Differential Expression

BCB 5200 Introduction Bioinformatics I

Fall 2017

Tae-Hyuk (Ted) Ahn

Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY™

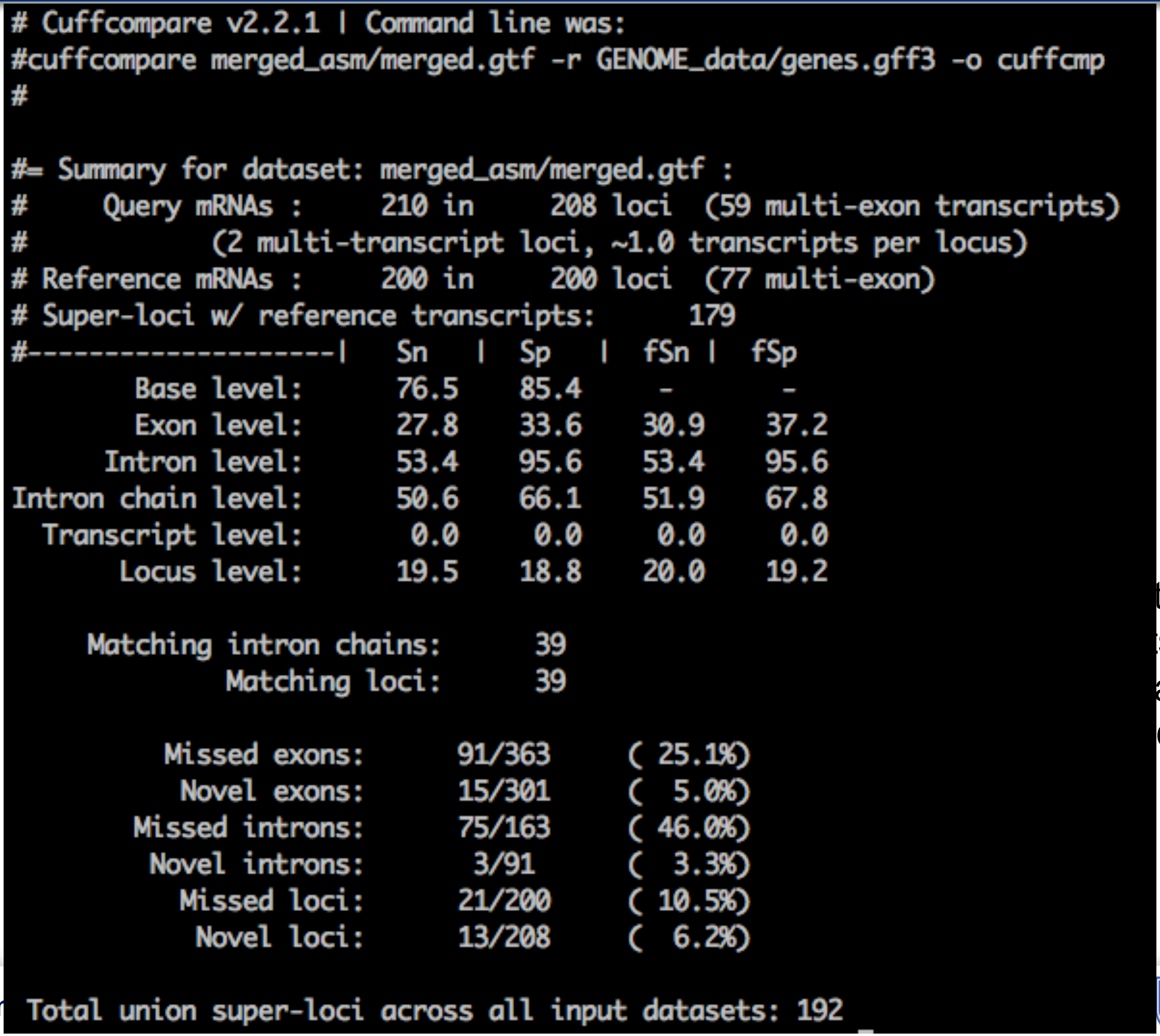
EST. 1818

Cuffcompare output files

- <outprefix>.stats
  - Various statistics related to the accuracy of the transcripts in each sample when compared to the reference annotation data
- <outprefix>.combined.gtf
  - The “union” of all transfrags in all assemblies.
- <cuff_in>.refmap
  - This tab-delimited file lists, for each reference transcript, which Cufflinks transcripts either fully or partially match it
- <cuff_in>.tmap
  - This tab-delimited file lists the most closely matching reference transcript for each Cufflinks transcript
- <outprefix>.tracking
  - This file matches transcripts between samples

http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/

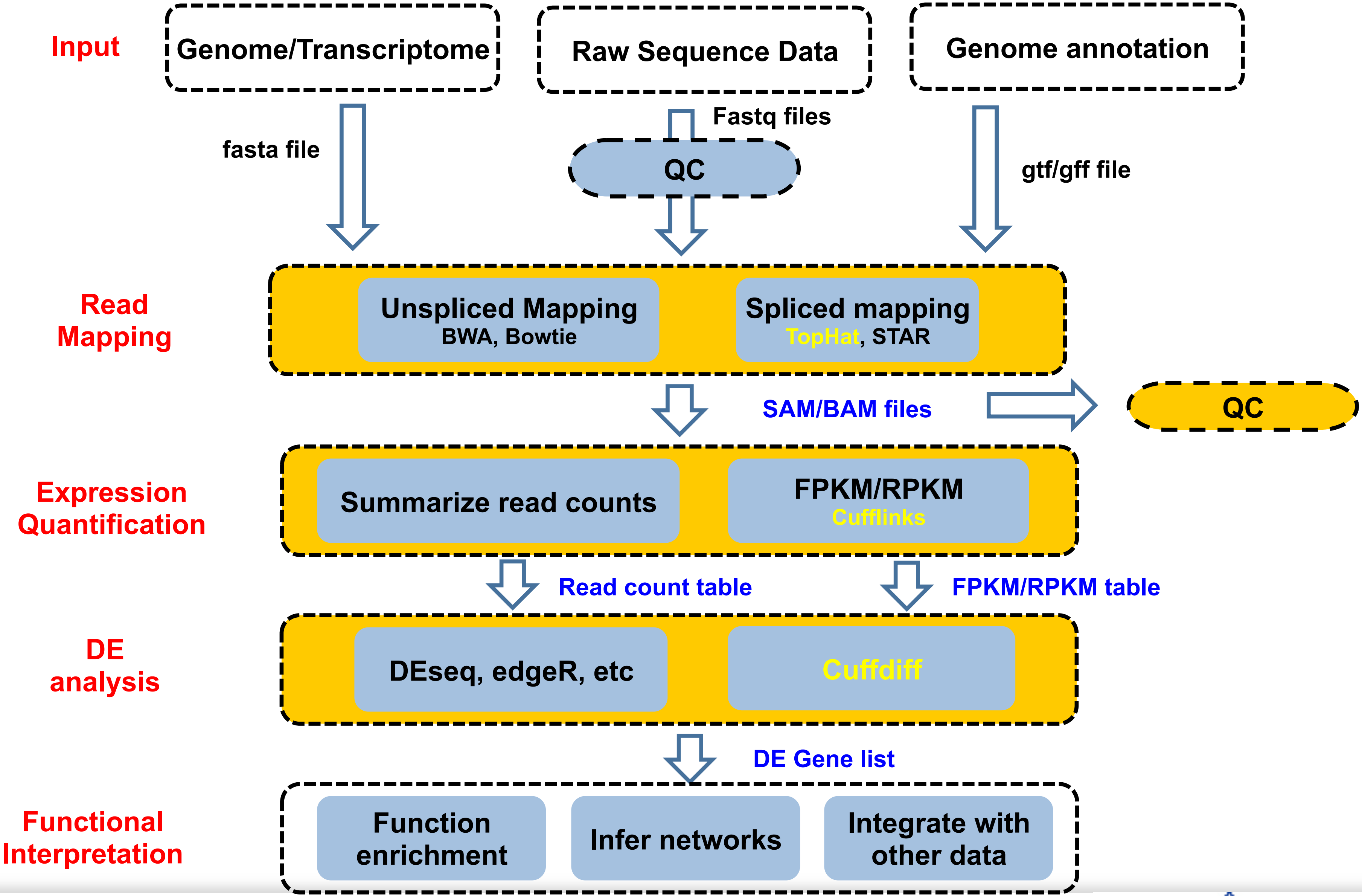
<outprefix>.stats



```
# Cuffcompare v2.2.1 | Command line was:
#cuffcompare merged_asm/merged.gtf -r GENOME_data/genes.gff3 -o cuffcmp
#

#= Summary for dataset: merged_asm/merged.gtf :
#     Query mRNAs :     210 in     208 loci  (59 multi-exon transcripts)
#            (2 multi-transcript loci, ~1.0 transcripts per locus)
# Reference mRNAs :     200 in     200 loci  (77 multi-exon)
# Super-loci w/ reference transcripts:      179
#-------------------|   Sn   |  Sp   |  fSn |  fSp
        Base level:     76.5    85.4     -       -
        Exon level:     27.8    33.6    30.9    37.2
      Intron level:     53.4    95.6    53.4    95.6
Intron chain level:     50.6    66.1    51.9    67.8
  Transcript level:      0.0     0.0     0.0     0.0
       Locus level:     19.5    18.8    20.0    19.2

     Matching intron chains:      39
              Matching loci:      39

          Missed exons:     91/363     ( 25.1%)
           Novel exons:     15/301     (  5.0%)
        Missed introns:     75/163     ( 46.0%)
         Novel introns:      3/91      (  3.3%)
           Missed loci:     21/200     ( 10.5%)
            Novel loci:     13/208     (  6.2%)

 Total union super-loci across all input datasets: 192
```

to the
s in
ared to
data

From reads to differential expression: reference based



Input

Genome/Transcriptome

Raw Sequence Data

Genome annotation

fasta file

Fastq files

QC

gtf/gff file

Read Mapping

Unspliced Mapping
BWA, Bowtie

Spliced mapping
TopHat, STAR

SAM/BAM files

QC

Expression Quantification

Summarize read counts

FPKM/RPKM
Cufflinks

Read count table

FPKM/RPKM table

DE analysis

DEseq, edgeR, etc

Cuffdiff

DE Gene list

Functional Interpretation

Function enrichment

Infer networks

Integrate with other data

RPKM and FPKM

- RPKM : Reads per kilobase per million mapped reads

$$\mathrm{RPKM} = \frac{total\ exon\ reads}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$

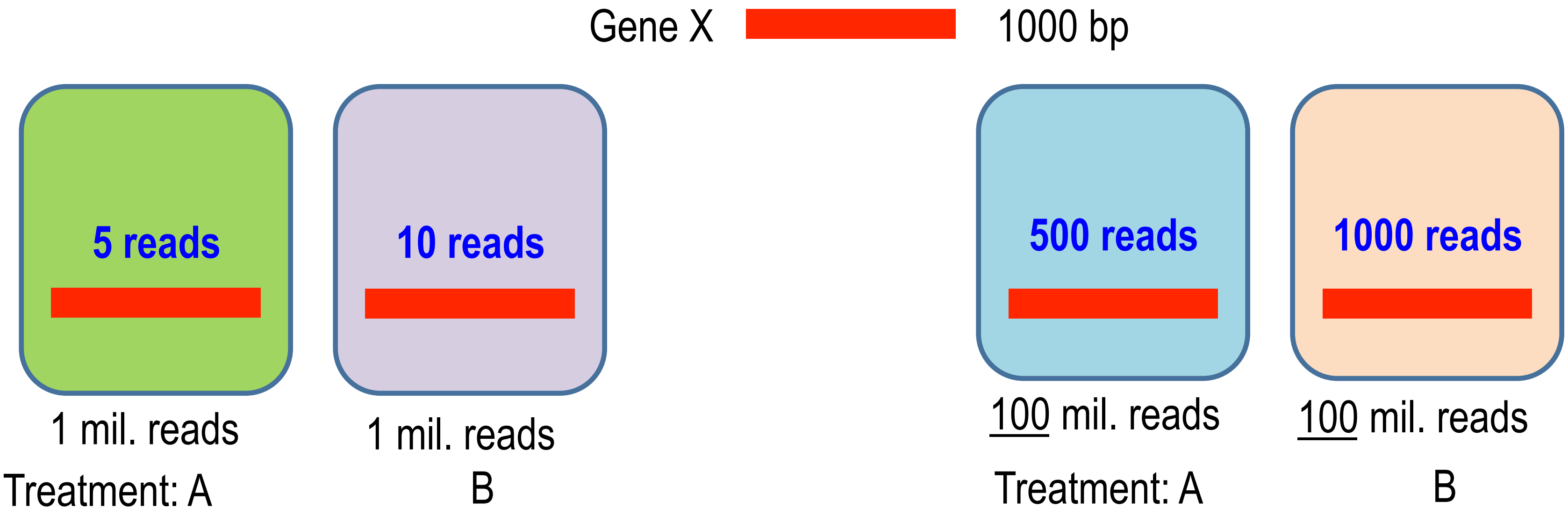
- FPKM : for paired-end sequencing

$$\mathrm{FPKM} = \frac{total\ fragments}{mapped\ reads\ (millions) * exon\ length\ (KB)}$$

RPKM vs FPKM

- People who use cufflinks end up with FPKM and ERANGE with RPKM. Cufflinks has nice explanation why FPKM save us from the skewed expression values called by other software tools. However, RPKM is a general metric.
- Differences between FPKM and RPKM are most likely due to the complicated procedure the cufflinks follows to estimate isoform abundance,
- Each region(gene) has multiple transcripts and each transcript has multiple exons, but the transcripts in a region share exons and that's why the reads don't map precisely, but probabilistically. That's why instead of giving you read count numbers, Cufflinks gives this fancy FPKM number.

Be Careful with RPKM/FPKM Values



RKPM of Gene X

In A, X_{FPKM}= 5 reads/(1kb*1mil)=5

In B, X_{FPKM}= 10 reads/(1kb*1mil)=10

RKPM of Gene X

In A, X_{FPKM}= 500 reads/(1kb*100mil)=5

In B, X_{FPKM}= 1000reads/(1kb*100mil)=10

the RPKM values would be the same for both scenarios

In the latter case, we can be much more confident that there is a true difference between the two treatments than in the first one

Thus, RPKM/FPKM are useful for reporting expression values, but NOT for statistical testing!

Why raw count?

- In principle, counting reads that map to a catalog of features is straightforward.
- Raw read counts is required to correctly model the Poisson component of the sample-to-sample variation
- raw read counts is required for statistical inference based on the negative binomial distribution.

Anders S, et al 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc 8:1765-1786.

Why raw count?

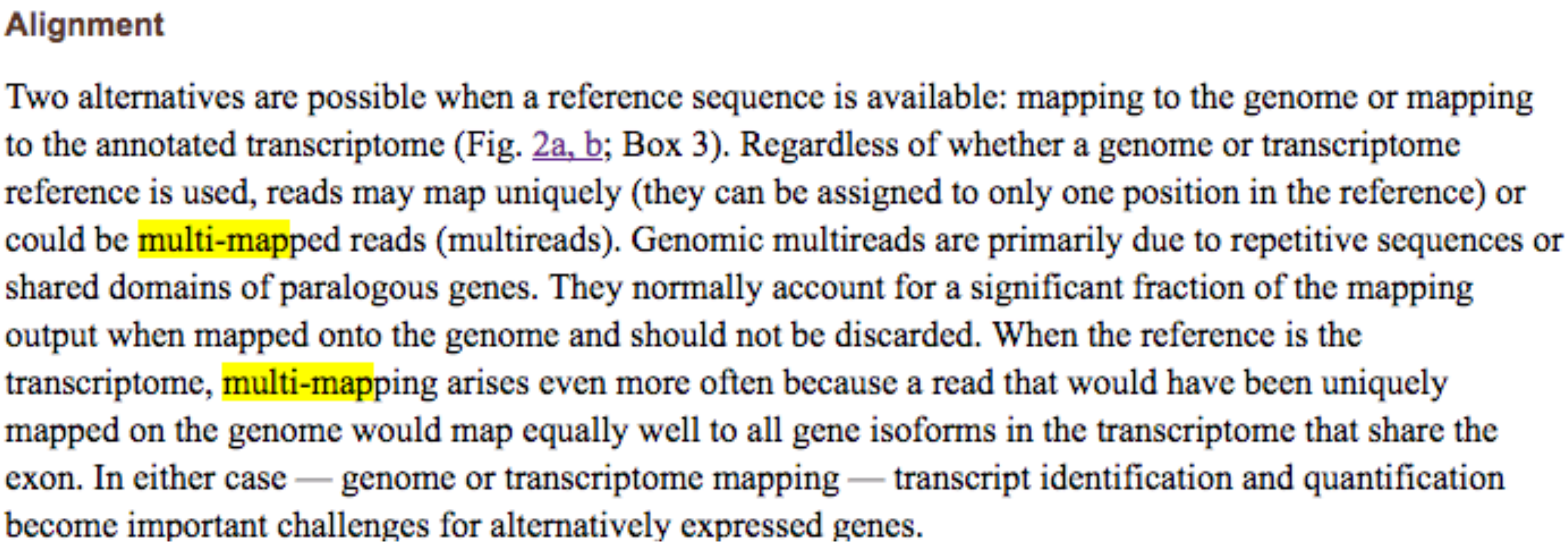
- Both DESeq and edgeR internally keep the raw counts and normalization factors separate, as this full information is needed to correctly model the data.
- No prior normalization or other transformation should be applied, including quantities such as RPKM, FPKM or otherwise depth-adjusted read counts.

Anders S, et al 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc 8:1765-1786.

Counting rules

- Count reads, not base-pairs
- Count each read at most once.
- Discard a read if
  - it cannot be uniquely mapped
  - its alignment overlaps with several genes
  - the alignment quality score is bad
  - (for paired-end reads) the mates do not map to the same gene

Multi-mapped reads

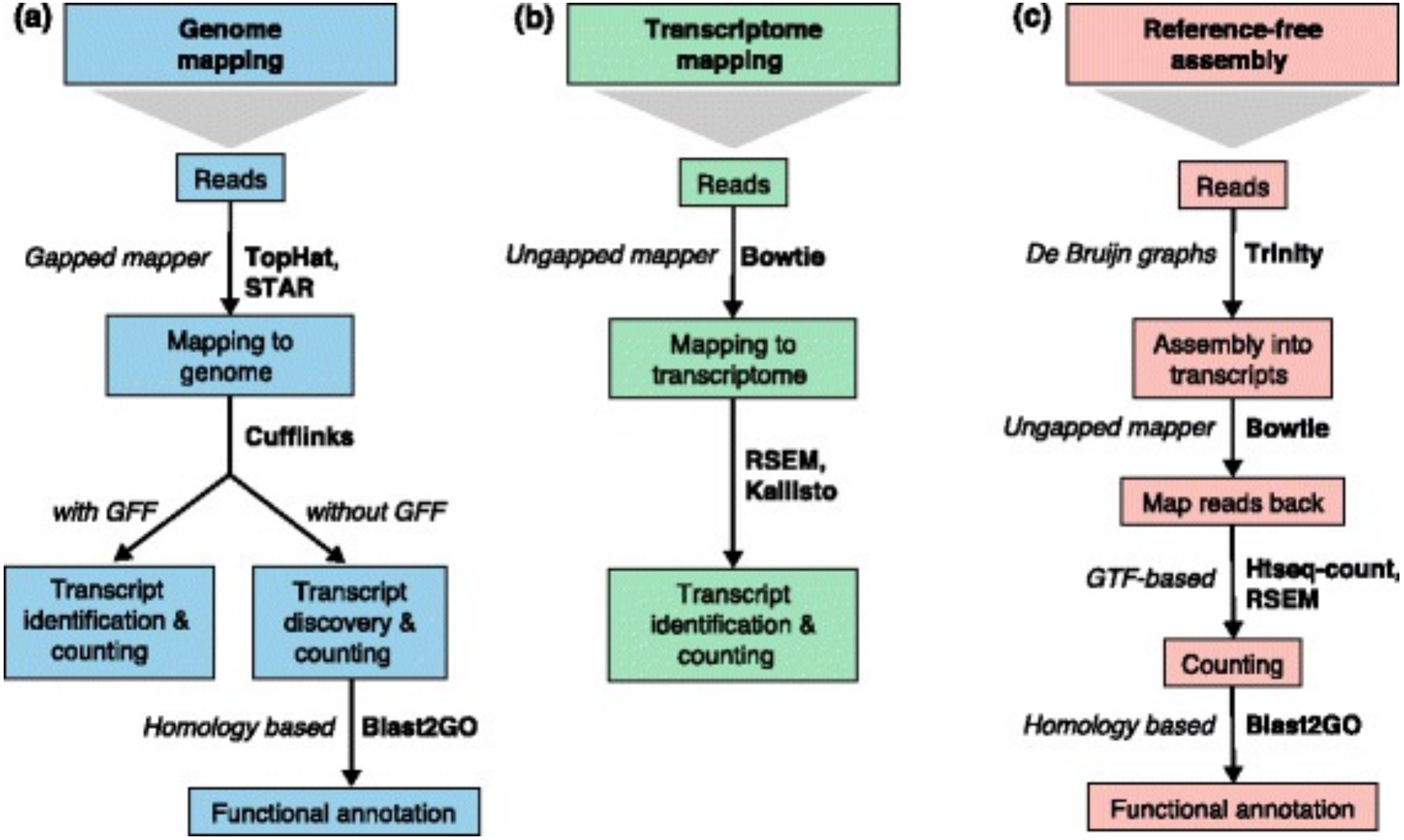


Alignment

Two alternatives are possible when a reference sequence is available: mapping to the genome or mapping to the annotated transcriptome (Fig. 2a, b; Box 3). Regardless of whether a genome or transcriptome reference is used, reads may map uniquely (they can be assigned to only one position in the reference) or could be multi-mapped reads (multireads). Genomic multireads are primarily due to repetitive sequences or shared domains of paralogous genes. They normally account for a significant fraction of the mapping output when mapped onto the genome and should not be discarded. When the reference is the transcriptome, multi-mapping arises even more often because a read that would have been uniquely mapped on the genome would map equally well to all gene isoforms in the transcriptome that share the exon. In either case — genome or transcriptome mapping — transcript identification and quantification become important challenges for alternatively expressed genes.

Conesa et al., A survey of best practices for RNA-seq data analysis, Genome Biol. 2016; 17: 13

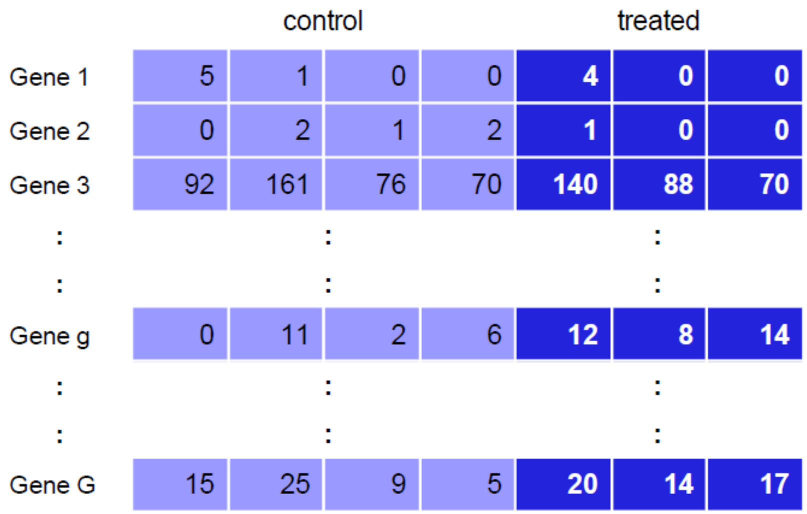
Alignment



Conesa et al., A survey of best practices for RNA-seq data analysis, Genome Biol. 2016; 17: 13

DE: How to quantify the difference?

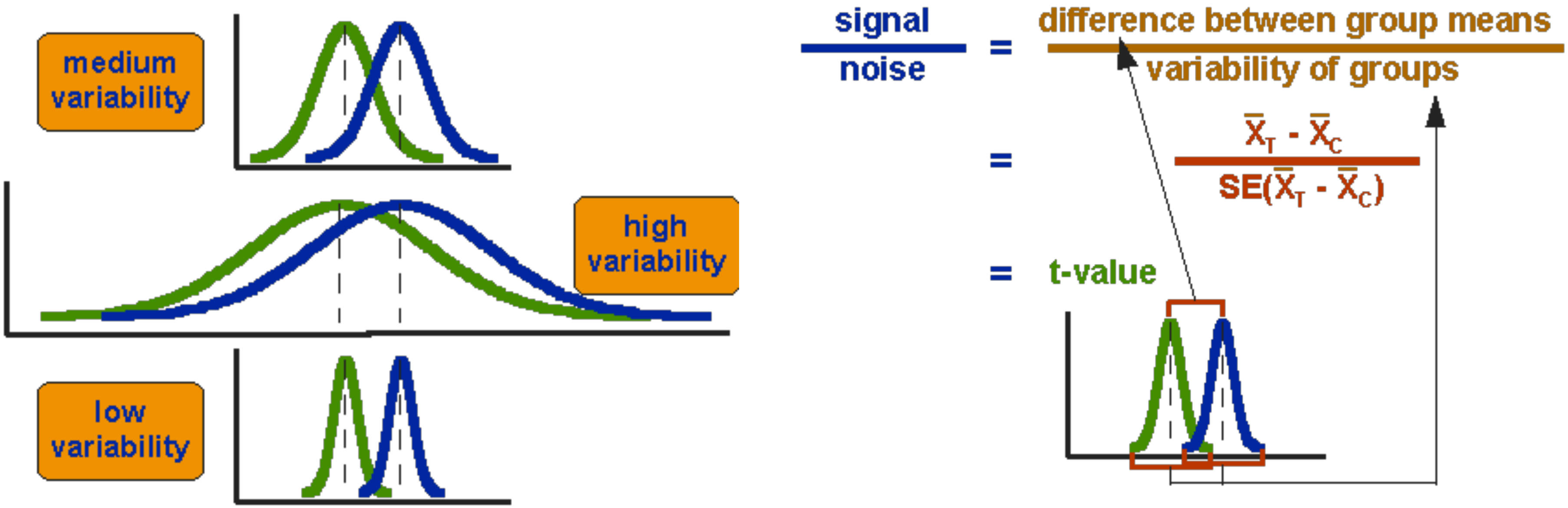
- Having quantified and normalized expression values, an important question is:
  - how to do statistical testing to decide whether, for a given gene, an observed difference in read counts is significant?



	control				treated		
Gene 1	5	1	0	0	4	0	0
Gene 2	0	2	1	2	1	0	0
Gene 3	92	161	76	70	140	88	70
:			:			:	
:			:			:	
Gene g	0	11	2	6	12	8	14
:			:			:	
:			:			:	
Gene G	15	25	9	5	20	14	17

Which genes have different expression level between control and treatment?

DE: How to quantify the difference?



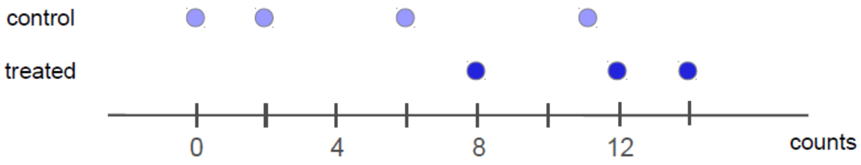
$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)} = \text{t-value}$$

In t-test, to quantify the difference between two groups of data, the variability of each group need to be calculated

http://www.socialresearchmethods.net/kb/stat_t.php

DE: How to quantify the difference?

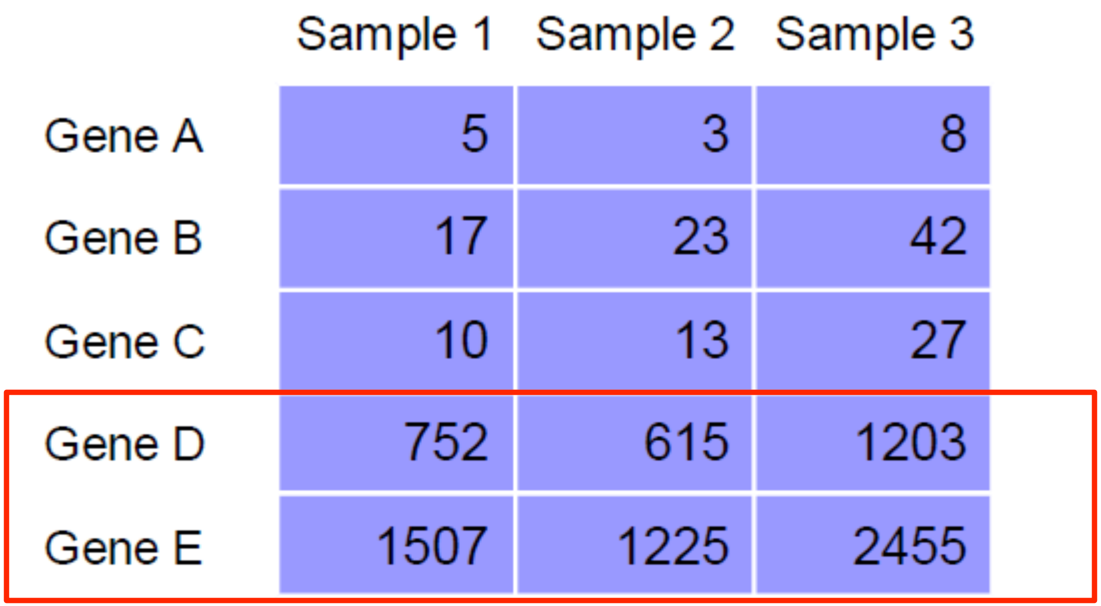
	control				treated		
Gene 1	5	1	0	0	**4**	**0**	**0**
Gene 2	0	2	1	2	**1**	**0**	**0**
Gene 3	92	161	76	70	**140**	**88**	**70**
:			:			:	
:			:			:	
Gene g	0	11	2	6	**12**	**8**	**14**
:			:			:	
:			:			:	
Gene G	15	25	9	5	**20**	**14**	**17**



A major challenge of DE analysis of RNA-seq data is limited number of replicates.

Solution: use statistical model to estimate the variability of each gene, based on limited number of replicates

Interpreting read counts

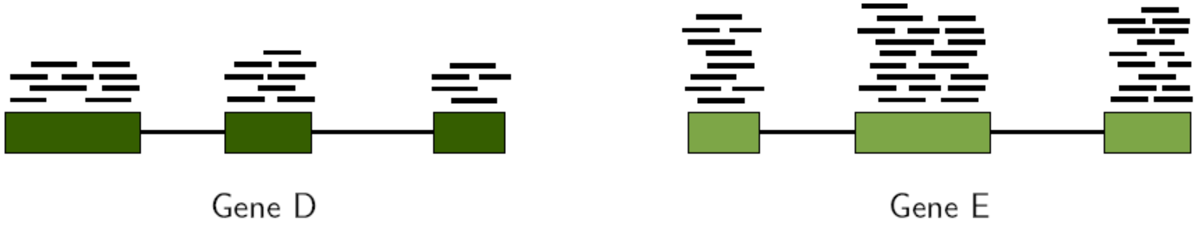


	Sample 1	Sample 2	Sample 3
Gene A	5	3	8
Gene B	17	23	42
Gene C	10	13	27
Gene D	752	615	1203
Gene E	1507	1225	2455

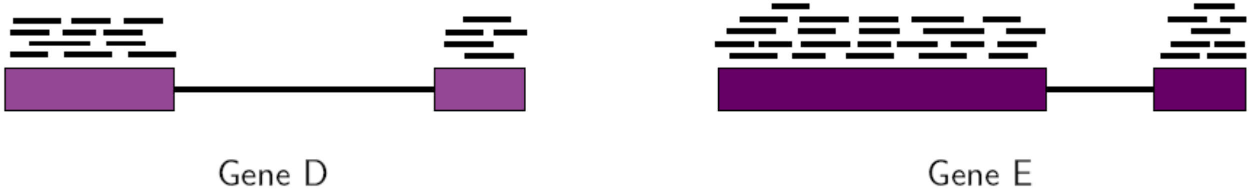
Gene E has about twice as many reads aligned to it as Gene D.

What does it mean?

This could mean..



1) Gene E is expressed with twice as many transcripts as Gene D



2) Both genes are expressed with the same number of transcripts but Gene E is twice as long as Gene D and produces twice as many fragments.

Conclusion: number of reads ≠expression level.

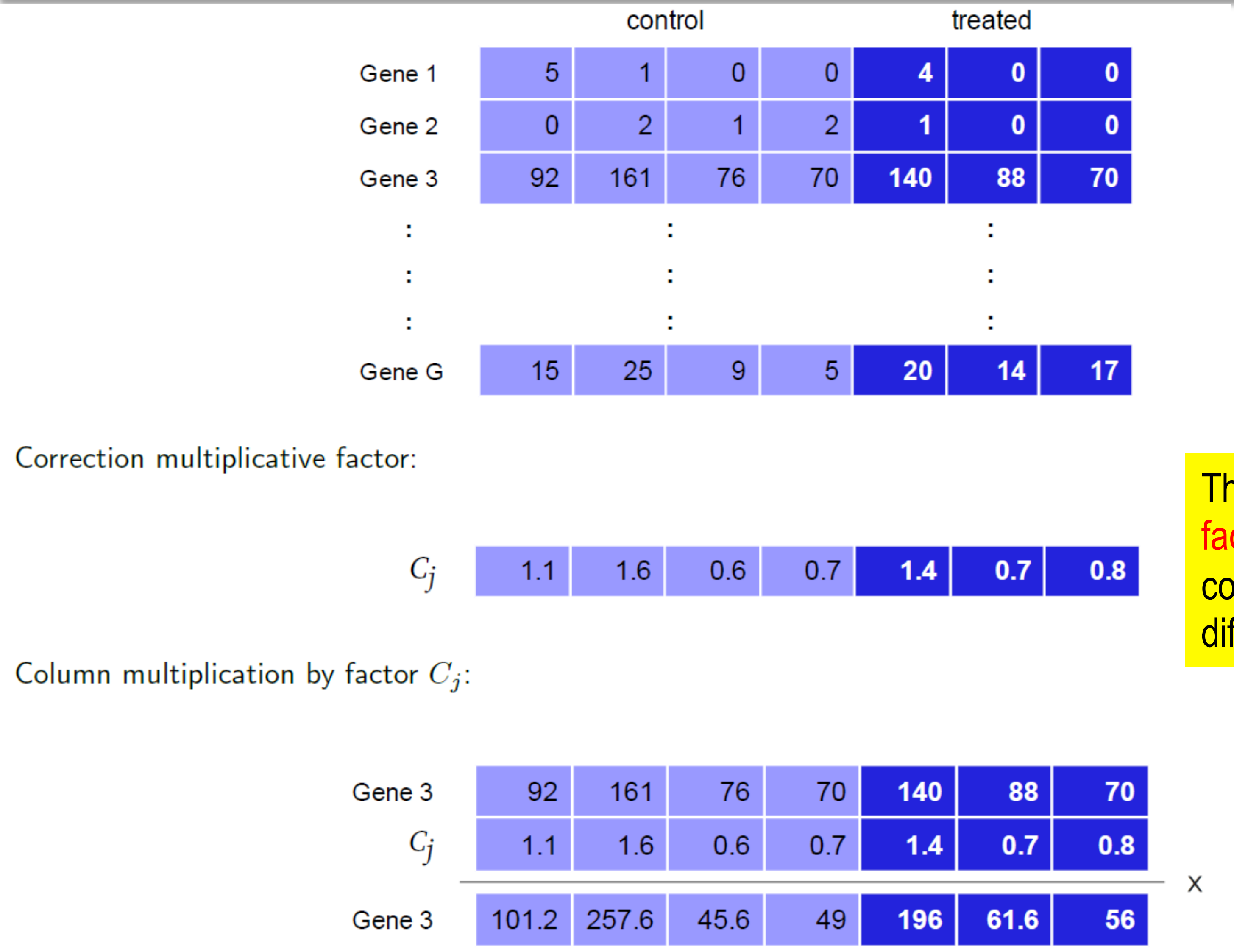
http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

Normalization for library size

- Differences in the total number of aligned reads need to be normalized before differential expression analysis
- It can be achieved by scaling raw read counts in each sample by a single sample-specific factor to scale the counts for each sample

Brief Bioinform (2013) 14 (6): 671-683.

Global normalization methods: basic idea



	control				treated		
Gene 1	5	1	0	0	4	0	0
Gene 2	0	2	1	2	1	0	0
Gene 3	92	161	76	70	140	88	70
:		:				:	
:		:				:	
:		:				:	
Gene G	15	25	9	5	20	14	17

Correction multiplicative factor:

C_j	1.1	1.6	0.6	0.7	1.4	0.7	0.8

Column multiplication by factor C_j:

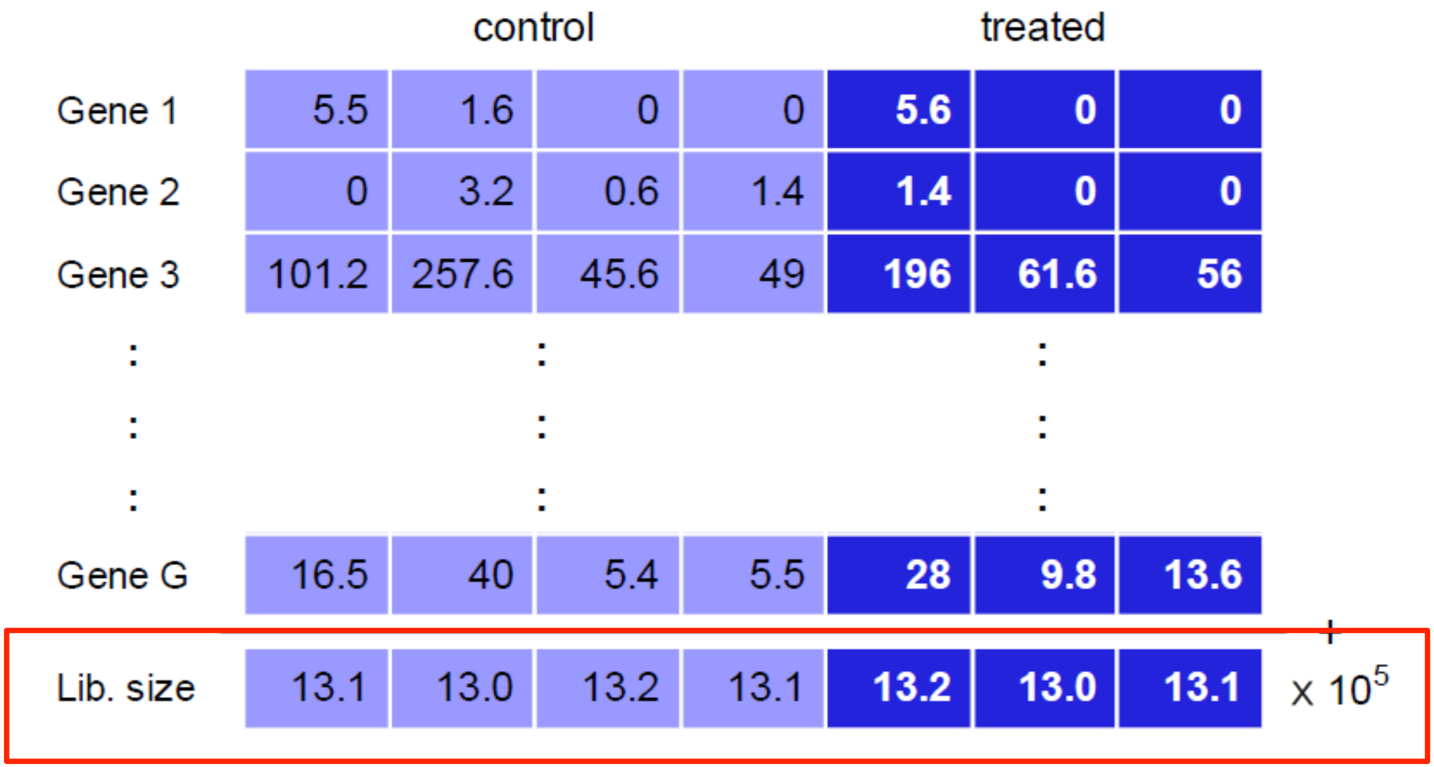
Gene 3	92	161	76	70	140	88	70
C_j	1.1	1.6	0.6	0.7	1.4	0.7	0.8
Gene 3	101.2	257.6	45.6	49	196	61.6	56

x

The purpose of the size factors C_j is to render comparable the counts of different samples

http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

Global normalization methods: basic idea



	control				treated			
Gene 1	5.5	1.6	0	0	**5.6**	**0**	**0**	
Gene 2	0	3.2	0.6	1.4	**1.4**	**0**	**0**	
Gene 3	101.2	257.6	45.6	49	**196**	**61.6**	**56**	
:		:				:		
:		:				:		
:		:				:		
Gene G	16.5	40	5.4	5.5	**28**	**9.8**	**13.6**	
Lib. size	13.1	13.0	13.2	13.1	**13.2**	**13.0**	**13.1**	x 10^5

Normalized library sizes are roughly equal.

http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

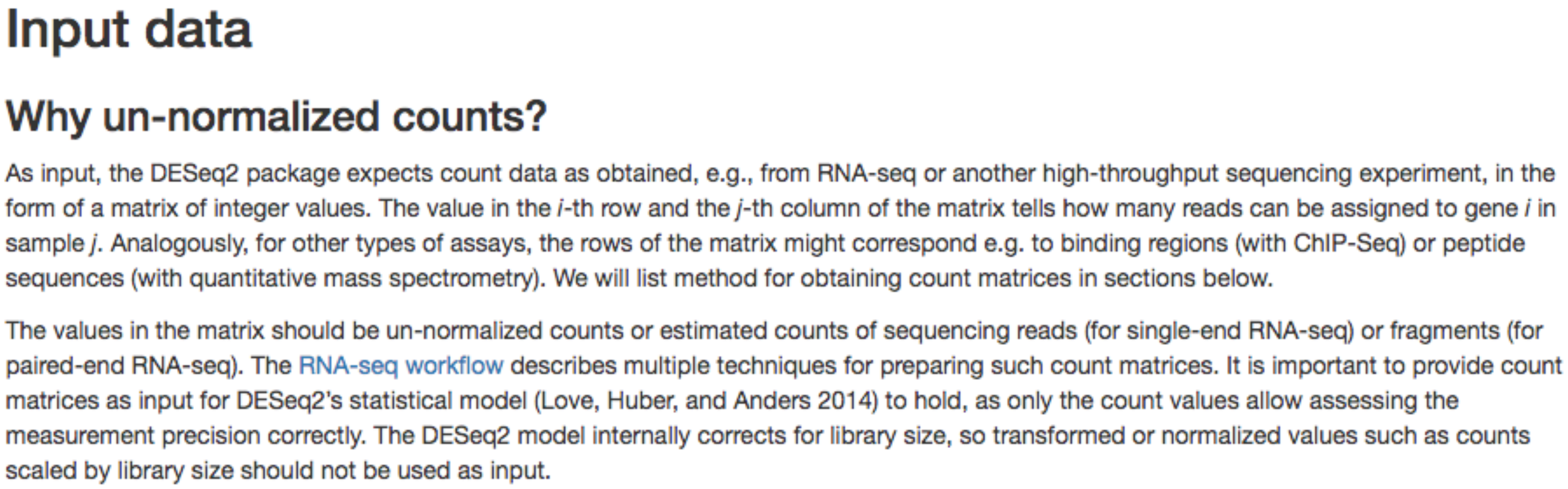
Normalization methods

- Ways to calculate normalization factor C_j associated with sample *j*:
  - Total Count (TC)
  - Upper Quartile (UQ)
  - Median (Med)
  - Trimmed Mean of M-values (TMM)
- Please study it with http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

Brief Bioinform (2013) 14 (6): 671-683.

Again, you should use RAW dataset in many tools

- For example, DESeq2
  - https://bioconductor.org/packages/3.7/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#why-un-normalized-counts



Input data

Why un-normalized counts?

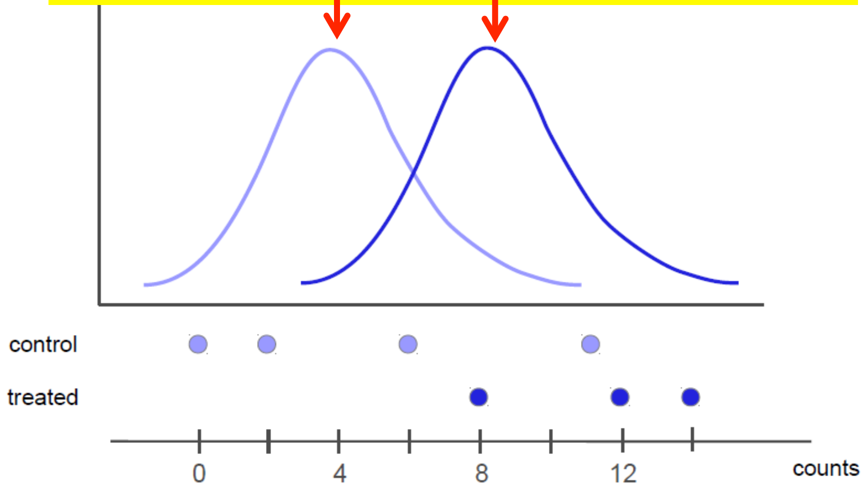
As input, the DESeq2 package expects count data as obtained, e.g., from RNA-seq or another high-throughput sequencing experiment, in the form of a matrix of integer values. The value in the *i*-th row and the *j*-th column of the matrix tells how many reads can be assigned to gene *i* in sample *j*. Analogously, for other types of assays, the rows of the matrix might correspond e.g. to binding regions (with ChIP-Seq) or peptide sequences (with quantitative mass spectrometry). We will list method for obtaining count matrices in sections below.

The values in the matrix should be un-normalized counts or estimated counts of sequencing reads (for single-end RNA-seq) or fragments (for paired-end RNA-seq). The RNA-seq workflow describes multiple techniques for preparing such count matrices. It is important to provide count matrices as input for DESeq2's statistical model (Love, Huber, and Anders 2014) to hold, as only the count values allow assessing the measurement precision correctly. The DESeq2 model internally corrects for library size, so transformed or normalized values such as counts scaled by library size should not be used as input.

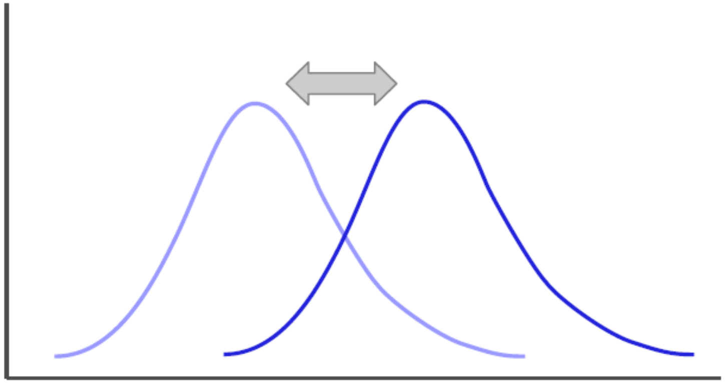
DE: How to quantify the difference

	control				treated		
Gene 1	5	1	0	0	4	0	0
Gene 2	0	2	1	2	1	0	0
Gene 3	92	161	76	70	140	88	70
:			:			:	
:			:			:	
Gene g	0	11	2	6	12	8	14
:			:			:	
:			:			:	
Gene G	15	25	9	5	20	14	17

NB model is estimated for each gene
2 parameters: mean and dispersion



Fit the model for each gene



Use statistics to quantify the difference

Difference is put into p-value

http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

p - value

- What a statistical test determines is how likely that null hypothesis is to be true
- The null hypothesis is the hypothesis that nothing is going on (H_0):

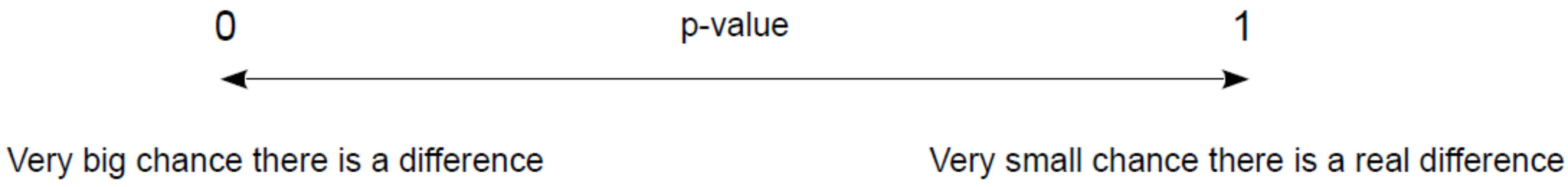
  the gene g is NOT differentially expressed between the conditions

- After a test statistic is computed, it is often converted to a “p-value“
  - If the p-value is small then the null hypothesis is deemed to be untrue and it is rejected in favor of the alternative

http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

p-value

- It is a usual convention in biology to use a critical p-value of 0.05 (often called alpha, α).
- There is nothing magical about p-value < 0.05, it is just a convention.
- This means that the probability of observing data as extreme as this if the null hypothesis is true is 0.05 (5% or 1 in 20)
- In other words, it indicates that the null hypothesis is unlikely to be true



0

p-value

1

Very big chance there is a difference

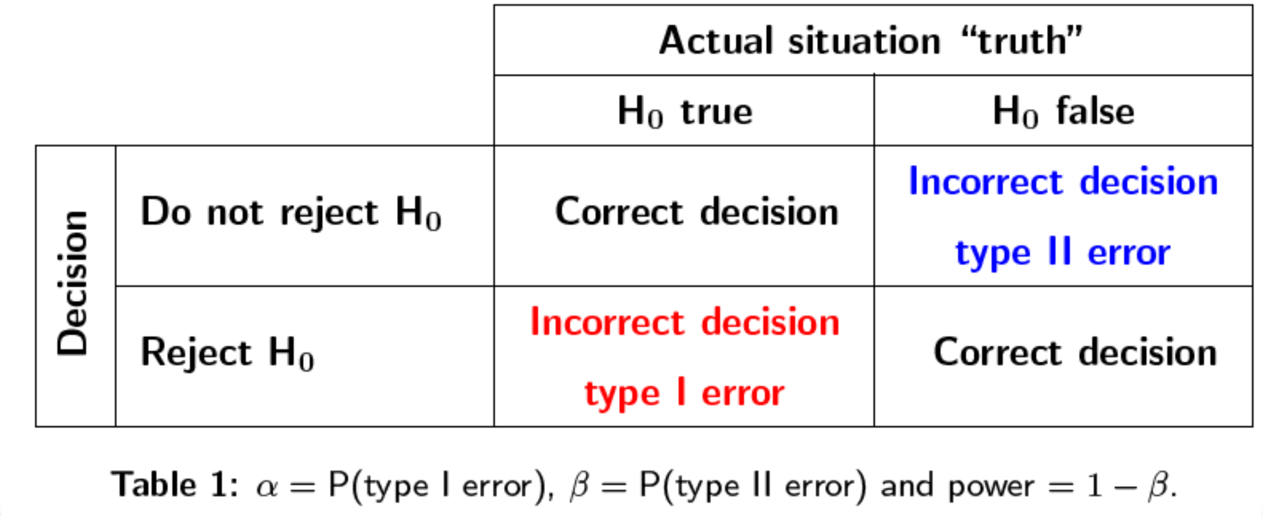
Very small chance there is a real difference

the smaller the p-value the more confident we can be in the conclusions drawn from it.

http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

Type I and Type II error

- Type I error (or false-positive)
  - the null hypothesis is really true (the gene is not differentially expressed) but the statistical test has led you to believe that it is false (there is a difference in expression).
- Type II error (or false-negative)
  - the null hypothesis is really false (the gene is differentially expressed) but the test has not picked up this difference.



		Actual situation "truth"	
		H_0 true	H_0 false
Decision	Do not reject H_0	Correct decision	Incorrect decision type II error
	Reject H_0	Incorrect decision type I error	Correct decision

Table 1: $\alpha = \mathrm{P}(\text{type I error})$, $\beta = \mathrm{P}(\text{type II error})$ and power $= 1 - \beta$.

Multiple testing issue

- The test for each gene has a probability of producing a type I error
- By performing a large number of hypothesis tests, a substantial number of false positives may accumulate

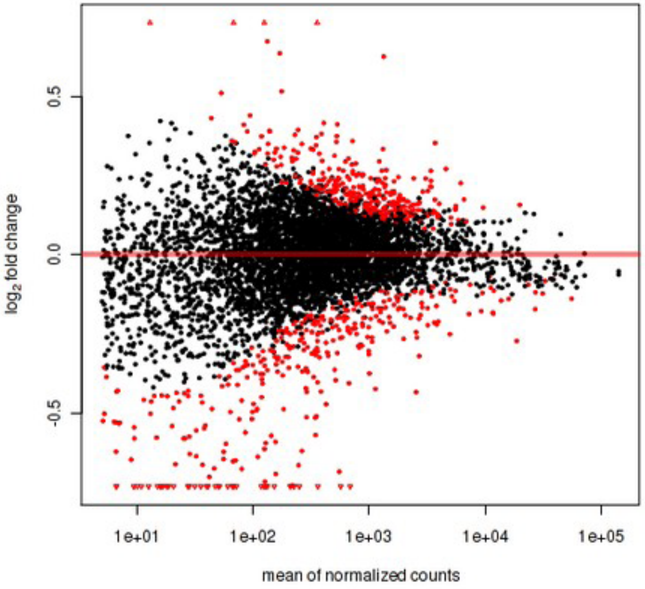
Why Multiple Testing Matters

- Genomics = Lots of Data = Lots of Hypothesis Tests
- Consider:
  - A genome with 10,000 genes (result in performing 10,000 separate hypothesis tests)
  - If we use a standard p-value cut-off of 0.05
  - How many genes are expected to be “significant” by chance?

0.05 × 10,000 = 500 genes.

  - If there are 1,500 genes have $p < 0.05$ under treament, % of false positive?

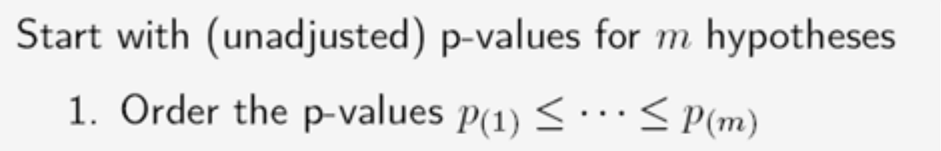
500 genes, i.e., 33%



Multiple testing correction

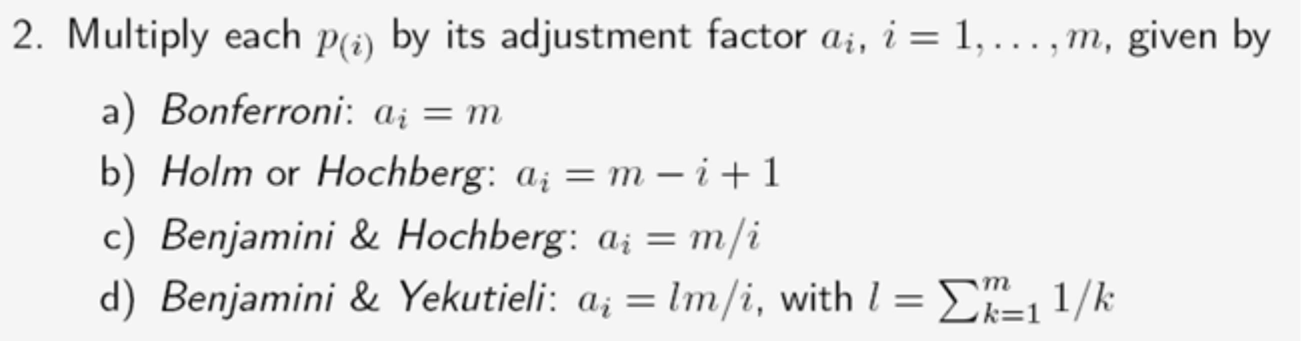
- Goal: to control false discovery rate (FDR)
- FDR is the fraction of false positive in the genes that are classified as DE
- If we set a threshold *α* of 0.05, 5% of the detected DE genes will be false positive
- Calculating adjusted p-values (q-values):
  - Bonferroni correction
  - Benjamini/Hochberg correction
  - Etc.

Calculating adjusted p-values

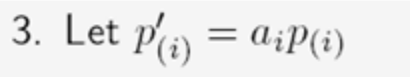


Start with (unadjusted) p-values for m hypotheses

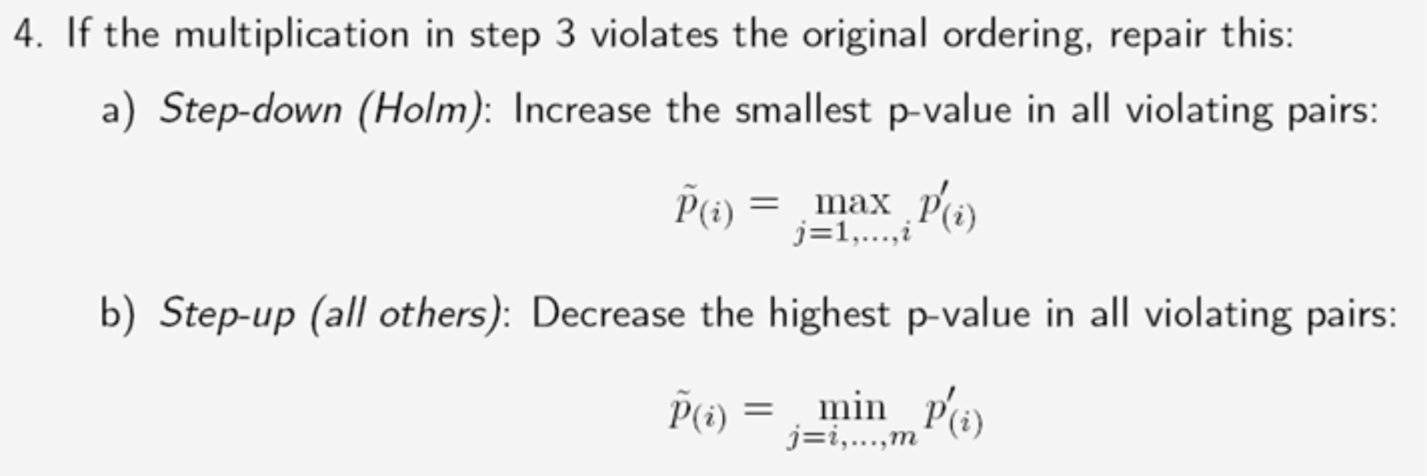
1. Order the p-values $p_{(1)} \leq \cdots \leq p_{(m)}$



2. Multiply each $p_{(i)}$ by its adjustment factor a_i, $i = 1, \ldots, m$, given by
   a) *Bonferroni*: $a_i = m$
   b) *Holm* or *Hochberg*: $a_i = m - i + 1$
   c) *Benjamini & Hochberg*: $a_i = m/i$
   d) *Benjamini & Yekutieli*: $a_i = lm/i$, with $l = \sum_{k=1}^{m} 1/k$



3. Let $p'_{(i)} = a_i p_{(i)}$

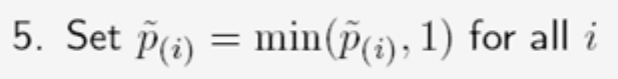


4. If the multiplication in step 3 violates the original ordering, repair this:
   a) *Step-down (Holm)*: Increase the smallest p-value in all violating pairs:

$$\tilde{p}_{(i)} = \max_{j=1,\ldots,i} p'_{(i)}$$

   b) *Step-up (all others)*: Decrease the highest p-value in all violating pairs:

$$\tilde{p}_{(i)} = \min_{j=i,\ldots,m} p'_{(i)}$$



5. Set $\tilde{p}_{(i)} = \min(\tilde{p}_{(i)}, 1)$ for all i

http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

Calculating adjusted p-values from Benjamini-Hochberg method

Benjamini & Hochberg: $a_i = m/i$

$p'_{(i)} = a_i p_{(i)}$

$\tilde{p}_{(i)} = \min(\tilde{p}_{(i)}, 1)$ for all i

- Suppose $m = 100$
- Consider the five genes with lowest p-values
- $\alpha = 0.05$

Gene	$p_{(i)}$	$a_{(i)}$	$p'_{(i)}$	$\tilde{p}_{(i)}$
1	0.00010	100	0.01000	0.00550
2	0.00011	50	0.00550	0.00550
3	0.00520	33	0.17333	0.17333
4	0.02400	25	0.60000	0.60000
5	0.06600	20	1.32000	1.00000

Gene 1 and 2 are declared differentially expressed.

DE: How to quantify the difference?

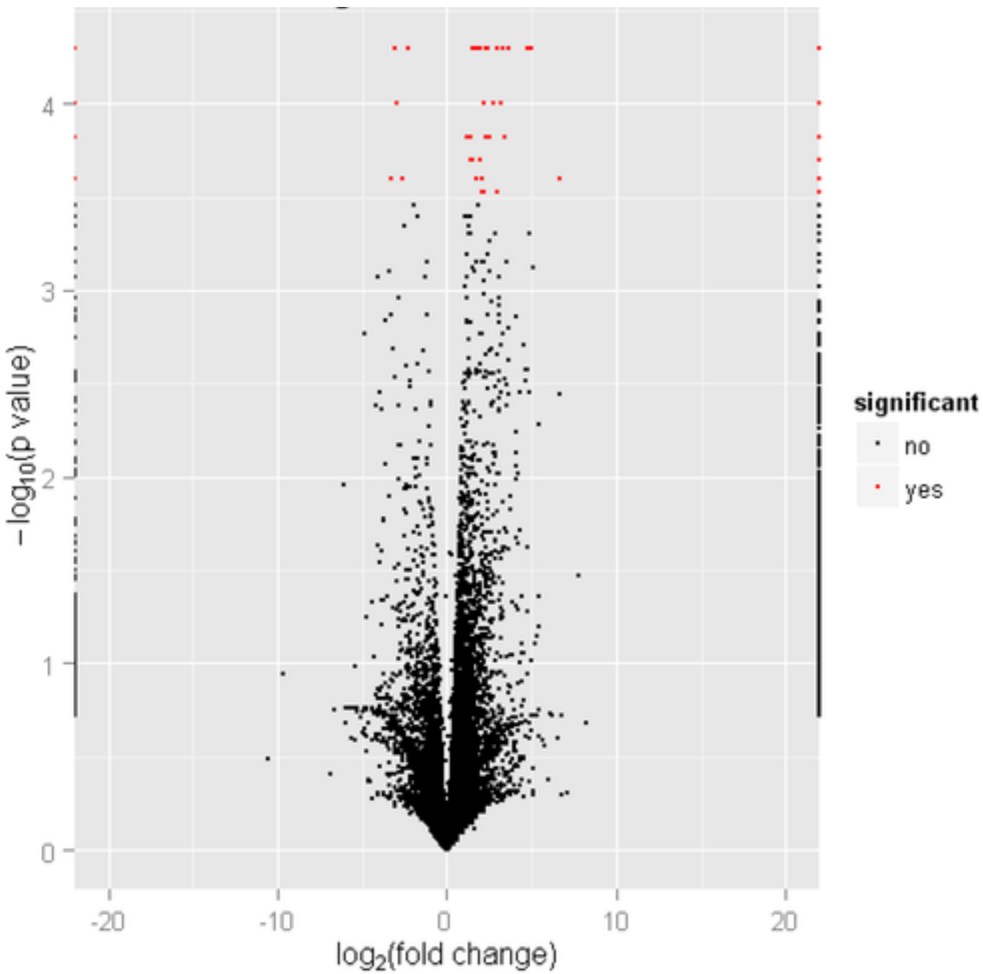
Why log transform?

Fold change = $X_{treatment}/X_{control}$

Original ratio		Log2-tranformed ratio	
Up	Down	Up	Down
2	0.5	1	-1
4	0.25	2	-2
8	0.125	3	-3
16	0.0625	4	-4

The main reason behind this is in order to be able to compare under expression and over expression on the same scale.

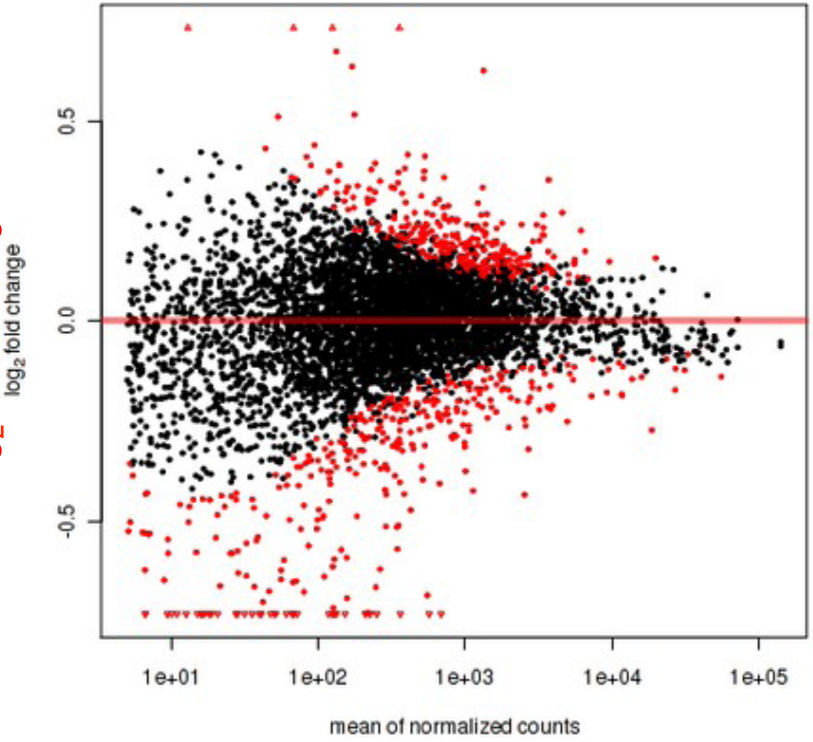
Illustrating DE results



$\log_2$ Fold change

volcano-plot: p value versus the log2 fold change between 2 conditions.

$\log_2$ Fold change



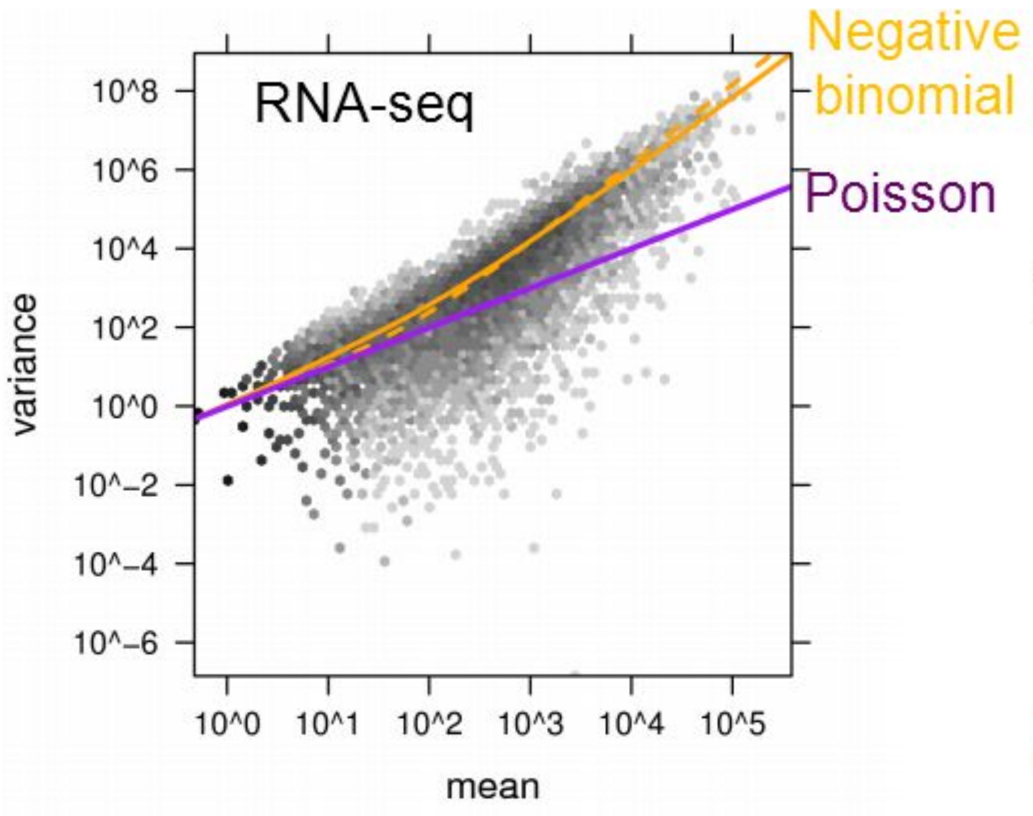
MA-plot: mean of counts versus the log2 fold change between 2 conditions.

Some popular DE tools

	DESeq2	edgeR	Cuffdiff / cummRbund
Normalization	DESeq sizeFactor/ geometric	TMM/upper quartile/RLE	geometric, upper, quartile, fpkm
Read count distribution assumption	Negative binomial	Negative binomial	Negative binomial
DE Test	exact test	exact test	t test
Ref	Genome Biol 2010;11:R106.	Bioinformatics 26, 139–140 (2010)	Nature biotechnology *31*, 46-53 (2013)

Brief Bioinform. 2015 Jan;16(1):59-70

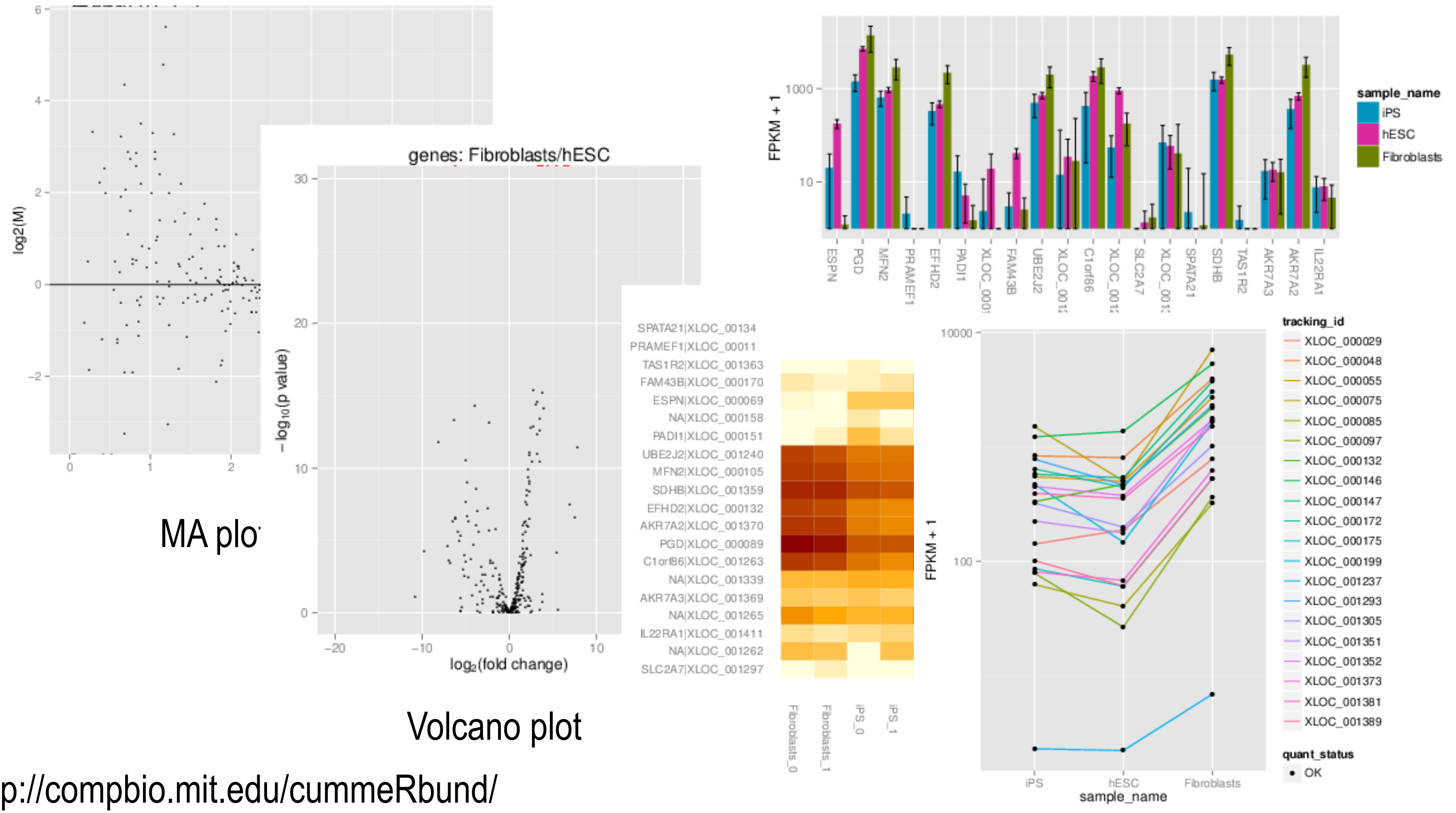
How to calculate variance



- RNA-Seq data was initially modeled as count data fitting a Poisson distribution like the microarray data.
- Issue: genes with high mean counts tend to show more variance between samples than genes with low mean counts (overdispersion)
- Solution: Negative binomial distribution (= Poisson distribution + local regression)

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2),$$

DE results visualization: cummeRbund



MA plo

Volcano plot

http://compbio.mit.edu/cummeRbund/manual_2_0.html

Heatmap

Mini Homework

Please install EdgeR and DESeq2

- https://web.stanford.edu/class/bios221/labs/rnaseq/lab_4_rnaseq.html

- RNA-Seq analysis homework will be announced if all of you are available to run DESeq.