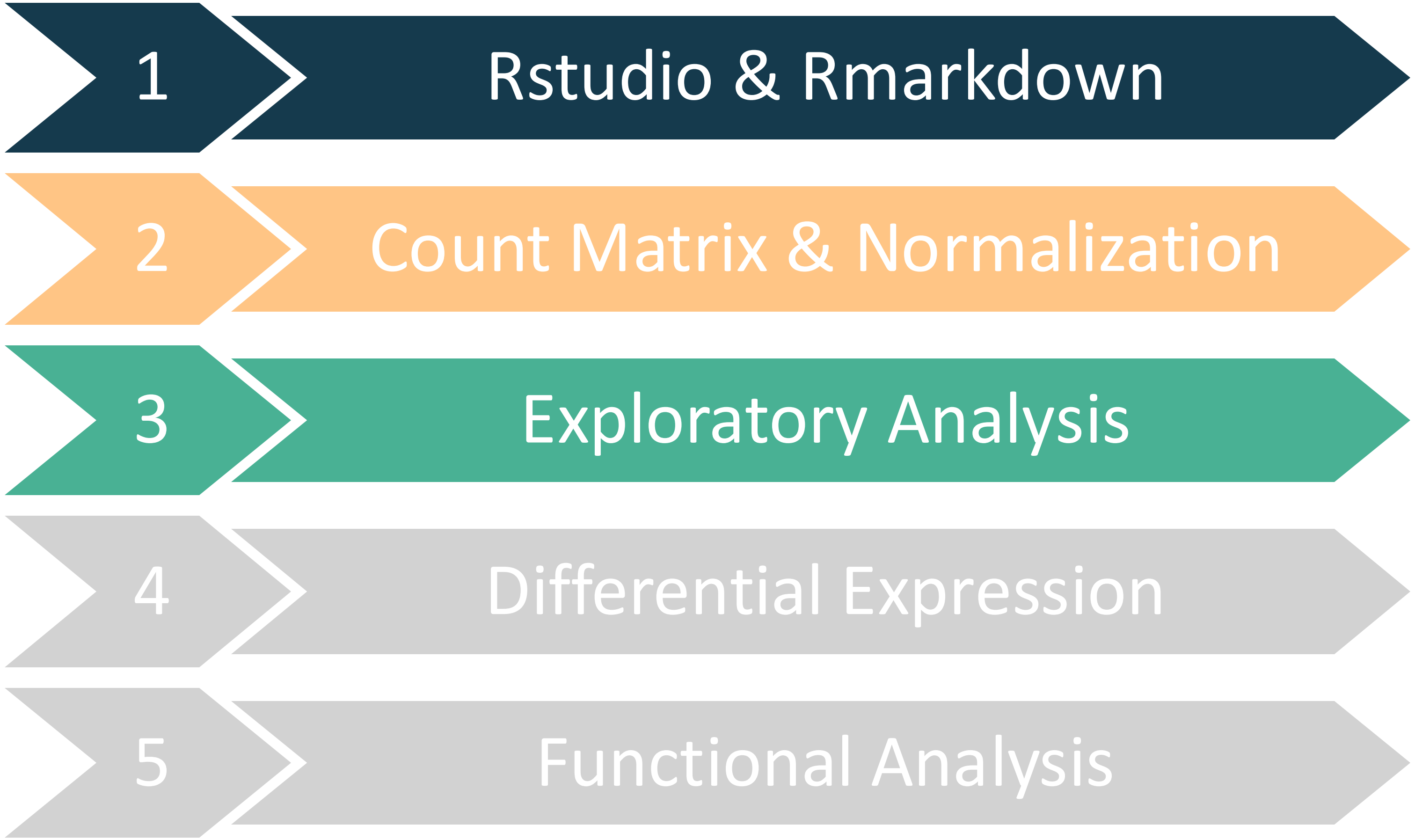
HeaDS

Exploratory analysis

Center for Health Data Science

Health Data Science Sandbox

Overview



1 Rstudio & Rmarkdown

2 Count Matrix & Normalization

3 Exploratory Analysis

4 Differential Expression

5 Functional Analysis

Exploratory analysis

Data is not a piñata:
always **LOOK** at your data!

Exploratory analysis

Helps to:

understand data

Inform decisions for downstream analysis

Exploratory analysis steps

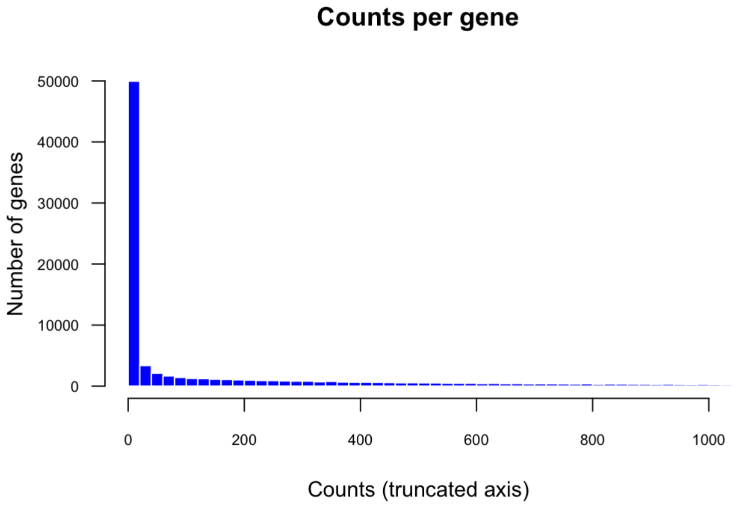
1. Pre-filtering: Remove low-expressed genes or outliers

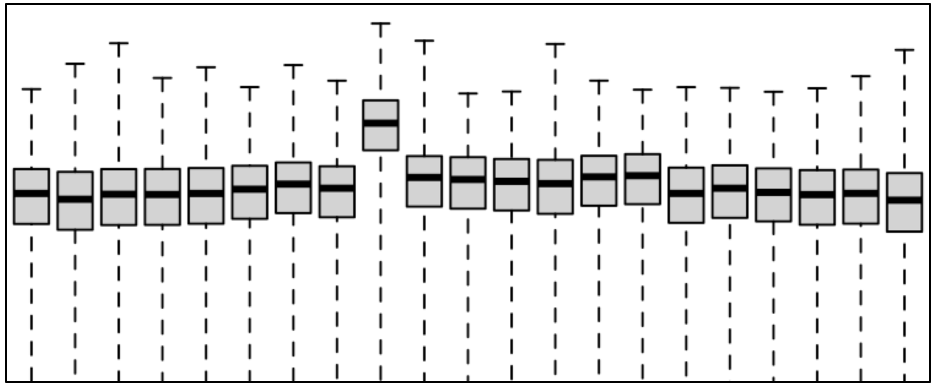
2. Transformation: Apply transformations to stabilize variance across samples.

3. Sample Distances & clustering: Calculate distances between samples (e.g., Euclidean distance)

4. Dimension reduction & clustering: see transcriptome-wide effects and sample relationships

Exploratory analysis – pre-filtering





Remove low-expressed genes: improve visualisations and save memory

Identify outliers: using Cook's distance

Exploratory analysis - transformation

Choose Distribution model that best fits the data

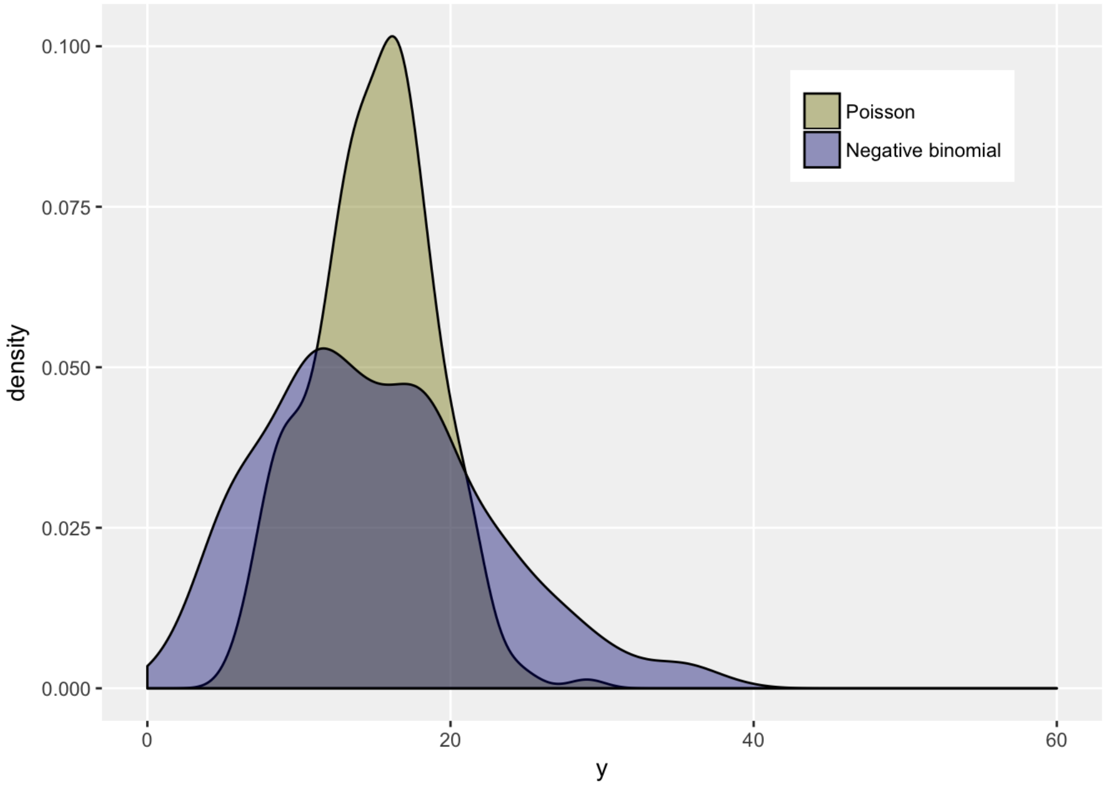
- if data fits negative binomial distribution, use **DESeq2** or **EdgeR**
- If data fits something else (e.g., *Poisson*), use **limma**

Exploratory analysis - transformation

RNAseq counts usually fit **Poisson** or **Negative Binomial** distribution:

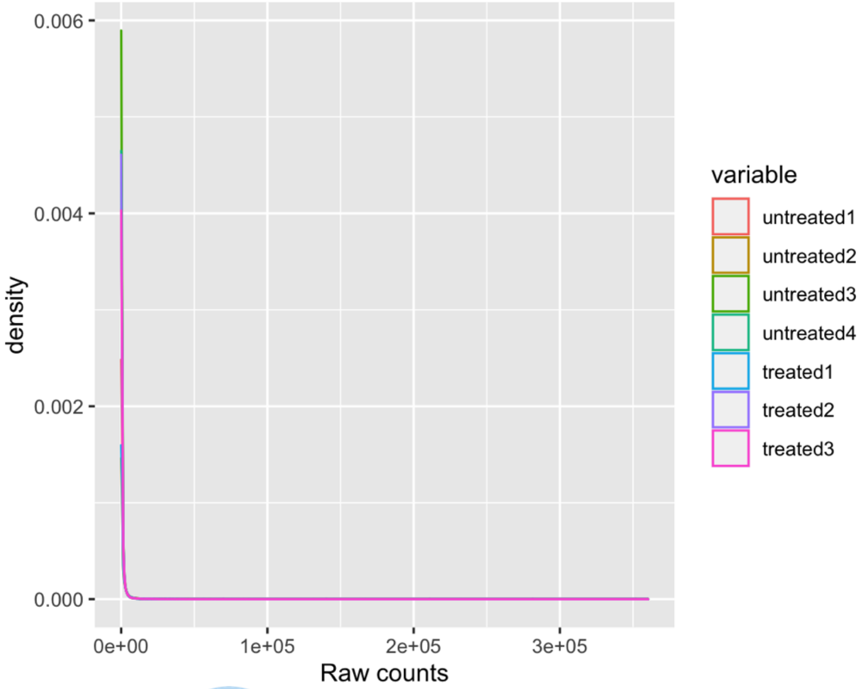
- Poisson distribution assumes *mean == variance* → count distributions are overdispersed

- Negative binomial distribution accounts for overdispersion

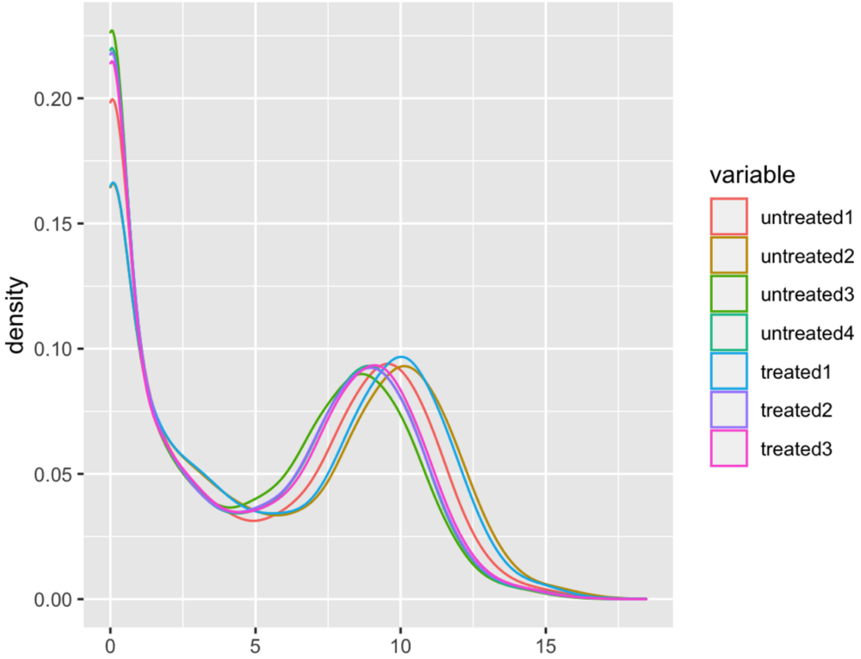


Exploratory analysis - transformation

raw counts is **NOT IDEAL** for clustering and visualisation



Transform to make it **NICER** to look at



log2(norm counts +1)

Exploratory analysis - transformation

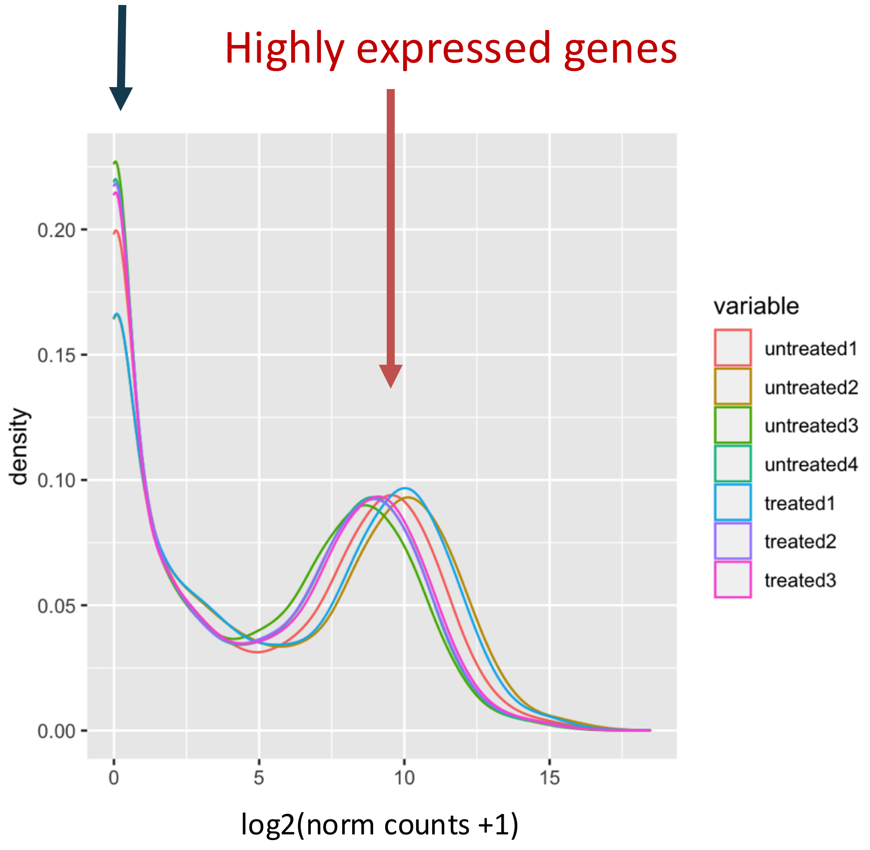
Genes with large mean counts **distort** sample relationship in low dimensional space

Data transformation **equalizes** the contribution to variance between high and low-expressed genes:

$$\underbrace{\texttt{log2(normalized counts}}_{\text{Log2-transformation}}\ \underbrace{\texttt{+ 1)}}_{\text{Add }\textbf{pseudocount}\text{ because }\texttt{log2(0) = Inf}}$$

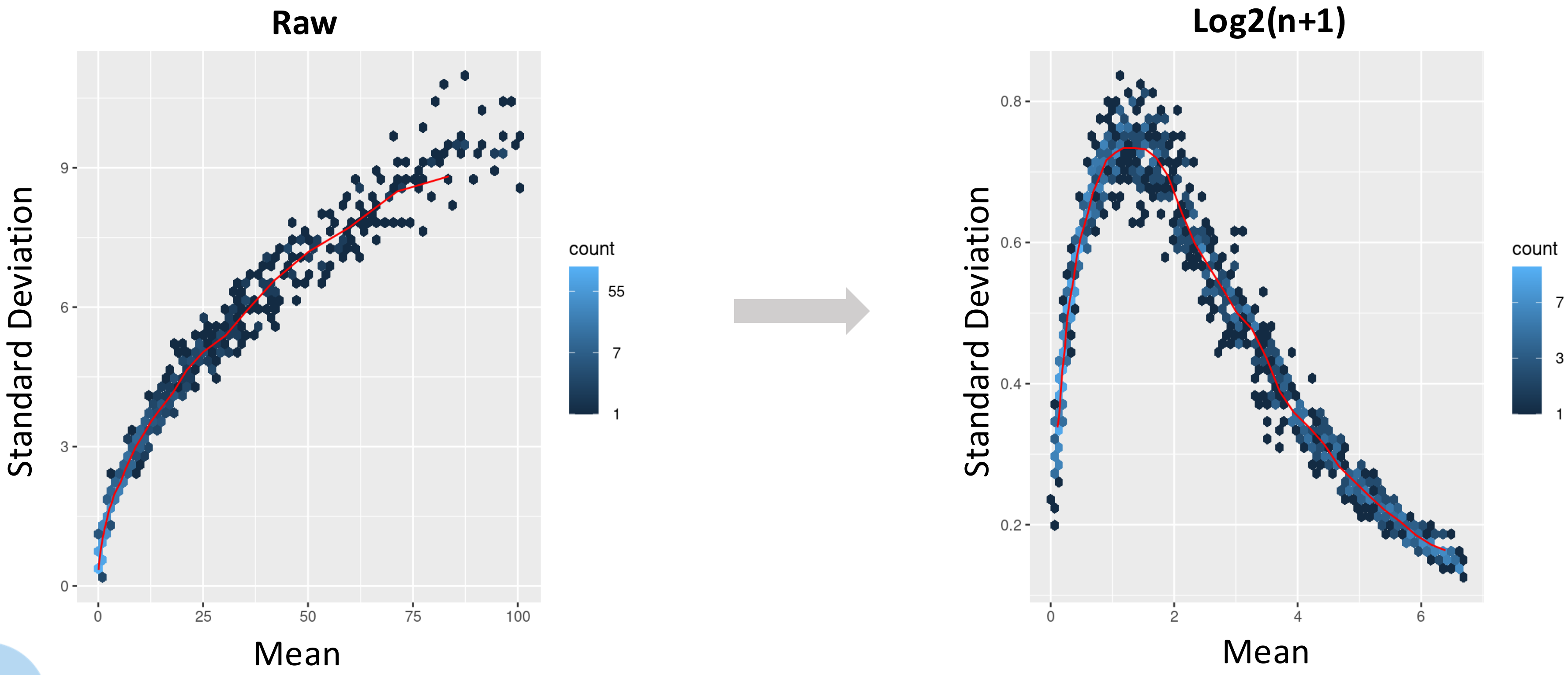
Lowly expressed genes

Highly expressed genes



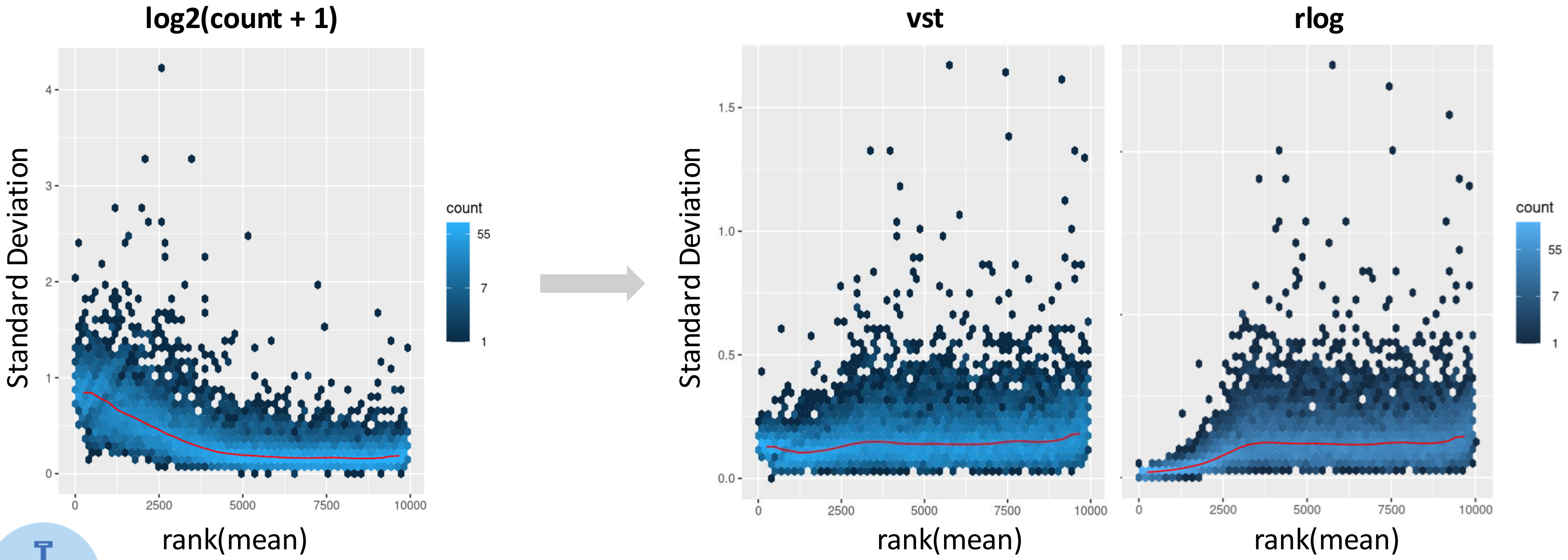
Exploratory analysis - transformation

$\log_2$(norm.counts + 1) fixes the issue of the genes with high expression (and variance), BUT introduces noise (variance) for lowly expressed genes.



Exploratory analysis - transformation

- Regularized logarithm (**rlog**) and variance stabilizing transformation (**vst**) remove the dependence of the variance on the expression mean
- For genes with low counts, values are shrunken towards the gene average across all samples



log2(count + 1)

vst

rlog

Exploratory analysis - transformation

⚠

DO NOT use **TRANSFORMED DATA**

(log2, vst, rlog)

for Differential Expression Analysis

Exploratory analysis - Dimension reduction

Use to visualize transcriptome-wide effects and sample relationships

Sources of variation

Desired:	Variance of variable of interest
Undesired:	Confounding variable variances, Technical & Batch effects

Dimension reduction methods

PCA, MDS, t-SNE, UMAP

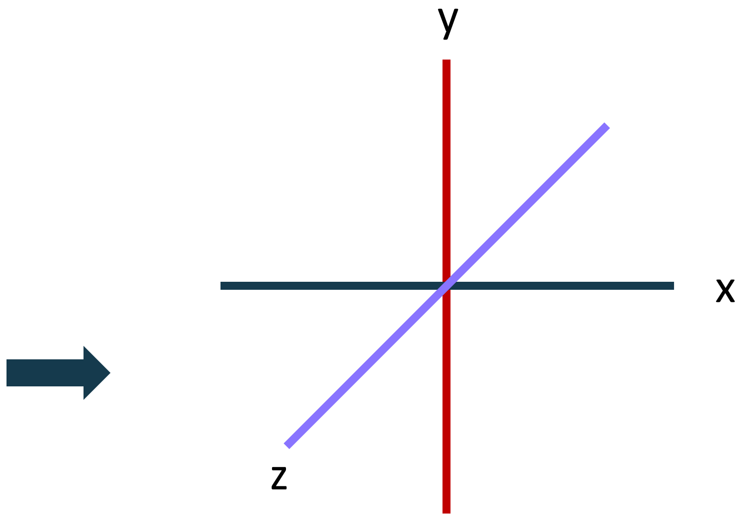
👍 **Use transformed data here**

Exploratory analysis – PCA

Principal Component Analysis

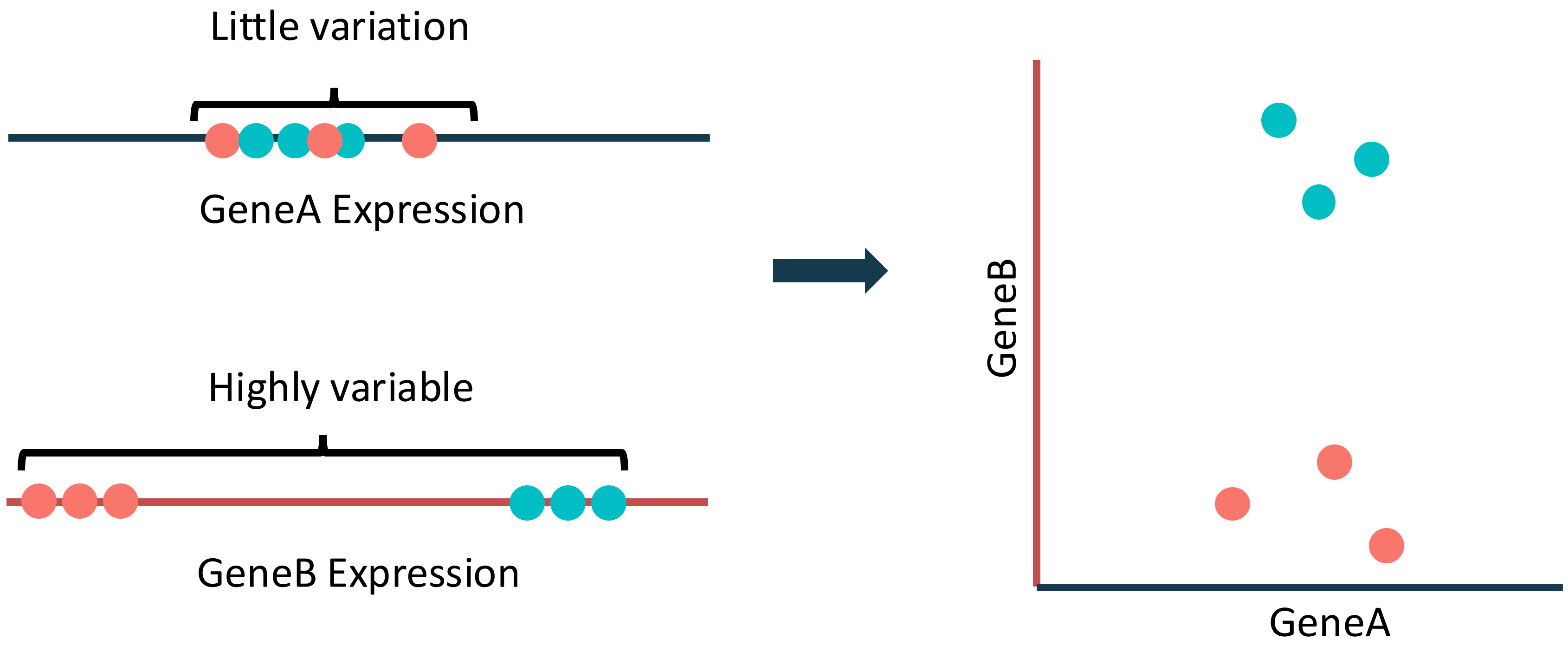
- Visualize variation dataset of high dimensionality
- Number of genes equals number of dimensions (d)
- We can only interpret 2 or 3 dimensions

Gene	Norm Sample A	Norm Sample B	n = 100
EF2A	1145.39	1176.62	…
ACBD1	16.92	16.88	…
d = 20000	…	…	…



Exploratory analysis – PCA

Variation of genes is collapsed into Principal Components (PCs)



Little variation

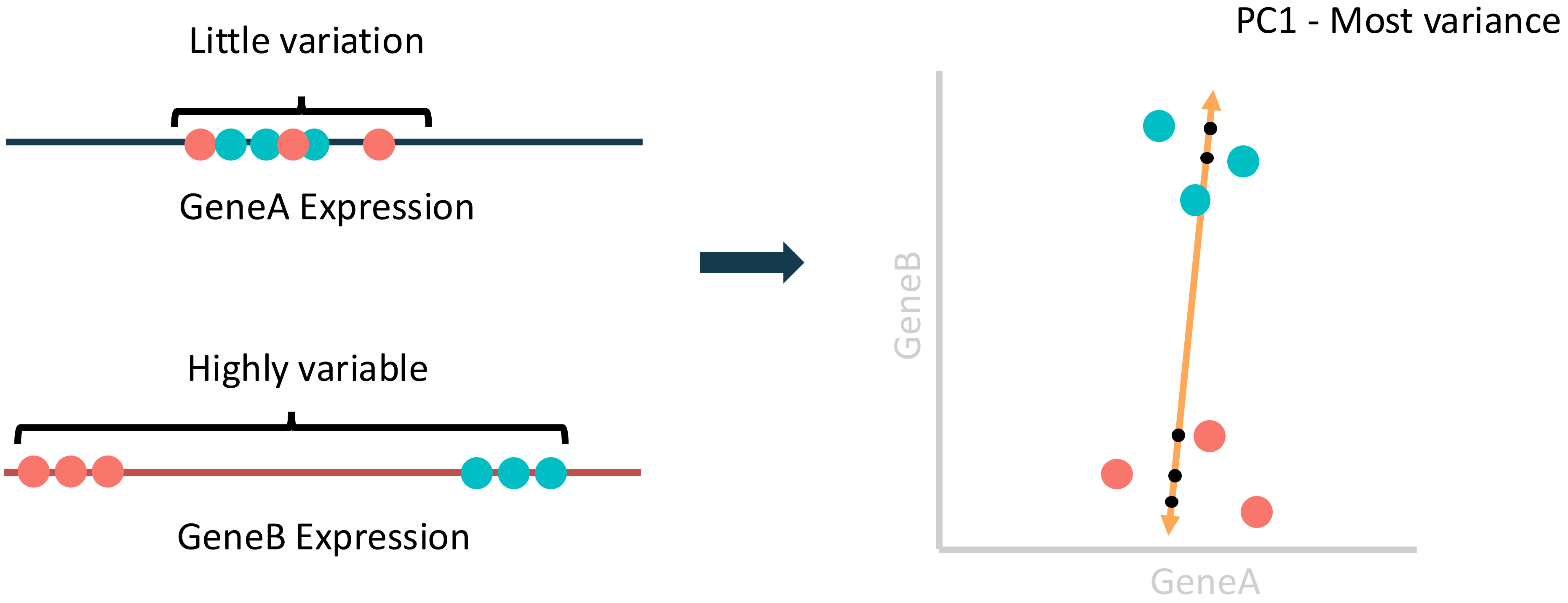
GeneA Expression

Highly variable

GeneB Expression

Exploratory analysis – PCA

Variation of genes is collapsed into Principal Components (PCs)



Little variation

GeneA Expression

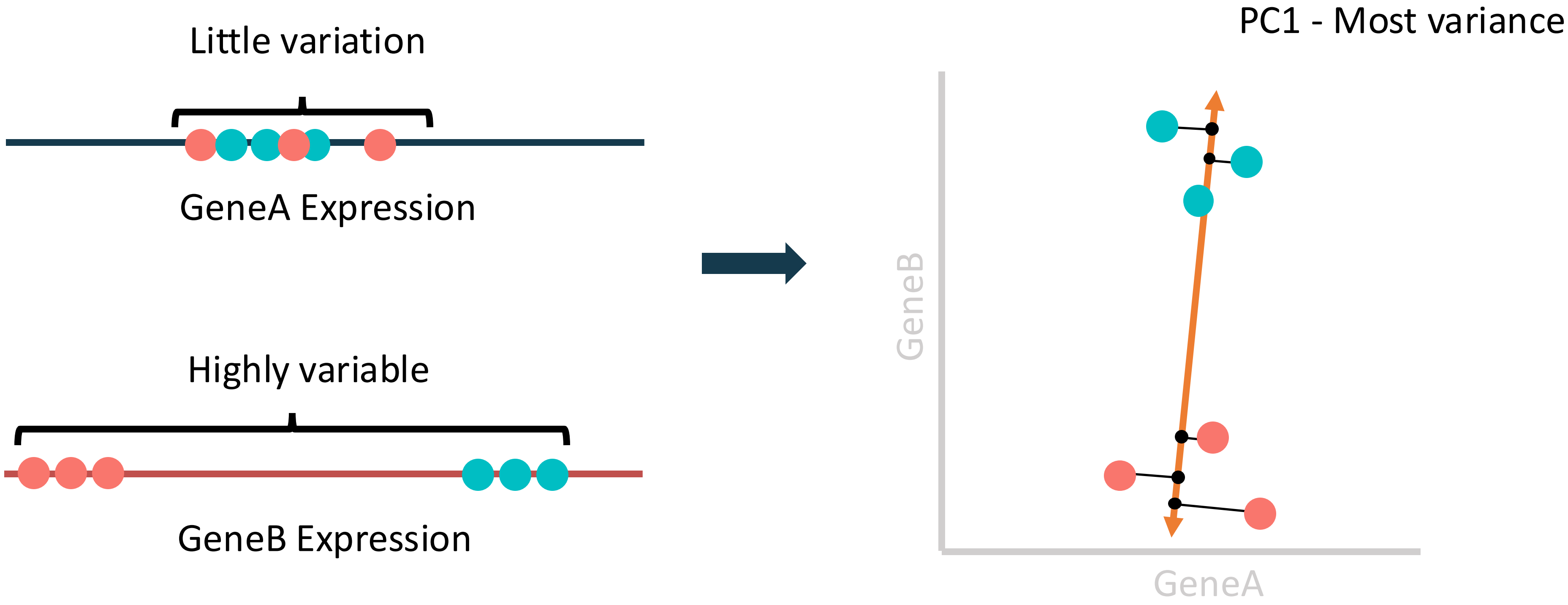
Highly variable

GeneB Expression

PC1 - Most variance

Exploratory analysis – PCA

Variation of genes is collapsed into Principal Components (PCs)



Little variation

GeneA Expression

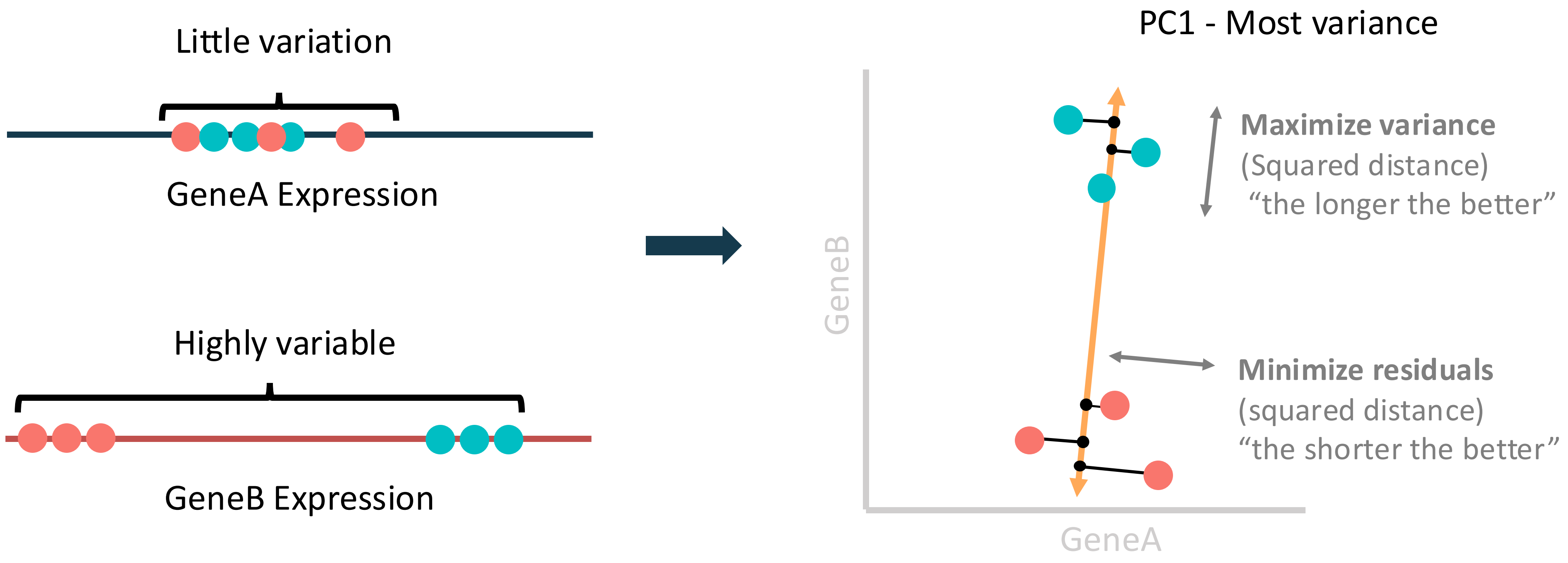
Highly variable

GeneB Expression

PC1 - Most variance

Exploratory analysis – PCA

Variation of genes is collapsed into Principal Components (PCs)



Little variation

GeneA Expression

Highly variable

GeneB Expression

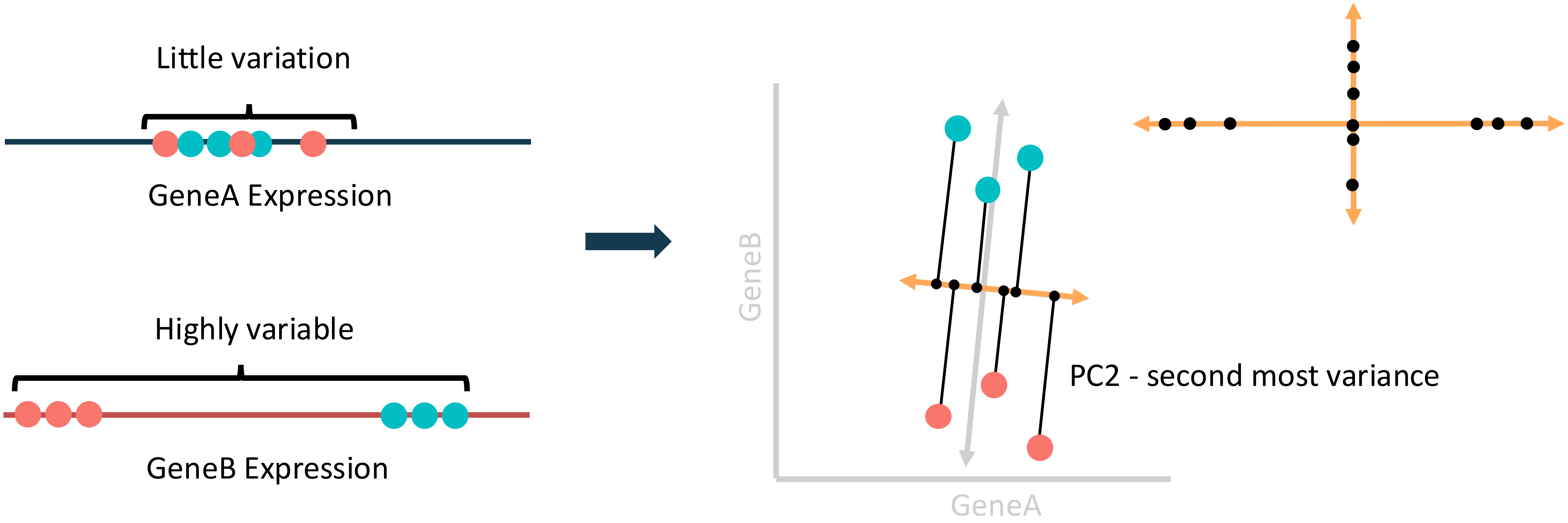
PC1 - Most variance

Maximize variance
(Squared distance)
"the longer the better"

Minimize residuals
(squared distance)
"the shorter the better"

Exploratory analysis – PCA

Variation of genes is collapsed into Principal Components (PCs)



Little variation

GeneA Expression

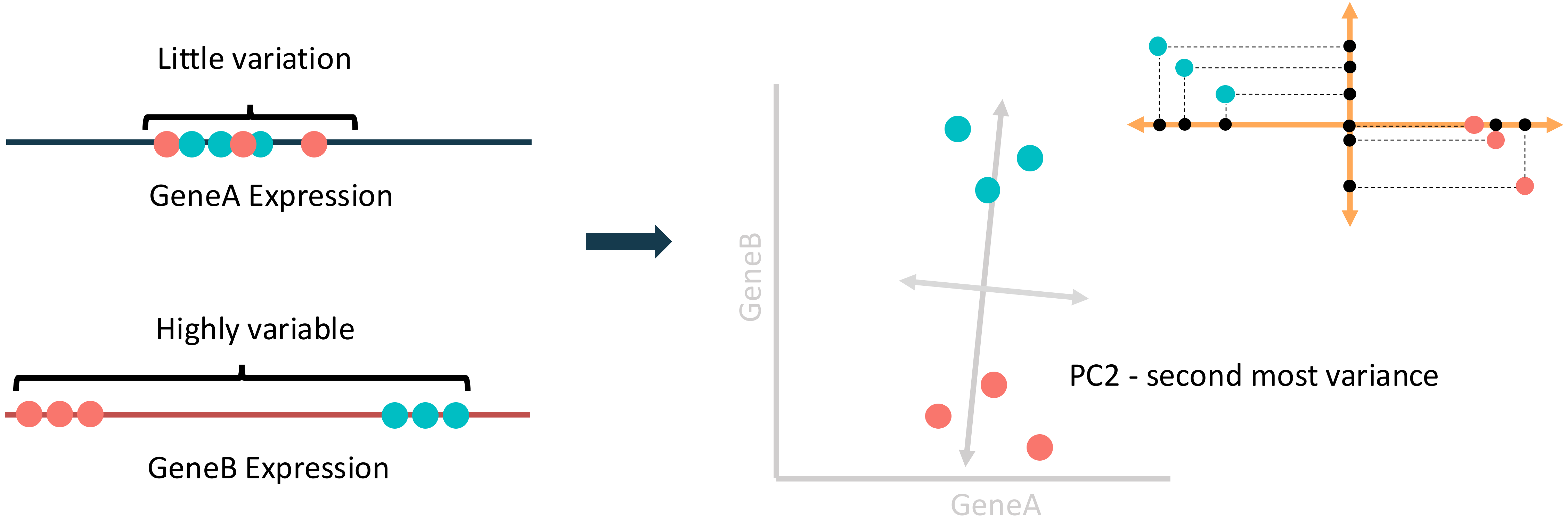
Highly variable

GeneB Expression

PC2 - second most variance

Exploratory analysis – PCA

Variation of genes is collapsed into Principal Components (PCs)



Little variation

GeneA Expression

Highly variable

GeneB Expression

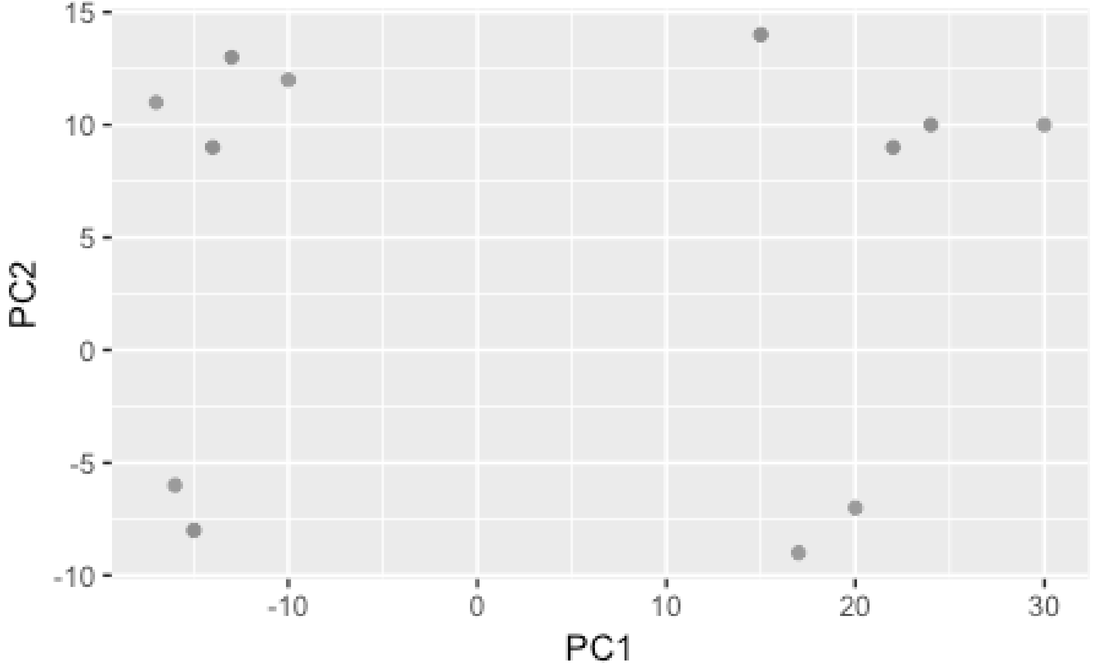
PC2 - second most variance

Exploratory analysis – PCA

Samples with **similar** gene expression related to Principal Components will be **together**

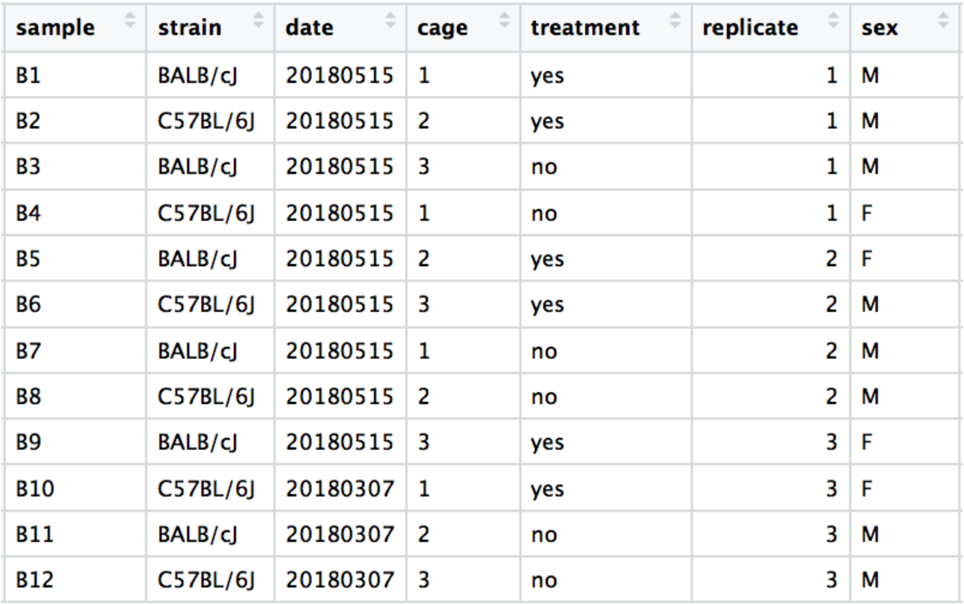
First Principal Components contain **most** variation: Usually PC1-PC4 are used

We need **metadata** to understand the source of variation, both biological and technical.

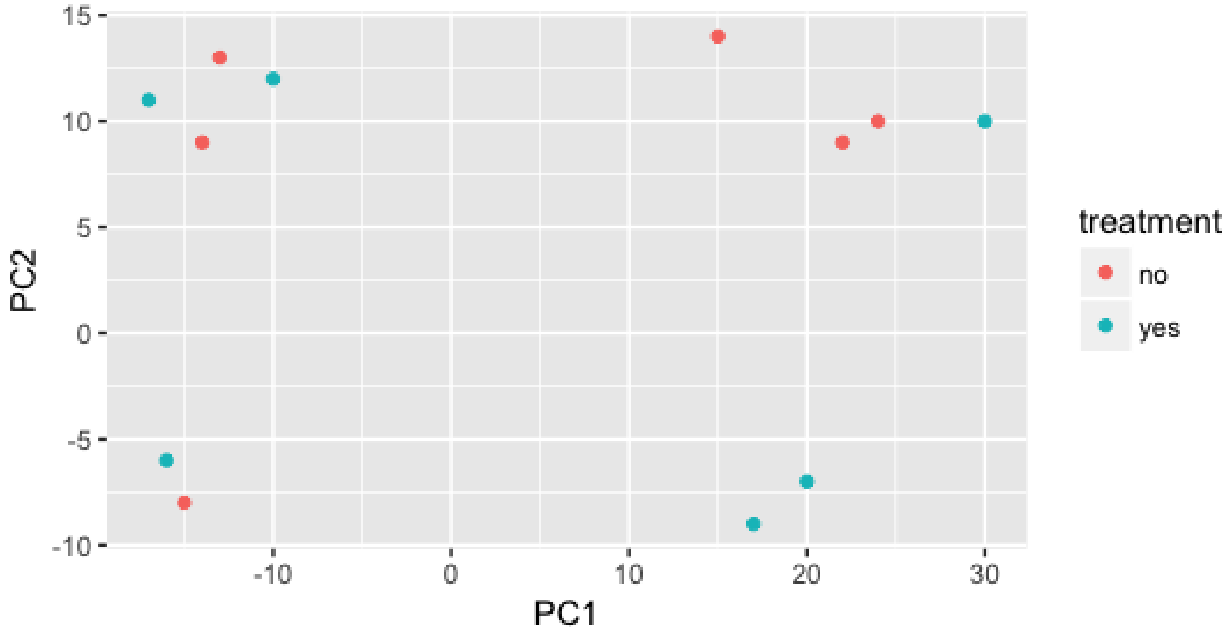


Exploratory analysis – PCA

metadata: colData()

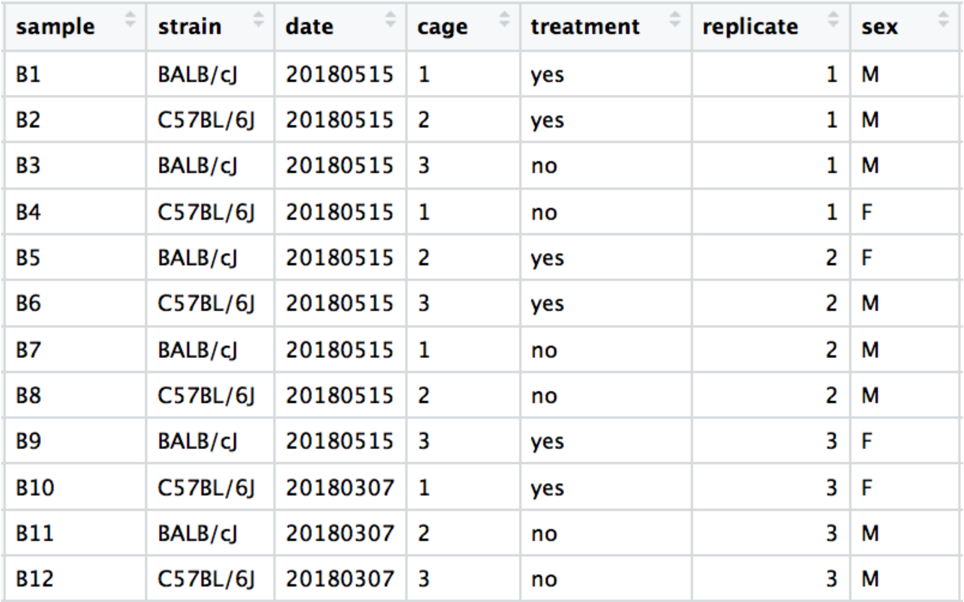


sample	strain	date	cage	treatment	replicate	sex
B1	BALB/cJ	20180515	1	yes	1	M
B2	C57BL/6J	20180515	2	yes	1	M
B3	BALB/cJ	20180515	3	no	1	M
B4	C57BL/6J	20180515	1	no	1	F
B5	BALB/cJ	20180515	2	yes	2	F
B6	C57BL/6J	20180515	3	yes	2	M
B7	BALB/cJ	20180515	1	no	2	M
B8	C57BL/6J	20180515	2	no	2	M
B9	BALB/cJ	20180515	3	yes	3	F
B10	C57BL/6J	20180307	1	yes	3	F
B11	BALB/cJ	20180307	2	no	3	M
B12	C57BL/6J	20180307	3	no	3	M

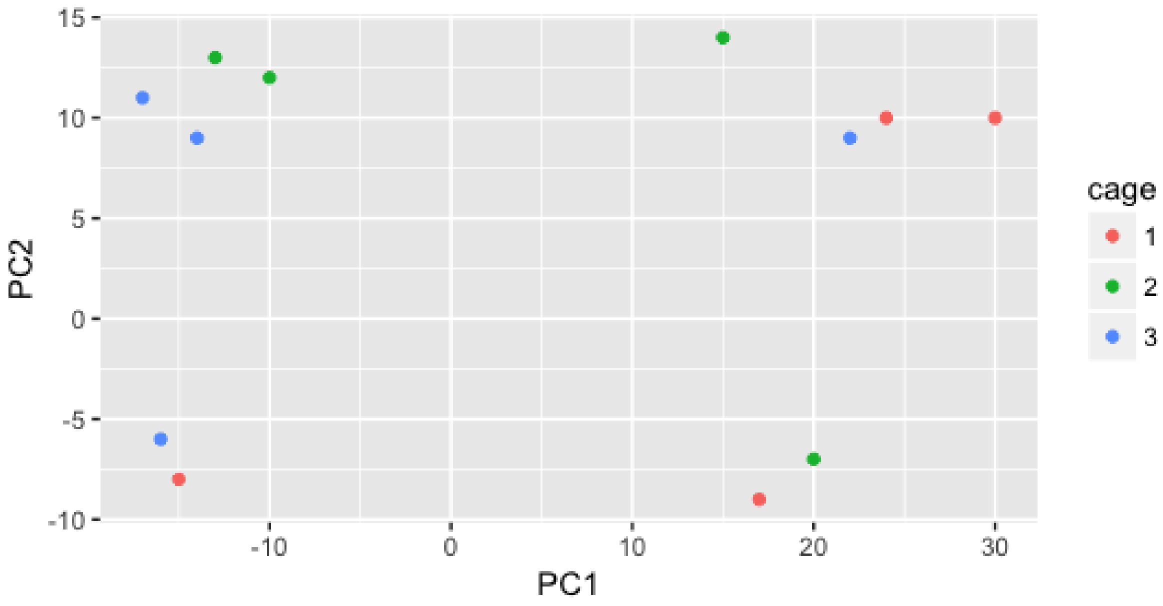


Exploratory analysis – PCA

metadata: colData()

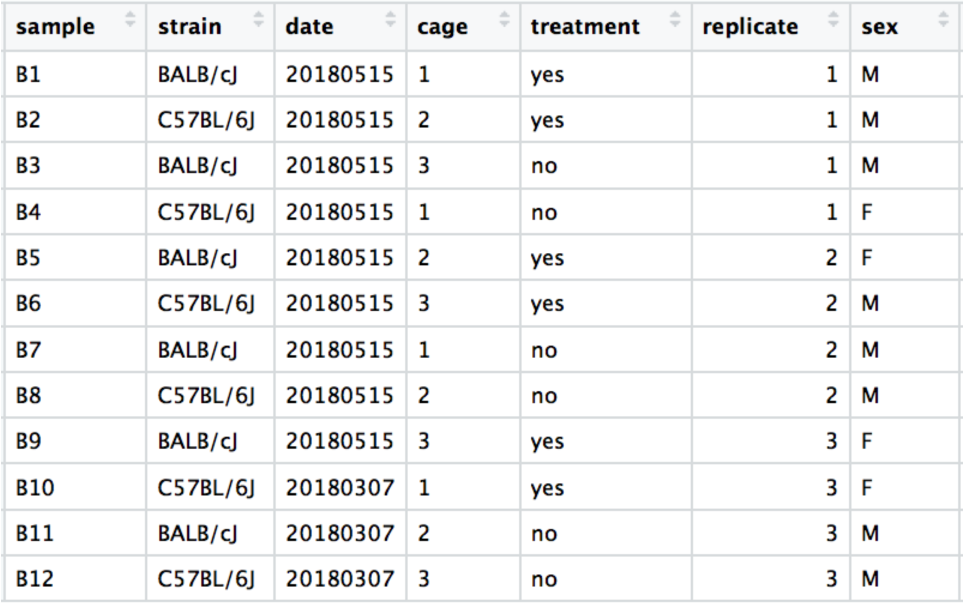


sample	strain	date	cage	treatment	replicate	sex
B1	BALB/cJ	20180515	1	yes	1	M
B2	C57BL/6J	20180515	2	yes	1	M
B3	BALB/cJ	20180515	3	no	1	M
B4	C57BL/6J	20180515	1	no	1	F
B5	BALB/cJ	20180515	2	yes	2	F
B6	C57BL/6J	20180515	3	yes	2	M
B7	BALB/cJ	20180515	1	no	2	M
B8	C57BL/6J	20180515	2	no	2	M
B9	BALB/cJ	20180515	3	yes	3	F
B10	C57BL/6J	20180307	1	yes	3	F
B11	BALB/cJ	20180307	2	no	3	M
B12	C57BL/6J	20180307	3	no	3	M

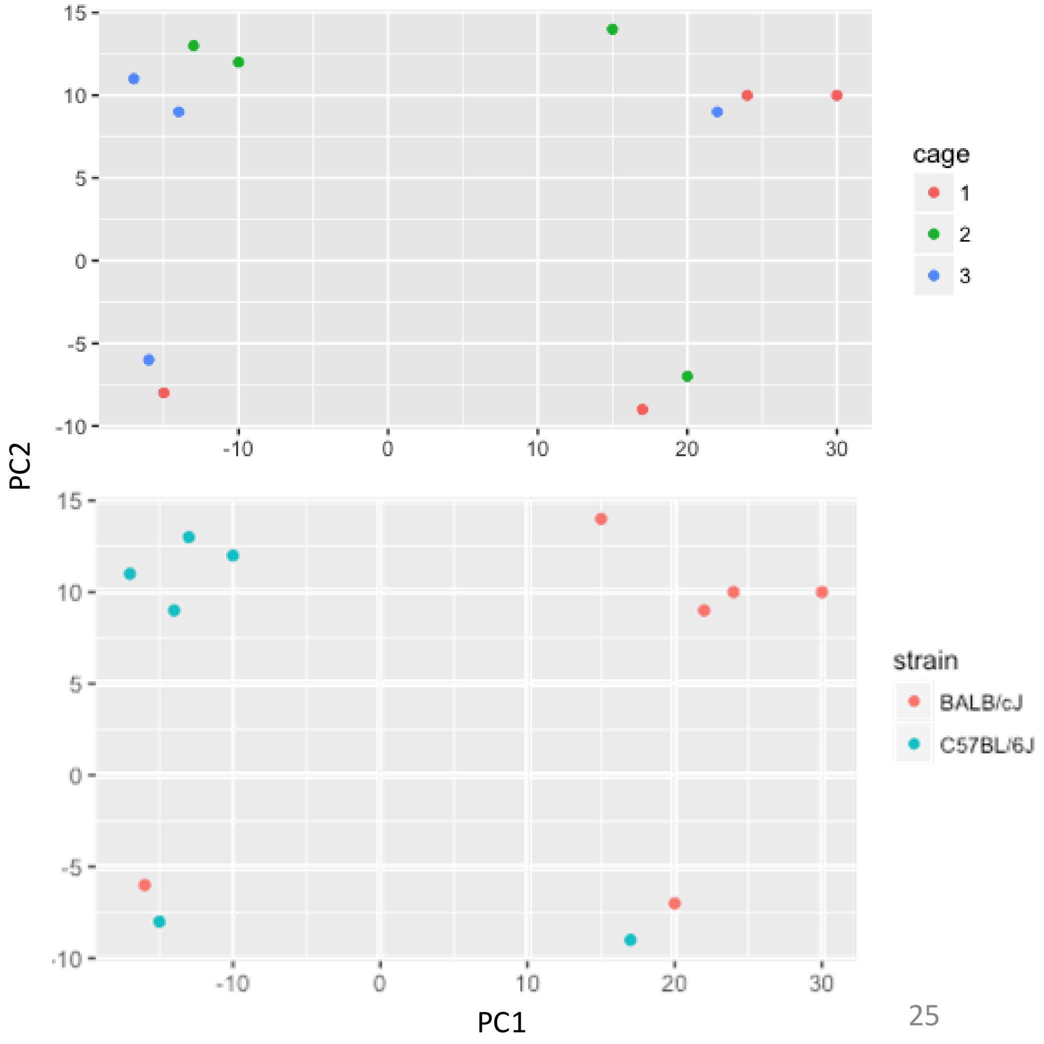


Exploratory analysis – PCA

metadata: colData()

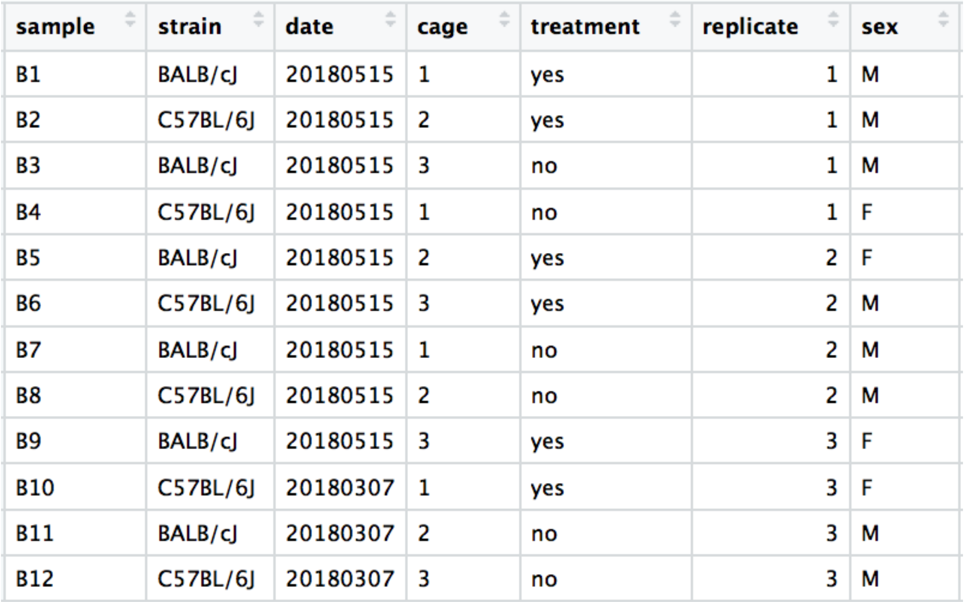


sample	strain	date	cage	treatment	replicate	sex
B1	BALB/cJ	20180515	1	yes	1	M
B2	C57BL/6J	20180515	2	yes	1	M
B3	BALB/cJ	20180515	3	no	1	M
B4	C57BL/6J	20180515	1	no	1	F
B5	BALB/cJ	20180515	2	yes	2	F
B6	C57BL/6J	20180515	3	yes	2	M
B7	BALB/cJ	20180515	1	no	2	M
B8	C57BL/6J	20180515	2	no	2	M
B9	BALB/cJ	20180515	3	yes	3	F
B10	C57BL/6J	20180307	1	yes	3	F
B11	BALB/cJ	20180307	2	no	3	M
B12	C57BL/6J	20180307	3	no	3	M

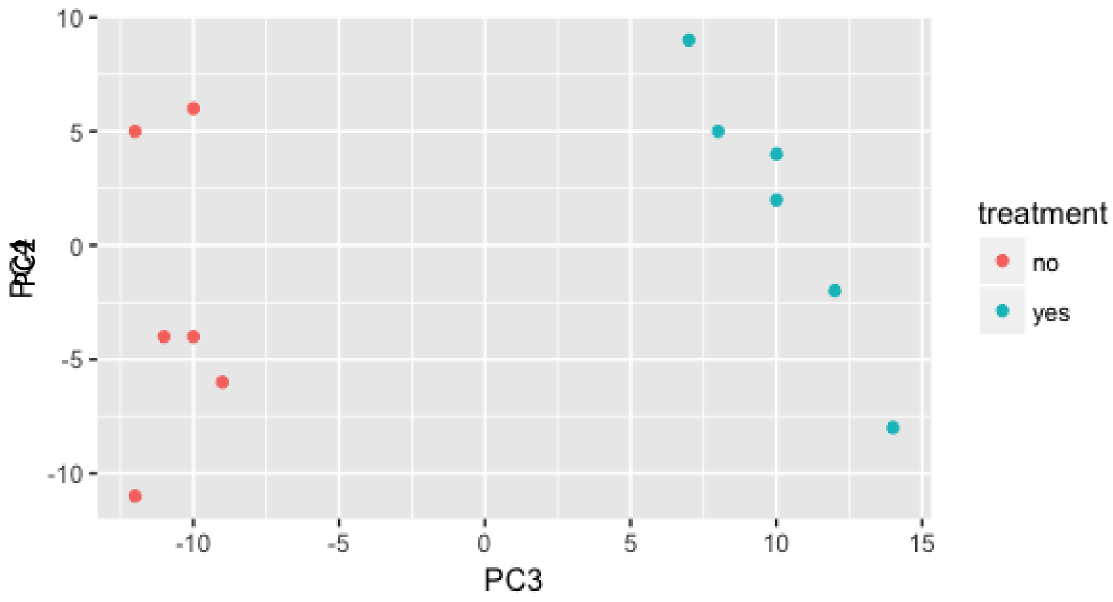


Exploratory analysis – PCA

metadata: colData()



sample	strain	date	cage	treatment	replicate	sex
B1	BALB/cJ	20180515	1	yes	1	M
B2	C57BL/6J	20180515	2	yes	1	M
B3	BALB/cJ	20180515	3	no	1	M
B4	C57BL/6J	20180515	1	no	1	F
B5	BALB/cJ	20180515	2	yes	2	F
B6	C57BL/6J	20180515	3	yes	2	M
B7	BALB/cJ	20180515	1	no	2	M
B8	C57BL/6J	20180515	2	no	2	M
B9	BALB/cJ	20180515	3	yes	3	F
B10	C57BL/6J	20180307	1	yes	3	F
B11	BALB/cJ	20180307	2	no	3	M
B12	C57BL/6J	20180307	3	no	3	M



Exploratory analysis - clustering

👍 Use transformed counts

1. Create a pairwise matrix for **samples**:

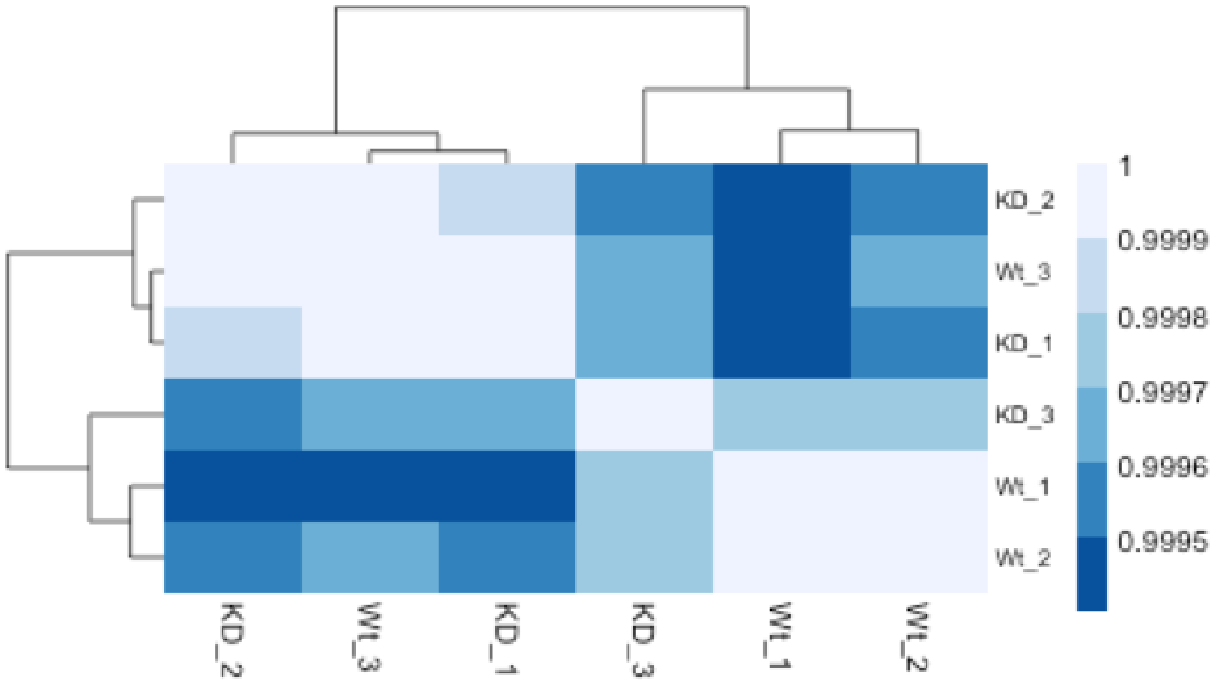
- Euclidean Distance
- Spearman Correlation

2. Apply a **clustering** approach to the distance matrix:

- hclust
- kmeans

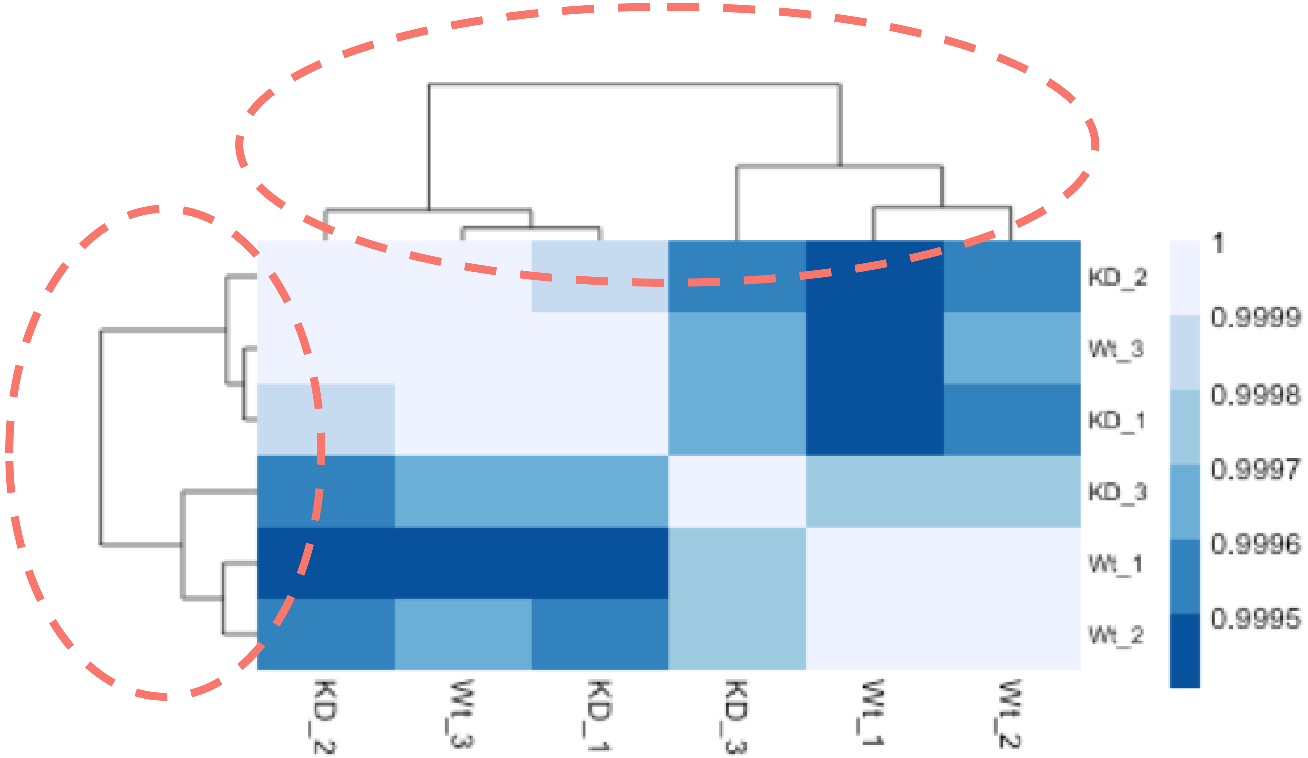
Exploratory analysis – Clustering

3. Visualise as heatmap + dendrogram



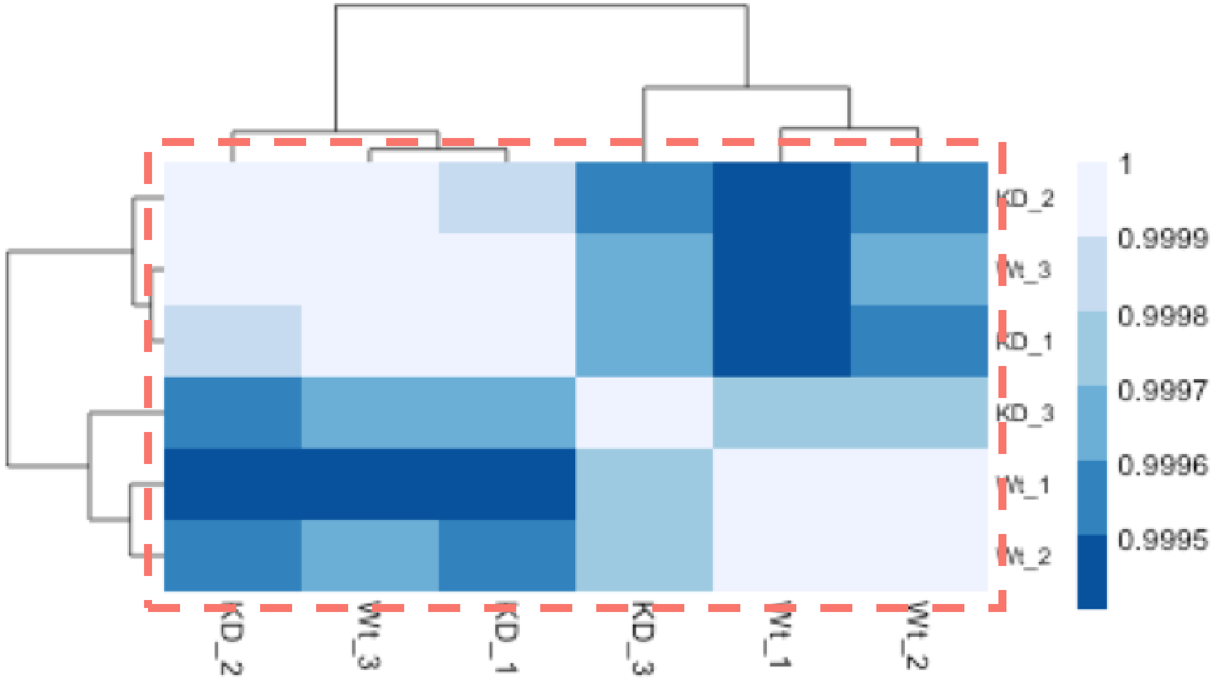
Exploratory analysis – Clustering

Dendrogram summarizes which samples are more similar



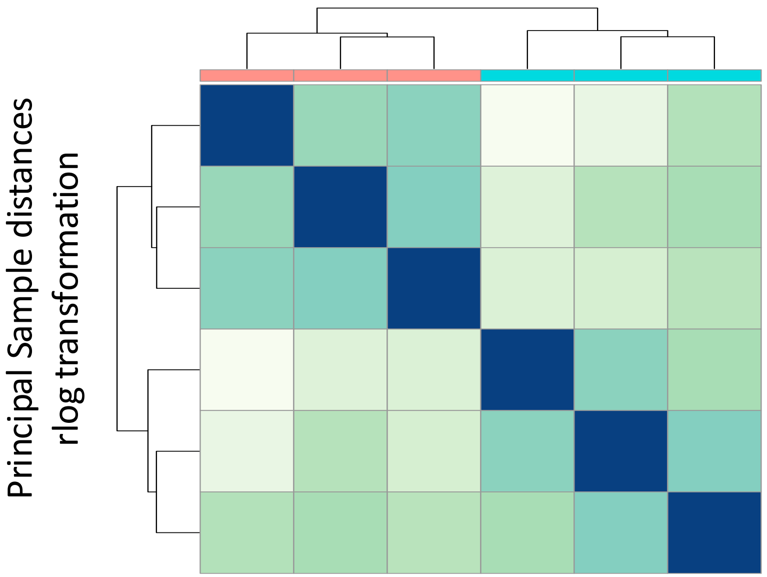
Exploratory analysis – Clustering

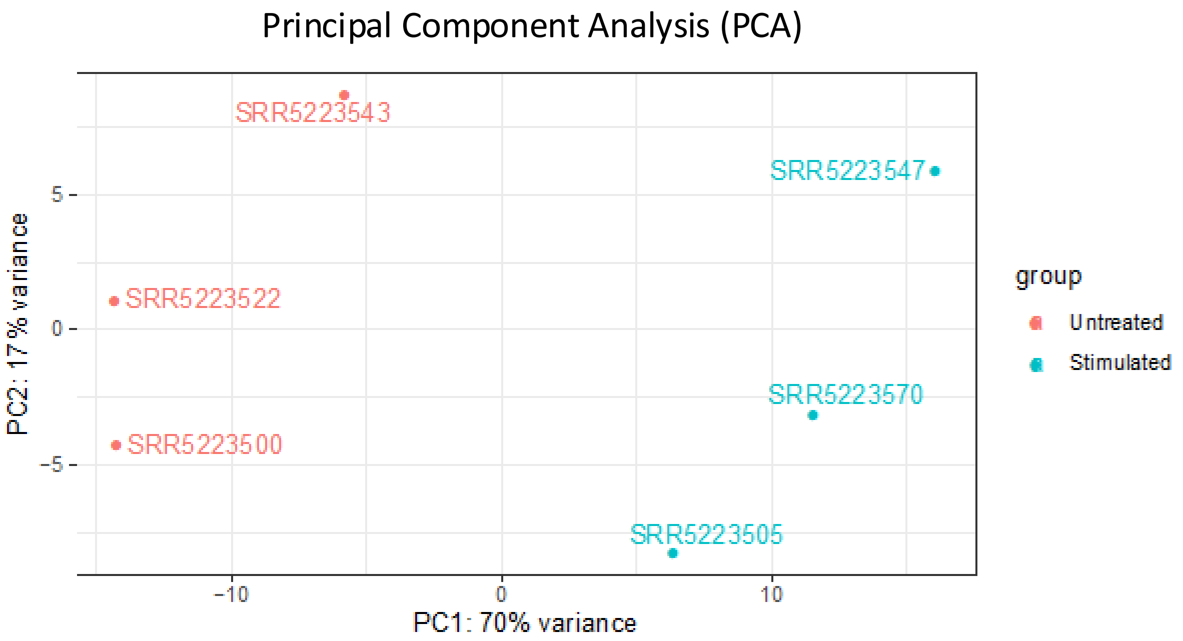
Heatmap shows pairwise distance or correlation as a color



Exploratory analysis – Summary

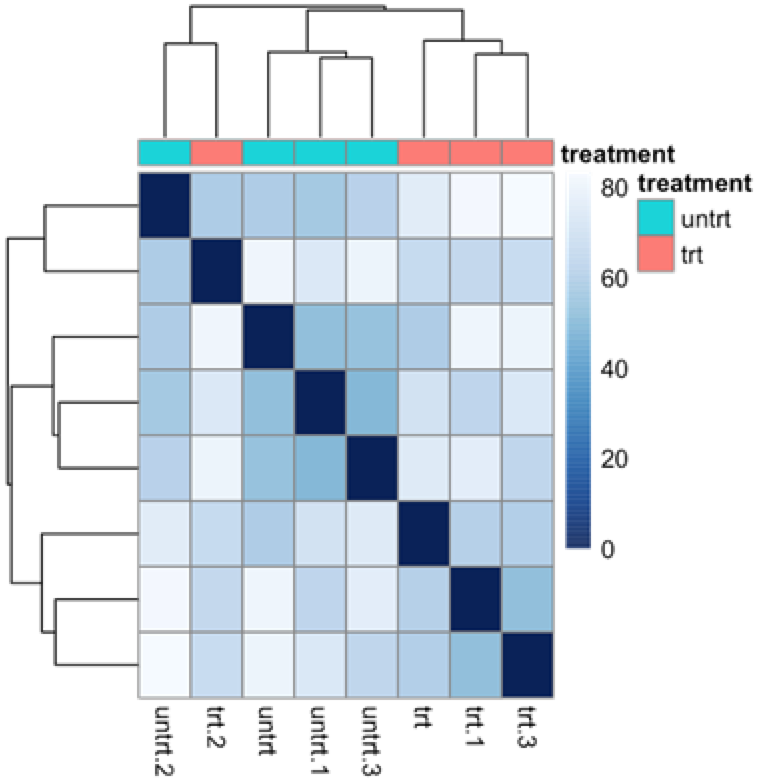
Dimension reduction, **clustering** and **heatmaps** of **transformed** counts help data exploration before further analysis: use these tools together to understand your data.

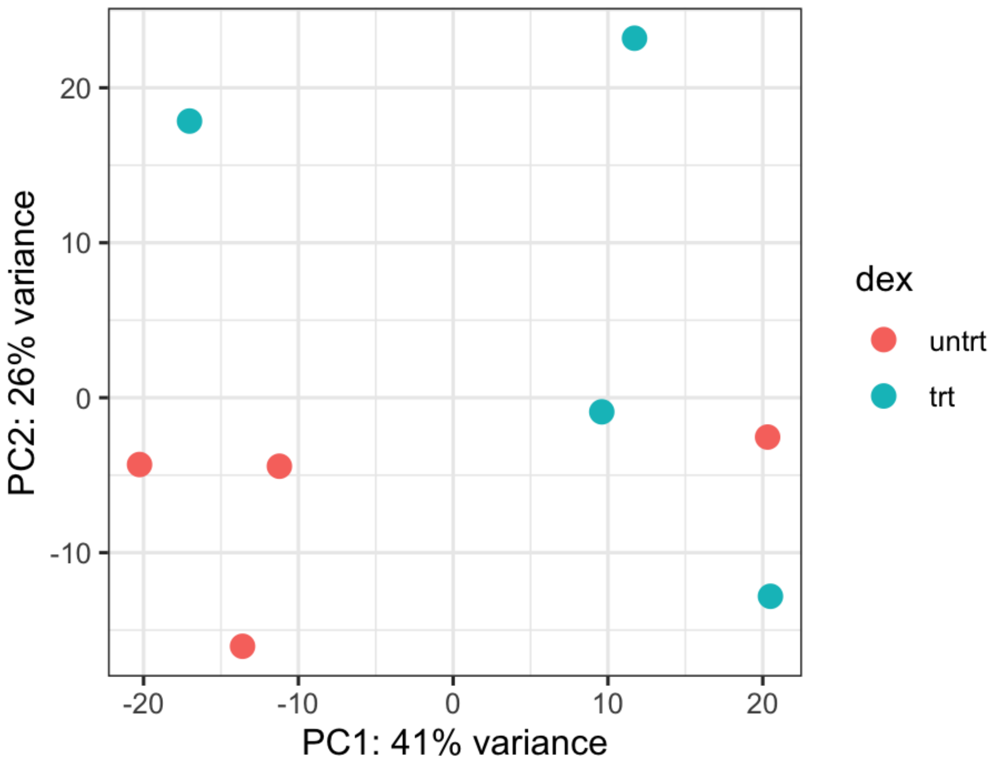




Exploratory analysis – Summary

An example of when something is not quite right…





Exploratory analysis

Let’s Do Some Exploratory Analysis:

Notebook:

- *06_exploratory_analysis.Rmd*