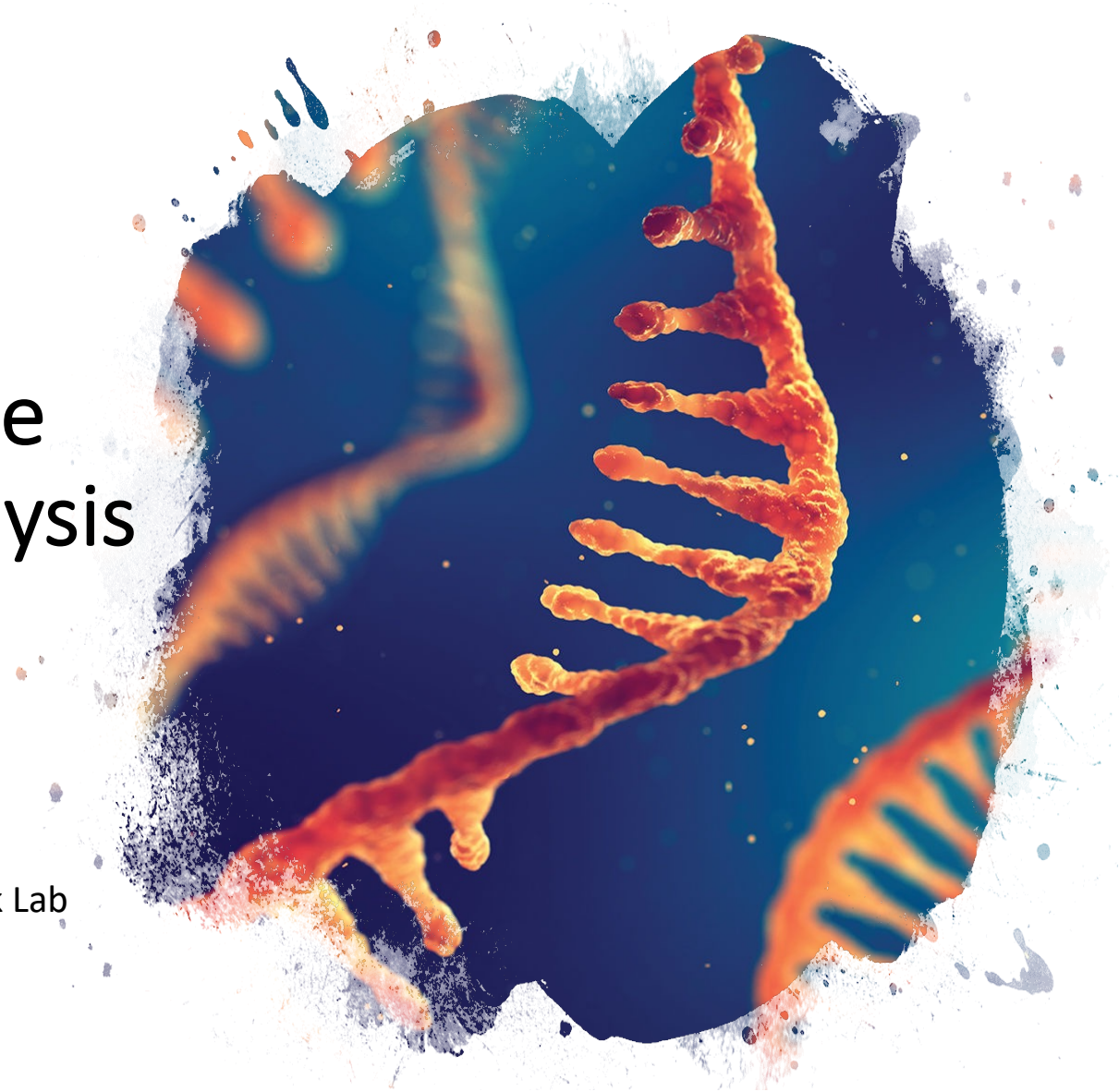
Differential gene expression analysis

2023 Workshop on Genomics,
Česky Krumlov

Rachel Steward
Postdoctoral researcher, Runemark Lab
Lund University



Today's activity

7:00 - 7:30 : Differential expression background

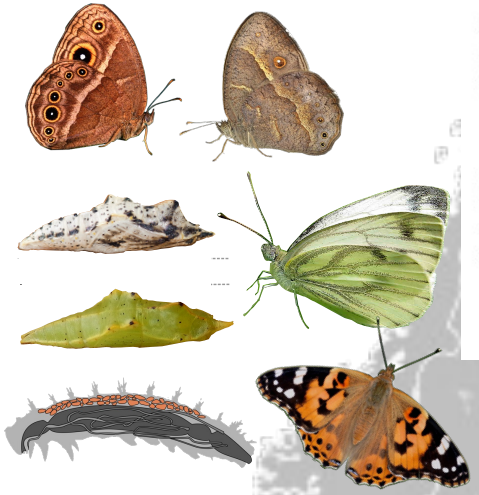
7:30 – 8:30 : Free work time
(take a break when you need/want it)

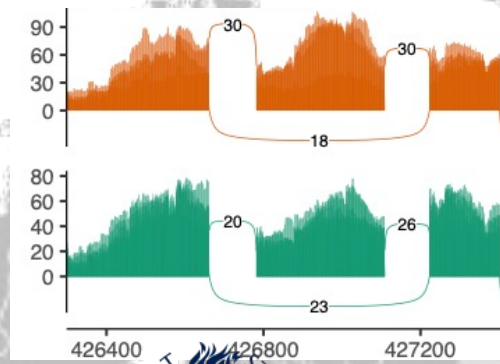
8:30 – 8:45 : Check-in, walk through some preliminary results

8:45 – 9:30 : Free work time
(take a break when you need/want it)

9:30 – 10 : Wrap-up and discussion

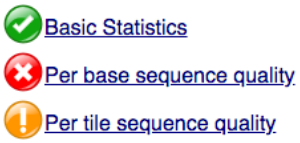
About me



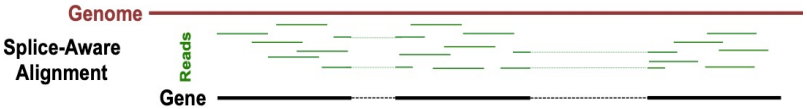


Lecture outline

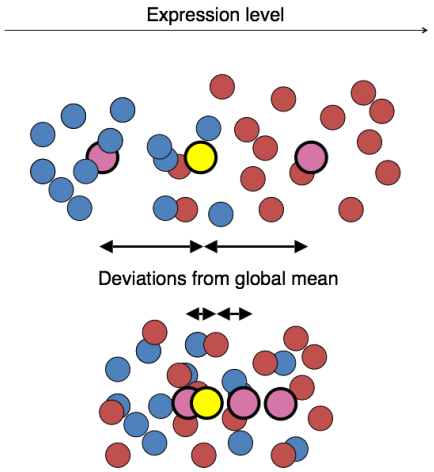
1. Ref-based differential expression overview



2. Trimming, mapping, counting



3. Differential expression analysis



   a. Normalization

   b. Dispersion estimates

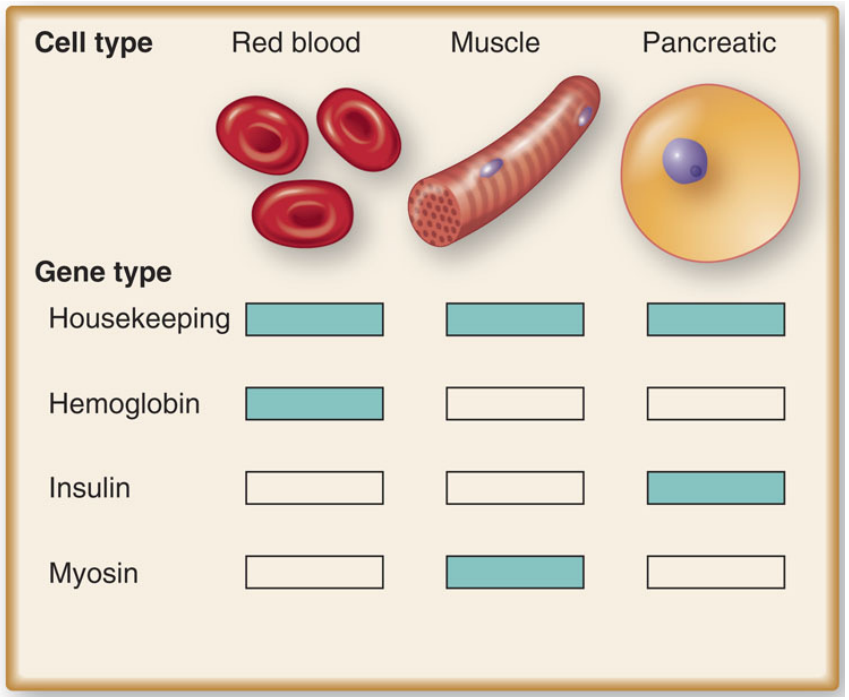
   c. Model fitting

   d. Hypothesis testing & output

Gene expression

The selective activity of certain genes is a highly regulated process called gene expression.

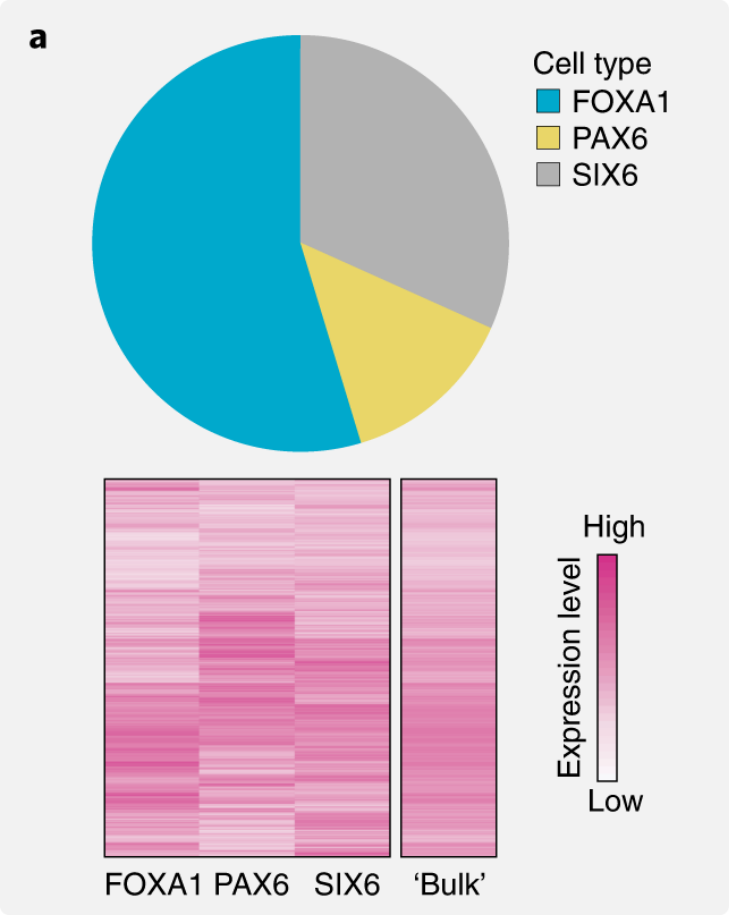
Gene expression is a characteristic of space (e.g., cell type, tissue, etc.) and time (e.g., developmental stage, time after event)



Gene expression

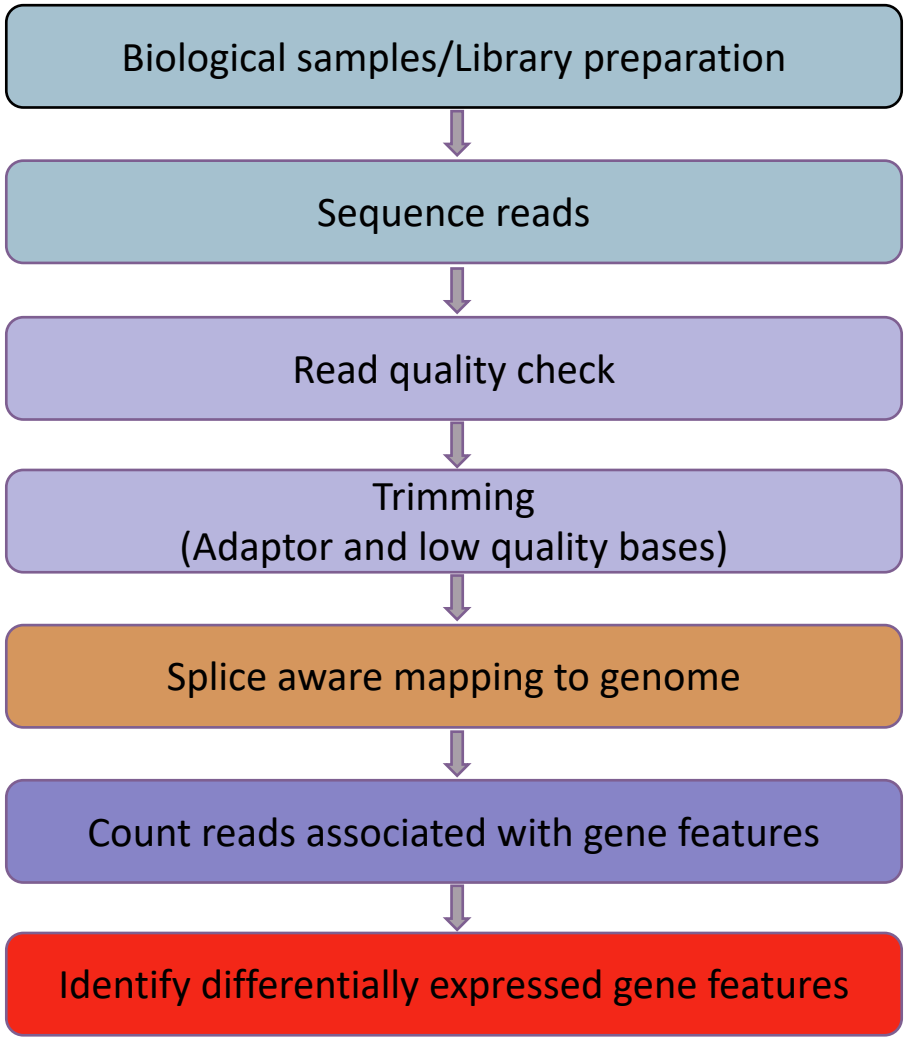
The selective activity of certain genes is a highly regulated process called gene expression.

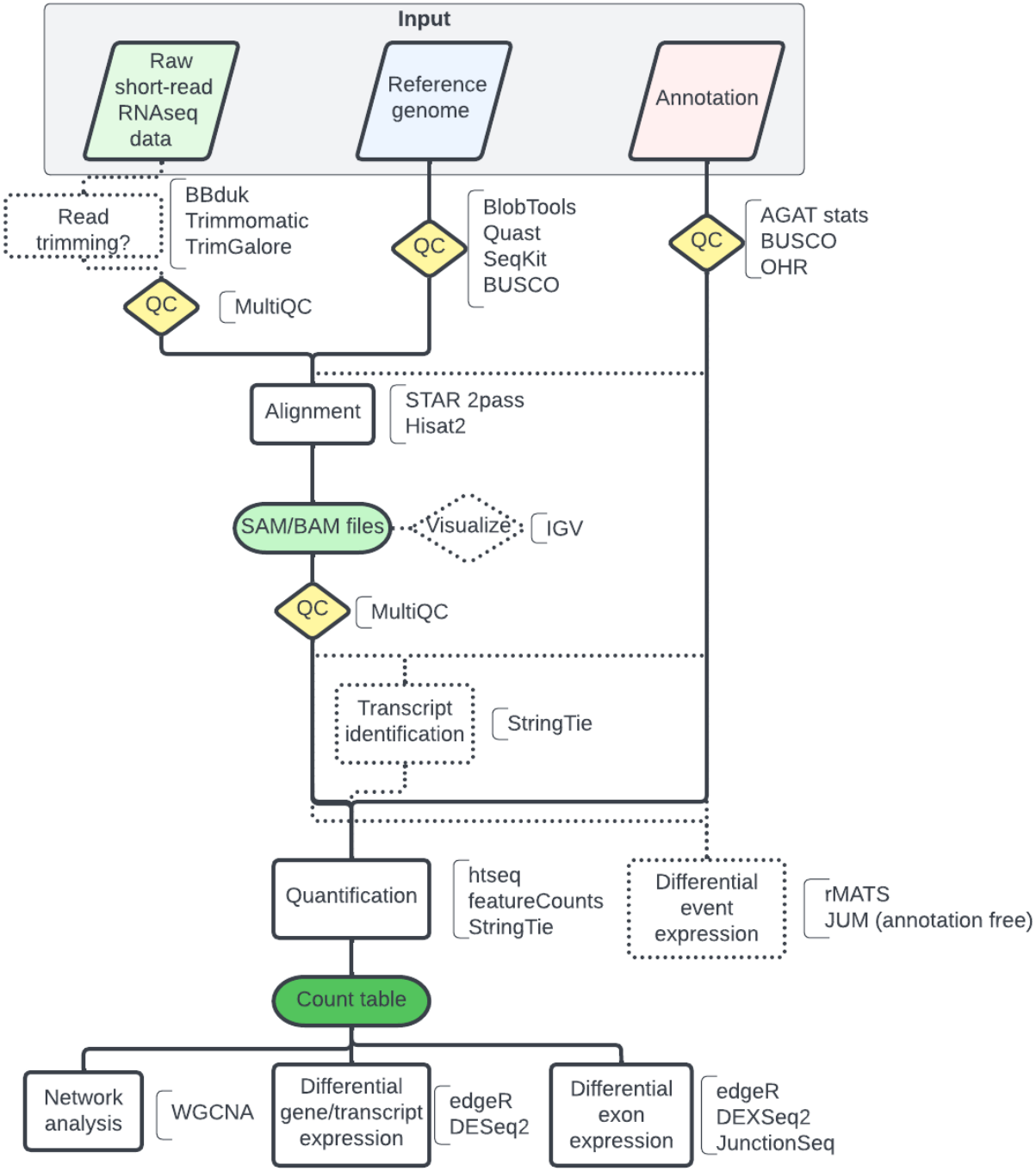
Gene expression is a characteristic of space (e.g., cell type, tissue, etc.) and time (e.g., developmental stage, time after event)

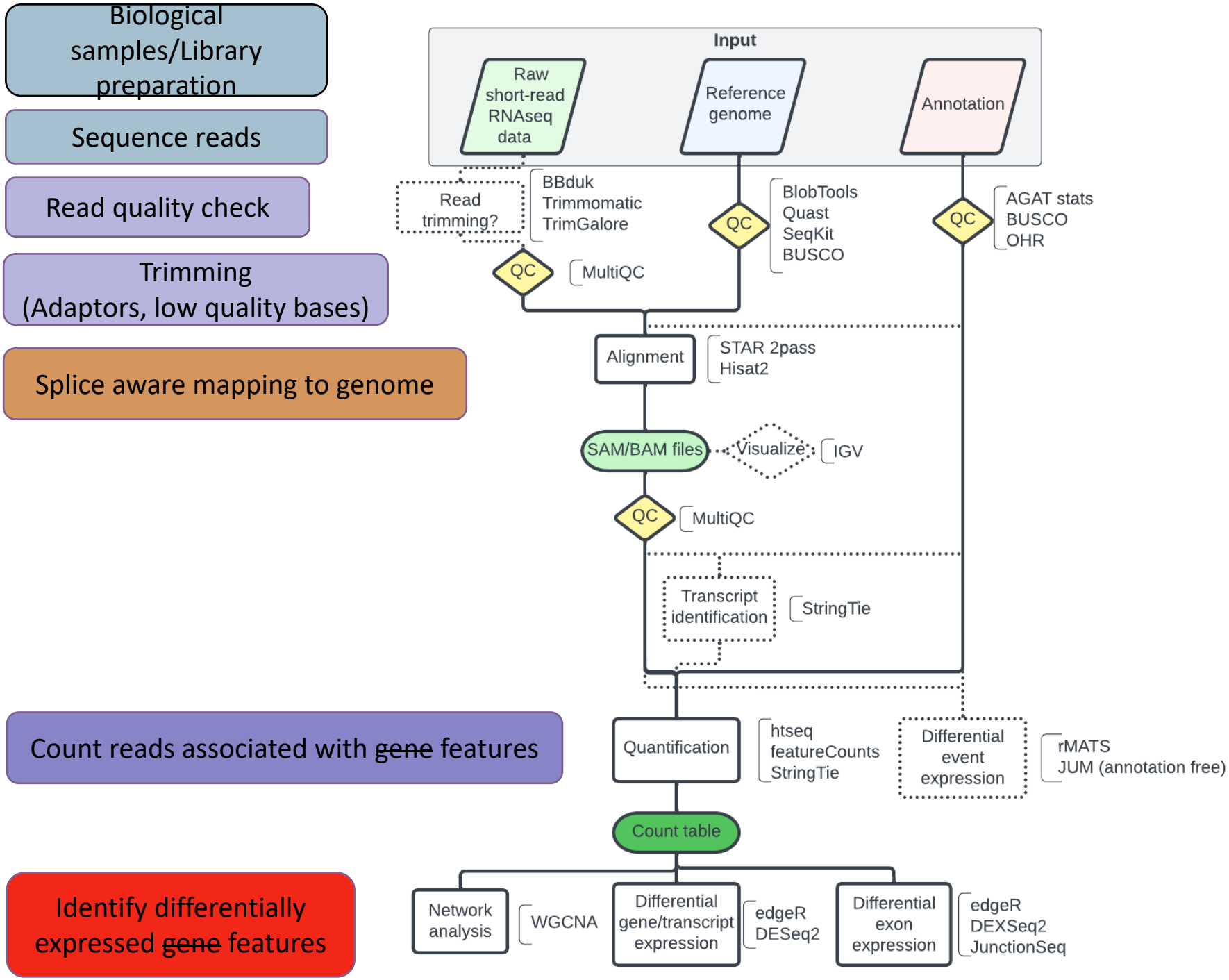


Price et al. 2022. Nature Ecology and Evolution

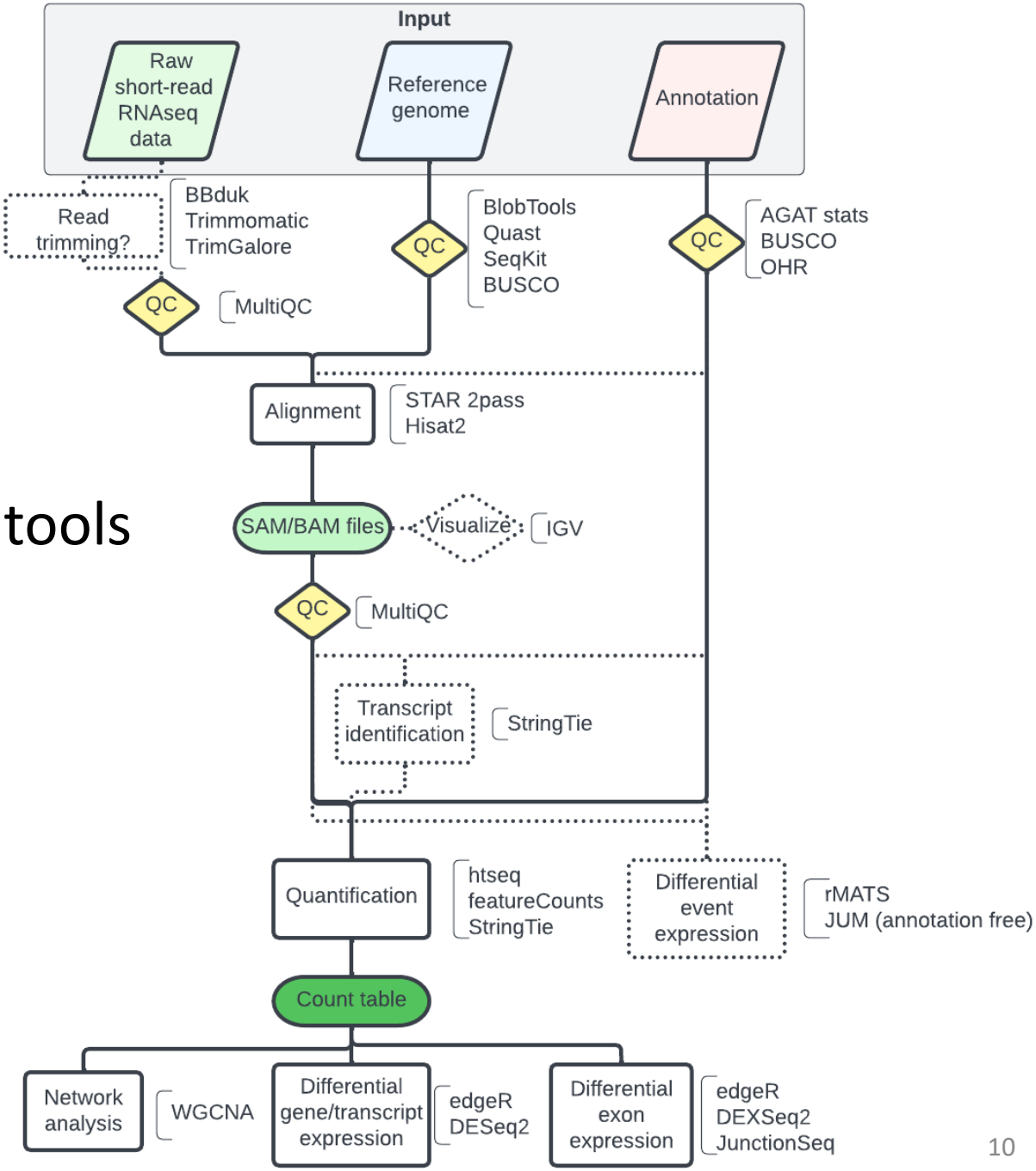
(Ref-based) DGE workflow




Input
Raw short-read RNAseq data
Reference genome
Annotation
Read trimming?
BBduk
Trimmomatic
TrimGalore
QC
MultiQC
QC
BlobTools
Quast
SeqKit
BUSCO
QC
AGAT stats
BUSCO
OHR
Alignment
STAR 2pass
Hisat2
SAM/BAM files
Visualize
IGV
QC
MultiQC
Transcript identification
StringTie
Quantification
htseq
featureCounts
StringTie
Differential event expression
rMATS
JUM (annotation free)
Count table
Network analysis
WGCNA
Differential gene/transcript expression
edgeR
DESeq2
Differential exon expression
edgeR
DEXSeq2
JunctionSeq

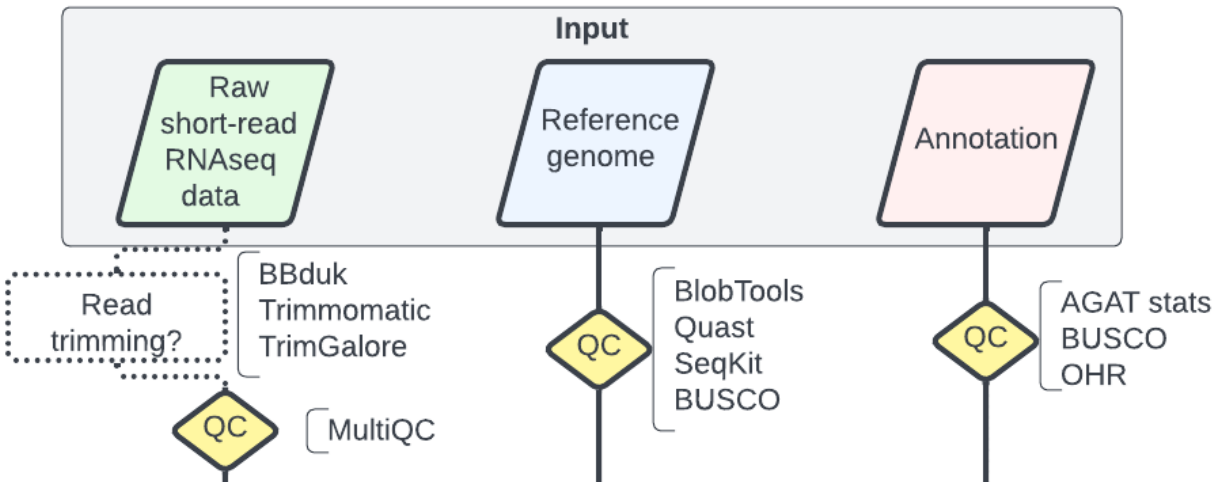

Biological samples/Library preparation
Sequence reads
Read quality check
Trimming (Adaptors, low quality bases)
Splice aware mapping to genome
Count reads associated with ~~gene~~ features
Identify differentially expressed ~~gene~~ features
Input
Raw short-read RNAseq data
Reference genome
Annotation
Read trimming?
BBduk
Trimmomatic
TrimGalore
QC
MultiQC
QC
BlobTools
Quast
SeqKit
BUSCO
QC
AGAT stats
BUSCO
OHR
Alignment
STAR 2pass
Hisat2
SAM/BAM files
Visualize
IGV
QC
MultiQC
Transcript identification
StringTie
Quantification
htseq
featureCounts
StringTie
Differential event expression
rMATS
JUM (annotation free)
Count table
Network analysis
WGCNA
Differential gene/transcript expression
edgeR
DESeq2
Differential exon expression
edgeR
DEXSeq2
JunctionSeq

Combinations of tools can matter.



(Baruzzo et al. 2015 Nature Methods)

Quality control

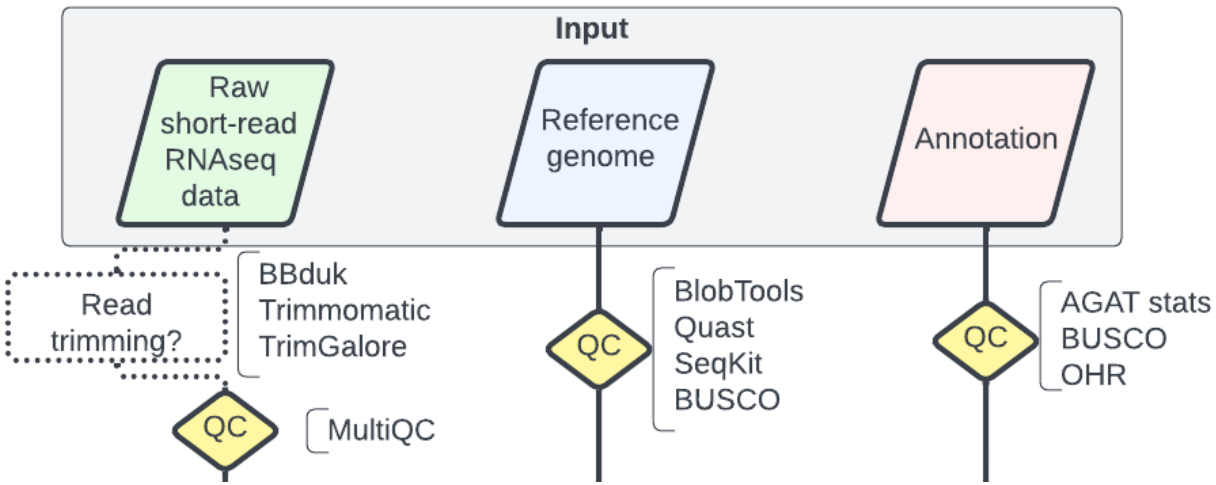


Reads: To trim or not to trim?

- genome annotation, variant calling, transcriptome assembly : Trim!
- Anything else, maybe trim lightly?
  - adapters + low quality score (Q10-15)

(Williams et al. 2016 BMC Bioinformatics,
Liao and Shi 2020 NAR Genomics and Bioinformatics)

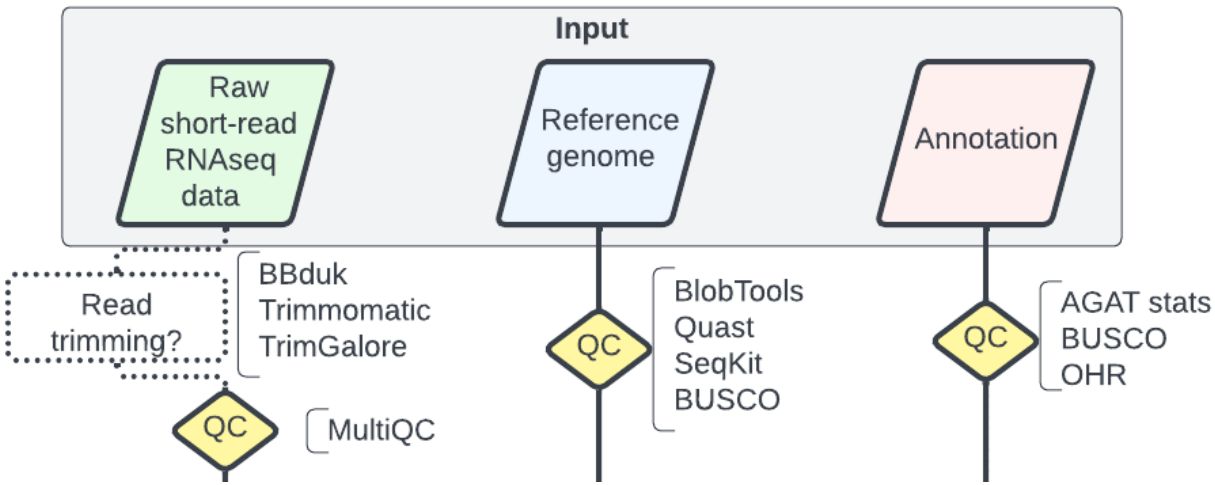
Quality control



Reference genome considerations:

- What maps where:
  - Recent duplications?
  - Highly repetetive content?
  - Missing content?

Quality control

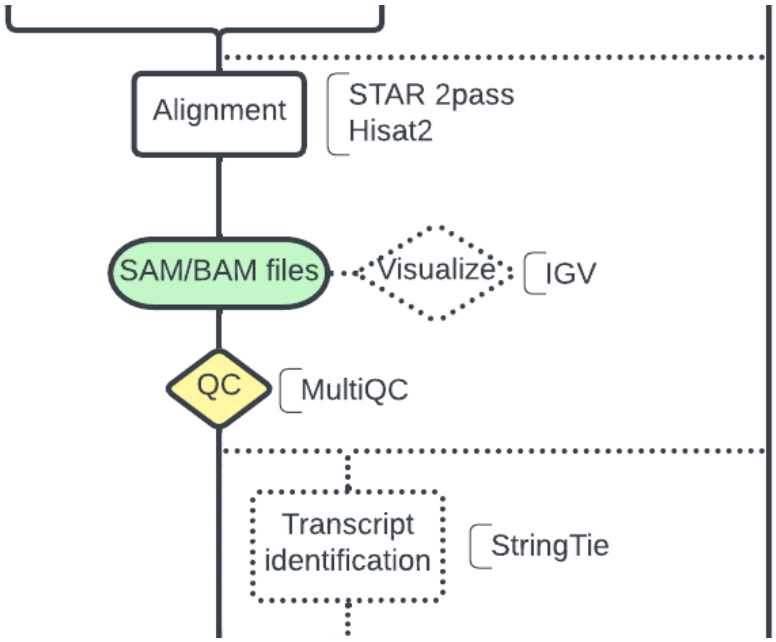


Annotation considerations:

- What features have been annotated?
- Was RNAseq data used in the annotation?
  - *What* RNA? Life stage? Sex?

- In the lab, we use a protein-based BRAKER2 annotation

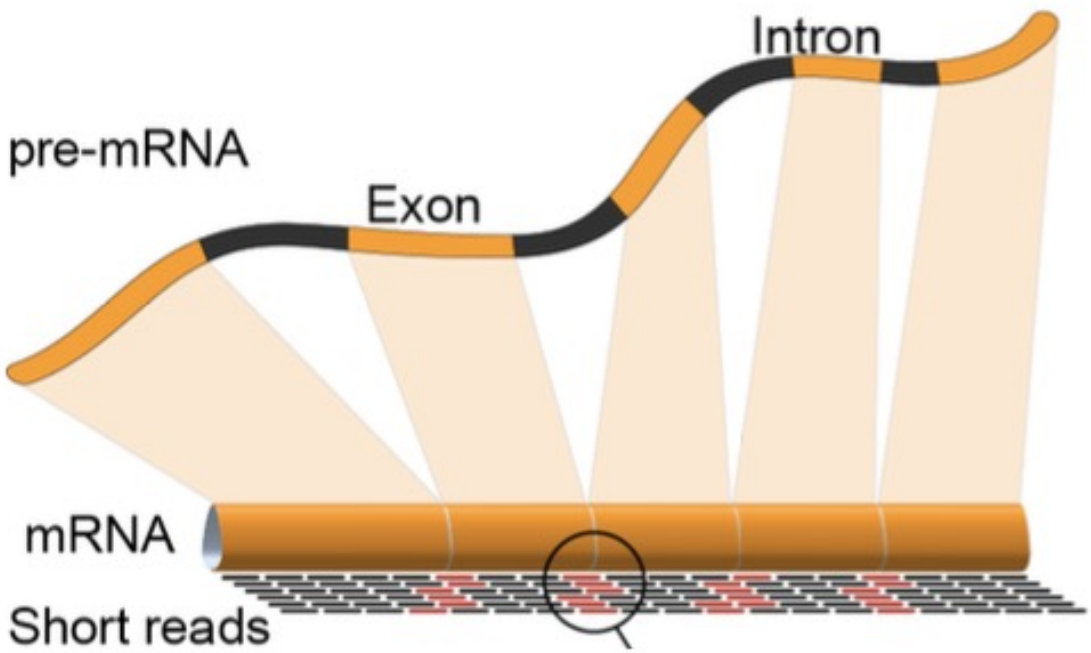
Sequence alignment



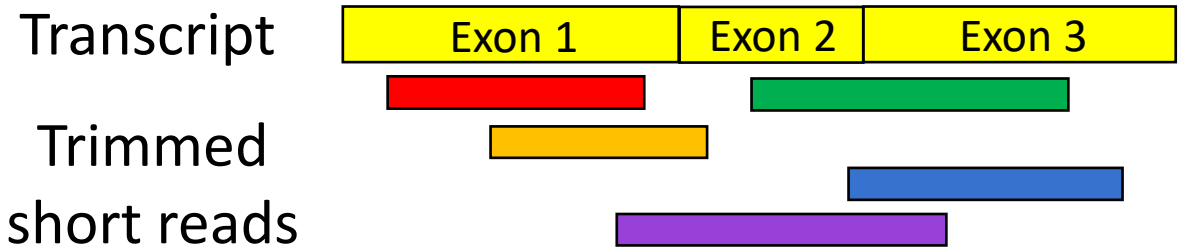
What are some challenges when aligning RNA-seq reads to the reference genome?

Sequence alignment

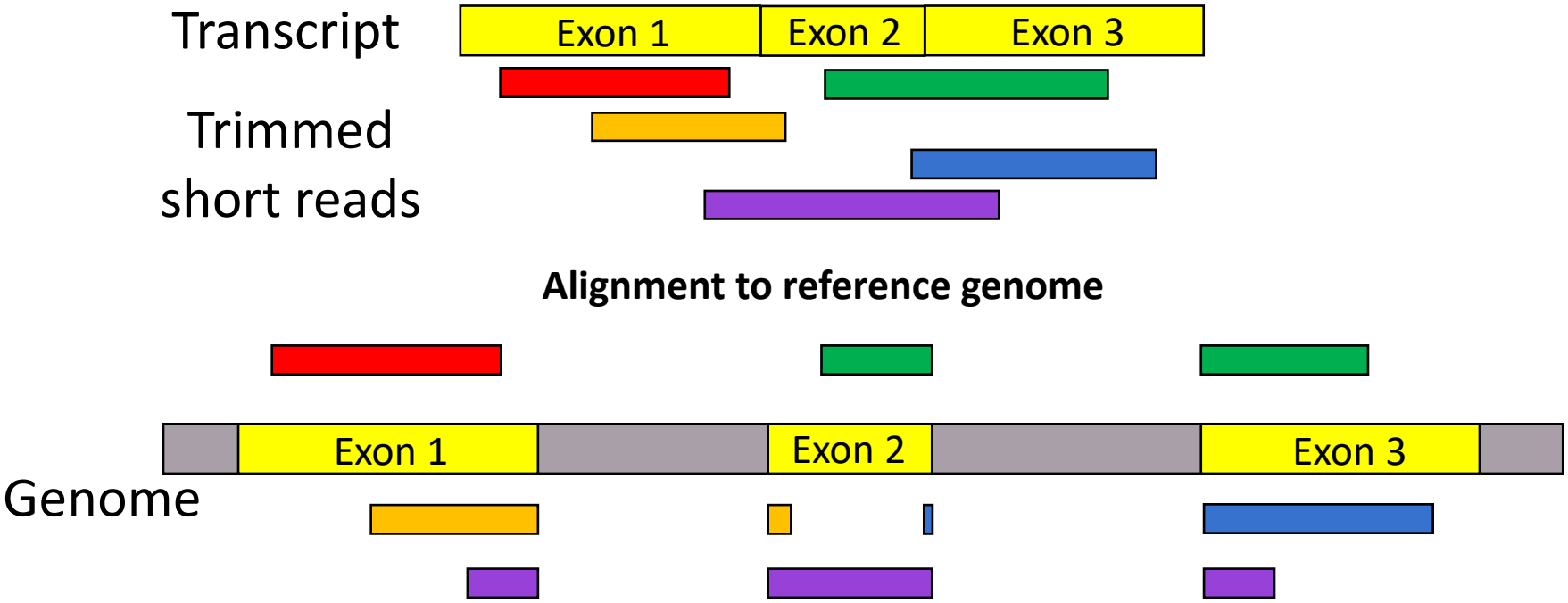
What are some challenges when aligning RNA-seq reads to the reference genome?



Splice-aware sequence alignment



Splice-aware sequence alignment



Counting reads as a measure of expression

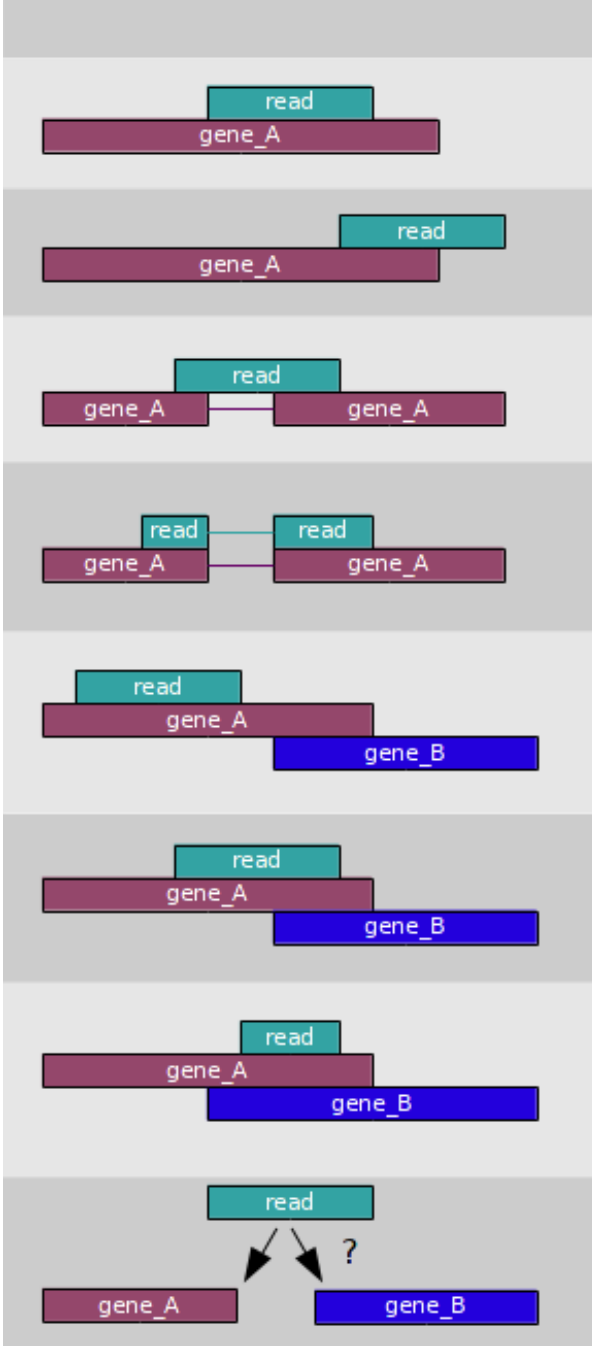
- Now we have our reads aligned to the genome, the next step is to count how many reads have been mapped to each features or metafeature.

- Two common counting tools are **featureCounts** and **htseq-count**.

- Total read count associated with a gene (*meta-feature*) = the sum of reads associated with each of the exons (*feature*) that "belong" to that gene.

```
genomics@ip-172-31-11-182:[~/workshop_materials/differential_expression/refs]$ head Pca_annotation.gtf
LG1     AUGUSTUS        transcript      22193   24413   .       -       .       transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1     AUGUSTUS        exon    22193   22320   .       -       .       transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1     AUGUSTUS        exon    23838   24048   .       -       .       transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1     AUGUSTUS        exon    24390   24413   .       -       .       transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1     AUGUSTUS        CDS     22193   22320   .       -       2       transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1     AUGUSTUS        CDS     23838   24048   .       -       0       transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1     AUGUSTUS        CDS     24390   24413   .       -       0       transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1     AUGUSTUS        transcript      79912   80136   .       -       .       transcript_id "Polcal_g2.t1"; gene_id "Polcal_g2";
LG1     AUGUSTUS        exon    79912   80136   .       -       .       transcript_id "Polcal_g2.t1"; gene_id "Polcal_g2";
LG1     AUGUSTUS        CDS     79912   80136   .       -       0       transcript_id "Polcal_g2.t1"; gene_id "Polcal_g2";
genomics@ip-172-31-11-182:[~/workshop_materials/differential_expression/refs]$
```

What should count??



HTSeq manual

Output of counting = A count matrix, with genes as rows and samples as columns

These are the "raw" counts and will be used in the downstream statistical program for differential gene expression.

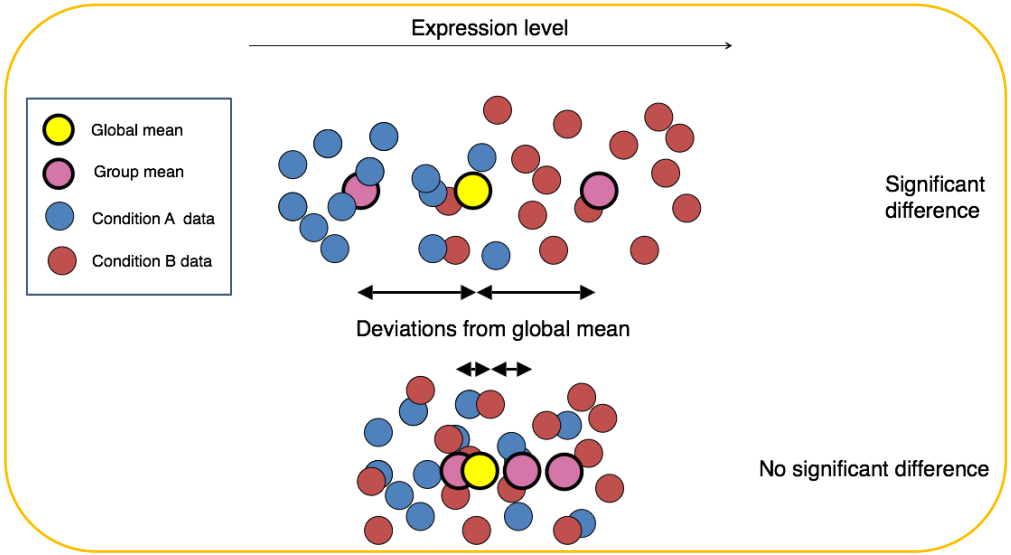
Each column is a sample

Each row is a gene

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

Differential expression analysis

Analysis



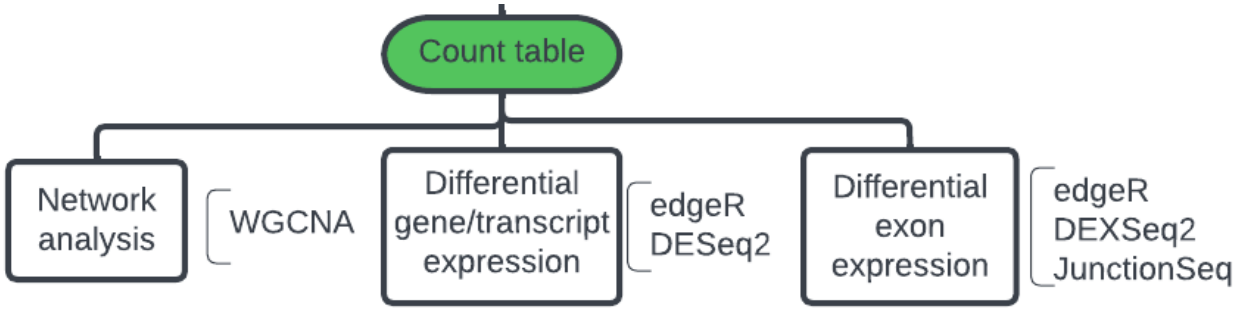
features (e.g. genes)

samples: want to see if differences across condition are significant

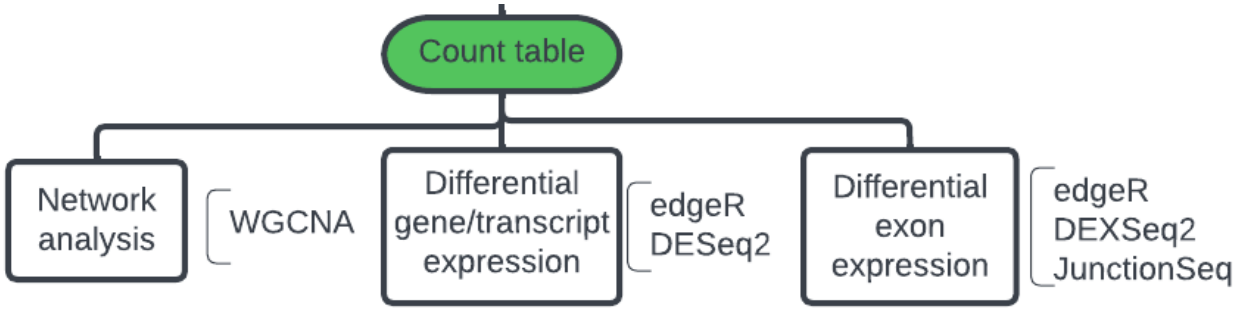
Input

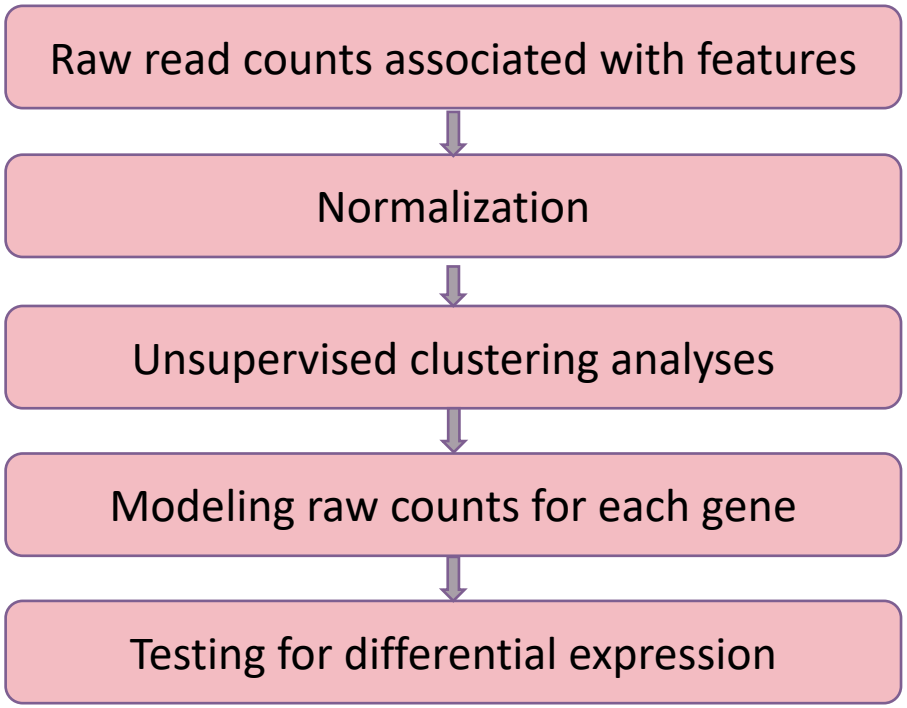
Gene_id	S1	S2	S3	S4	S5	S6
Polcal_g1	17	10	5	23	10	6
Polcal_g2	0	1	0	1	2	1
Polcal_g3	7	0	2	7	4	0
Polcal_g4	17	11	5	21	10	12

Differential expression analysisis



Differential expression analysisis





DESeq2 package

METHOD | Open Access | Published: 05 December 2014

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love, Wolfgang Huber & Simon Anders

Genome Biology **15**, Article number: 550 (2014) | Cite this article

450k Accesses | **34853** Citations | **131** Altmetric | Metrics

Normalization

- **Normalization is NOT** fitting a normal distribution or transforming data transformation.
- **Normalization aims to** identify the nature and magnitude of **systematic biases**, and take them into account in our model-based analysis of the data.

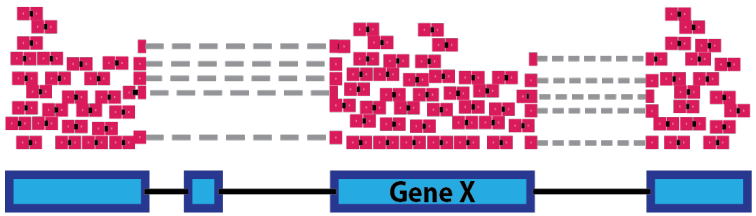
The main factors often considered during normalization are:

- Sequencing depth
- RNA composition
- Gene length (some methods)

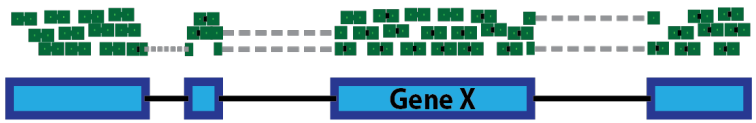
Normalization

Sequencing depth

Sample A Reads

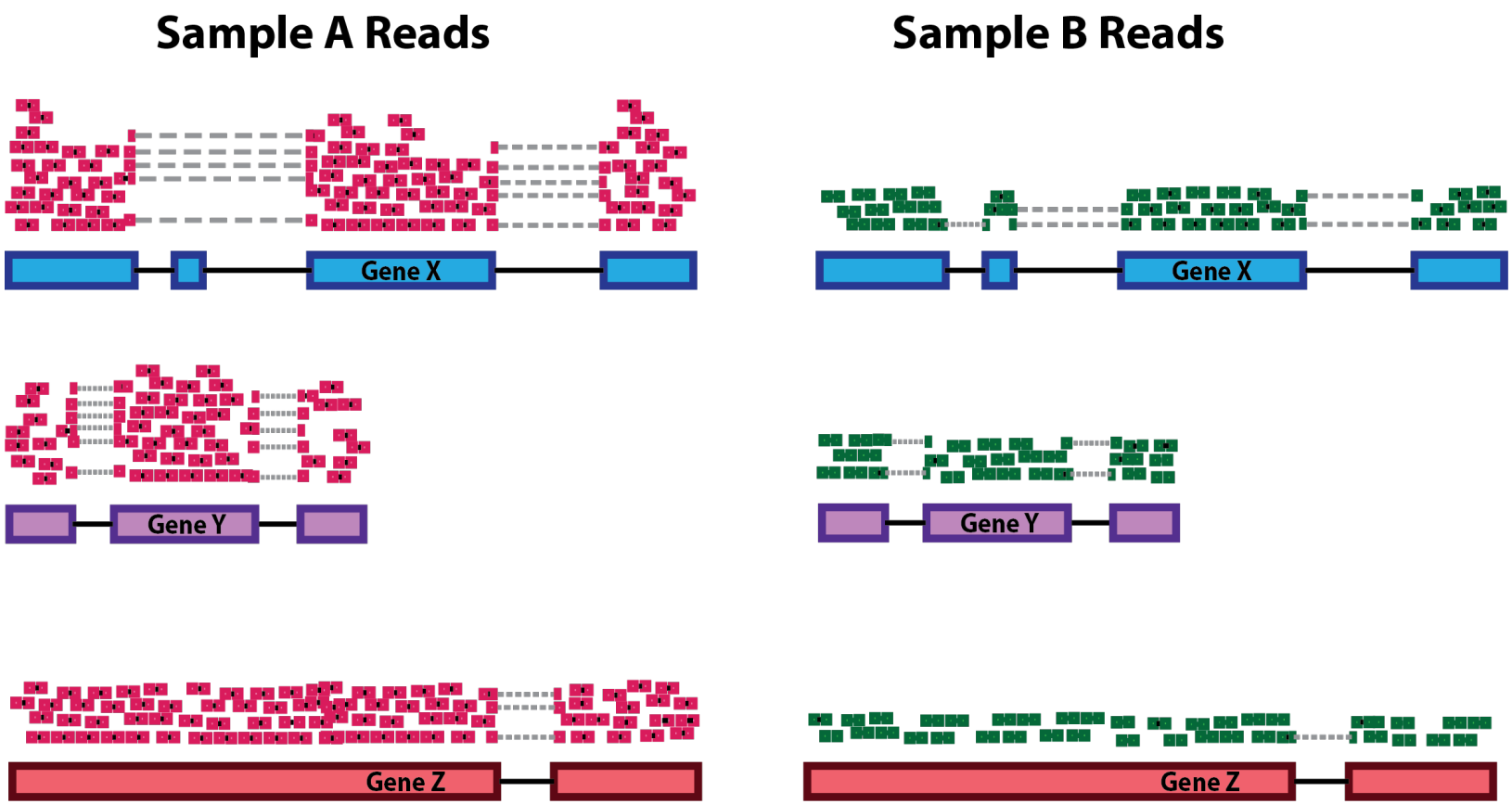


Sample B Reads



Normalization

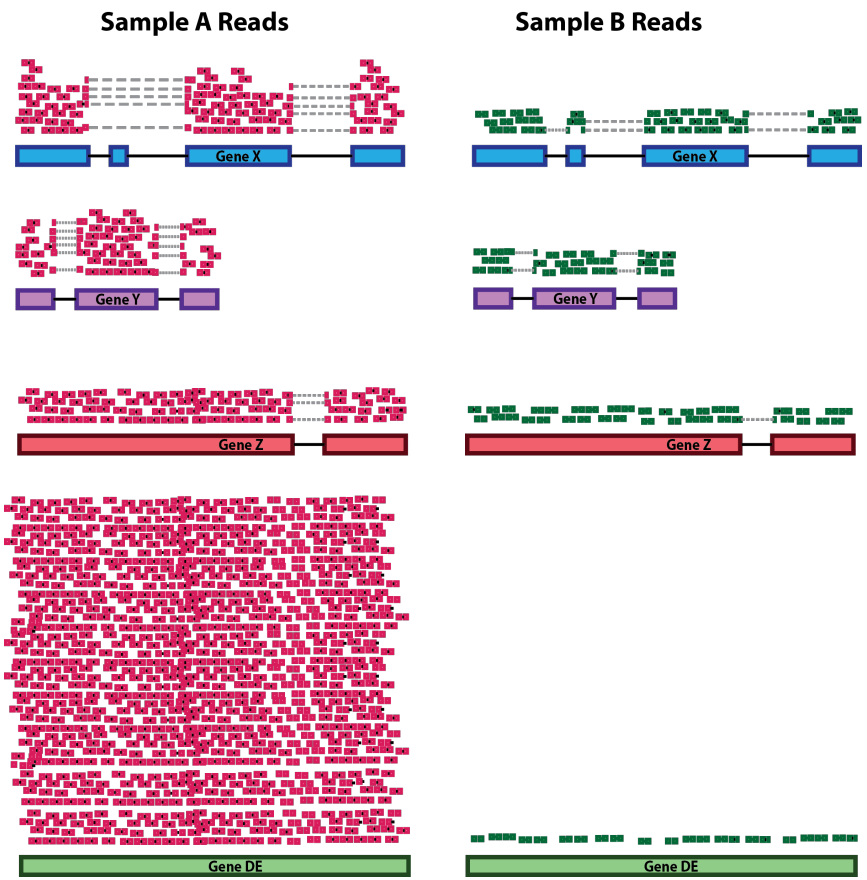
Sequencing depth



Normalization

RNA composition

- A few highly differentially expressed genes
- Can skew some normalization methods



Median of ratios (MRN) normalization

- Used by DESeq2 (DGE analysis tool we will use today)

Let’s see how the normalization works…

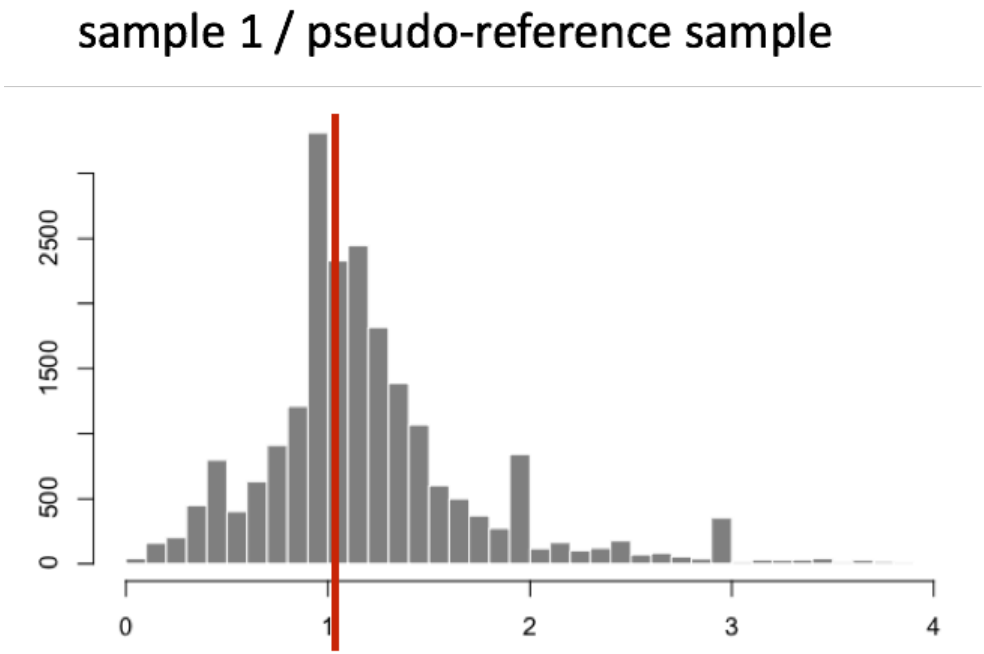
Step 1. Create a pseudo-reference sample for each gene (row-wise geometric mean)

Gene	sampleA	sampleB	Pseudo-reference sample
EF2A	1489	906	sqrt(1489*906) = 1161.5
ABCD1	22	13	sqrt(22*13) = 16.9
…	…	…	…

Step 2. Calculates ratio of each sample to the reference

Gene	sampleA	sampleB	Pseudo-reference sample	Ratio of sampleA/ref	Ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = 1.28	906/1161.5 = 0.78
ABCD1	22	13	16.9	22/16.9 = 1.30	13/16.9 = 0.77
MEFV	793	410	570.2	793/570.2 = 1.39	410/570.2 = 0.72
…	…	…	…	…	…

The figure below illustrates the median value for the distribution of all gene ratios for a single sample (frequency is on the y-axis).



The median of ratio methods makes the assumption that not ALL genes are differentially expressed; therefore, the normalization factors should account for sequencing depth and RNA composition of the sample (large outlier genes will not represent the median ratio values).

Step 3. Calculate the normalization factor for each sample (size factor)

Gene	sampleA	sampleB	Pseudo-reference sample	Ratio of sampleA/ref	Ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = 1.28	906/1161.5 = 0.78
ABCD1	22	13	16.9	22/16.9 = 1.30	13/16.9 = 0.77
MEFV	793	410	570.2	793/570.2 = 1.39	410/570.2 = 0.72
…	…	…	…	…	…

```
median(c(1.28, 1.3, 1.39, 1.35, 0.59))
=1.3
```

```
median(c(1.28, 1.3, 1.39, 1.35, 0.59))
=1.3
```

Step 4: calculate the normalized count values using the normalization factor

Raw counts:

Gene	sampleA	sampleB
EF2A	1489	906
ABCD1	22	13
…	…	…

Normalized counts

Gene	sampleA	sampleB
EF2A	1489/1.3 = 1145.39	906/0.77 = 1176.62
ABCD1	22/1.3 = 16.92	13/0.77 = 16.88
…	…	…

Normalized counts are not whole numbers!

Modeling raw counts for each gene

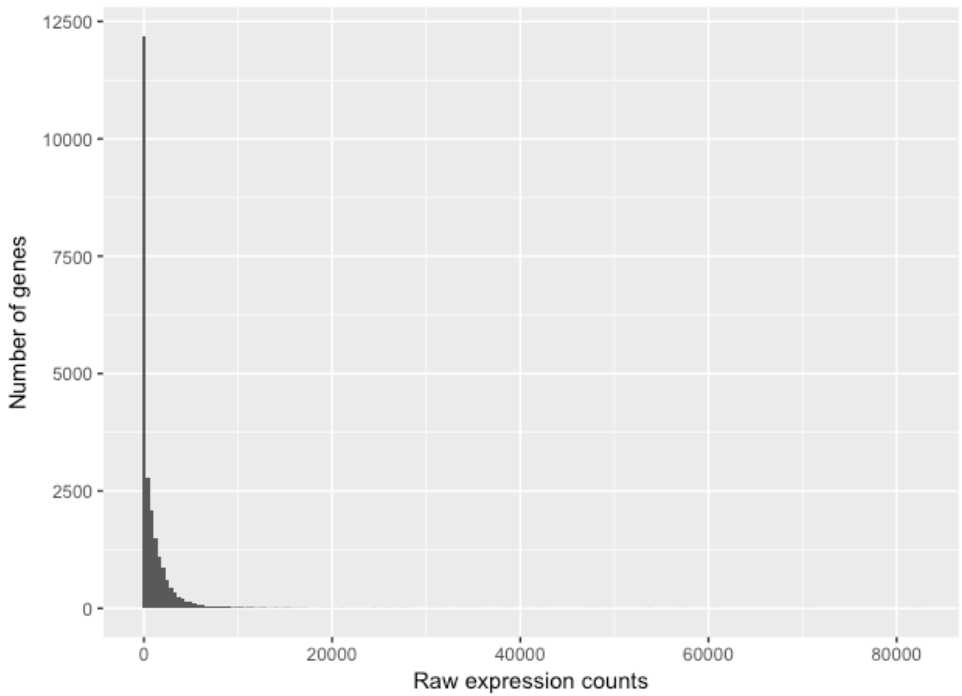
Step 1. Normalization (aka estimation of size factors) → done!

Step 2. Estimate gene-wise dispersion

- To accurately model sequencing counts, we need to generate accurate estimates of **within-group variation** for each gene (aka dispersion)
  - need to choose the right distribution

Properties of RNA-seq count data

The distribution of RNA-seq counts for a single sample looks as below:



Low number of counts associated with a large proportion of genes and a long right tail due to the lack of any upper limit for expression.

Statistical modeling of count data

Statistical modeling of count data

Which probability distributions are suitable for modeling count data?

Poisson distribution?

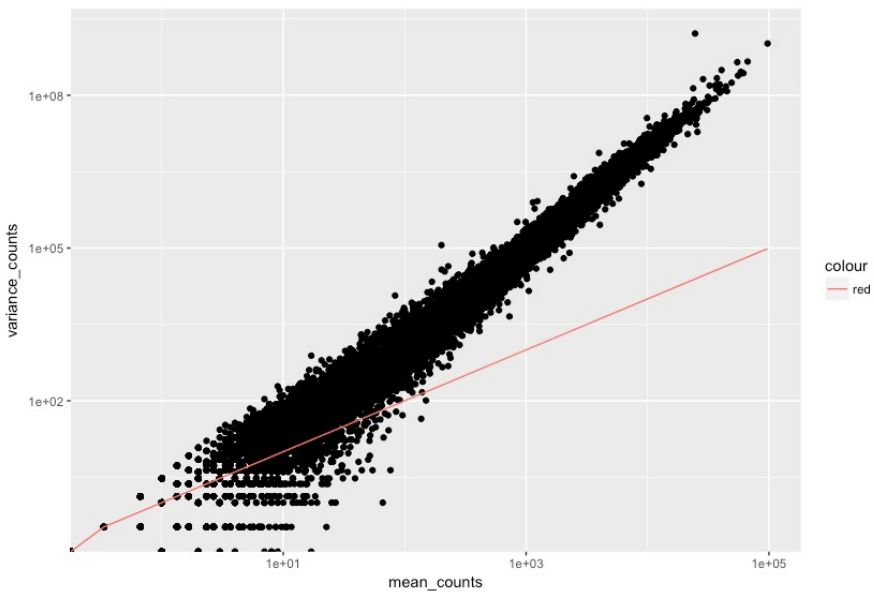
- Used when the number of cases is very large but the chance of a particular event is very low.
- **A property of Poission distribution is that the mean = variance.**

Statistical modeling of count data

Which probability distributions are suitable for modeling count data?

Poisson distribution?

- Used when the number of cases is very large but the chance of a particular event is very low.
- **A property of Poission distribution is that the mean = variance.**



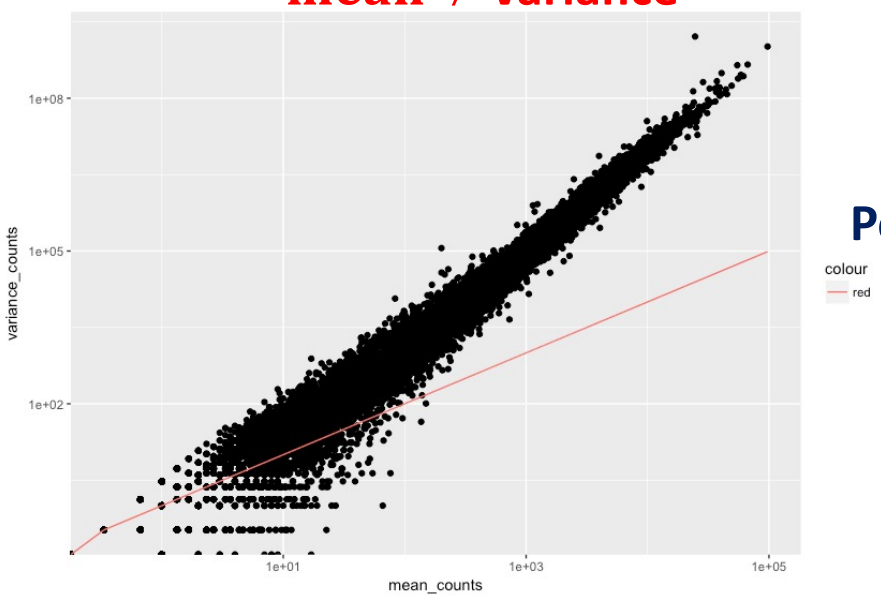
Statistical modeling of count data

Which probability distributions are suitable for modeling count data?

Poisson distribution?

- Used when the number of cases is very large but the chance of a particular event is very low.
- **A property of Poission distribution is that the mean = variance.**

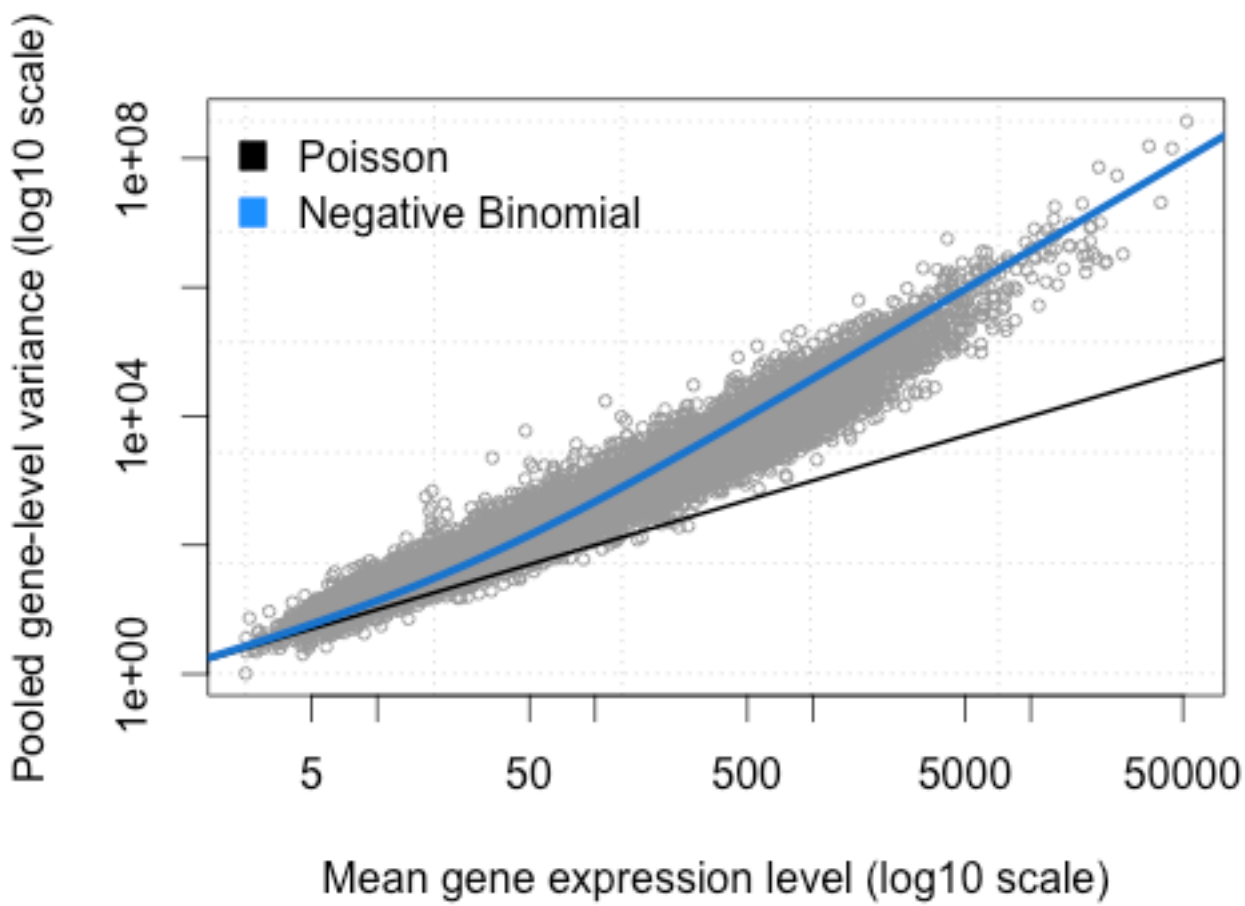
mean ≠ variance



Poisson distribution is not suitable to model count data across the biological samples.

The distribution that fits best is the Negative Binomial (NB) distribution.

- two parameters, one for the mean and one for the variance
- fexlibility to estimate the amount of dispersion for each gene across samples.



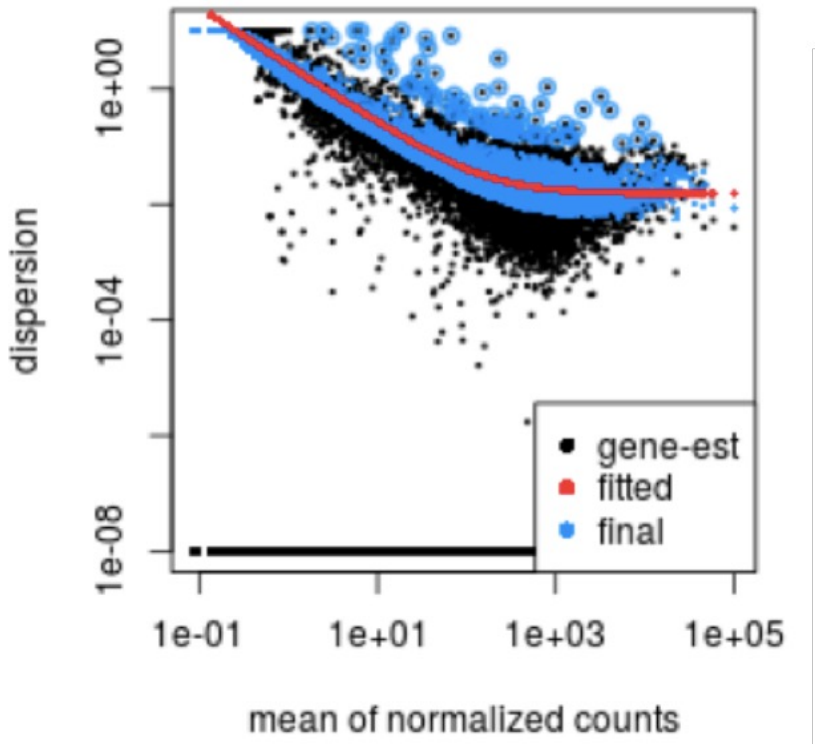
How does the dispersion relate to our model?

- the **estimates of variation for each gene are often unreliable**.

- DESeq2 **shares information across genes** to generate more accurate estimates of variation : **'shrinkage**'.
  - assumes that genes with similar expression levels have similar dispersion.

Step 3: Fit curve to gene-wise dispersion estimates

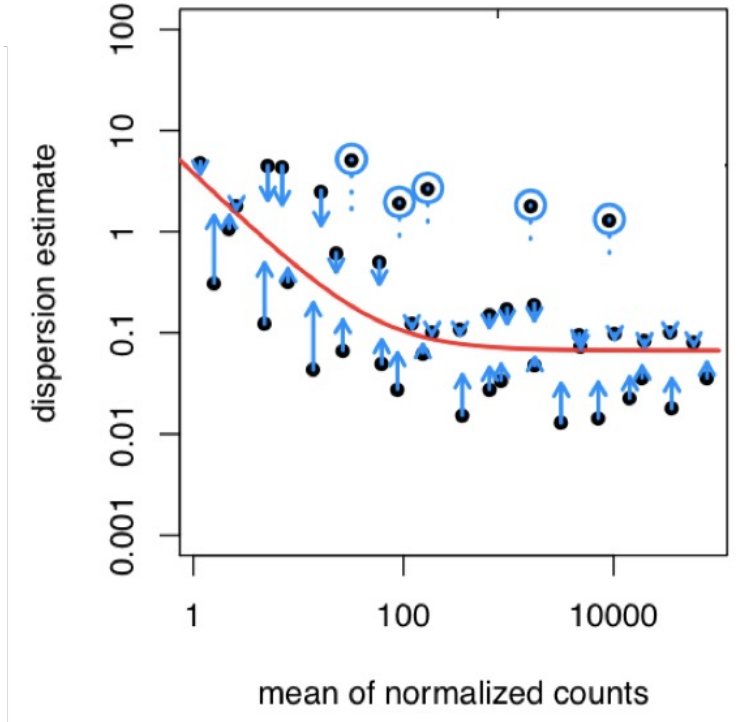
- Different genes will have different scales of biological variability
- However, we make the assumption that **DESeq2 assumes that genes with similar expression levels have similar dispersion.**
- Fitted dispersion curve = **expected dispersion for genes of a given level of expression (e.g., mean normalized count)**



Step 4: Shrink dispersion estimates for each gene toward the values predicted by the curve

- Genes with low dispersion estimates are shrunken towards the curve
- Genes with high dispersion estimates do not follow model assumptions, and are their dispersion is not shrunken

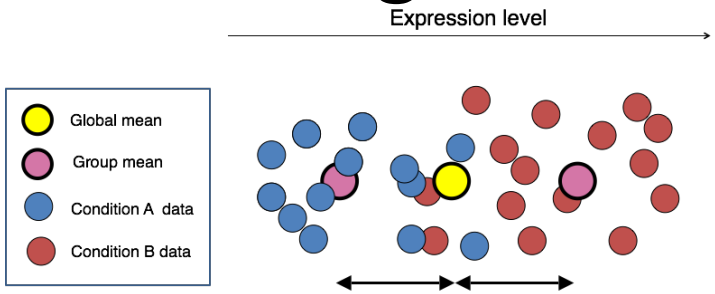
This shrinkage method is particularly important to reduce false positives in the differential expression analysis.



Model fitting and hypothesis testing

Blue: base level group, control group

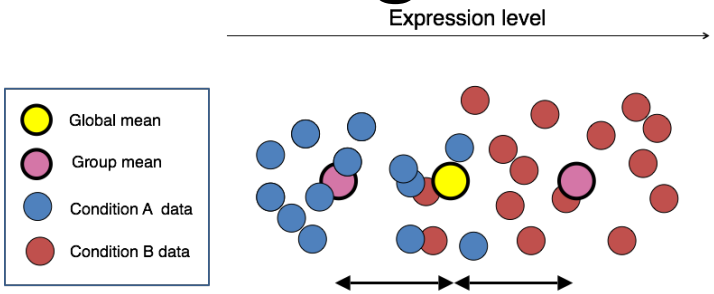
Red: treatment group



Model fitting and hypothesis testing

Blue: base level group, control group

Red: treatment group



Step 5. Generalized Linear Model fit for each gene

$$y = \beta_0 + x_1\beta_1$$

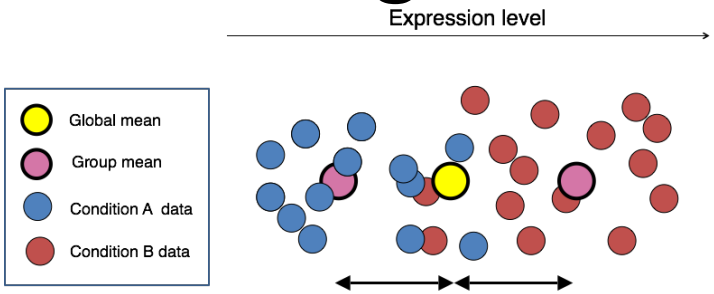
- y = transformed **expression level**
- β_0 = **intercept** (the estimated expression for the base level group, expression in the **blue** group)
- x_1 = a binary indicator variable for (0 if part of the blue group, 1 if part of the **red** group)
- β_1 = coefficient for the treatment group (**red**)
  - represents the **difference** between **red** and **blue**

$$y = \beta_0 + \beta_1$$

Model fitting and hypothesis testing

Blue: base level group, control group

Red: treatment group



Step 5. Generalized Linear Model fit for each gene

$$y = \beta_0 + x_1\beta_1$$

- y = transformed **expression level**
- β_0 = **intercept** (the estimated expression for the base level group, expression in the **blue** group)
- x_1 = a binary indicator variable for (0 if part of the blue group, 1 if part of the **red** group)
- β_1 = coefficient for the treatment group (**red**)
  - represents the **difference** between **red** and **blue**

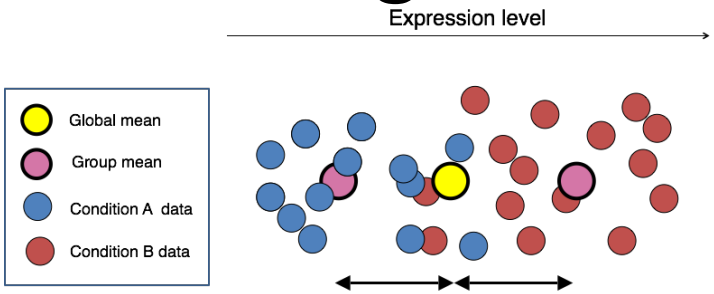
$$y = \beta_0 + \beta_1$$

$$\beta_1 = y - \beta_0 = \log_2(expression_{\boldsymbol{red}}) - \log_2(expression_{\boldsymbol{blue}})$$

Model fitting and hypothesis testing

Blue: base level group, control group

Red: treatment group



Step 5. Generalized Linear Model fit for each gene

$$y = \beta_0 + x_1\beta_1$$

- y = transformed **expression level**
- β_0 = **intercept** (the estimated expression for the base level group, expression in the **blue** group)
- x_1 = a binary indicator variable for (0 if part of the blue group, 1 if part of the **red** group)
- β_1 = coefficient for the treatment group (**red**)
  - represents the **difference** between **red** and **blue**

$$y = \beta_0 + \beta_1$$

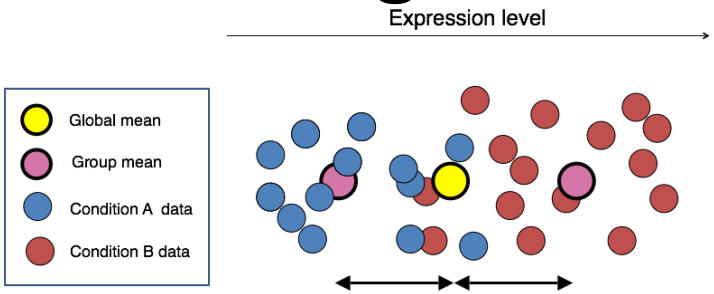
$$\beta_1 = y - \beta_0 = \log_2(expression_{red}) - \log_2(expression_{blue})$$

$$= \frac{\log_2(expression_{red})}{\log_2(expression_{blue})}$$

Model fitting and hypothesis testing

Blue: base level group, control group

Red: treatment group



Step 5. Generalized Linear Model fit for each gene

$$y = \beta_0 + x_1\beta_1$$

- y = transformed **expression level**
- β_0 = **intercept** (the estimated expression for the base level group, expression in the **blue** group)
- x_1 = a binary indicator variable for (0 if part of the blue group, 1 if part of the **red** group)
- β_1 = coefficient for the treatment group (**red**)
  - represents the **difference** between **red** and **blue**

$$y = \beta_0 + \beta_1$$

$$\beta_1 = y - \beta_0 = \log_2(expression_{red}) - \log_2(expression_{blue})$$

$$\log_2\left(\frac{expression_{red}}{expression_{blue}}\right)$$

$$= \boldsymbol{log_2}\ \textbf{Fold Change}$$

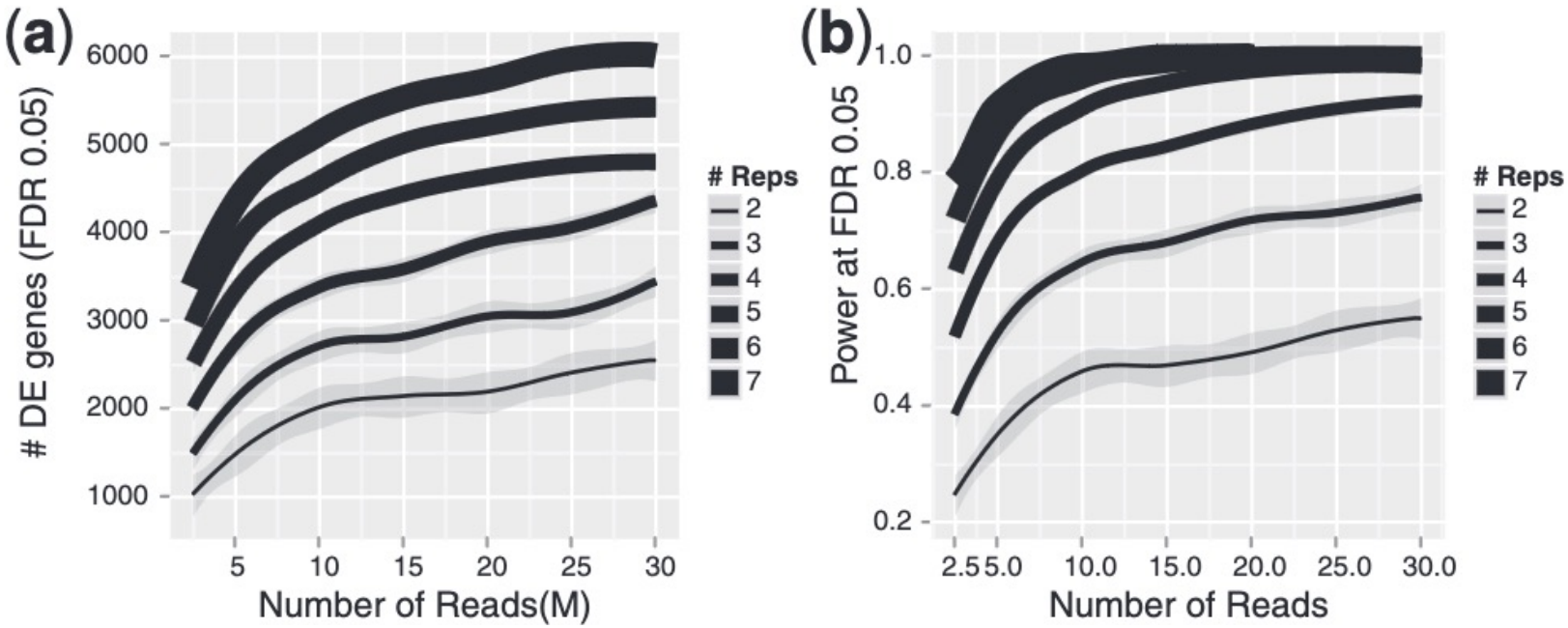
$$\boxed{\begin{aligned} log_2 1 &= 0 \\ log_2 2 &= 1 \\ log_2 4 &= 2 \end{aligned}}$$

Output of DESeq2

```
log2 fold change (MAP): sampletype MOV10_overexpression vs control
Wald test p-value: sampletype MOV10_overexpression vs control
DataFrame with 6 rows and 6 columns
               baseMean log2FoldChange      lfcSE       stat    pvalue       padj
              <numeric>      <numeric>  <numeric>  <numeric> <numeric>  <numeric>
1/2-SBSRNA4  45.6520399     0.26976764 0.18775752  1.4367874 0.1507784 0.25242910
A1BG         61.0931017     0.20999700 0.17315013  1.2128030 0.2252051 0.34444163
A1BG-AS1    175.6658069    -0.05197768 0.12366259 -0.4203185 0.6742528 0.77216278
A1CF          0.2376919     0.02237286 0.04577046  0.4888056 0.6249793         NA
A2LD1        89.6179845     0.34598540 0.15901426  2.1758136 0.0295692 0.06725157
A2M           5.8600841    -0.27850841 0.18051805 -1.5428286 0.1228724 0.21489067
```

1. baseMean: mean of normalized counts for all samples
2. log2FoldChange: log2 fold change
3. lfcSE: standard error
4. stat: Wald statistic
5. pvalue: Wald test p-value
6. padj: BH adjusted p-values

When can we detect differential expression?



Liu et al. 2014. Bioinformatics

What do we do with DE genes?

- Visualize expression levels, log fold changes, and significance
- Identify up- and down-regulated genes
- Compare sets of DE genes
- Test for functional enrichment of DE gene sets

Today's activity

Focus on Differential Gene Expression Analysis:

- Evaluating our genomic resources
- Unsupervised clustering of samples based on expression
- Identifying differentially expressed (DE) genes
- Evaluating functional enrichment of DE gene sets

When you finish, you can run the steps for trimming, mapping and counting.

Today's activity

NOTES:

1. **Skip counting the genes in the annotation (Rachel's mistake)**
2. **Using ggsave on guacamole: not compatible with .png, use .pdf.**

7:00 - 7:30 : Differential expression background

7:30 – 8:30 : Free work time
(take a break when you need/want it)

8:45: Check-in

9: – 9:40 : Free work time
(take a break when you need/want it)

9:40 – 10 : Wrap-up and discussion

Links to other DE/DS tools

Tool	Use	Link to best resource
WGCNA (R package)	Weighted gene coexpression analysis groups genes into modules/clusters by expression patterns across samples	Horvath lab website: https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/
DEXSeq (R package)	Differential exon expression within the DESeq2 framework from exon count data	Vignette: https://bioconductor.org/packages/release/bioc/vignettes/DEXSeq/inst/doc/DEXSeq.html
EdgeR (R package)	Differential expression analysis with differential exon expression functions from exon count data	User guide: https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf
LeafCutter (python & R scripts)	Differential splicing analysis specifically focused on differential intron retention from junction count data	Github page: https://davidaknowles.github.io/leafcutter/
IsoformSwitchAnalyzer (R package)	Differential isoform usage from transcript count data	Vignette: https://bioconductor.org/packages/release/bioc/vignettes/IsoformSwitchAnalyzeR/inst/doc/IsoformSwitchAnalyzeR.html
EBSeq	Bayesian differenital expression framework	Vignette: https://bioconductor.org/packages/release/bioc/vignettes/EBSeq/inst/doc/EBSeq_Vignette.pdf Github page: https://github.com/lengning/EBSeq