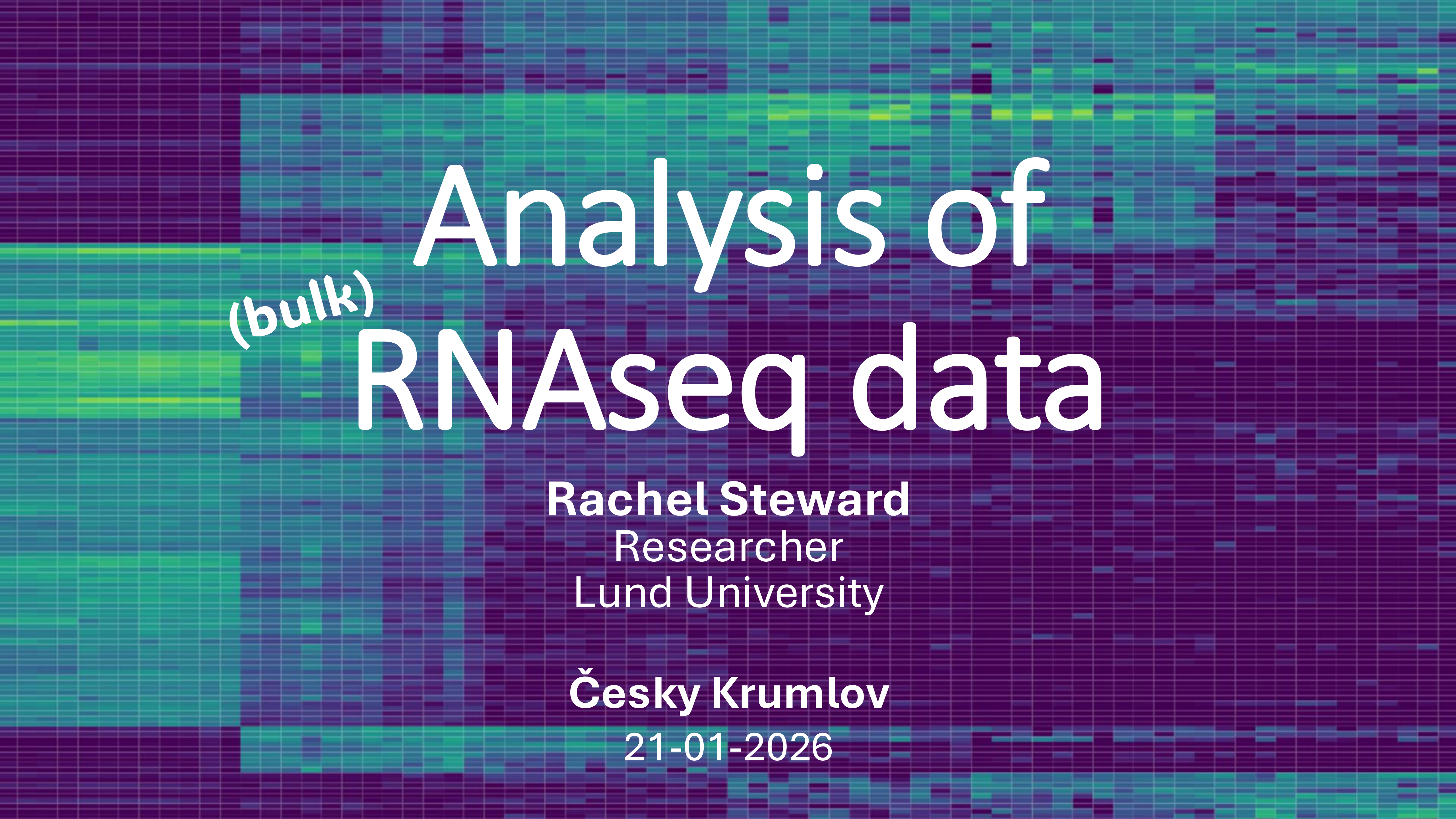


Analysis of (bulk) RNAseq data

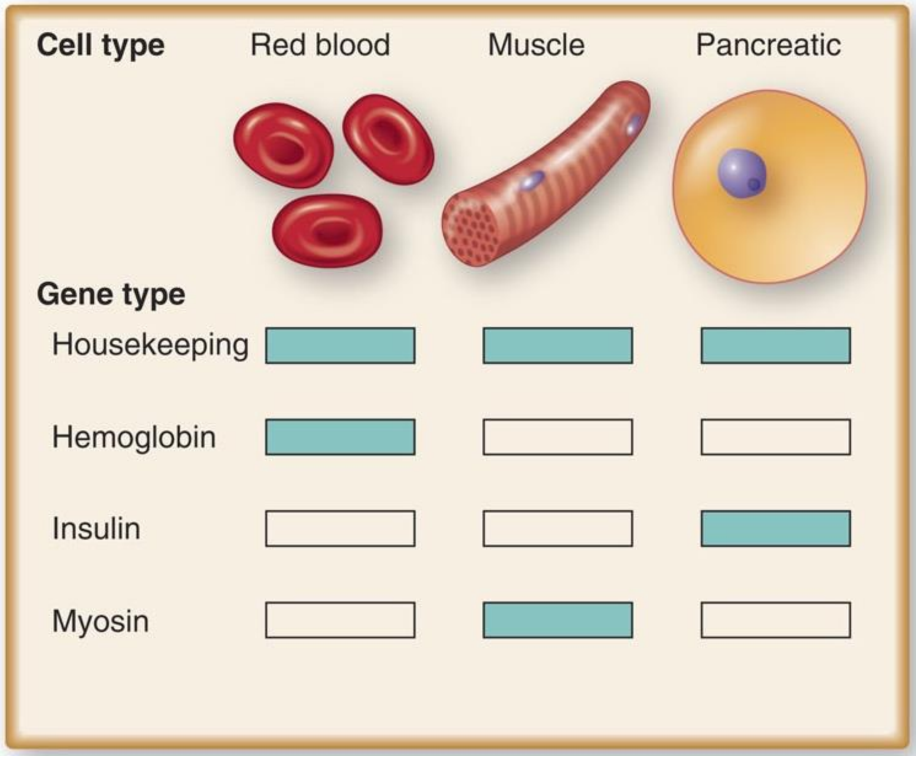
Rachel Steward
Researcher
Lund University

Česky Krumlov
21-01-2026

Gene expression

The selective activity of certain genes is a highly regulated process

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

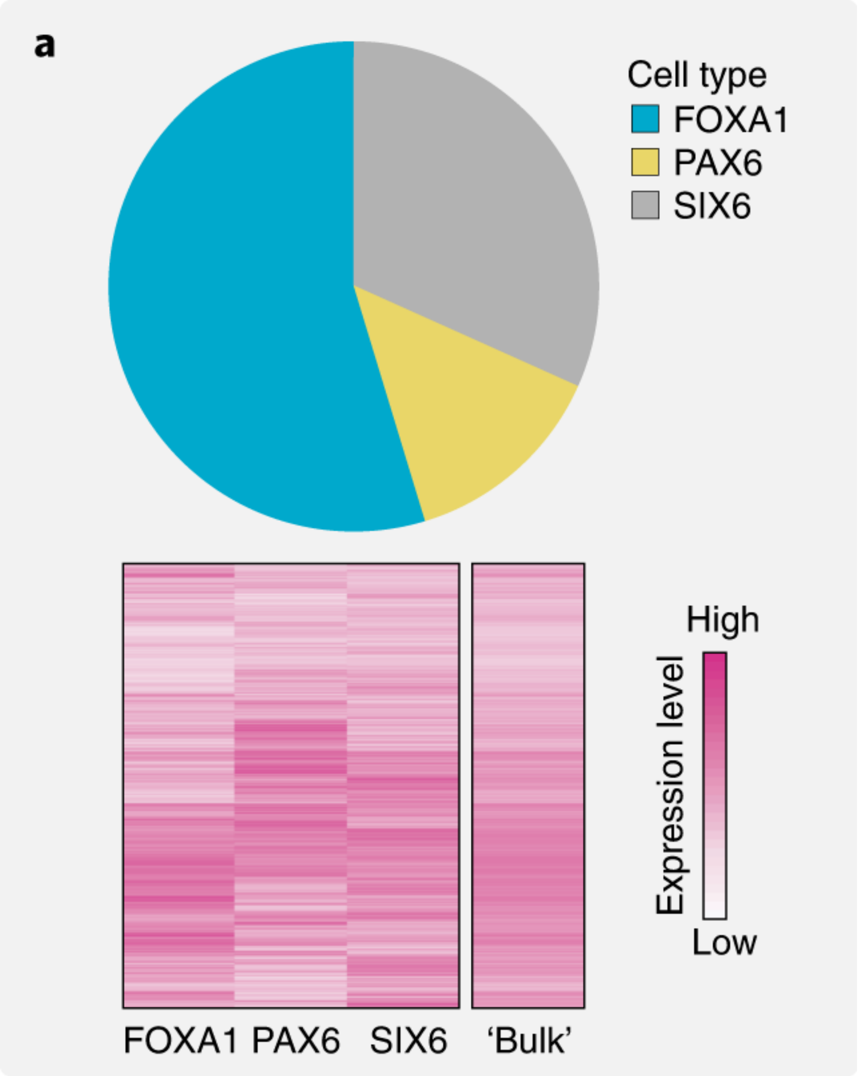


Cell type	Red blood	Muscle	Pancreatic
Gene type			
Housekeeping	🟩	🟩	🟩
Hemoglobin	🟩	□	□
Insulin	□	□	🟩
Myosin	□	🟩	□

Gene expression

The selective activity of certain genes is a highly regulated process

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

What are some questions we can answer with bulk RNAseq data?

How many genes are being expressed?

Which genes are uniquely expressed?

Does gene expression differ? between groups? According to a certain variable?

Are patterns of gene expression different among samples?

Are patterns of expression different among genes?

Which genes are co-expressed?

What are the functional roles of groups of differently expressed genes?

Lab activities

Part 1

- Exploring patterns in RNAseq data

Part 2

- Differential gene expression analysis

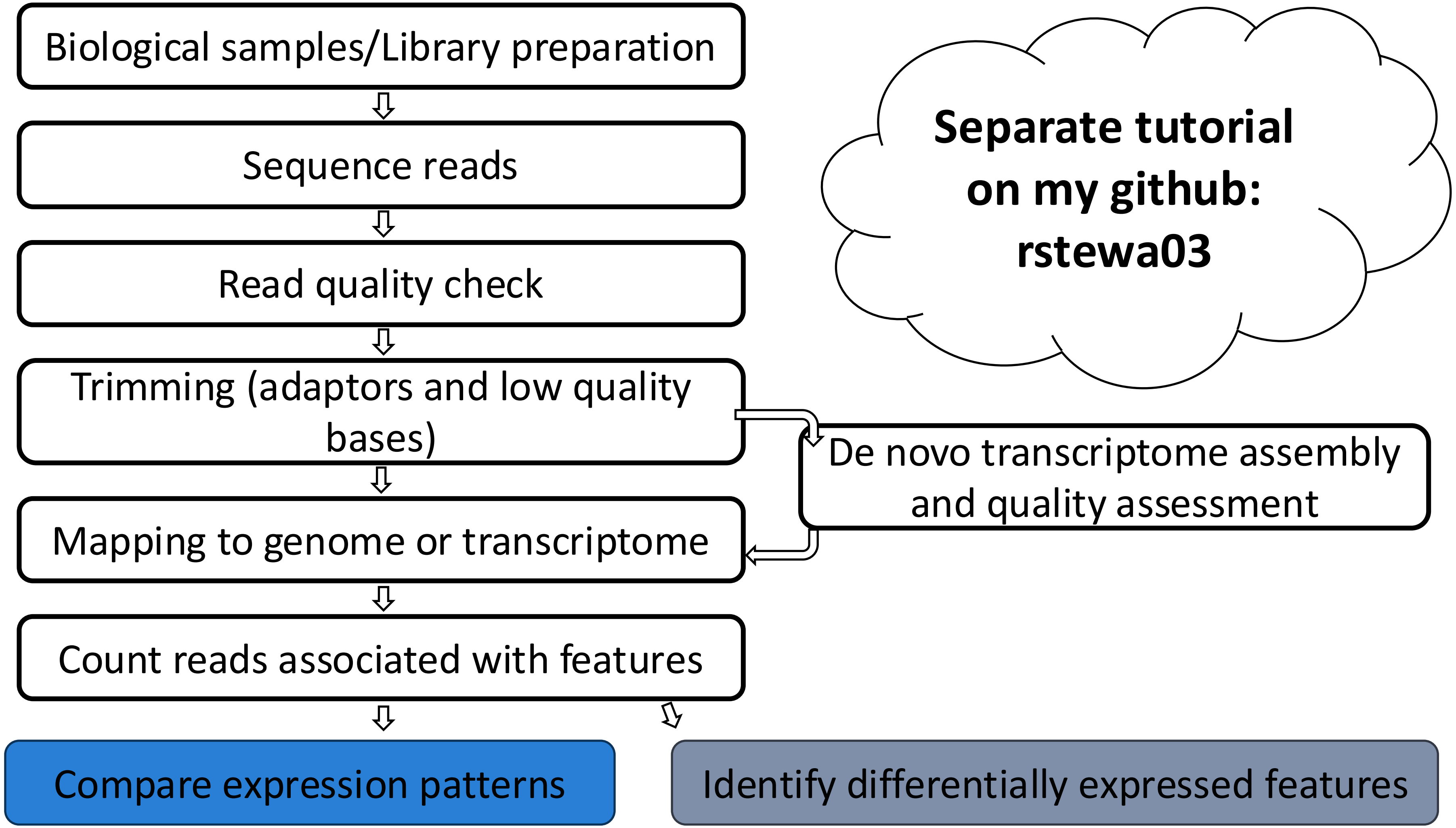
Part 3

- Functional enrichment of gene sets

Structure:

Short background
Open work time
Review

Gene expression analysis



Biological samples/Library preparation

Sequence reads

Read quality check

Trimming (adaptors and low quality bases)

De novo transcriptome assembly and quality assessment

Mapping to genome or transcriptome

Count reads associated with features

Compare expression patterns

Identify differentially expressed features

Separate tutorial on my github: rstewa03

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)

Reference genome considerations:

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

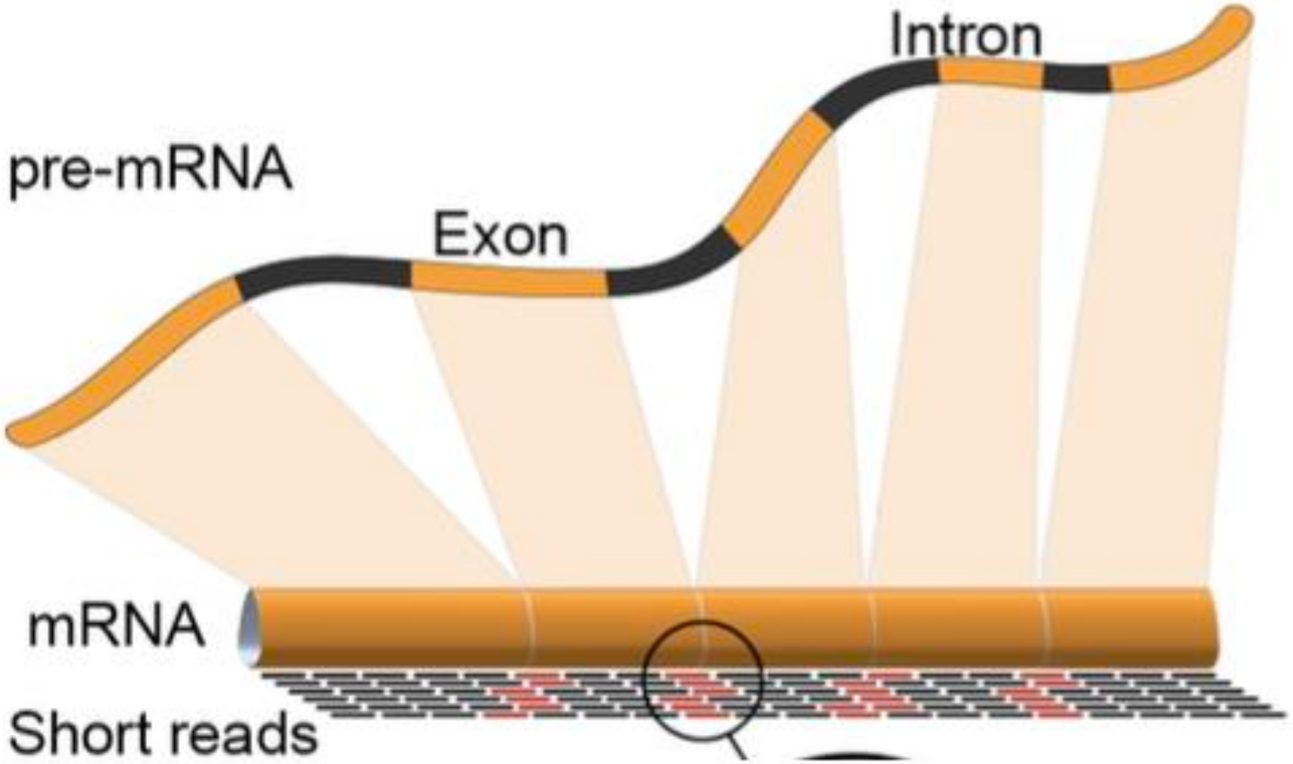
Annotation considerations:

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

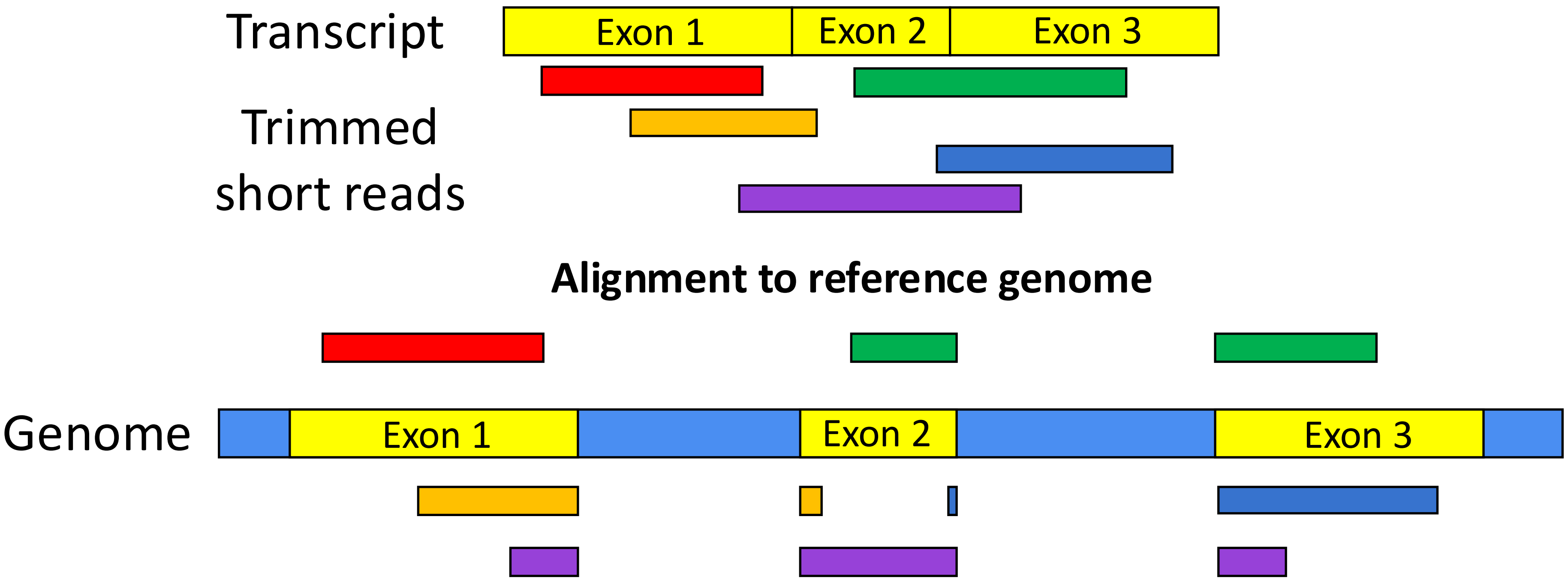
Williams et al. 2016 BMC Bioinformatics,
Liao and Shi 2020 NAR Genomics and Bioinformatics

RNA sequence alignment to a reference

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

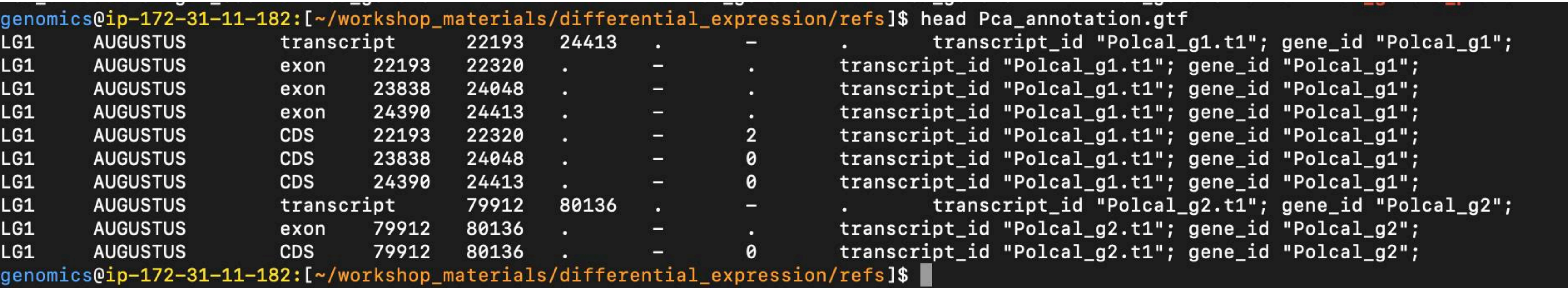


Splice-aware sequence alignment



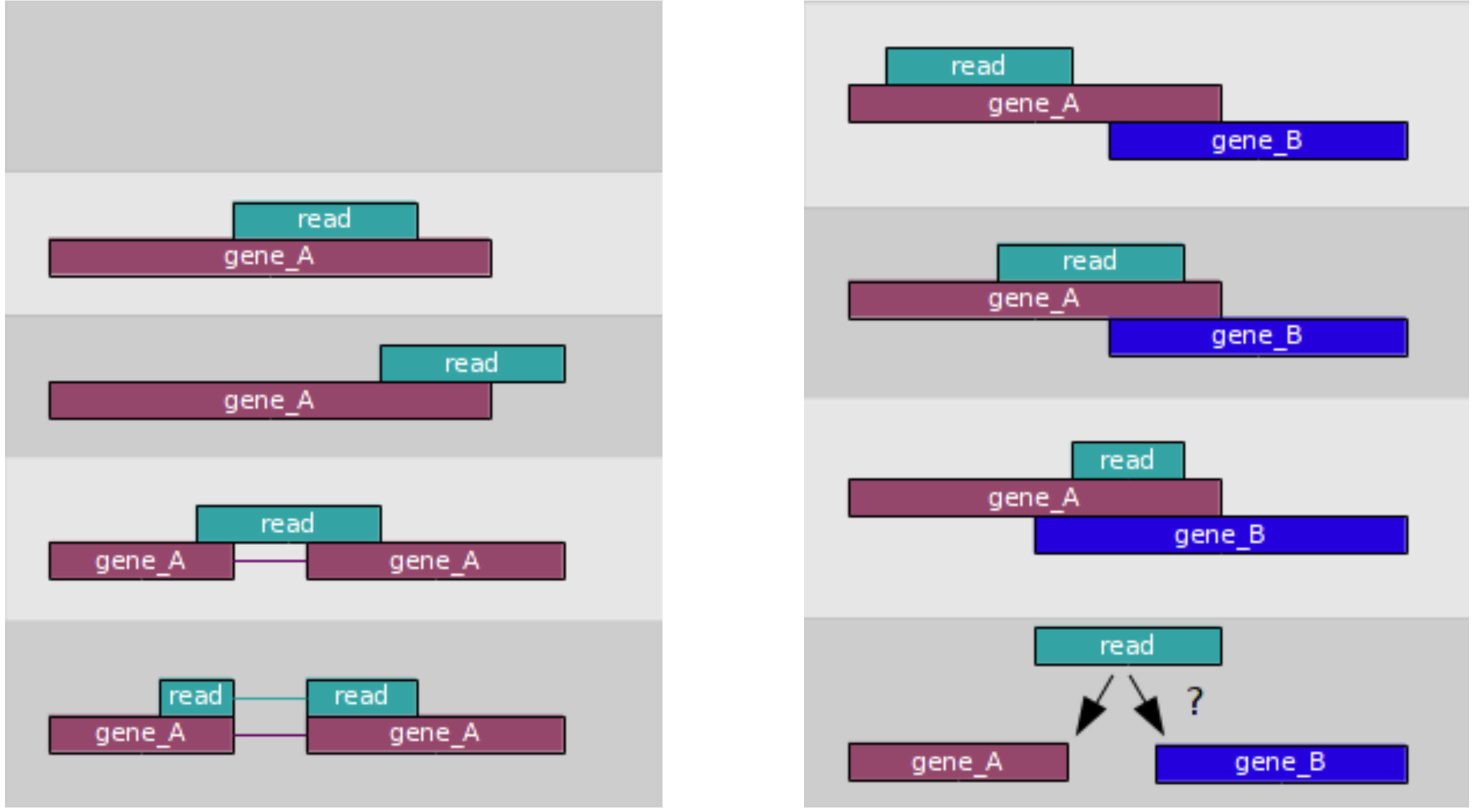
Counting reads as a measure of expression

- Two common counting tools are **featureCounts** and **htseq**.
- Total read count associated with a gene (meta-feature)
  - sum of reads mapping to exons (features) that are a part of that gene.
- Alternative approach: pseudoalignment and quantification of transcripts like **Salmon** and **Kallisto**



```
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??



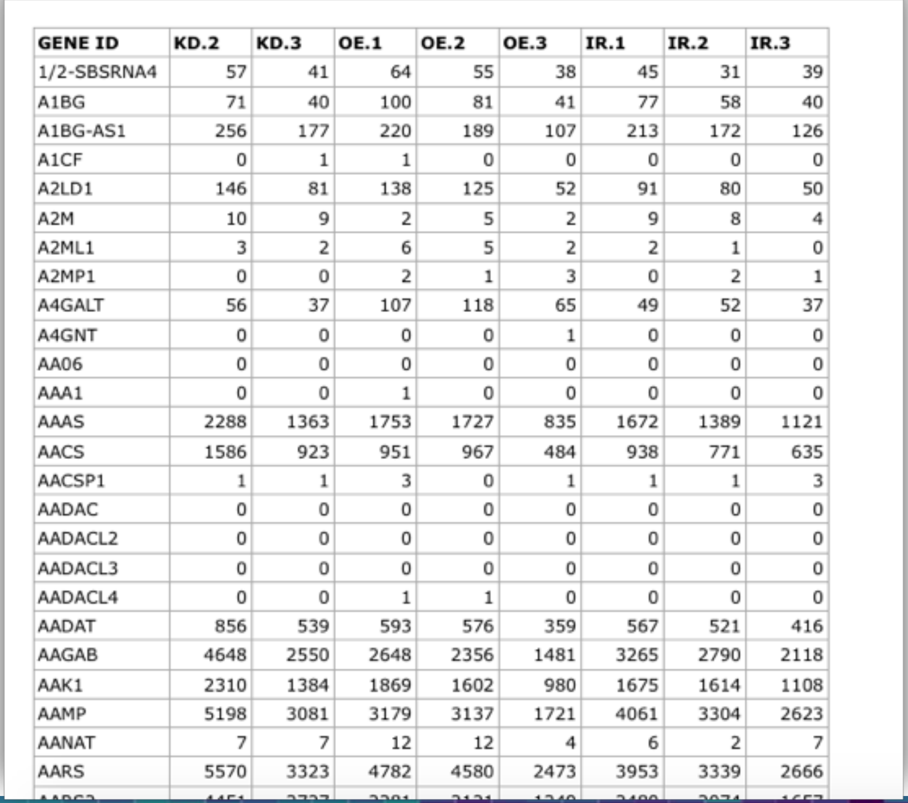
HTSeq manual

Read count matrix

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

Each column is a sample

Each row is a ~~gene~~ feature



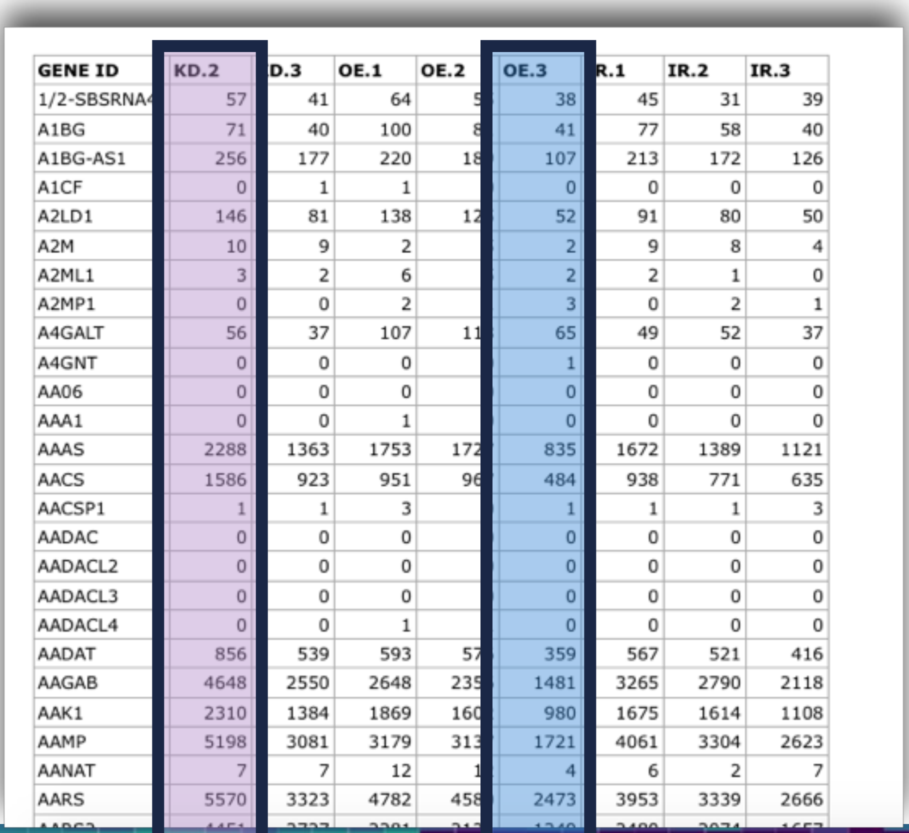
GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	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

Some problems with raw counts…

- Some samples consistently have more reads, some have fewer: **systematic biases**

Each column is a sample

Each row is a ~~gene~~ feature



GENE ID	KD.2	D.3	OE.1	OE.2	OE.3	R.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	5	38	45	31	39
A1BG	71	40	100	8	41	77	58	40
A1BG-AS1	256	177	220	18	107	213	172	126
A1CF	0	1	1		0	0	0	0
A2LD1	146	81	138	12	52	91	80	50
A2M	10	9	2		2	9	8	4
A2ML1	3	2	6		2	2	1	0
A2MP1	0	0	2		3	0	2	1
A4GALT	56	37	107	11	65	49	52	37
A4GNT	0	0	0		1	0	0	0
AA06	0	0	0		0	0	0	0
AAA1	0	0	1		0	0	0	0
AAAS	2288	1363	1753	172	835	1672	1389	1121
AACS	1586	923	951	96	484	938	771	635
AACSP1	1	1	3		1	1	1	3
AADAC	0	0	0		0	0	0	0
AADACL2	0	0	0		0	0	0	0
AADACL3	0	0	0		0	0	0	0
AADACL4	0	0	1		0	0	0	0
AADAT	856	539	593	57	359	567	521	416
AAGAB	4648	2550	2648	235	1481	3265	2790	2118
AAK1	2310	1384	1869	160	980	1675	1614	1108
AAMP	5198	3081	3179	313	1721	4061	3304	2623
AANAT	7	7	12	1	4	6	2	7
AARS	5570	3323	4782	458	2473	3953	3339	2666

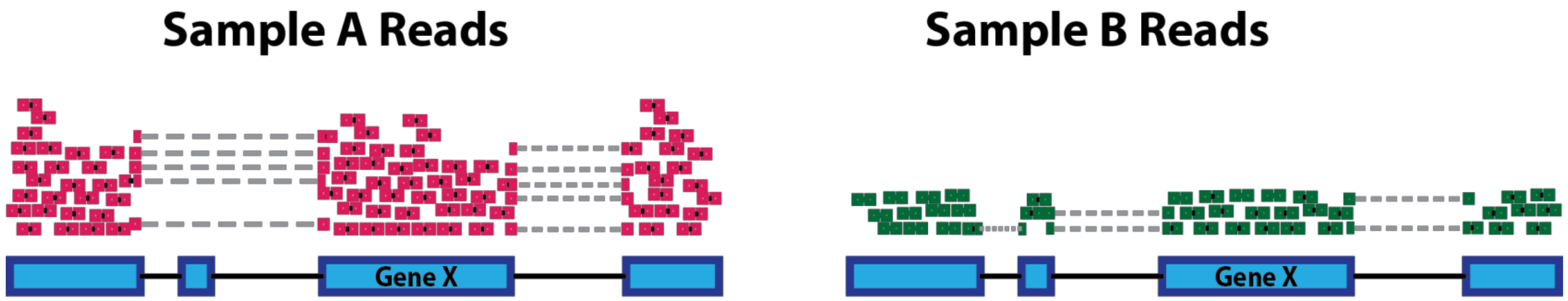
Solution: normalization

- Normalization is NOT fitting a normal distribution or transforming data.
- Normalization aims to identify and account for the nature and magnitude of **systematic biases**

- The main factors often considered during normalization:
  - **Sequencing depth (aka library size)**
  - **RNA composition**
  - Gene length (some methods)

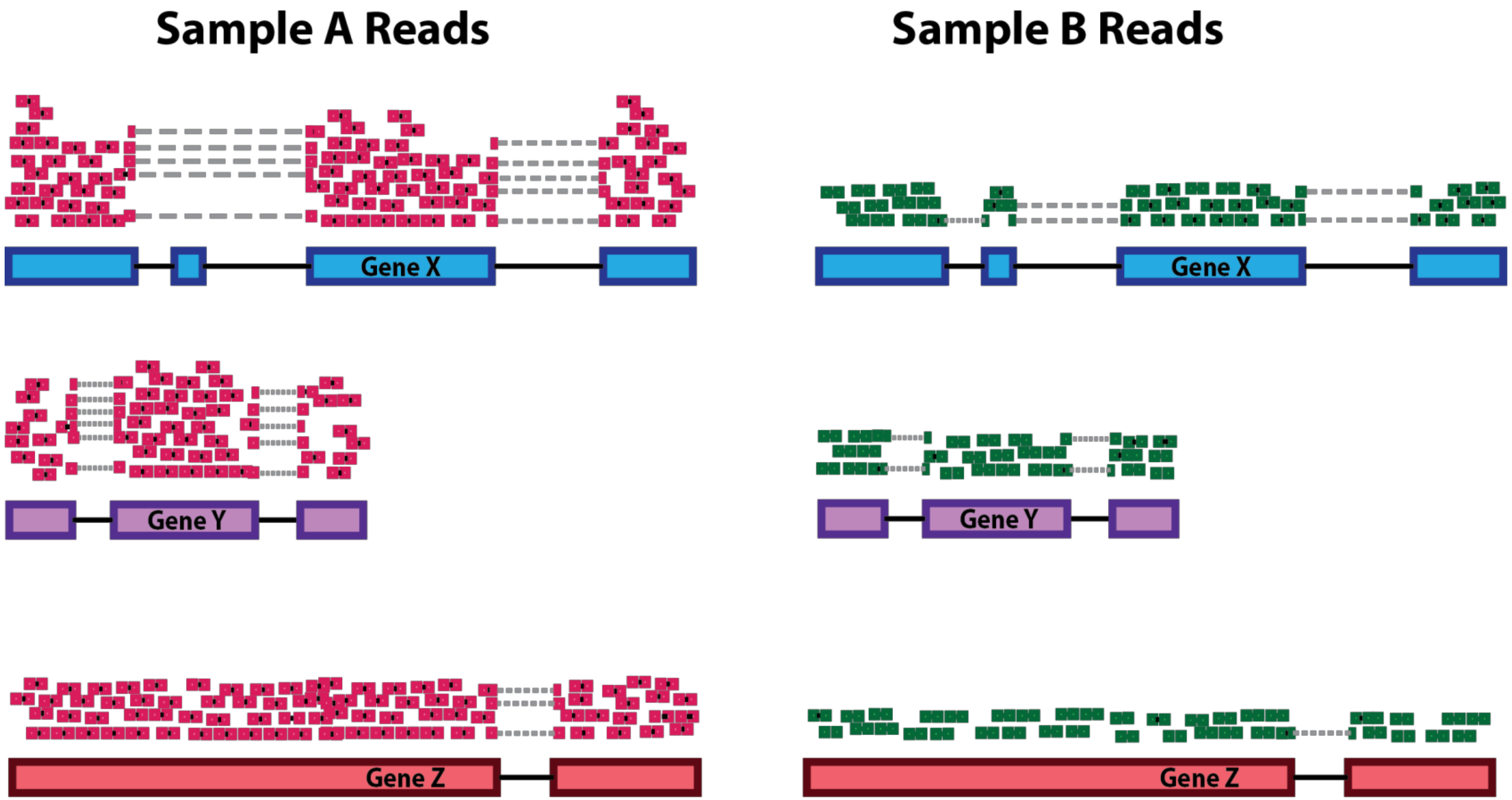
Normalization

Sequencing depth



Normalization

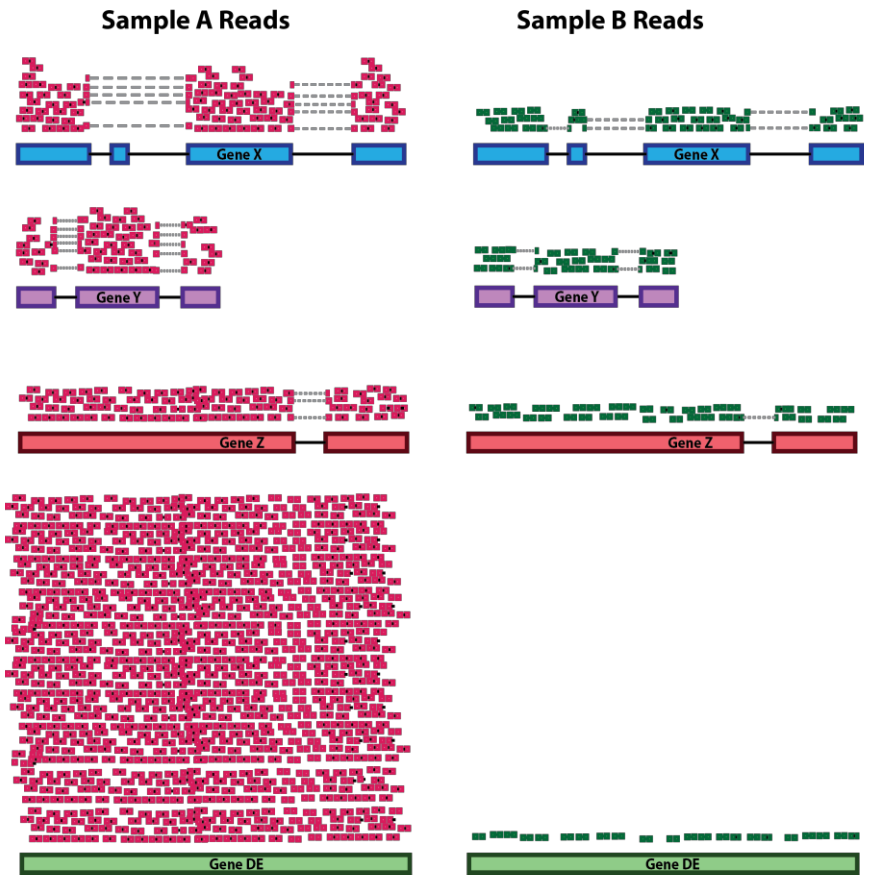
Sequencing depth



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)
- Generates a **size factor** **(also called scaling factor)** for each sample to account for variation in library s

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!

Why don’t we normalize by gene length?

- Most of the time we are comparing WITHIN gene, ACROSS samples
    - Dividing by the gene length is superfluous
    - Common tools like edgeR, limma, Deseq2

- Some metrics facilitate comparison ACROSS genes, WITHIN sample
    - These should be length-normalized
    - TPM (Transcripts Per Kilobase Million): normalize by length, then by depth
    - RPKM (Reads Per Kilobase Million): normalize by depth, then by length

Exploring patterns in RNAseq data

Clustering of samples

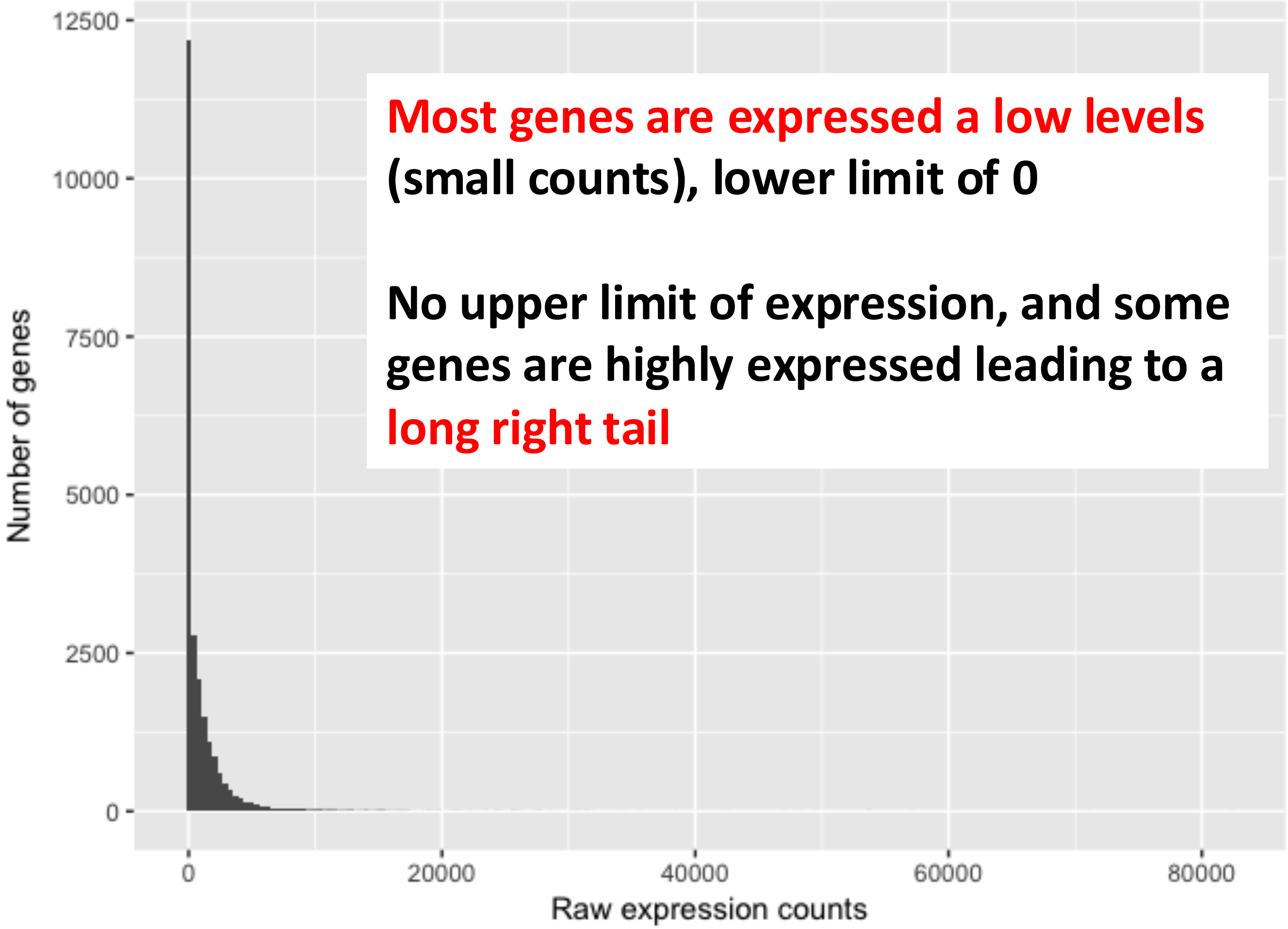
- Dimension reduction analysis (e.g., PCA, PLS, MDS)
- Clustering (e.g., hierarchical clustering, k-means clustering)

Clustering of features

- Same as above, just focusing on features
- Weighted co-expression analysis (WGCNA, correlation among features)

Properties of RNA-seq count data

The distribution of RNA-seq counts for **a single sample**:



Most genes are expressed a low levels (small counts), lower limit of 0

No upper limit of expression, and some genes are highly expressed leading to a long right tail

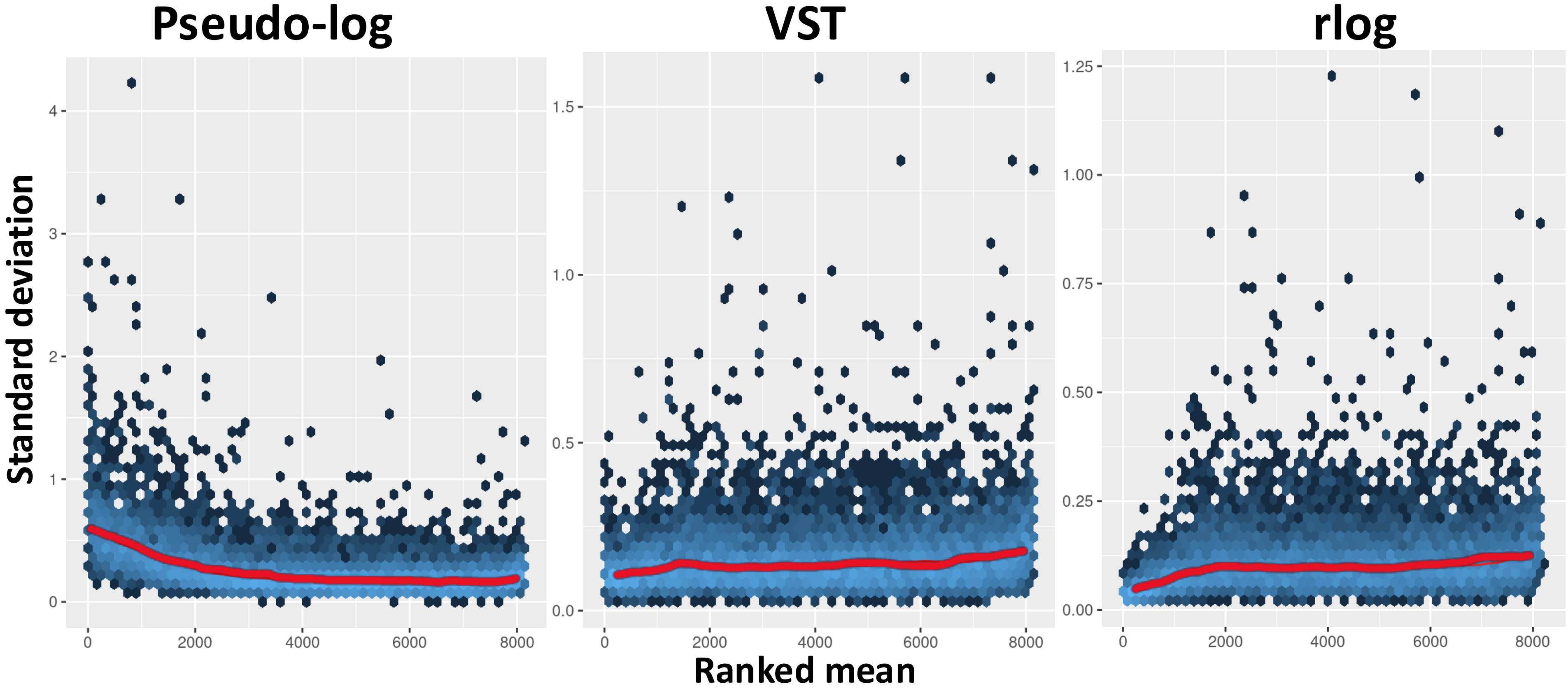
Data transformations for clustering and visualization

- Pseudo-log: $y = \ log_2(n + \ n_0)$
  - n_0 is a constant, like 1
  - Variance not stable at low values (does not scale with expression)

- Instead, we want to transform the data to remove the trend (variances roughly similar across mean values)
- Variance stabilizing transformation *DESeq2::vst()*
- Regularized log transformation *DESeq2::rlog()*

Huber et al. 2003 Stat. Appl. Genet. Mol. Biol.,
Anders & Huber 2010 Nature,
Love et al. 2023 “Analyzing RNA-seq data with DESeq2”

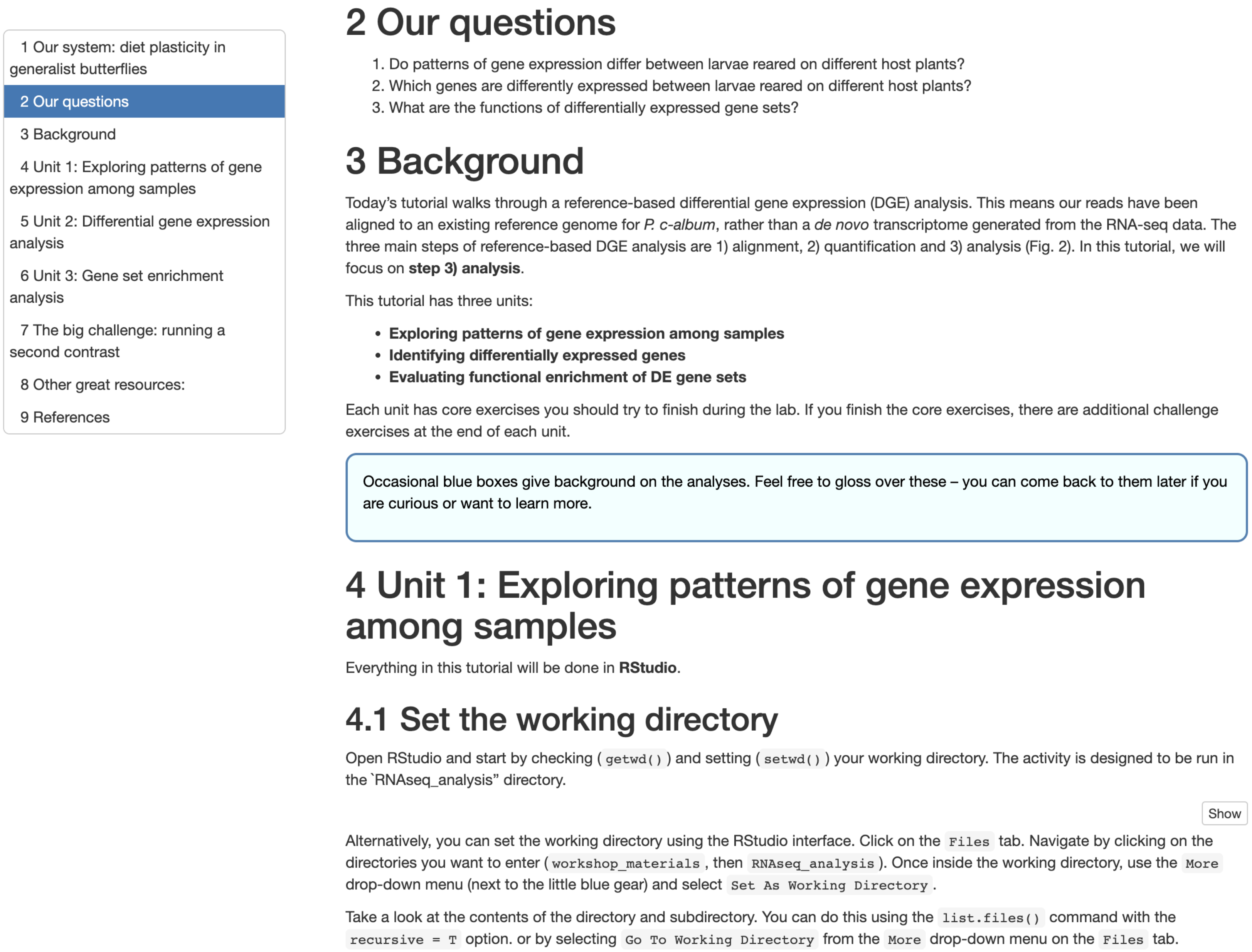
Effect of transformations on variance



Love et al. 2023 “Analyzing RNA-seq data with DESeq2”

Today’s lab: Polygonia c-album

Orientation to the tutorial



2 Our questions

1. Do patterns of gene expression differ between larvae reared on different host plants?
2. Which genes are differently expressed between larvae reared on different host plants?
3. What are the functions of differentially expressed gene sets?

3 Background

Today’s tutorial walks through a reference-based differential gene expression (DGE) analysis. This means our reads have been aligned to an existing reference genome for *P. c-album*, rather than a *de novo* transcriptome generated from the RNA-seq data. The three main steps of reference-based DGE analysis are 1) alignment, 2) quantification and 3) analysis (Fig. 2). In this tutorial, we will focus on **step 3) analysis**.

This tutorial has three units:

- **Exploring patterns of gene expression among samples**
- **Identifying differentially expressed genes**
- **Evaluating functional enrichment of DE gene sets**

Each unit has core exercises you should try to finish during the lab. If you finish the core exercises, there are additional challenge exercises at the end of each unit.

> Occasional blue boxes give background on the analyses. Feel free to gloss over these – you can come back to them later if you are curious or want to learn more.

4 Unit 1: Exploring patterns of gene expression among samples

Everything in this tutorial will be done in **RStudio**.

4.1 Set the working directory

Open RStudio and start by checking (`getwd()`) and setting (`setwd()`) your working directory. The activity is designed to be run in the `RNAseq_analysis” directory.

Alternatively, you can set the working directory using the RStudio interface. Click on the `Files` tab. Navigate by clicking on the directories you want to enter (`workshop_materials` , then `RNAseq_analysis`). Once inside the working directory, use the `More` drop-down menu (next to the little blue gear) and select `Set As Working Directory` .

Take a look at the contents of the directory and subdirectory. You can do this using the `list.files()` command with the `recursive = T` option. or by selecting `Go To Working Directory` from the `More` drop-down menu on the `Files` tab.

Exploring patterns in RNAseq data

Part 1

Core tasks:

- Load raw count matrix
- Transform for visualization
- PCA of samples
- Hierarchical clustering of samples

Challenge exercises

Open work time (25 min)

Five more minutes!

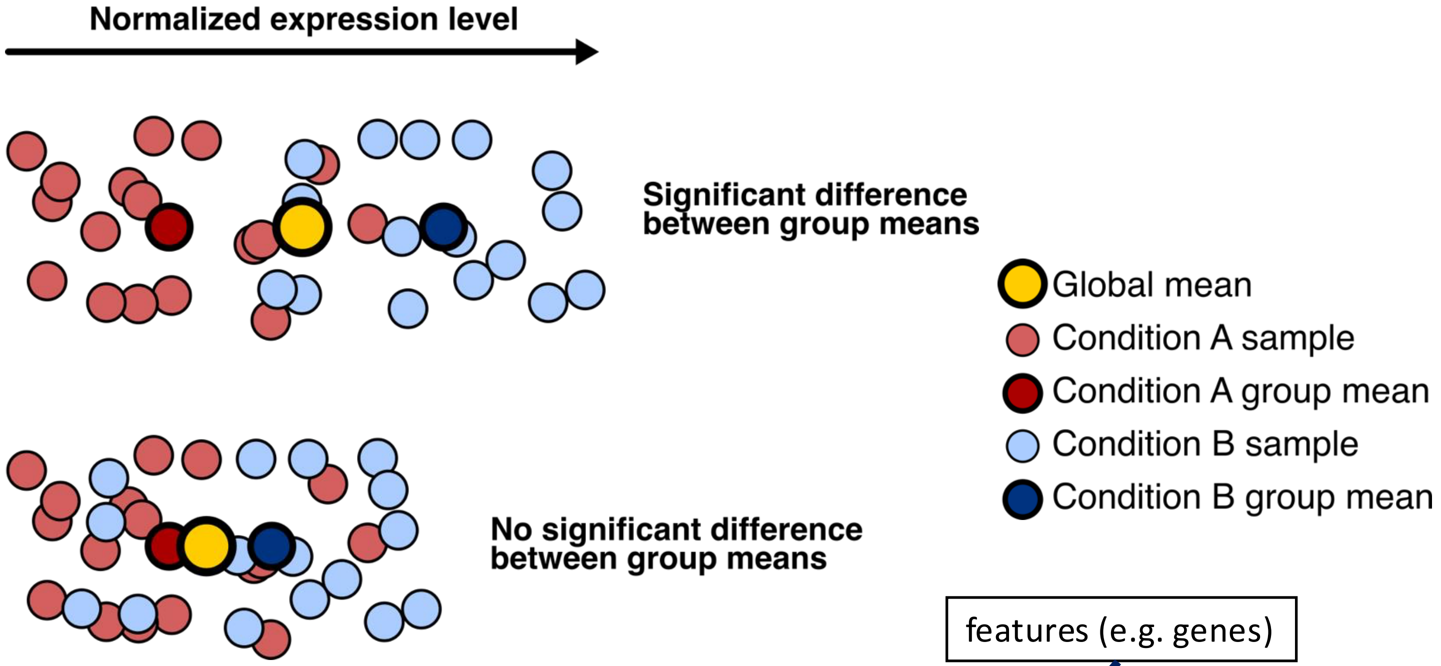
Review

Lab activities

- **Part 1** Exploring patterns in RNAseq data
- **Part 2** Differential gene expression analysis
- **Part 3** Functional enrichment of gene sets

Differential expression analysis

Analysis


Normalized expression level
Significant difference between group means
No significant difference between group means
Global mean
Condition A sample
Condition A group mean
Condition B sample
Condition B group mean

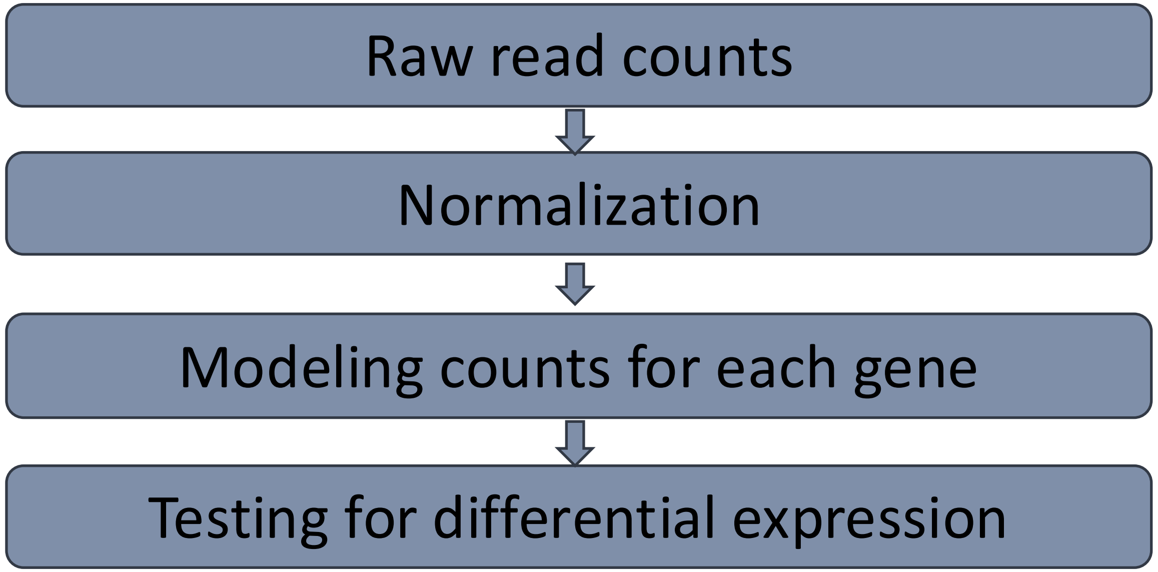
Input

features (e.g. genes)

samples

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



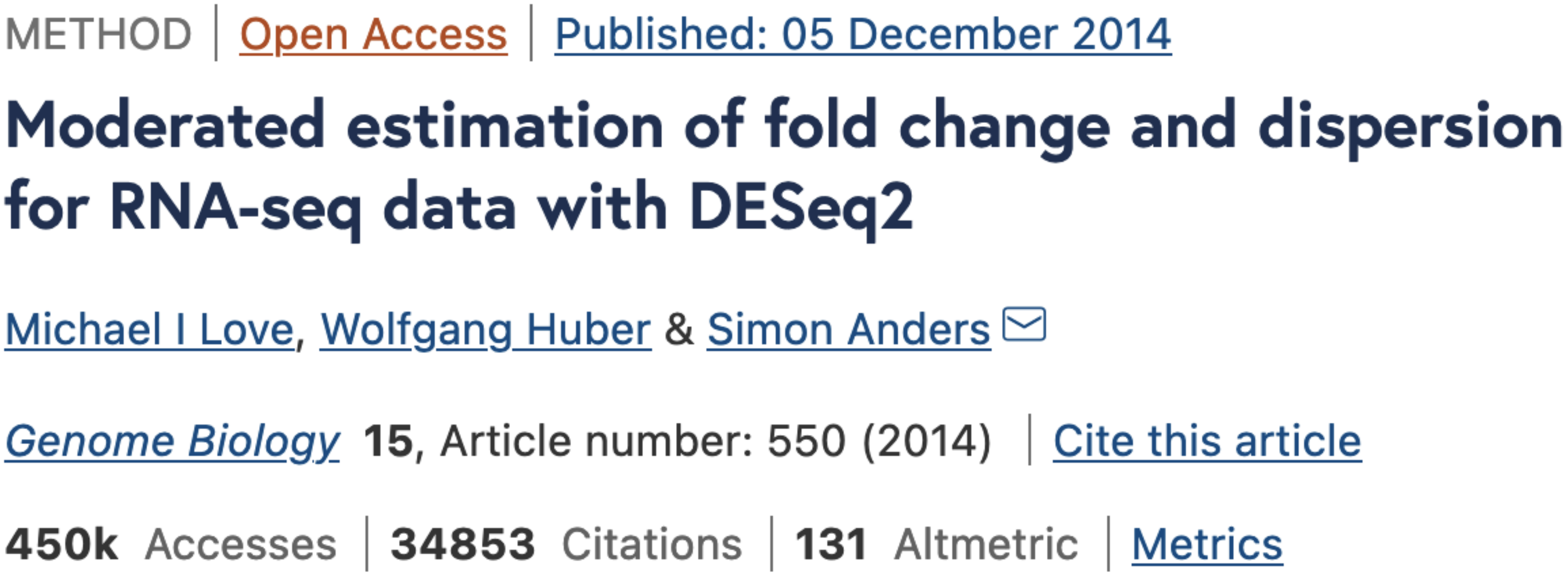
Raw read counts

Normalization

Modeling counts for each gene

Testing for differential expression

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

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

Statistical modeling of count data

Which probability distributions are suitable for modeling count data?

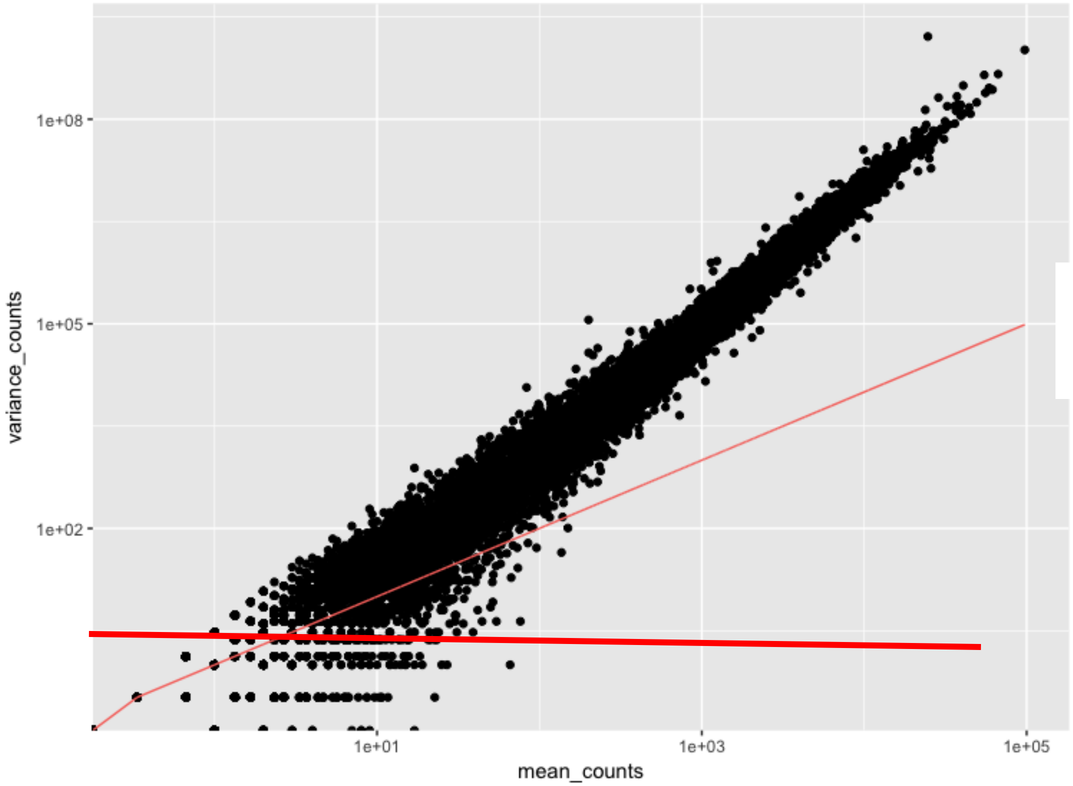
Poisson distribution?

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?



Fitted poisson distribution

mean = variance?

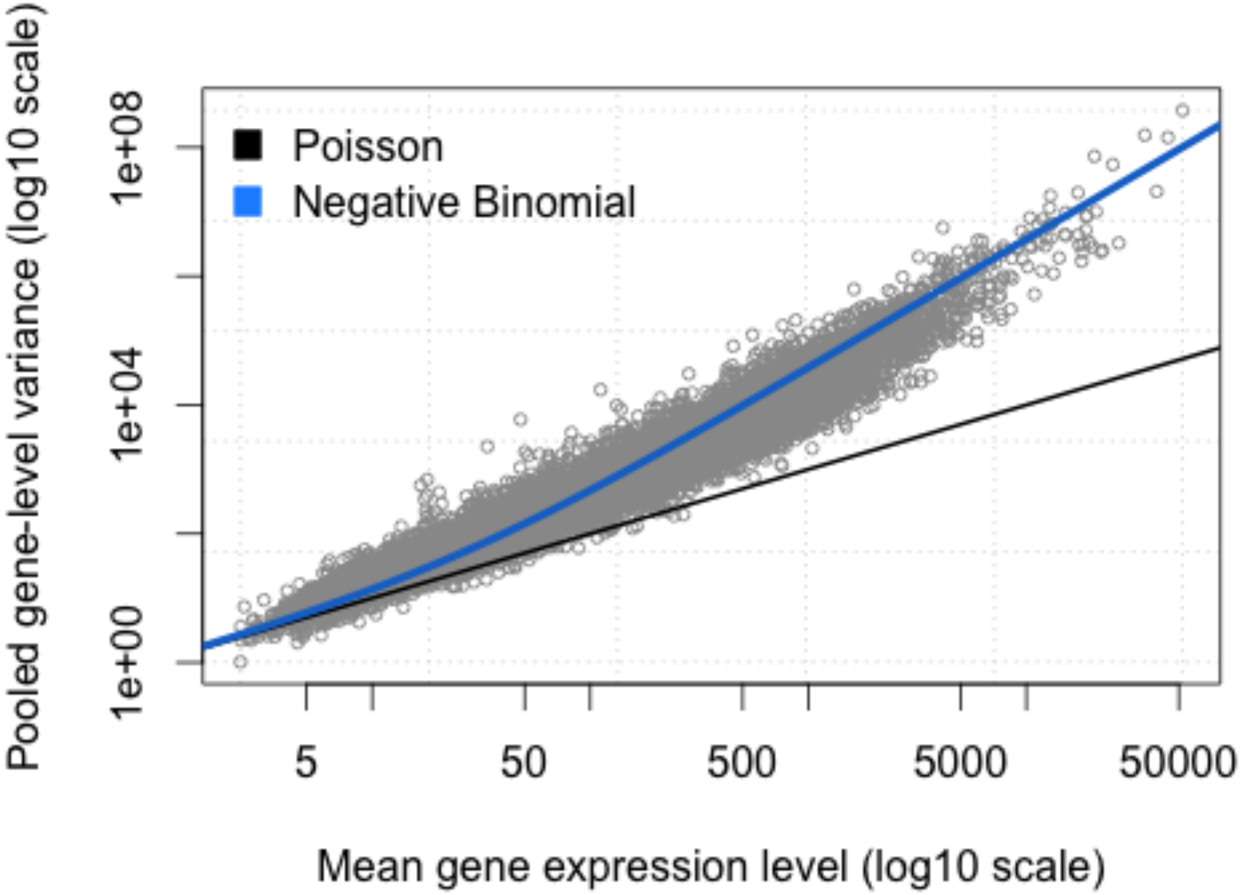
Poisson distribution is not suitable to model count data across the biological samples.

Statistical modeling of count data

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

- two parameters, one for the mean and one for the variance

- fexlibility to estimate the amount of **dispersion** for each gene across samples.

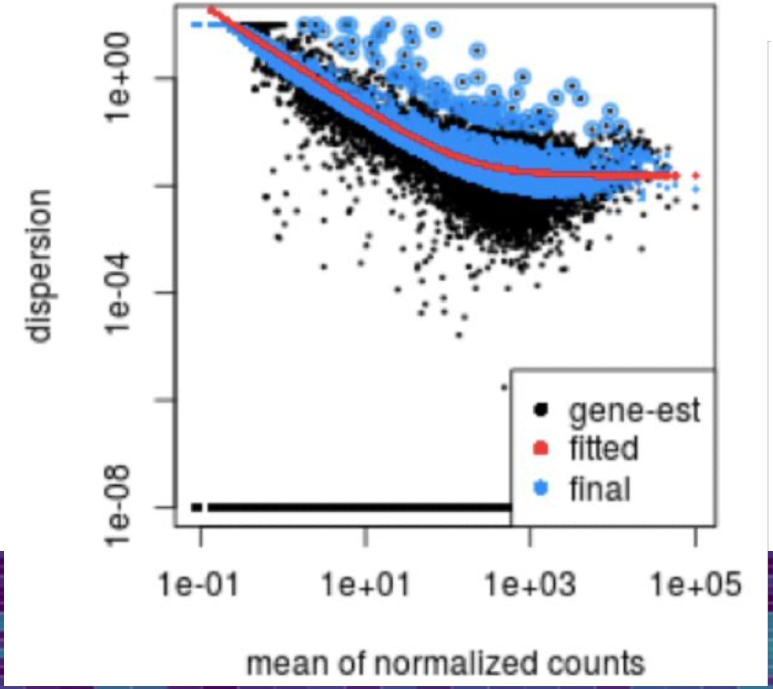


How does the dispersion relate to our model?

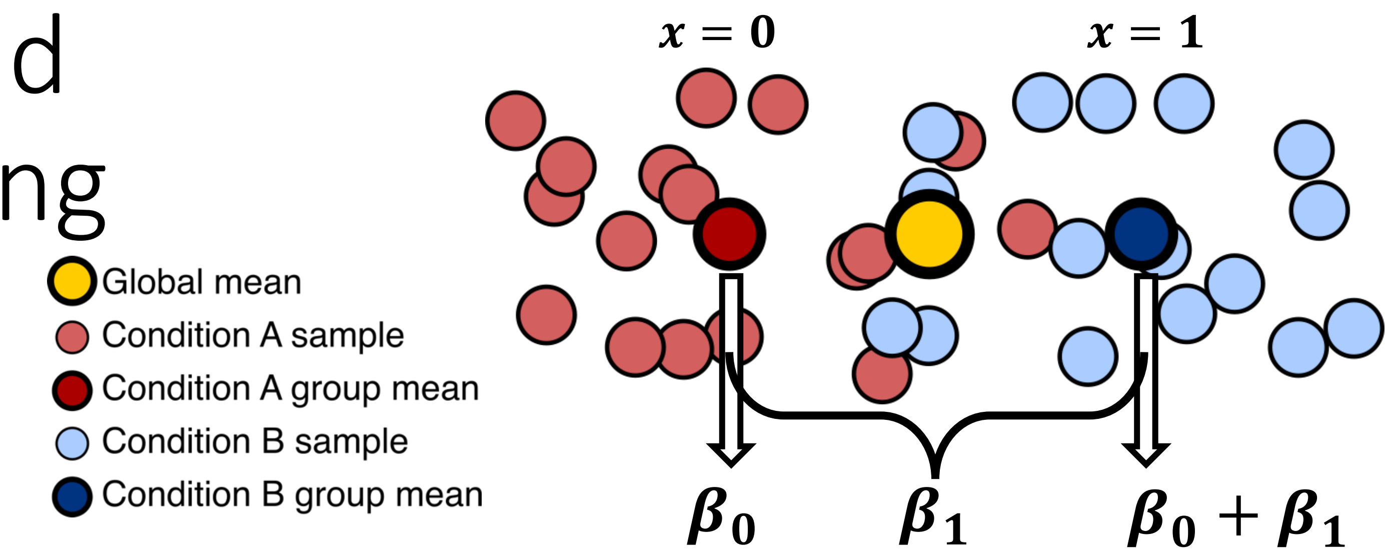
- Variation is an important part of model fitting and hypothesis testing.
- Estimates of variation for individual genes are often unreliable.

- Tools like DESeq2 share information across genes to generate more accurate estimates of variation:

Fitted dispersion curve = **expected dispersion for genes of a given level of expression (e.g., mean normalized count)**



Model fitting and hypothesis testing

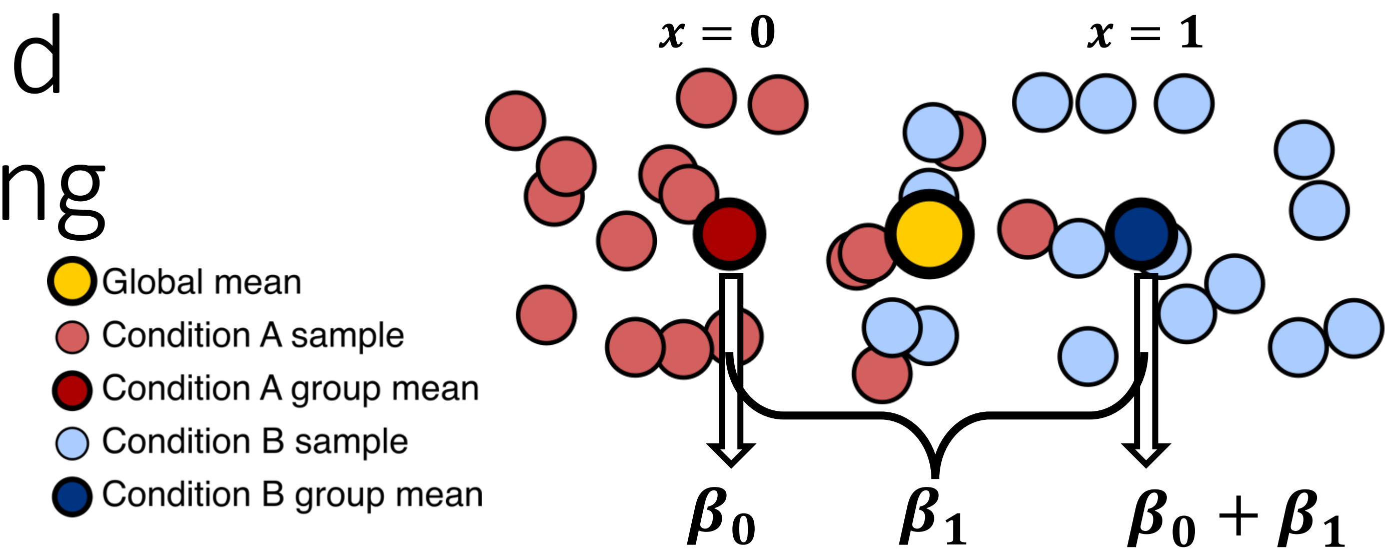


Step 4. Generalized Linear Model fit for each gene

$$y = \beta_0 + x_1\beta_1$$
$$y = \beta_0 + 0 * \beta_1$$
$$y = \beta_0$$
$$y = \beta_0 + 1 * \beta_1$$
$$y = \beta_0 + \beta_1$$

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

Model fitting and hypothesis testing



Step 4. Generalized Linear Model fit for each gene

$$y = \beta_0 + \beta_1$$

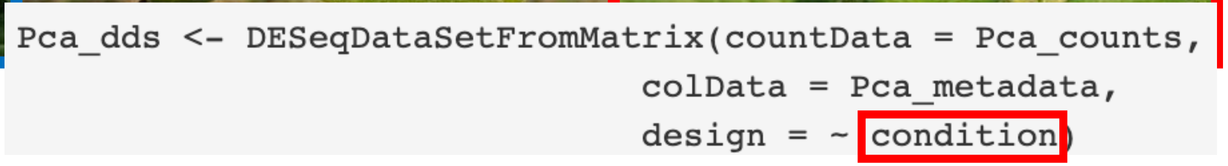
$$y - \beta_0 = \beta_1$$

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

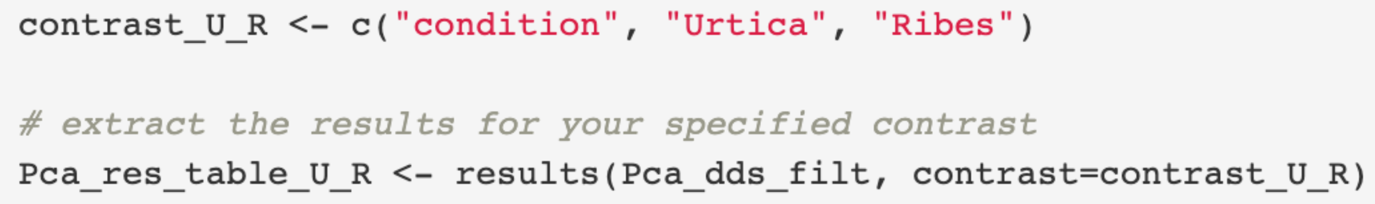
$$\log_2\left(\frac{expression_{blue}}{expression_{red}}\right) = \beta_1 \quad \text{“}\mathbf{log_2}\ \textbf{Fold Change}\text{”}$$

$$\boxed{\begin{array}{l} log_2 1 = 0 \\ log_2 2 = 1 \\ log_2 4 = 2 \end{array}}$$

Specifying contrasts



```
Pca_dds <- DESeqDataSetFromMatrix(countData = Pca_counts,
                                  colData = Pca_metadata,
                                  design = ~ condition)
```

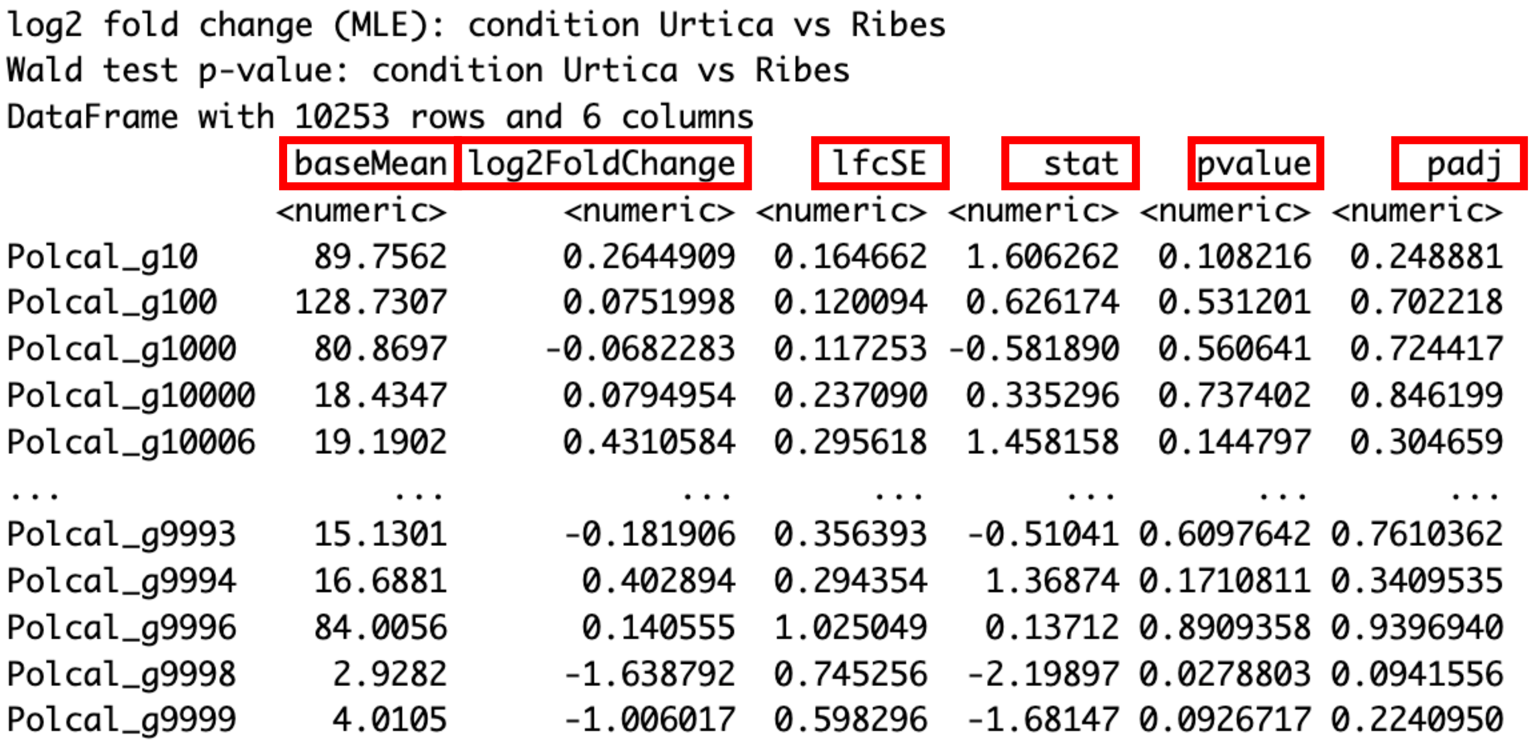


```
contrast_U_R <- c("condition", "Urtica", "Ribes")

# extract the results for your specified contrast
Pca_res_table_U_R <- results(Pca_dds_filt, contrast=contrast_U_R)
```

$$\log_2\left(\frac{expression_{\boldsymbol{Ribes}}}{expression_{\boldsymbol{Urtica}}}\right) = \boldsymbol{\beta}_1 \quad \text{“}\mathbf{log_2\ Fold\ Change}\text{”}$$

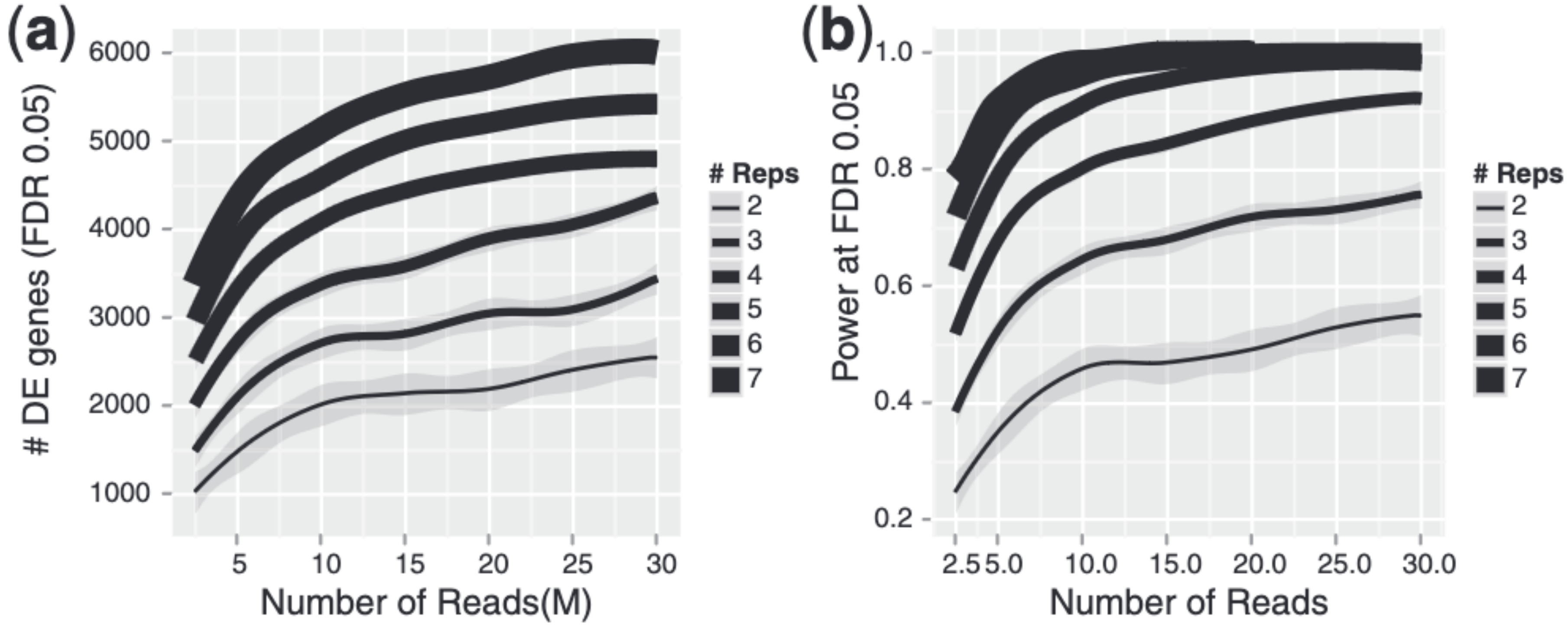
Output of DESeq2



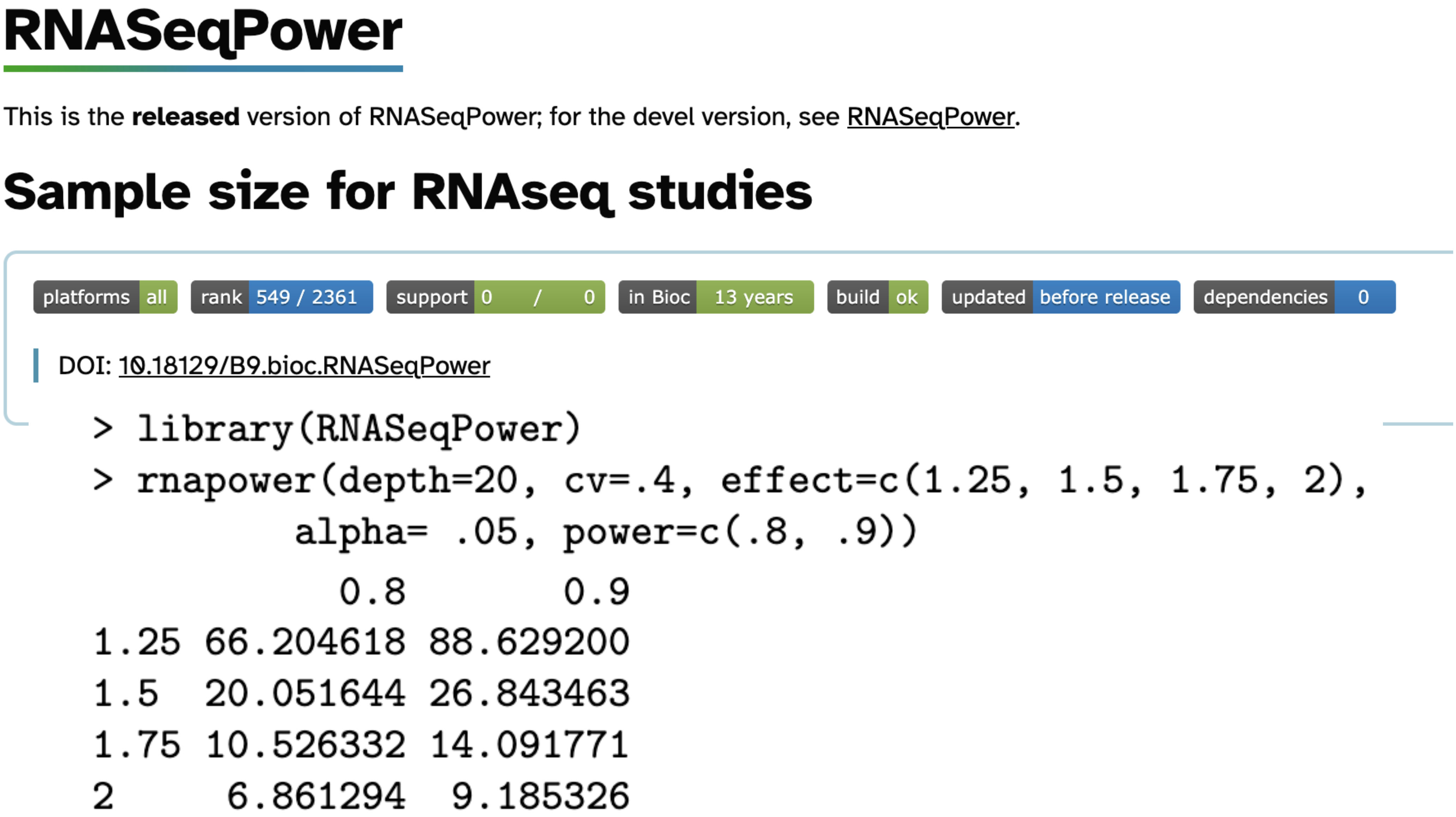
```
log2 fold change (MLE): condition Urtica vs Ribes
Wald test p-value: condition Urtica vs Ribes
DataFrame with 10253 rows and 6 columns
              baseMean log2FoldChange     lfcSE      stat    pvalue      padj
             <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
Polcal_g10     89.7562      0.2644909  0.164662  1.606262  0.108216  0.248881
Polcal_g100   128.7307      0.0751998  0.120094  0.626174  0.531201  0.702218
Polcal_g1000   80.8697     -0.0682283  0.117253 -0.581890  0.560641  0.724417
Polcal_g10000  18.4347      0.0794954  0.237090  0.335296  0.737402  0.846199
Polcal_g10006  19.1902      0.4310584  0.295618  1.458158  0.144797  0.304659
...                ...            ...       ...       ...       ...       ...
Polcal_g9993   15.1301      -0.181906  0.356393  -0.51041 0.6097642 0.7610362
Polcal_g9994   16.6881       0.402894  0.294354   1.36874 0.1710811 0.3409535
Polcal_g9996   84.0056       0.140555  1.025049   0.13712 0.8909358 0.9396940
Polcal_g9998    2.9282      -1.638792  0.745256  -2.19897 0.0278803 0.0941556
Polcal_g9999    4.0105      -1.006017  0.598296  -1.68147 0.0926717 0.2240950
```

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 – use a pre-defined cutoff for significance

When can we detect differential expression?



Liu et al. 2014. Bioinformatics



RNASeqPower

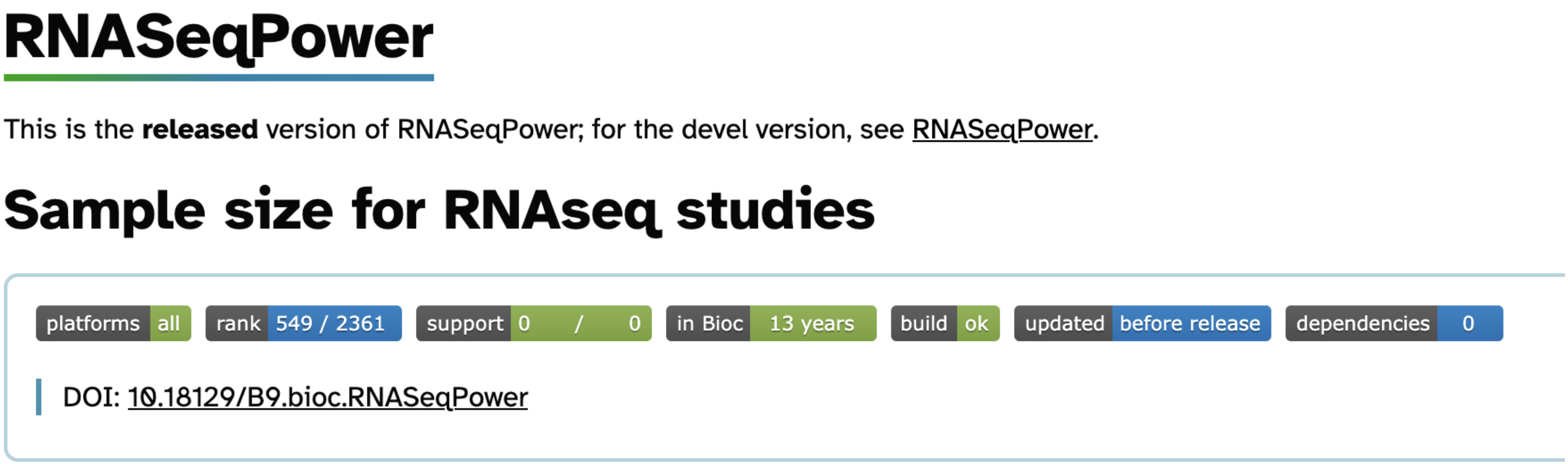
This is the **released** version of RNASeqPower; for the devel version, see RNASeqPower.

Sample size for RNAseq studies

platforms all | rank 549 / 2361 | support 0 / 0 | in Bioc 13 years | build ok | updated before release | dependencies 0

DOI: 10.18129/B9.bioc.RNASeqPower

```
> library(RNASeqPower)
> rnapower(depth=20, cv=.4, effect=c(1.25, 1.5, 1.75, 2),
           alpha= .05, power=c(.8, .9))
           0.8       0.9
1.25 66.204618 88.629200
1.5  20.051644 26.843463
1.75 10.526332 14.091771
2     6.861294  9.185326
```



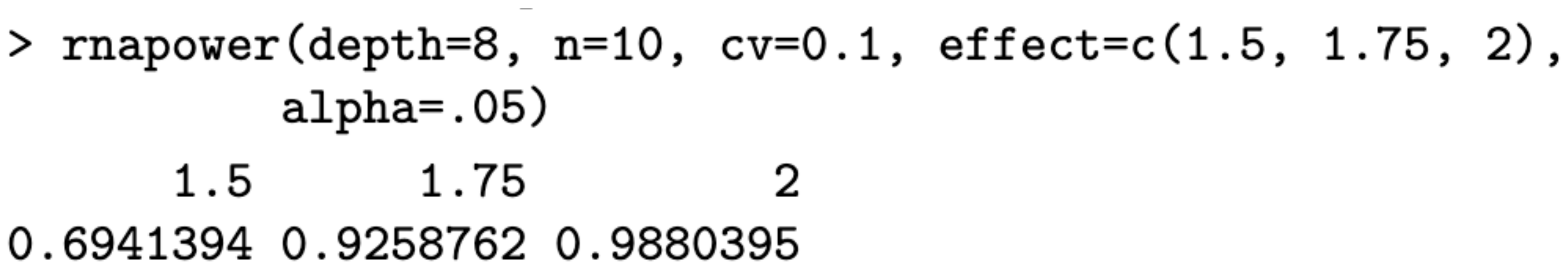
RNASeqPower

This is the **released** version of RNASeqPower; for the devel version, see RNASeqPower.

Sample size for RNAseq studies

platforms all | rank 549 / 2361 | support 0 / 0 | in Bioc 13 years | build ok | updated before release | dependencies 0

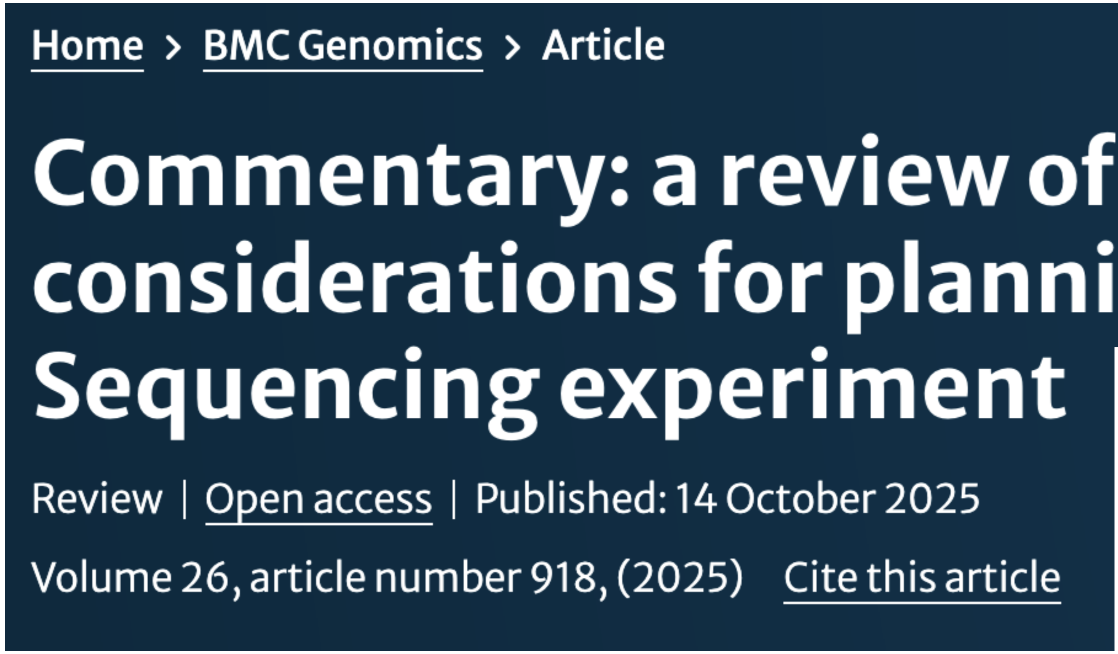
DOI: 10.18129/B9.bioc.RNASeqPower



```
> rnapower(depth=8, n=10, cv=0.1, effect=c(1.5, 1.75, 2),
            alpha=.05)
      1.5      1.75         2
0.6941394 0.9258762 0.9880395
```

Additional tools for study design (or post-hoc evaluation)

- ssizeRNA (2016, updated 2025)

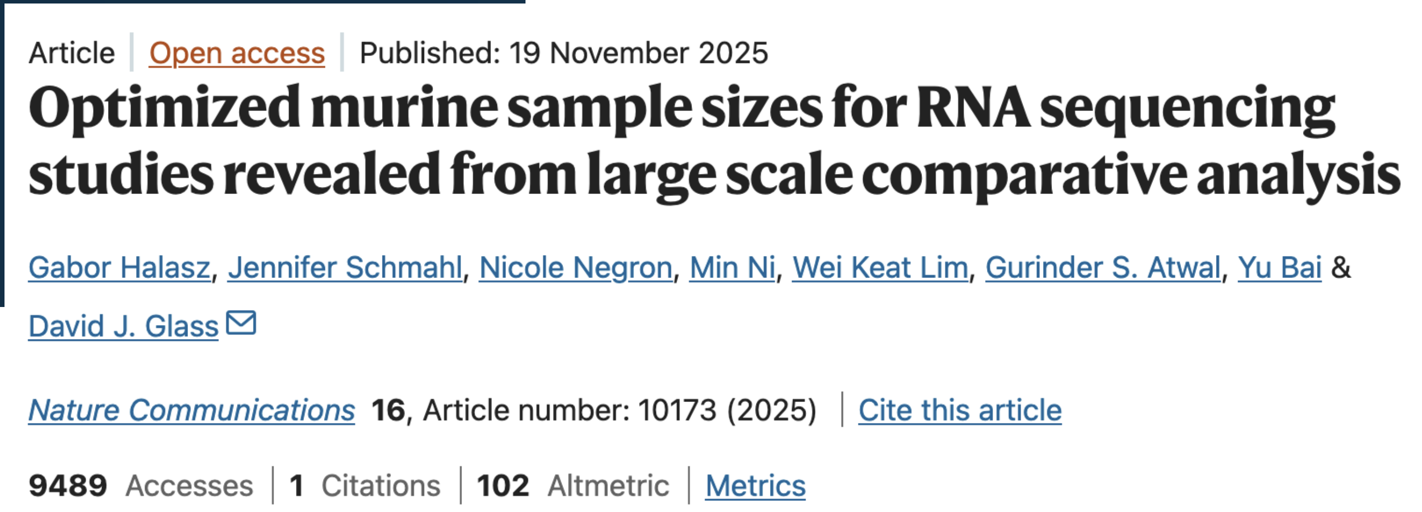


Home > BMC Genomics > Article

Commentary: a review of technical considerations for planning an RNA-Sequencing experiment

Review | Open access | Published: 14 October 2025

Volume 26, article number 918, (2025) Cite this article



Article | Open access | Published: 19 November 2025

Optimized murine sample sizes for RNA sequencing studies revealed from large scale comparative analysis

Gabor Halasz, Jennifer Schmahl, Nicole Negron, Min Ni, Wei Keat Lim, Gurinder S. Atwal, Yu Bai & David J. Glass

Nature Communications **16**, Article number: 10173 (2025) | Cite this article

9489 Accesses | **1** Citations | **102** Altmetric | Metrics

What do we do with DE genes?

- Visualize expression levels, $\log_2$ fold changes, and significance
- Identify up- and down-regulated genes
- Compare sets of DE genes
- Test for functional enrichment of DE gene sets

Differential gene expression

- Part 2 Core tasks:
  - Run a pairwise contrast
  - Visualize differential expression with a volcano plot
  - Extract the list of DE genes
  - Visualize DE genes in a heatmap

- Challenge exercises
- Challenge questions

Open work time

Five more minutes!

Review

Part 3: Functional enrichment

Differential expression or clustering analysis can produce large gene sets.

How can we figure out the functional consequences of these differences?

Gene set enrichment analysis:

Do functional terms occur in the target gene set more than expected by chance?

- GO terms
- KEGG pathways
- Reactome pathways

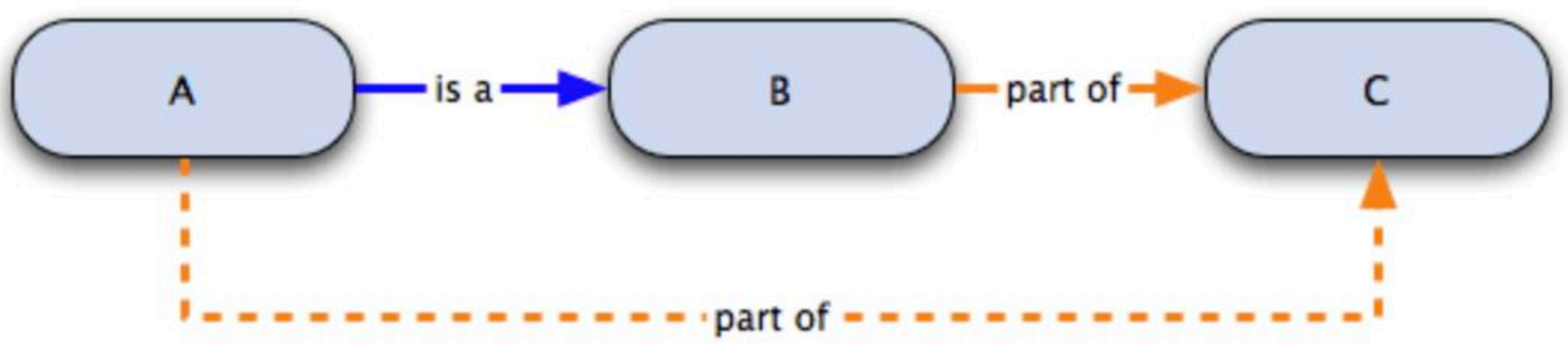
GO term enrichment

$$background\ frequency: \frac{genes\ annotated\ with\ the\ GO\ term\ in\ full\ annotation}{total\ genes\ in\ full\ annotation}$$

$$sample\ frequency: \frac{genes\ annotated\ with\ GO\ term\ in\ sample}{total\ genes\ in\ sample}$$

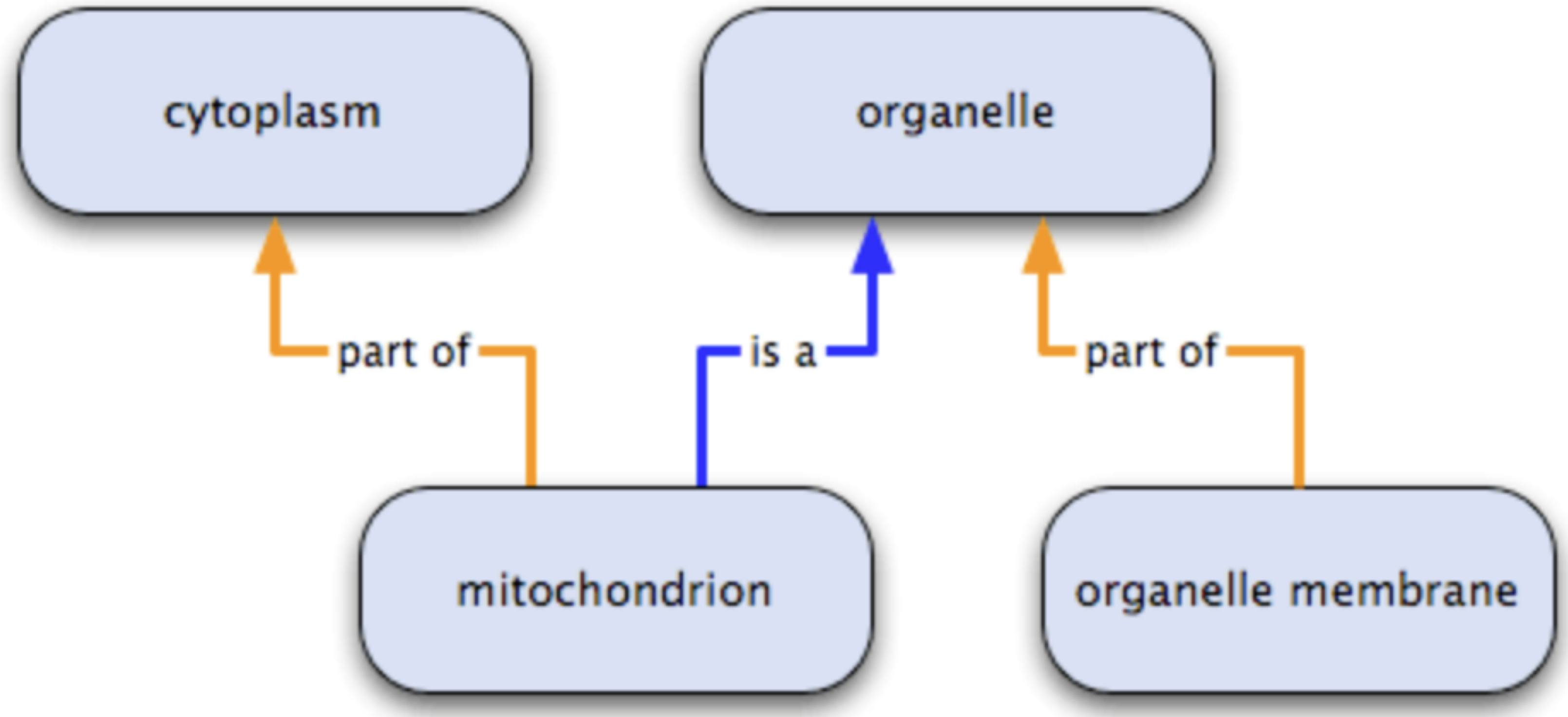
P-value is the probability or chance of seeing the sample frequency, given the background frequency.

GO relations



Relations in the Gene Ontology
https://geneontology.org/docs/ontology-relations/

GO relations



Relations in the Gene Ontology
https://geneontology.org/docs/ontology-relations/

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	Tutorial: https://fuzzyatelin.github.io/bioanth-stats/module-F21-Group1/module-F21-Group1.html#Weighted_Gene_Correlation_Network_Analysis
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 Good when: few reps, lowly expressed genes	User guide: https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf
Limma (R package)	Differential expression analysis with complex experimental designs, including batch and random effects	Vignette:https://www.bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/intro.html
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

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)

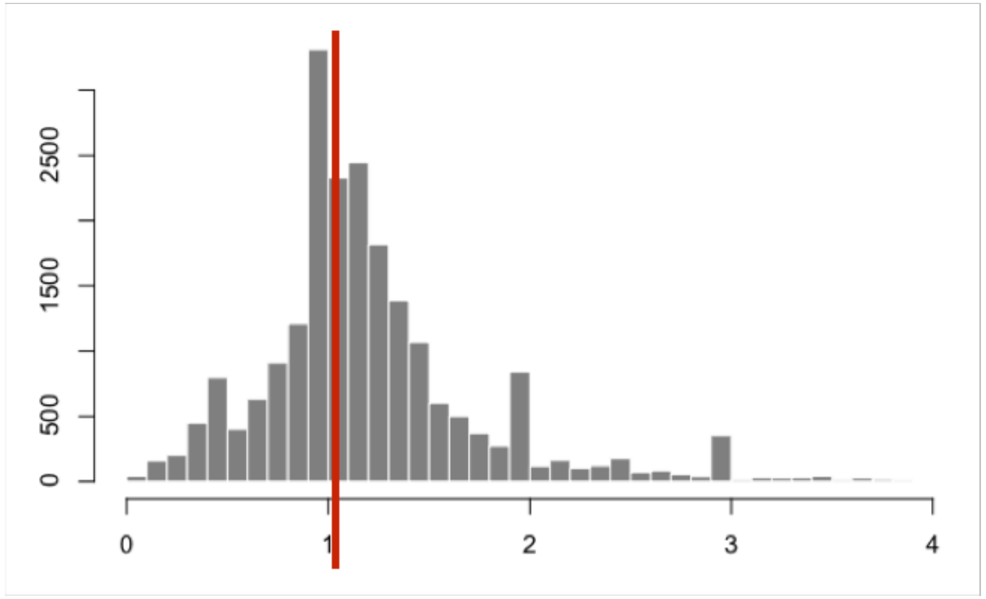
Gene	sampleA	sampleB	Pseudo-reference sample
EF2A	1489	906	sqrt(1489*906) = 1161.5
ABCD1	22	13	sqrt(22*13) = 16.9
…	…	…	…

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).

sample 1 / pseudo-reference sample



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(0.78, 0.77, 0.72, 0.8, 0.73, …))
=0.77
```

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!