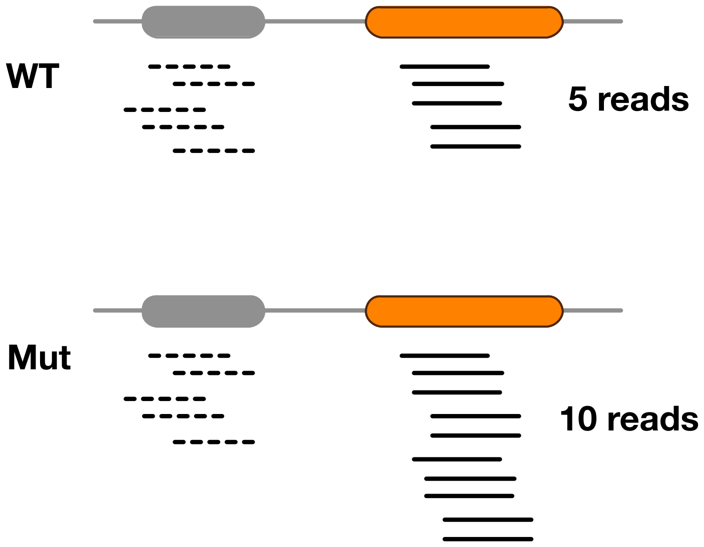
WT

5 reads

Mut

10 reads

RNA-seq: Source of Variance

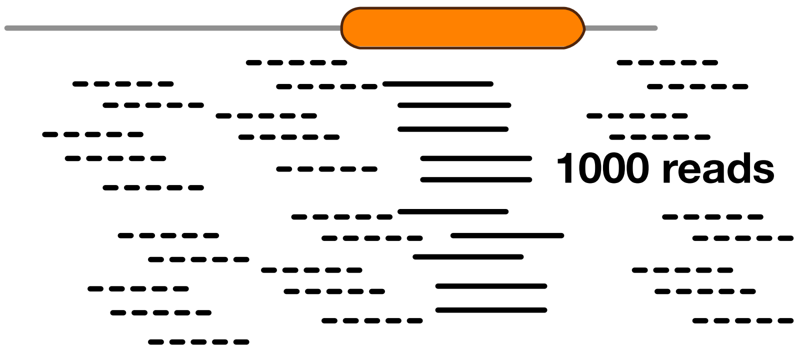


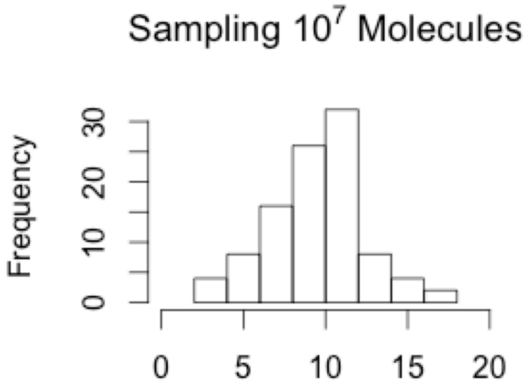
- Biological variance
  - **Treatment effect (WT vs. Mut)**
  - Difference between two plants
- Technical variance
  - RNA isolation difference
  - Sequencing library preparation difference
  - Sequencing difference
- Sampling variance
  - Sampling issue

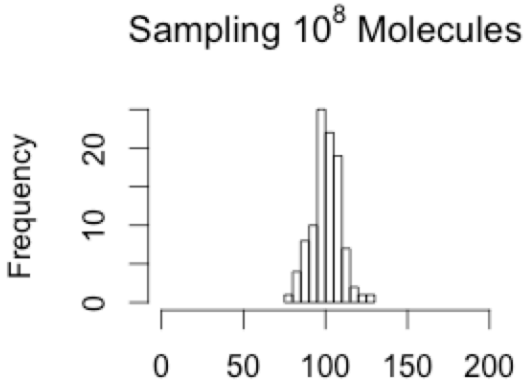
Experimental Design:
Control Source of Variance

Sampling variance is the inherent nature of a counting experiment

1000 molecules / 10^9 in total



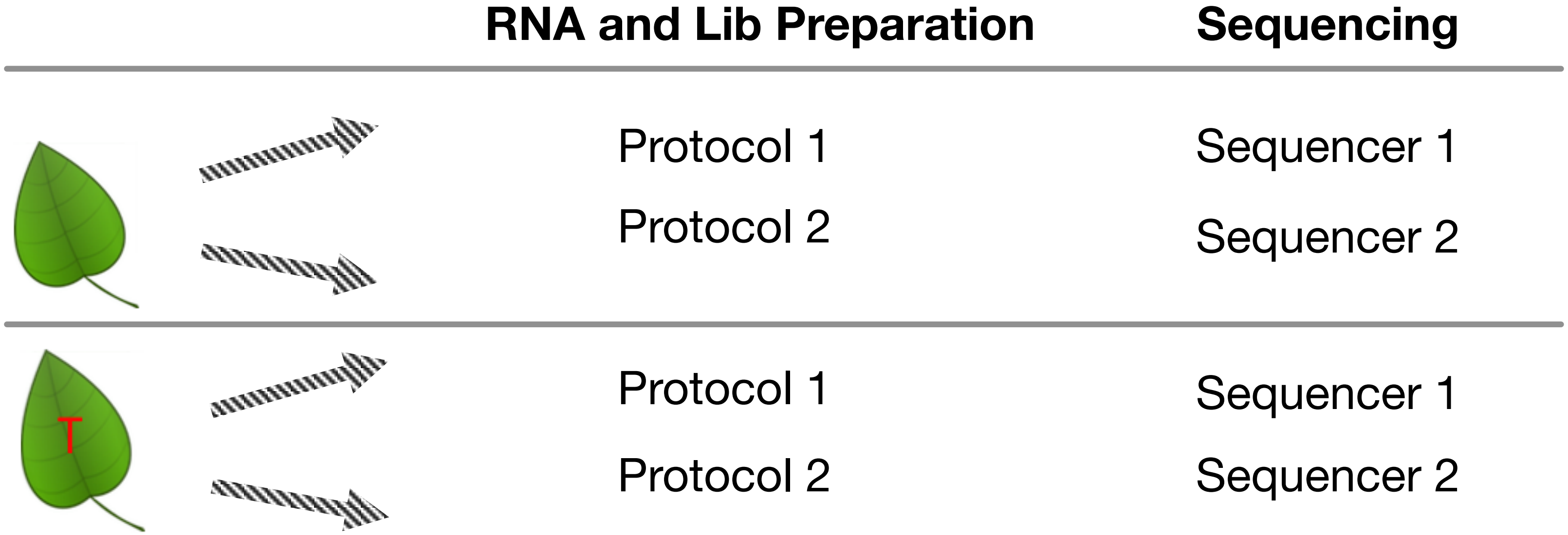

Sampling 10^7 Molecules


Sampling 10^8 Molecules

Sequencing depth is critical

Technical Replication

Technical replication refers to sequencing multiple libraries derived from the same biological sample

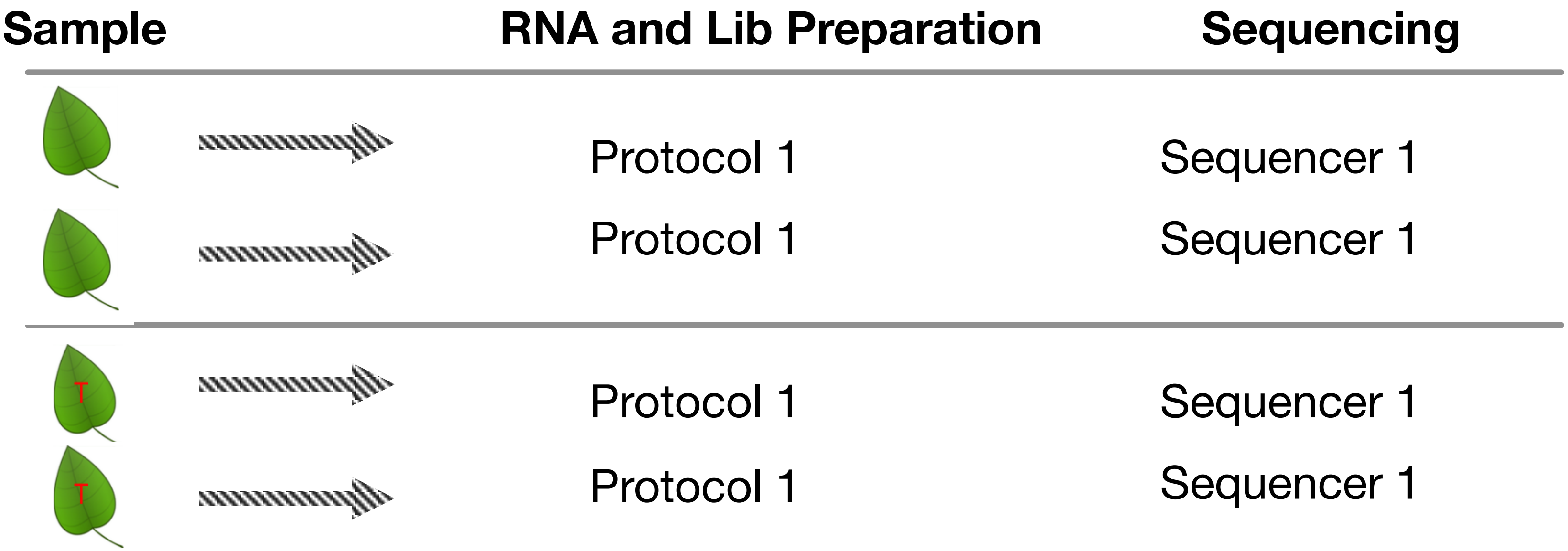


RNA and Lib Preparation	Sequencing
Protocol 1	Sequencer 1
Protocol 2	Sequencer 2
Protocol 1	Sequencer 1
Protocol 2	Sequencer 2

Technical differences are not limited to above factors.
WHO, WHEN and HOW can all be considered as the source of tech variance.

Biological Replication

Biological replication refers to sequencing libraries derived from the different biological individuals or tissues.



Sample	RNA and Lib Preparation	Sequencing
	Protocol 1	Sequencer 1
	Protocol 1	Sequencer 1
	Protocol 1	Sequencer 1
	Protocol 1	Sequencer 1

Take Home Message

- It is better to have more depth of sequencing (sampling variance)
  - About 10 million reads per sample is a benchmark to start. (Wang *et al*. 2011)

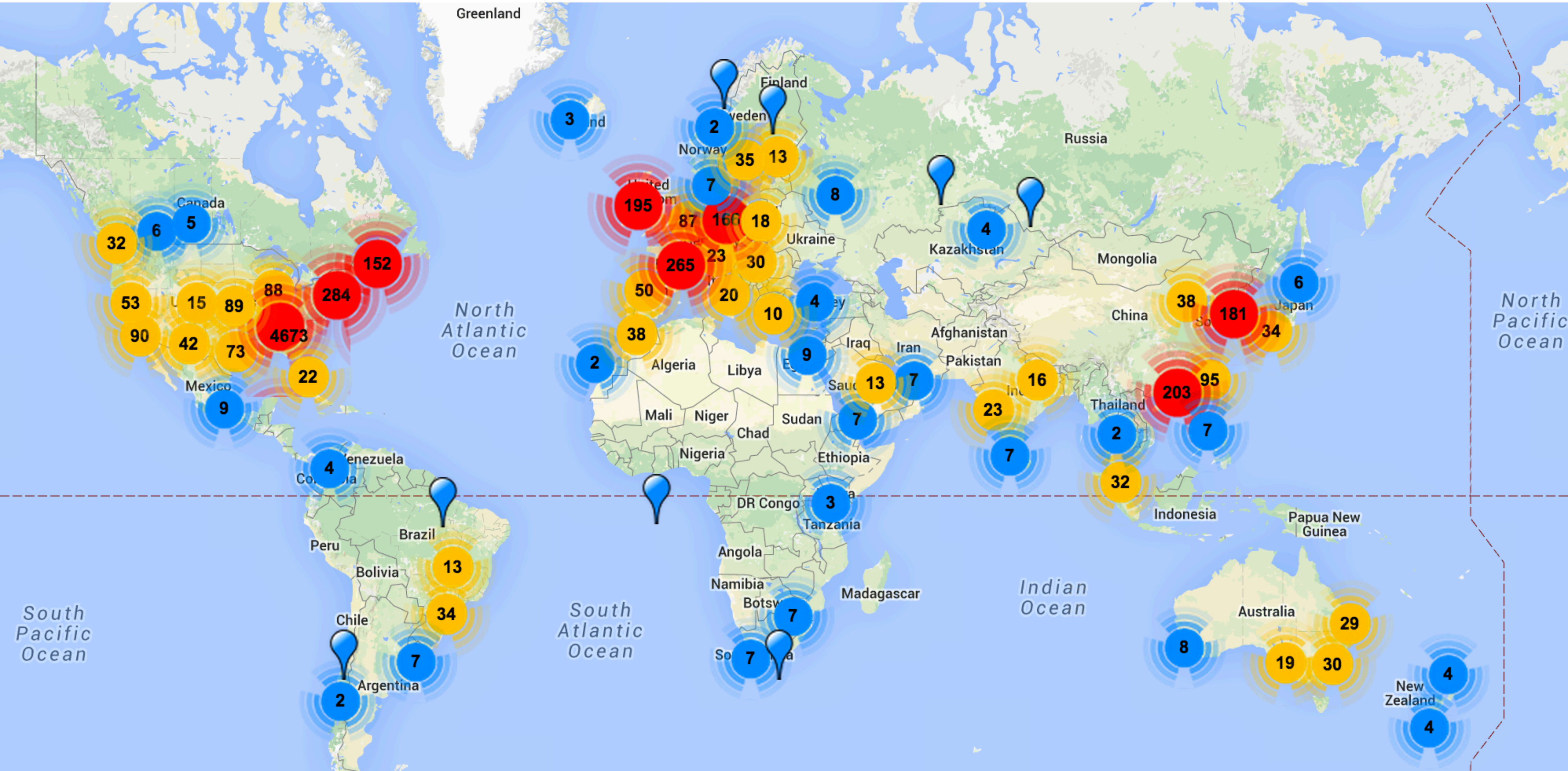
- Replication and randomization helps to control confounding variance

- Limited resources are probably better to allocate to biological rep than technical rep

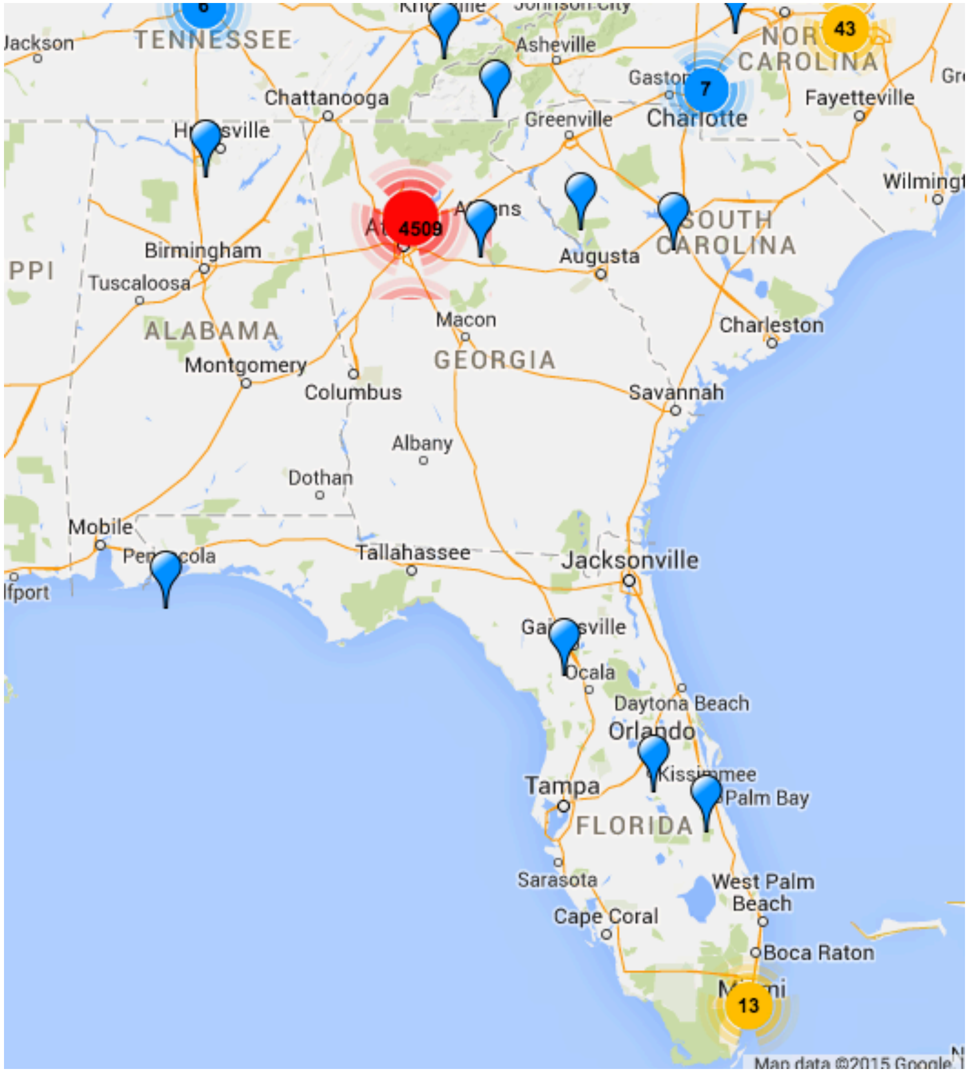
RNA-seq Platforms

Illumina, Ion Torrent, PacBio and Oxford Nanopore

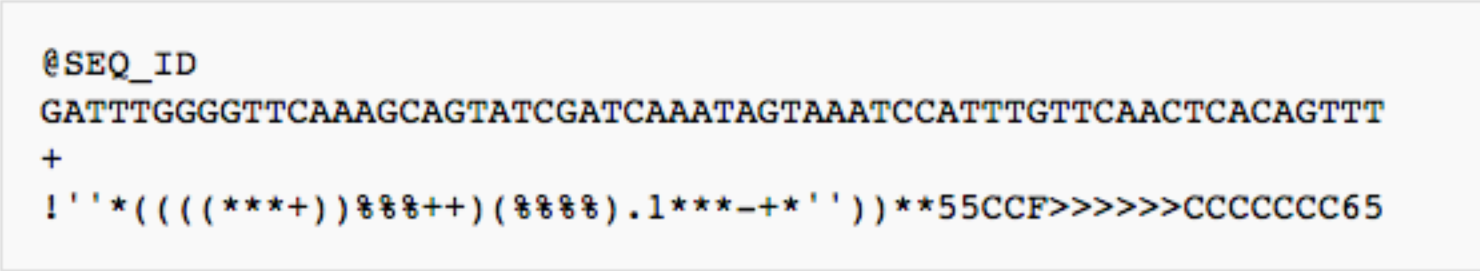
http://omicsmaps.com/



Ion PGM Sequencer?



RNA-seq Data



```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- line 1: sequence id
- line 2: nucleotide sequence
- line 3: a "+" sign separator, optionally with the read identifier repeated
- line 4: a corresponding ASCII string of quality characters

$$Q_{\text{sanger}} = -10 \log_{10} p$$

p represents the probability that a given base is incorrectly called.

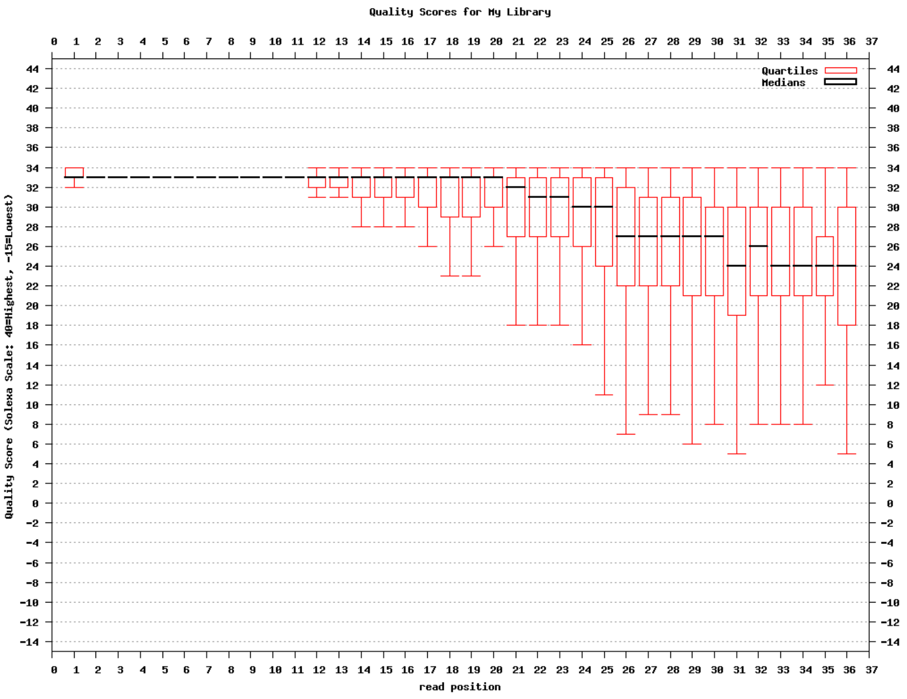
RNA-seq: Data Analysis Outline

- Quality checking and data cleaning
- Aligning RNA-seq reads to reference
- Count reads in gene models
- Differential Gene Expression Study

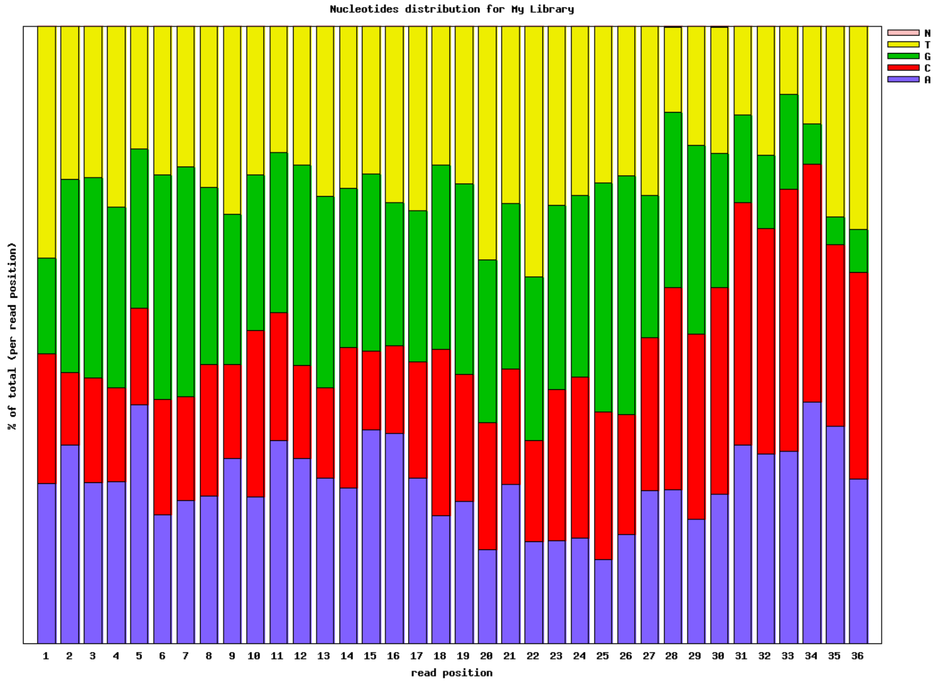
Pipeline: https://github.com/yangjl/Demo
Or /group/jrigrp5/ECL298/Demo

RNA-seq: Quality Checking

Quality Score



Nucleotide Distribution

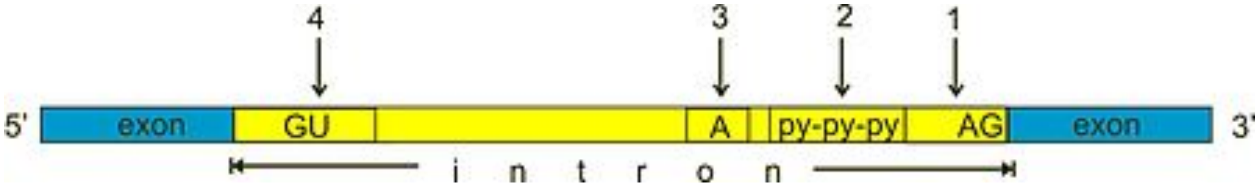


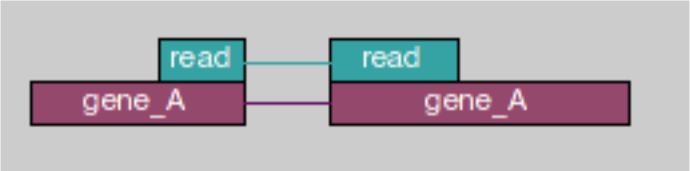
FASTX-Toolkit

Data Cleaning

- 1. Demultiplex by barcode
  - HiSeq 2000 normally yields ~150 million reads
  - 6nt barcode or index
- 2. Remove adapter sequences
- 3. Trim basepairs by quality
- 4. Discard reads by quality/ambiguity

Aligning Reads to a Reference Genome





- RNA-seq aligner:
  - Efficiency and splicing awareness
  - Widely used: GSNAP, BWA-mem, Bowtie2

- For differential expression study:
  - Reference genome
  - Gene annotation (typically the intron/exon annotations are available in GFF3 or GTF file)

Summary Statistics of Alignment

File	# of total reads	# of uniquely aligned reads	Mapping Rate (100%)
1.fq	4,413,034	4,147,652	94
2.fq	4,879,212	4,619,978	94.7
3.fq	6,924,929	6,618,966	95.6
4.fq	6,848,552	6,469,461	94.5
5.fq	5,438,538	4,962,816	91.3
6.fq	3,772,703	3,507,526	93

Uniquely aligned reads will be used for differential expression analysis

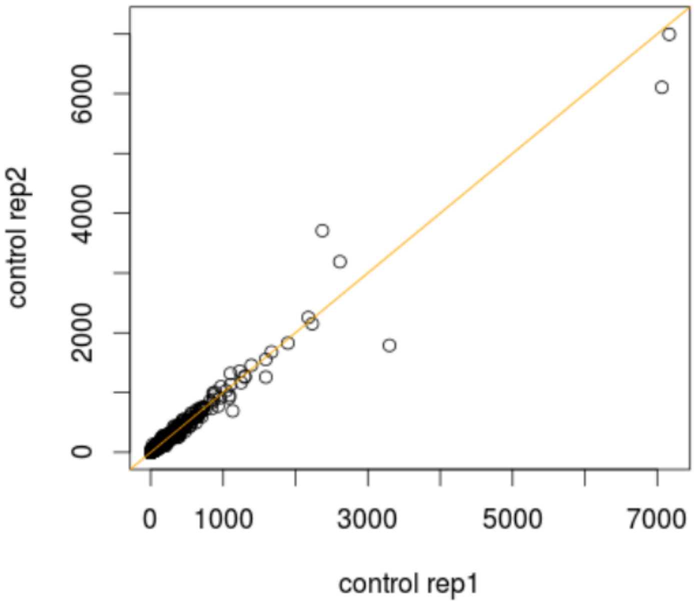
Read Counts and Normalization

RPKM: reads per kilobase per one million mapped reads. Adjust gene length and library size.

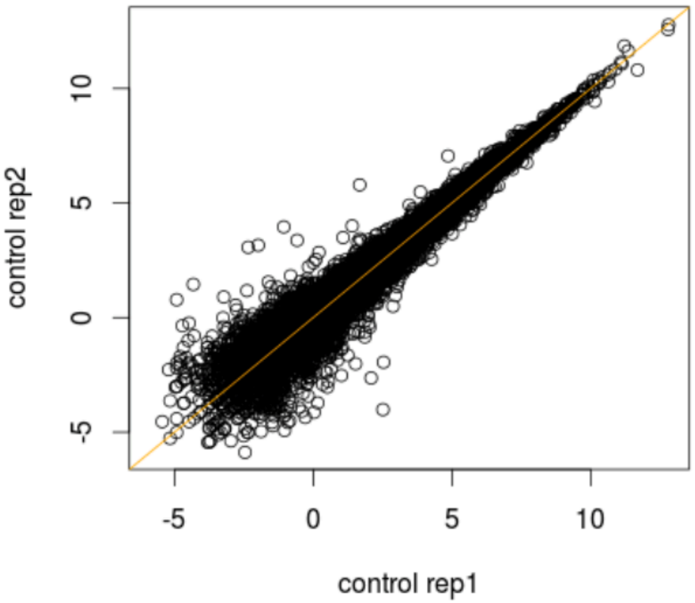
Gene	Control rep1	Control rep2
1	2679	2360
2	177	161
3	381	371
…		

Gene	Control rep1 RPKM	Control rep2 RPKM
1	3.4	3.3
2	1.3	1.2
3	2.0	2.0
…		

Raw counts scatter plot

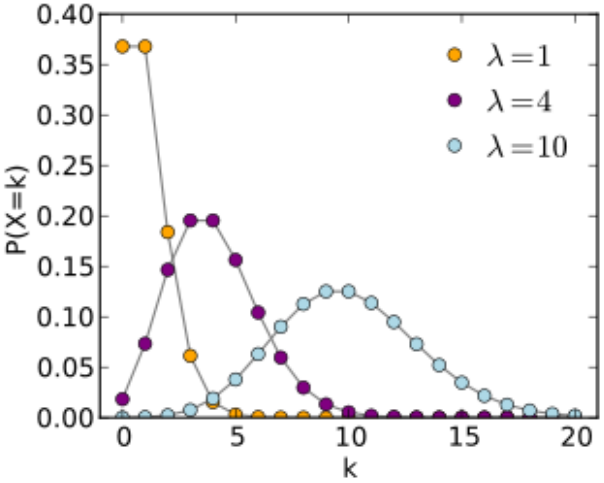


RPKM scatter plot



Statistical Model for Differential Gene Expression Study

- Poisson distribution
  - Approximate the random draw reads from a population with given, fixed fraction of genes
  - Overdispersion issue
    - A poisson distribution assumes mean = variance
    - But, RNA-seq data variance > mean



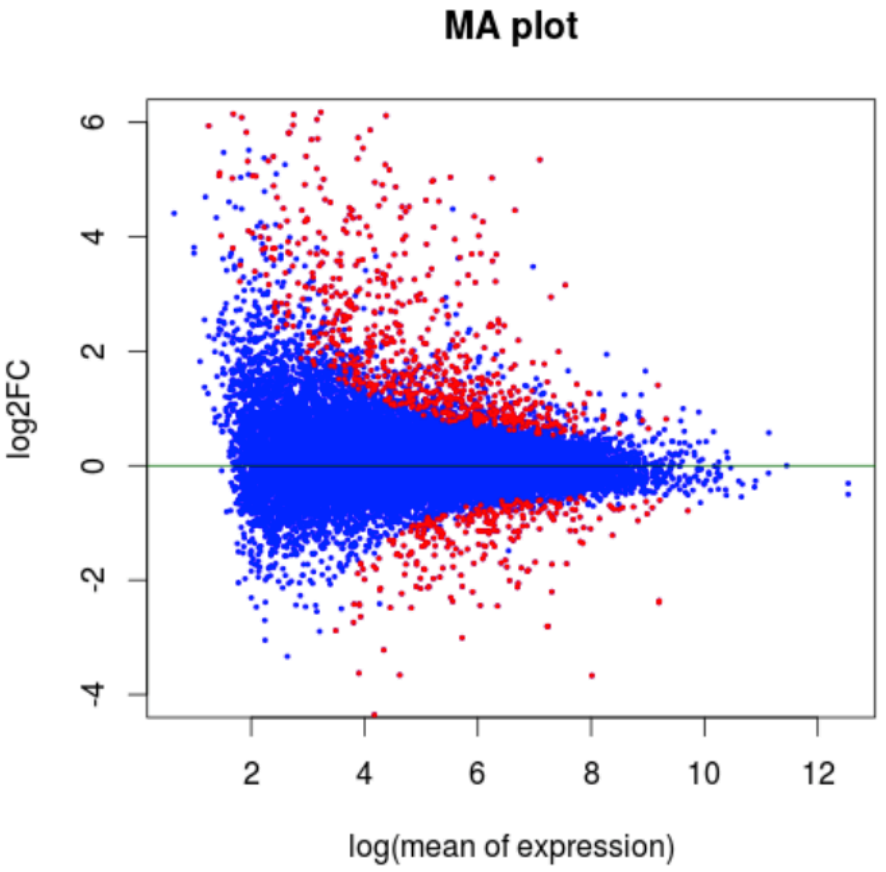
- **Negative binomial (NB)** distribution
  - Characterized by an additional dispersion parameter
  - R packages: edgeR, DESeq2

MA plot

M (Log ratios) and A (mean value)

Fold change: measure how much the expression change from one treatment to the other, for example WT to Mutant.

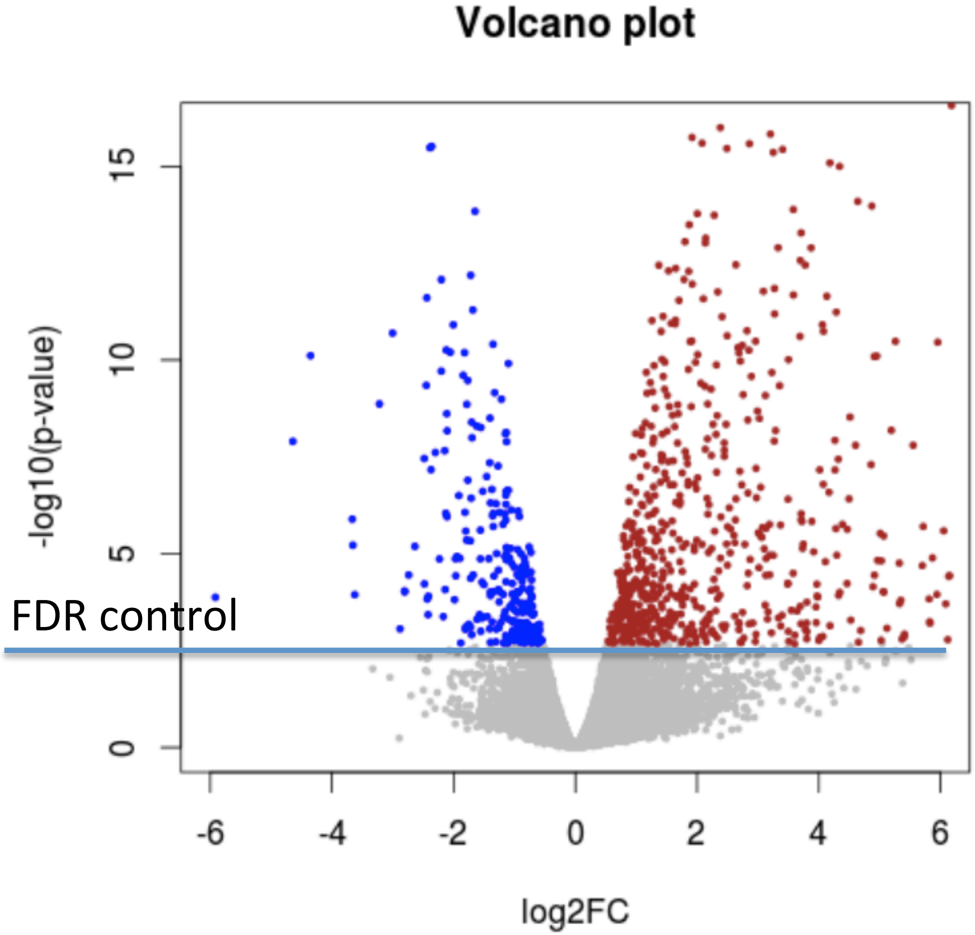
GeneID	Mean RPKM	log mean	log2FC
1	0.51	-0.29	-0.40
2	1.25	0.10	0.03
3	3.52	0.55	-0.89
4	0.19	-0.72	0.30
5	2.34	0.37	-0.36
6	6.14	0.79	-0.07
...			



Volcano Plot

False Discovery Rate: is a statistical method to correct multiple hypothesis testing problem.

DE Result			
GeneID	Log2FC	p-value	-log10(pvalue)
1	-0.40	0.037	1.43
2	0.03	0.916	0.04
3	-0.89	2.42E-05	4.62
4	0.30	0.130	0.89
5	-0.36	0.140	0.85
6	-0.07	0.811	0.09
...			



Take Home Message

- Experimental Design is critical
- Conduct quality checking and then determine how to clean the data
- NB distribution is a widely accepted distribution for modeling read counts data
- Use FDR and FC to clarify differentially expressed genes