Differential expression analysis

RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

datacamp

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home | Install | Help

Home » Bioconductor 3.6 » Software Packages » DESeq2

DESeq2

platforms all | downloads top 5% | posts 353 / 1 / 3 / 60 | in Bioc 5 years
build warnings

DOI: 10.18129/B9.bioc.DESeq2

Differential gene expression analysis based on the negative binomial distribution

Bioconductor version: Release (3.6)

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.

Author: Michael Love, Simon Anders, Wolfgang Huber

Maintainer: Michael Love <michaelisaiahlove at gmail.com>

Citation (from within R, enter `citation("DESeq2")`):

Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq

Differential expression analysis: DESeq2 vignette

```
vignette(DESeq2)
```

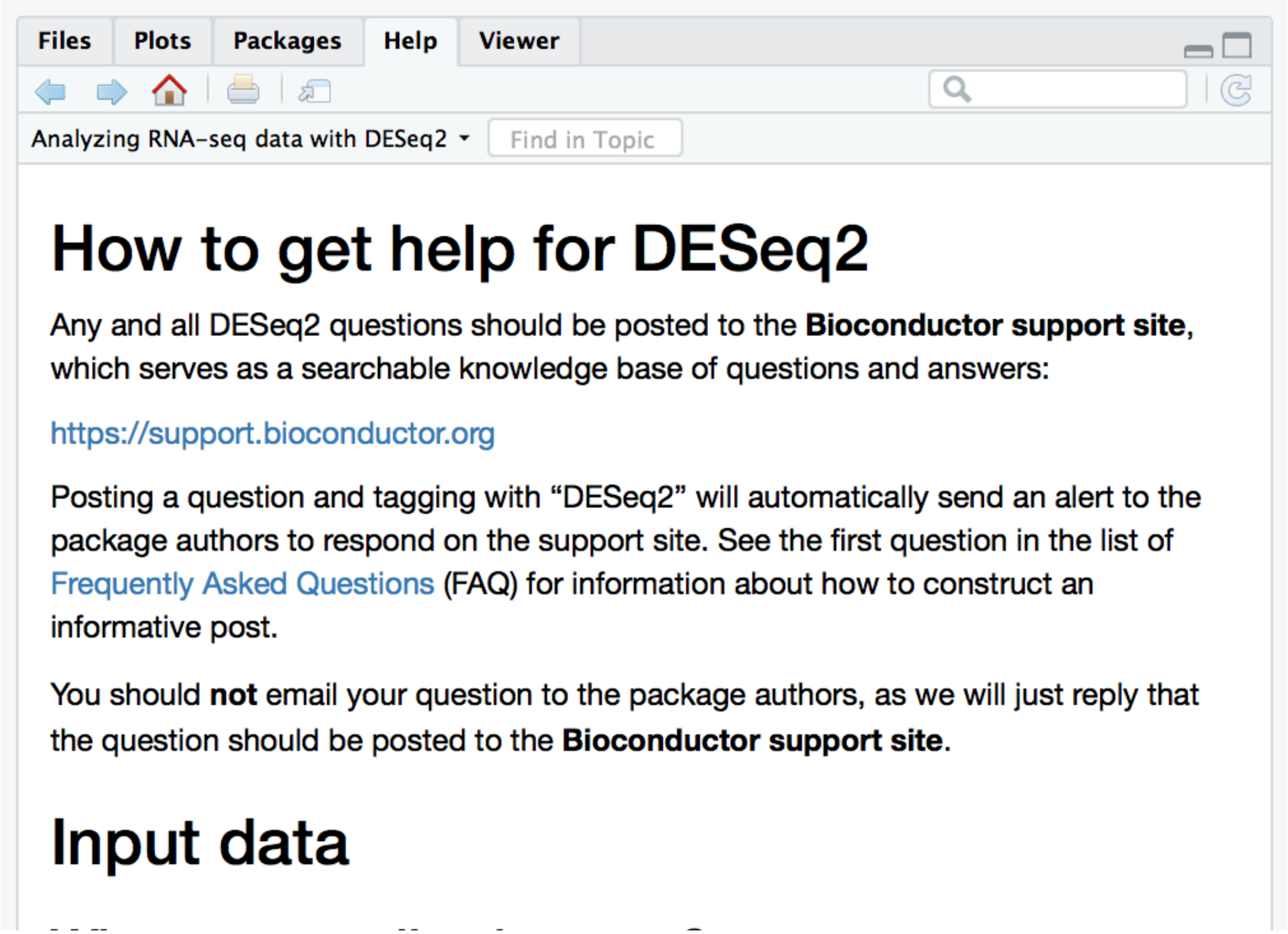
Analyzing RNA-seq data with DESeq2

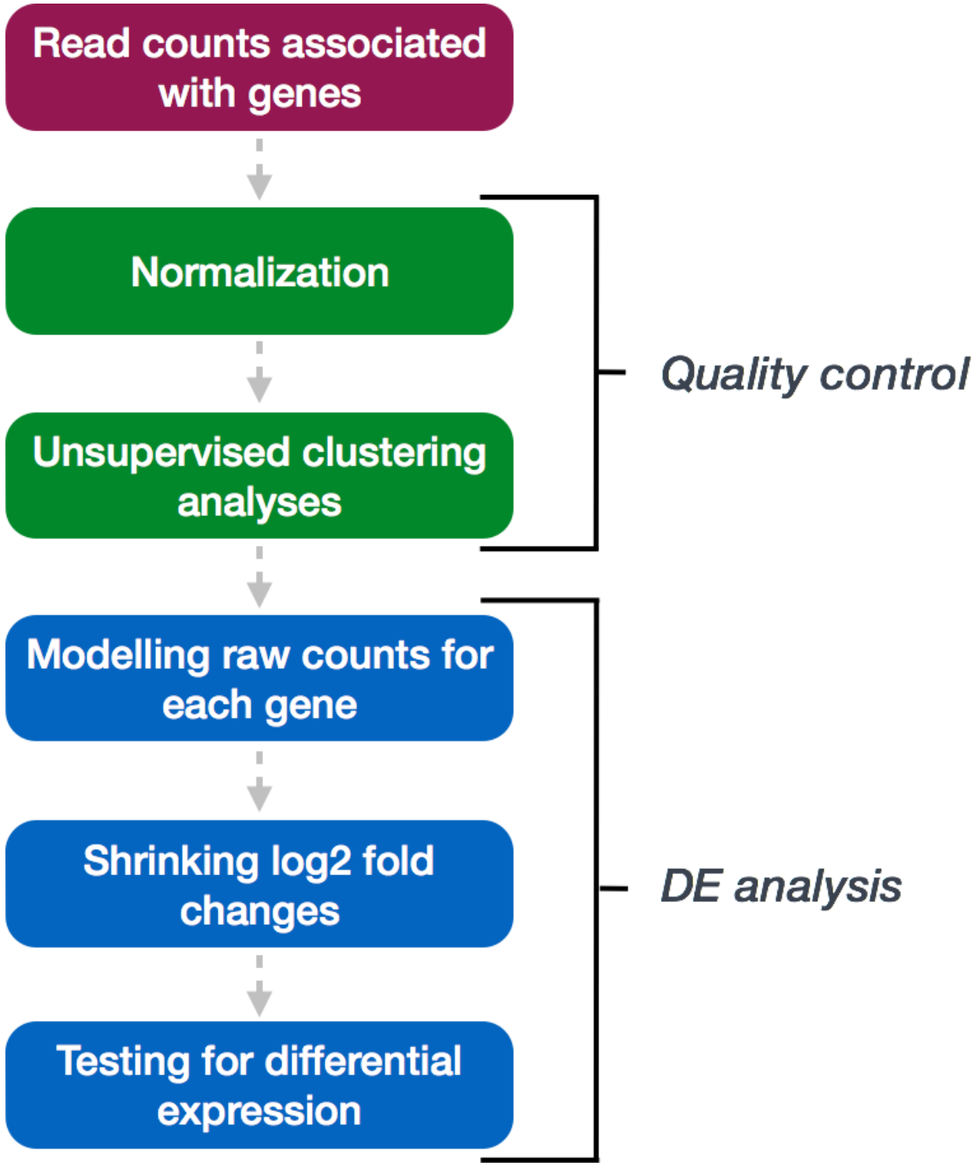
Michael I. Love, Simon Anders, and Wolfgang Huber

11 November 2017

Abstract

A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions This vignette explains the use of the package and demonstrates typical workflows. An RNA-seq workflow on the Bioconductor website covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files. DESeq2 package version: 1.18.1


Files
Plots
Packages
Help
Viewer
Analyzing RNA-seq data with DESeq2
Find in Topic
How to get help for DESeq2
Any and all DESeq2 questions should be posted to the Bioconductor support site, which serves as a searchable knowledge base of questions and answers:
https://support.bioconductor.org
Posting a question and tagging with "DESeq2" will automatically send an alert to the package authors to respond on the support site. See the first question in the list of Frequently Asked Questions (FAQ) for information about how to construct an informative post.
You should not email your question to the package authors, as we will just reply that the question should be posted to the Bioconductor support site.
Input data


Read counts associated with genes
Normalization
Unsupervised clustering analyses
Quality control
Modelling raw counts for each gene
Shrinking log2 fold changes
Testing for differential expression
DE analysis

Bringing in data for DESeq2

```
# Read in raw counts
wt_rawcounts <- read.csv("fibrosis_wt_rawcounts.csv")
View(wt_rawcounts)
```

	wt_normal1	wt_normal2	wt_normal3	wt_fibrosis1	wt_fibrosis2	wt_fibrosis3	wt_fibrosis4
ENSMUSG00000102693	0	0	0	0	0	0	0
ENSMUSG00000064842	0	0	0	0	0	0	0
ENSMUSG00000051951	3	1	1	42	52	16	35
ENSMUSG00000102851	0	0	0	0	0	0	0
ENSMUSG00000103377	0	0	0	0	0	0	0
ENSMUSG00000104017	0	0	0	0	0	0	0
ENSMUSG00000103025	0	0	0	1	0	0	0
ENSMUSG00000089699	0	0	0	0	0	0	0
ENSMUSG00000103201	0	0	0	0	0	0	0
ENSMUSG00000103147	0	0	0	0	1	1	1

Bringing in data for DESeq2: metadata

```
# Read in metadata
wt_metadata <- read.csv("fibrosis_wt_metadata_unordered.csv")
View(wt_metadata)
```

	genotype	condition
wt_normal3	wt	normal
wt_fibrosis3	wt	fibrosis
wt_normal1	wt	normal
wt_fibrosis2	wt	fibrosis
wt_normal2	wt	normal
wt_fibrosis4	wt	fibrosis
wt_fibrosis1	wt	fibrosis

Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R

Organizing the data for DESeq2

RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

datacamp

Bringing in data for DESeq2: sample order

Metadata

	genotype	condition
wt_normal3	wt	normal
wt_fibrosis3	wt	fibrosis
wt_normal1	wt	normal
wt_fibrosis2	wt	fibrosis
wt_normal2	wt	normal
wt_fibrosis4	wt	fibrosis
wt_fibrosis1	wt	fibrosis

Raw counts

	wt_normal1	wt_normal2	wt_normal3	wt_fibrosis1	wt_fibrosis2	wt_fibrosis3	wt_fibrosis4
ENSMUSG00000102693	0	0	0	0	0	0	0
ENSMUSG00000064842	0	0	0	0	0	0	0
ENSMUSG00000051951	3	1	1	42	52	16	35
ENSMUSG00000102851	0	0	0	0	0	0	0
ENSMUSG00000103377	0	0	0	0	0	0	0
ENSMUSG00000104017	0	0	0	0	0	0	0
ENSMUSG00000103025	0	0	0	1	0	0	0

Bringing in data for DESeq2: sample order

```
rownames(wt_metadata)
```

```
 [1] "wt_normal3"  "smoc2_fibrosis2" "wt_fibrosis3"  "smoc2_fibrosis3" "smoc2_normal3"   "wt_normal1"
 [7] "smoc2_normal4"  "wt_fibrosis2"  "wt_normal2"  "smoc2_normal1"  "smoc2_fibrosis1" "smoc2_fibrosis4"
[13] "wt_fibrosis4"  "wt_fibrosis1"
```

```
colnames(wt_rawcounts)
```

```
 [1] "wt_normal1"  "wt_normal2"  "wt_normal3"  "wt_fibrosis1"   "wt_fibrosis2"    "wt_fibrosis3"
 [7] "wt_fibrosis4"  "smoc2_normal1"  "smoc2_normal3"  "smoc2_normal4"   "smoc2_fibrosis1" "smoc2_fibrosis2"
[13] "smoc2_fibrosis3" "smoc2_fibrosis4"
```

Bringing in data for DESeq2: sample order

```
all(rownames(wt_metadata) == colnames(wt_rawcounts))
```

```
FALSE
```

Matching order between vectors

Using the `match()` function:

```
match(vector1, vector2)
```

vector1: vector of values with the desired order

vector2: vector of values to reorder

output: the indices for how to rearrange vector2 to be in the same order as vector1

```
match(colnames(wt_rawcounts), rownames(wt_metadata)
```

```
6   9   1   14  8   3   13  10  5   7   11  2   4   12
```

Reordering using `match()` output:

```
idx <- match(colnames(wt_rawcounts), rownames(wt_metadata))
reordered_wt_metadata <- wt_metadata[idx, ]
View(reordered_wt_metadata)
```

	genotype	condition
wt_normal1	wt	normal
wt_normal2	wt	normal
wt_normal3	wt	normal
wt_fibrosis1	wt	fibrosis
wt_fibrosis2	wt	fibrosis
wt_fibrosis3	wt	fibrosis
wt_fibrosis4	wt	fibrosis

```
all(rownames(reordered_wt_metadata) == colnames(wt_rawcounts))
```

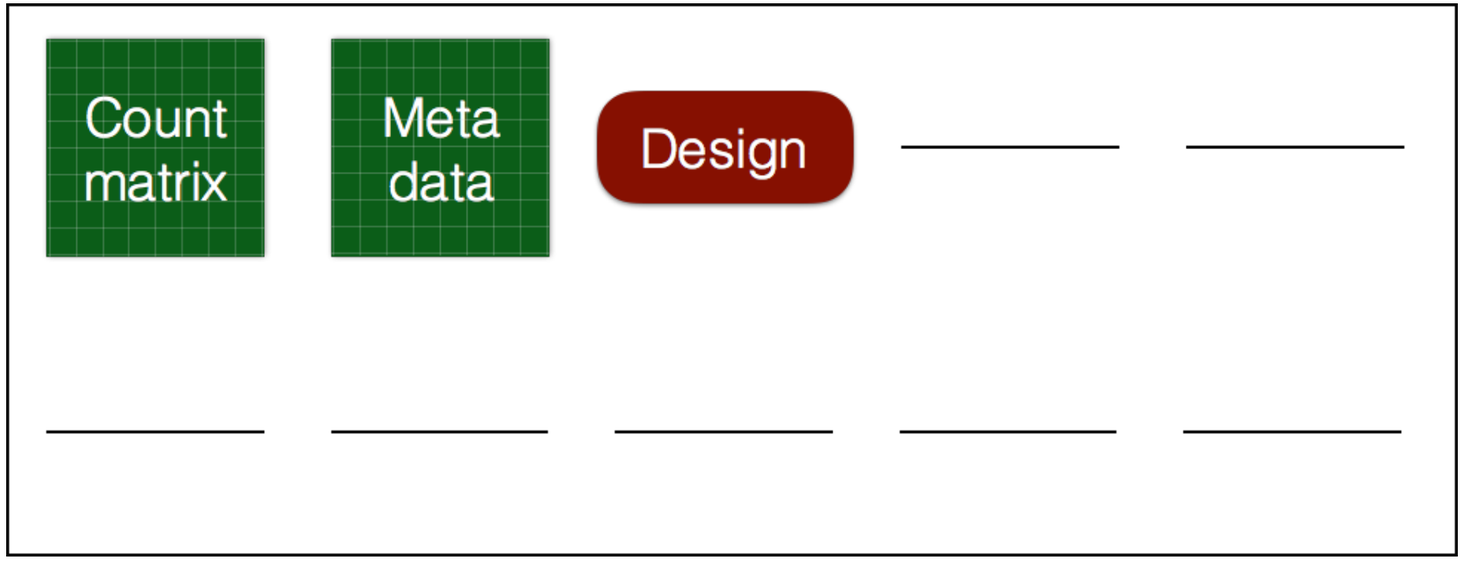
```
TRUE
```

	genotype	condition
wt_normal1	wt	normal
wt_normal2	wt	normal
wt_normal3	wt	normal
wt_fibrosis1	wt	fibrosis
wt_fibrosis2	wt	fibrosis
wt_fibrosis3	wt	fibrosis
wt_fibrosis4	wt	fibrosis

	wt_normal1	wt_normal2	wt_normal3	wt_fibrosis1	wt_fibrosis2	wt_fibrosis3	wt_fibrosis4
ENSMUSG00000102693	0	0	0	0	0	0	0
ENSMUSG00000064842	0	0	0	0	0	0	0
ENSMUSG00000051951	3	1	1	42	52	16	35
ENSMUSG00000102851	0	0	0	0	0	0	0
ENSMUSG00000103377	0	0	0	0	0	0	0
ENSMUSG00000104017	0	0	0	0	0	0	0
ENSMUSG00000103025	0	0	0	1	0	0	0
ENSMUSG00000089699	0	0	0	0	0	0	0
ENSMUSG00000103201	0	0	0	0	0	0	0
ENSMUSG00000103147	0	0	0	0	1	1	1

Creating the DESeq2 object

```
# Create DESeq object
dds_wt <- DESeqDataSetFromMatrix(countData = wt_rawcounts,
                                 colData = reordered_wt_metadata,
                                 design = ~ condition)
```



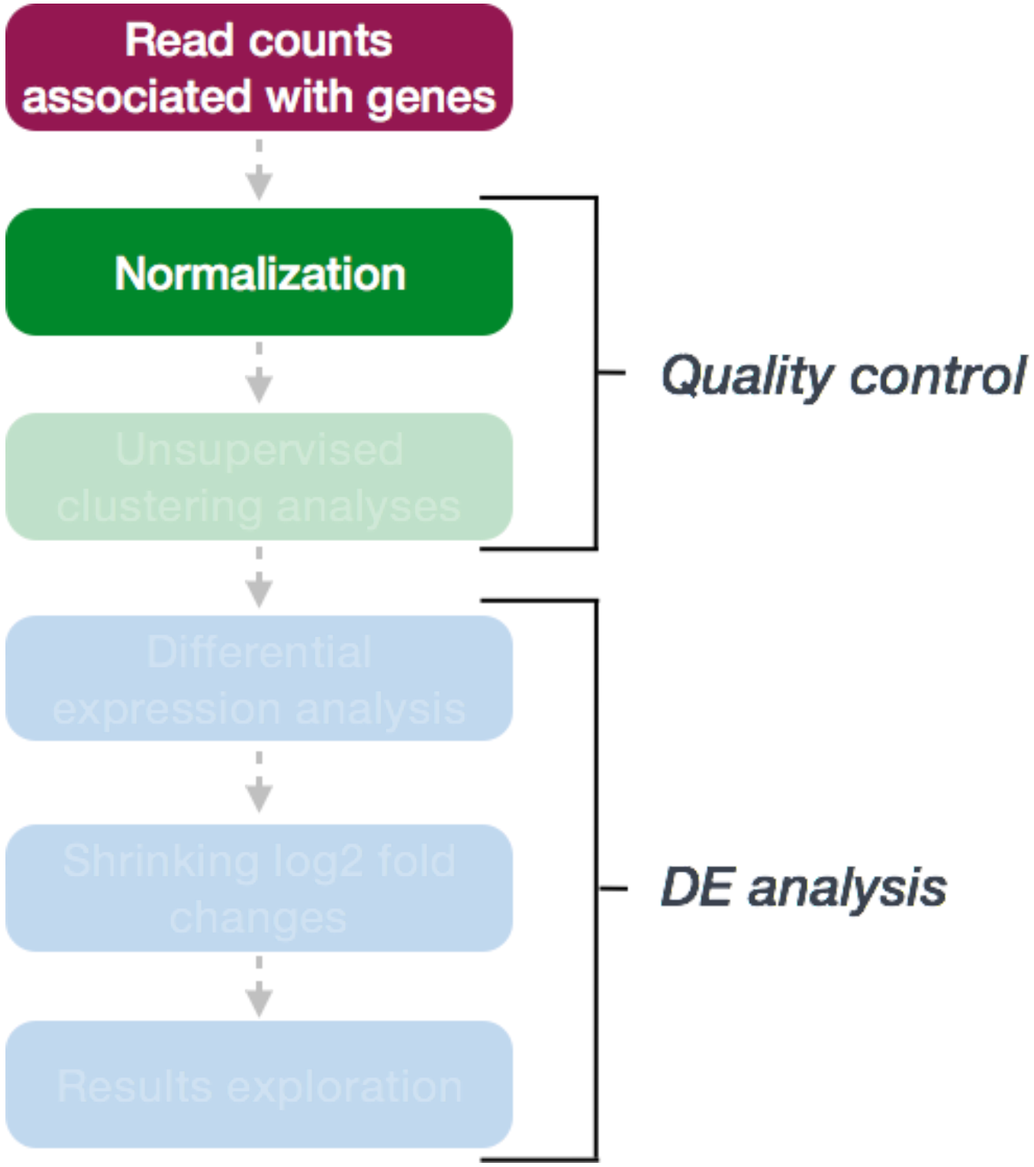
Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R

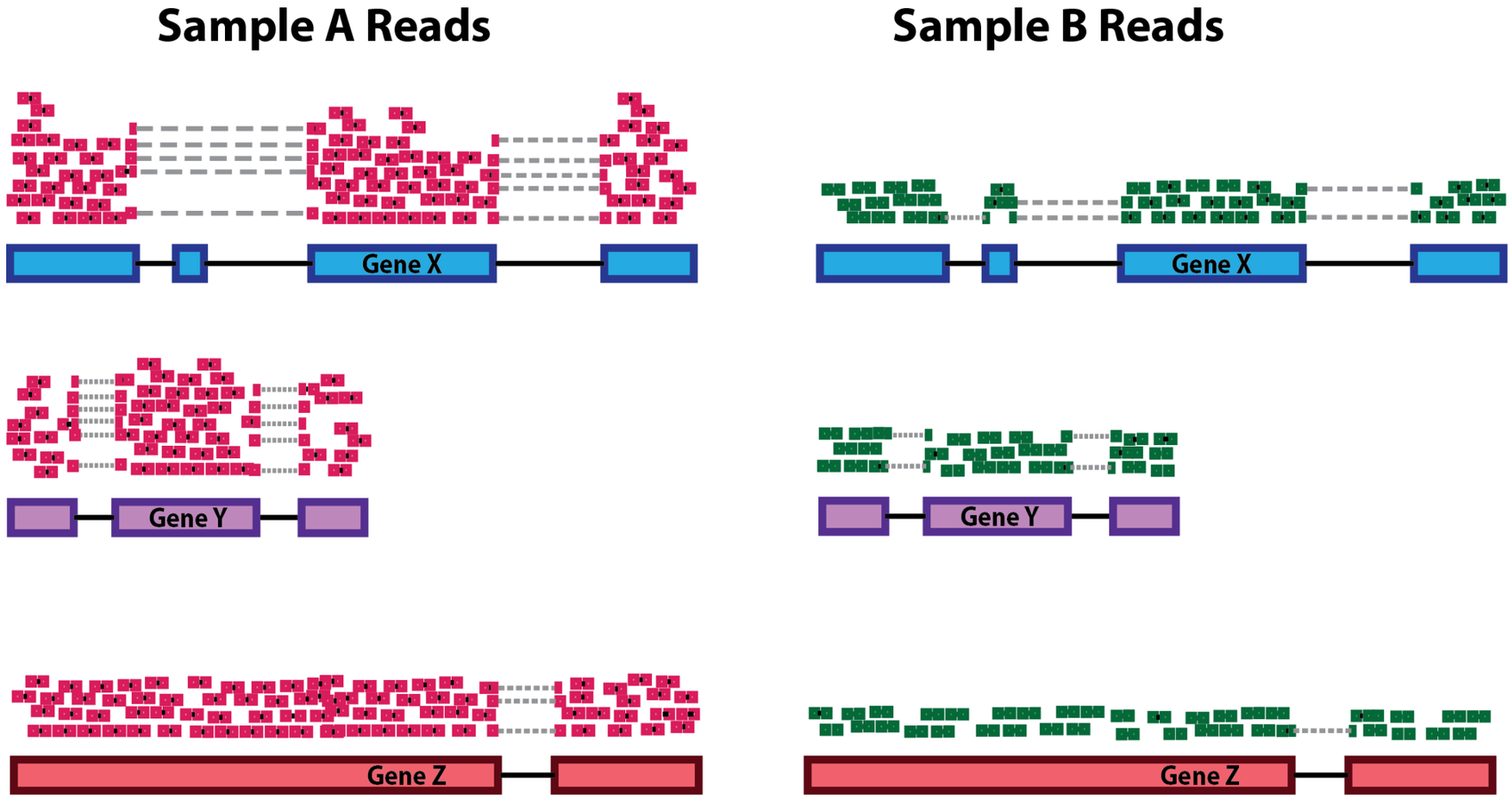
Count normalization

RNA-SEQ WITH BIOCONDUCTOR IN R

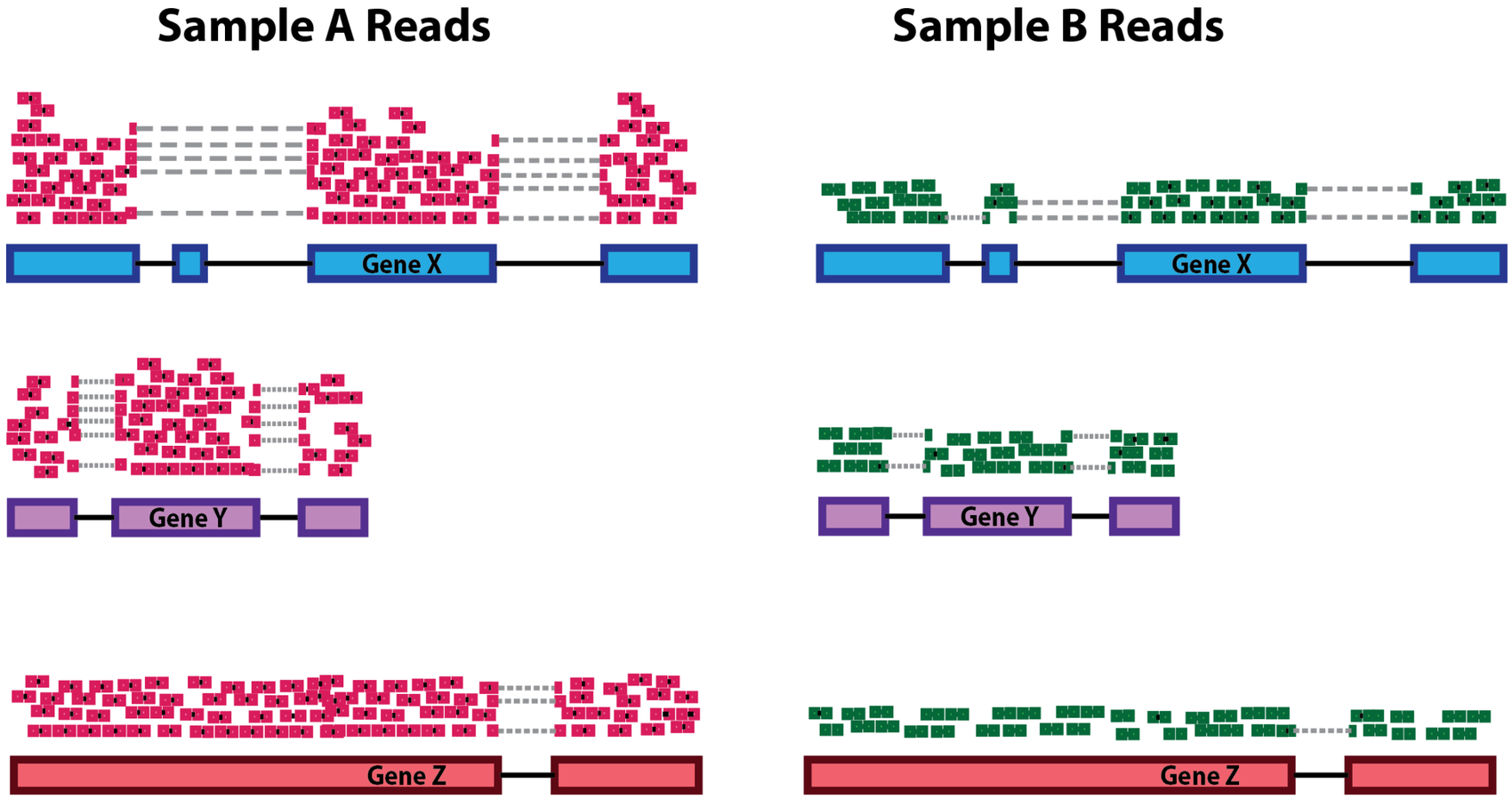
Mary Piper
Bioinformatics Consultant and Trainer

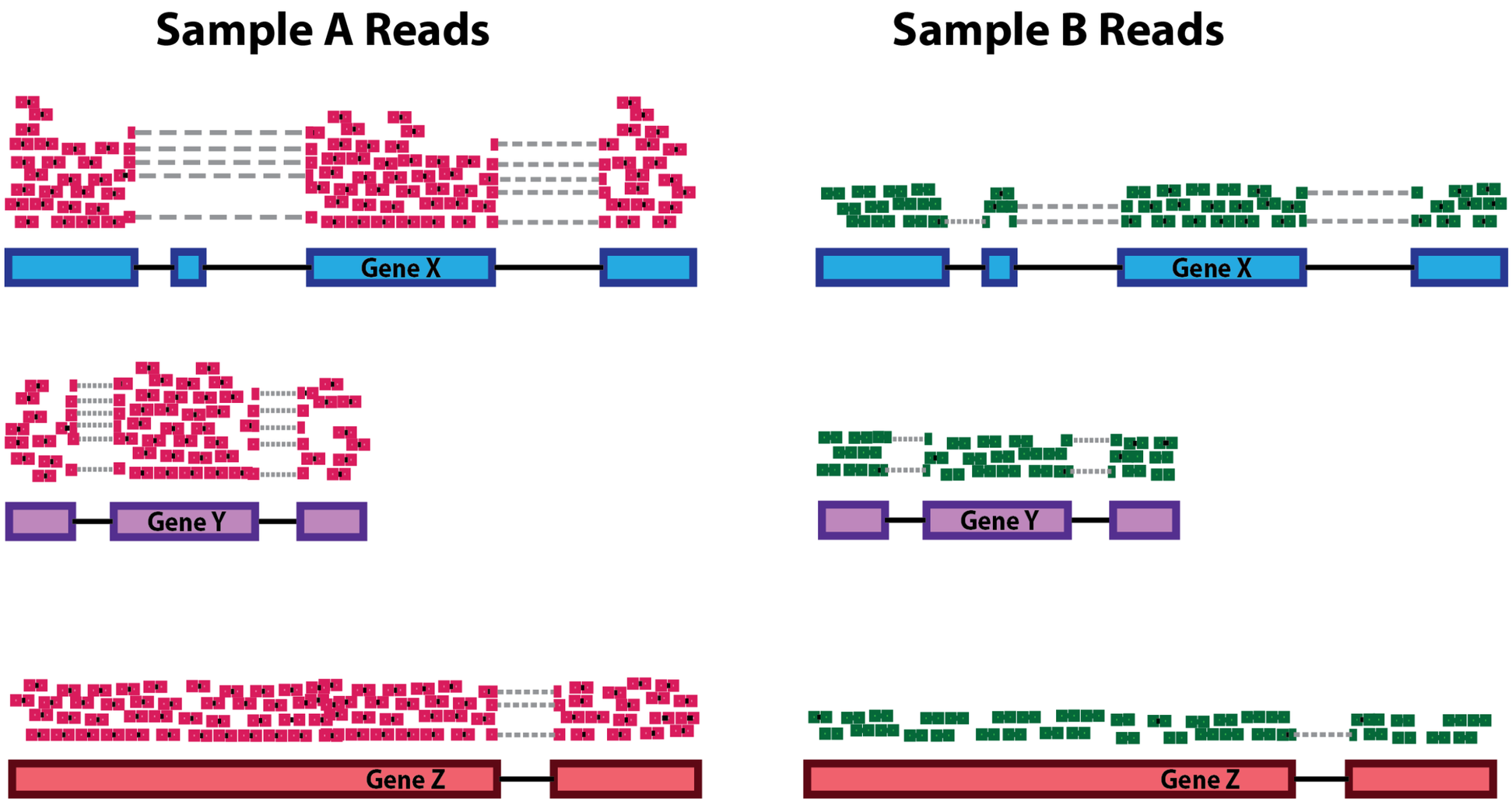

Read counts associated with genes
Normalization
Unsupervised clustering analyses
Differential expression analysis
Shrinking log2 fold changes
Results exploration
Quality control
DE analysis

Count normalization

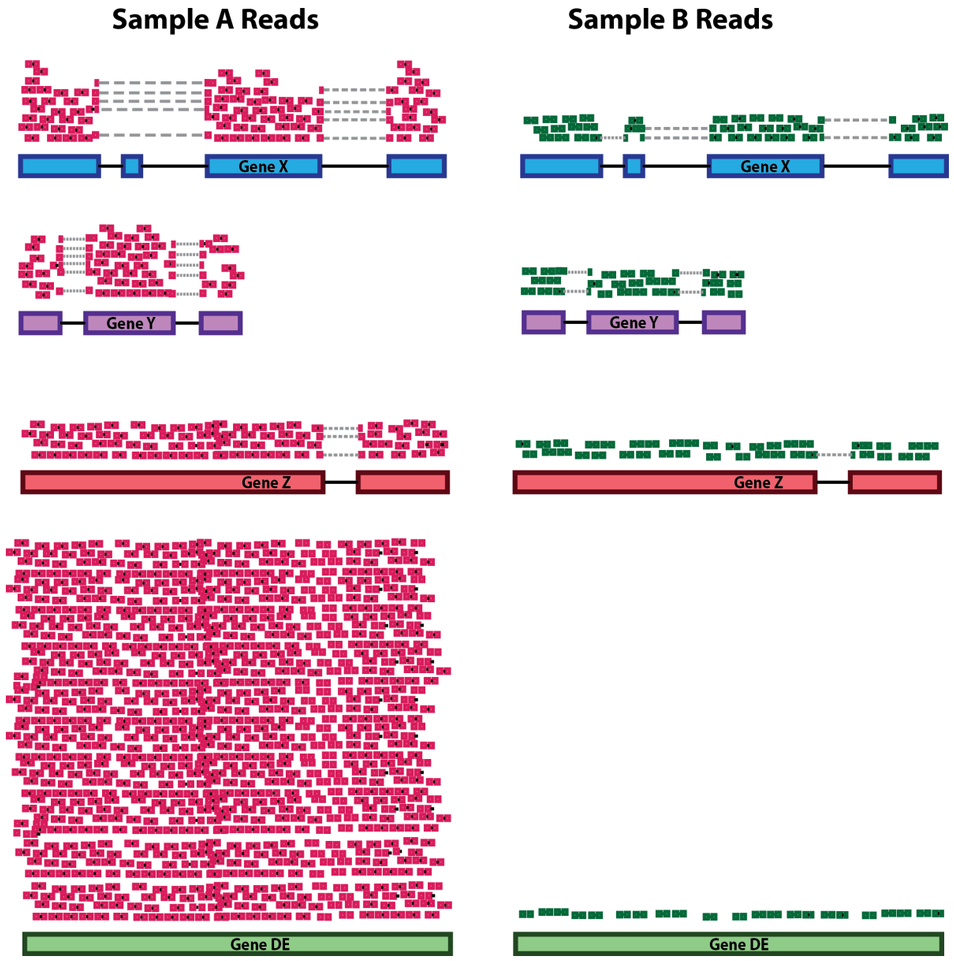


Library depth normalization

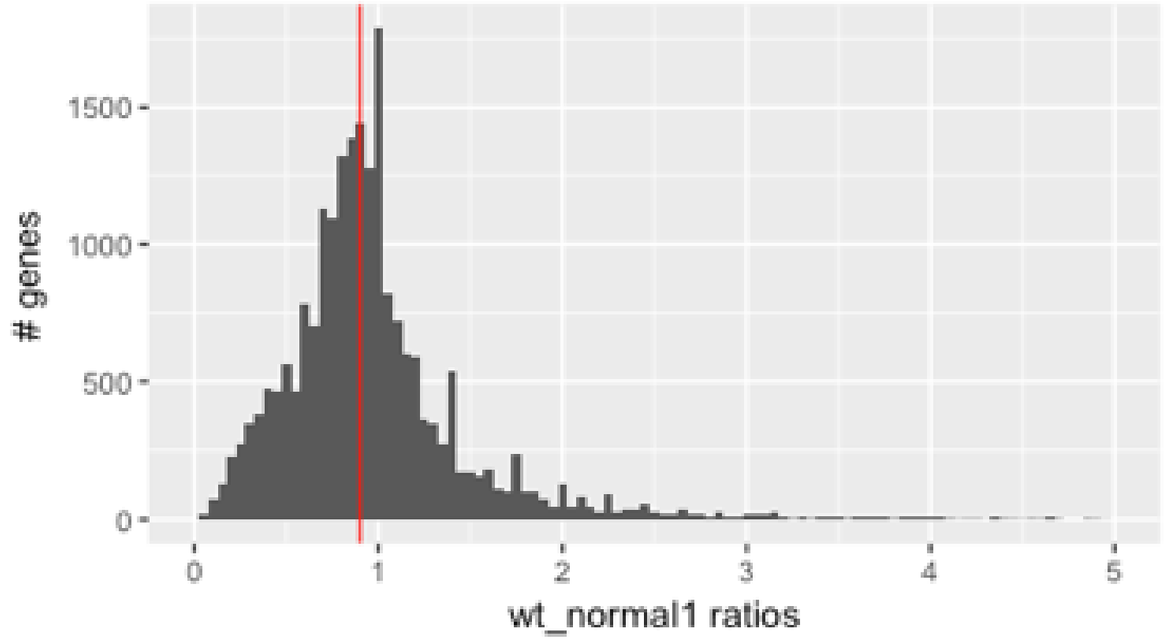


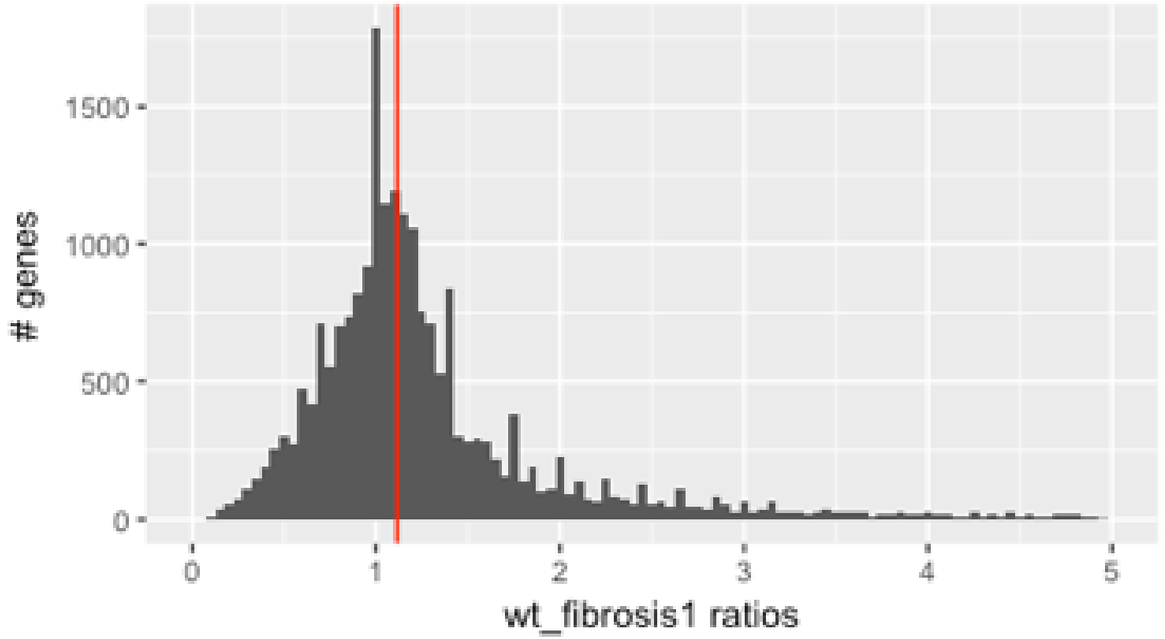

Sample A Reads
Sample B Reads
Gene X
Gene X
Gene Y
Gene Y
Gene Z
Gene Z

Library composition effect



DESeq2 normalization





Normalized counts: calculation

```
dds_wt <- estimateSizeFactors(dds_wt)
sizeFactors(dds_wt)
```

```
wt_normal1   wt_normal2   wt_normal3 wt_fibrosis1 wt_fibrosis2 wt_fibrosis3 wt_fibrosis4
 0.9131884    0.7250234    1.0441118    1.1346070    1.2059020    1.1731687    0.9418653
```

Normalized counts: extraction

```
normalized_wt_counts <- counts(dds_wt, normalized=TRUE)
View(normalized_wt_counts)
```

	wt_normal1	wt_normal2	wt_normal3	wt_fibrosis1	wt_fibrosis2	wt_fibrosis3	wt_fibrosis4
ENSMUSG00000102693	0.000000	0.000000	0.000000e+00	0.0000000	0.0000000	0.0000000	0.000000
ENSMUSG00000064842	0.000000	0.000000	0.000000e+00	0.0000000	0.0000000	0.0000000	0.000000
ENSMUSG00000051951	3.285193	1.379266	9.577519e-01	37.0172230	43.1212477	13.6382769	37.160301
ENSMUSG00000102851	0.000000	0.000000	0.000000e+00	0.0000000	0.0000000	0.0000000	0.000000
ENSMUSG00000103377	0.000000	0.000000	0.000000e+00	0.0000000	0.0000000	0.0000000	0.000000
ENSMUSG00000104017	0.000000	0.000000	0.000000e+00	0.0000000	0.0000000	0.0000000	0.000000
ENSMUSG00000103025	0.000000	0.000000	0.000000e+00	0.8813625	0.0000000	0.0000000	0.000000
ENSMUSG00000089699	0.000000	0.000000	0.000000e+00	0.0000000	0.0000000	0.0000000	0.000000
ENSMUSG00000103201	0.000000	0.000000	0.000000e+00	0.0000000	0.0000000	0.0000000	0.000000
ENSMUSG00000103147	0.000000	0.000000	0.000000e+00	0.0000000	0.8292548	0.8523923	1.061723
ENSMUSG00000103161	0.000000	0.000000	0.000000e+00	0.0000000	0.0000000	0.0000000	0.000000
ENSMUSG00000102331	1.095064	0.000000	0.000000e+00	5.2881747	6.6340381	8.5239231	7.432060

Let's practice!

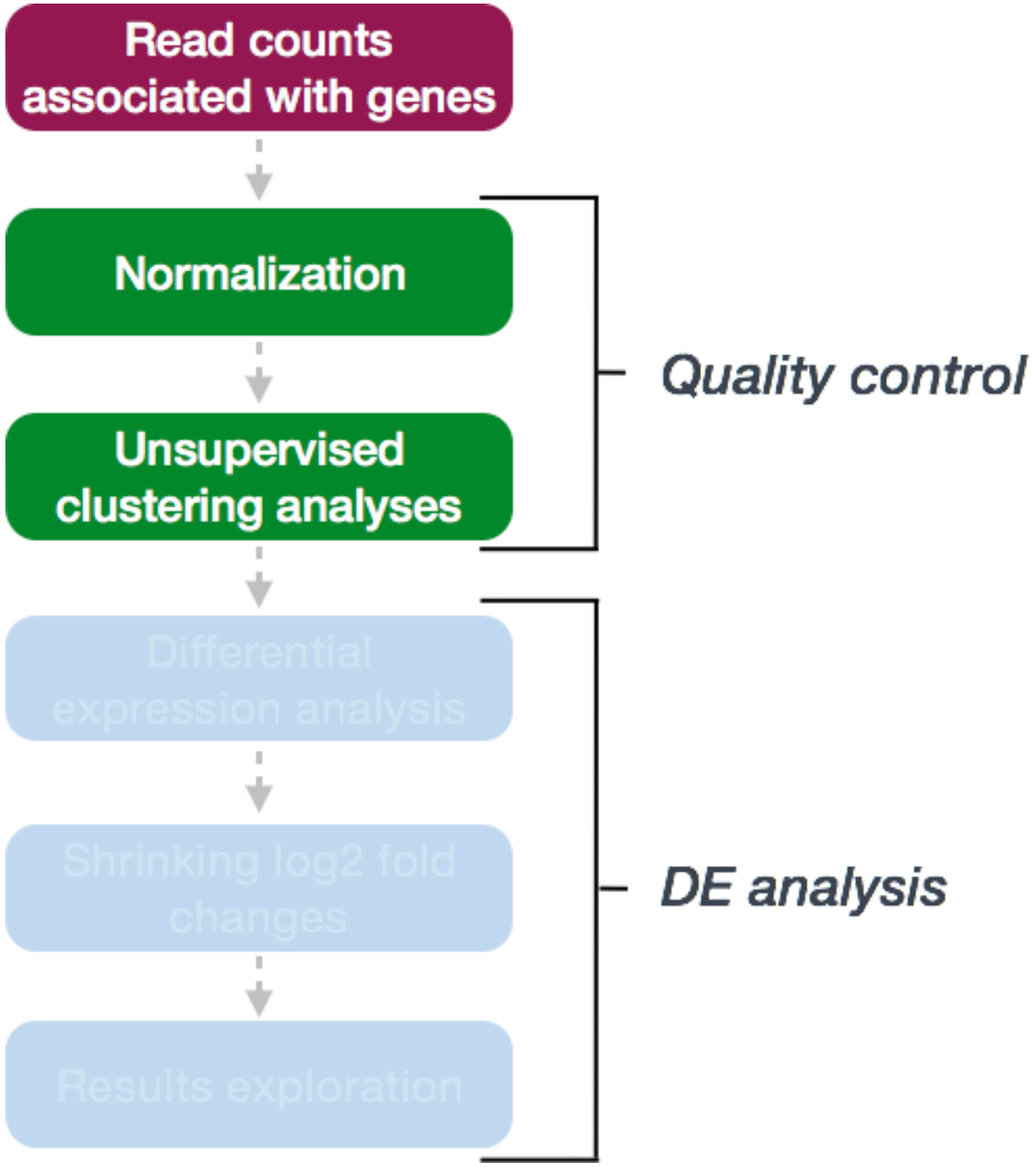
RNA-SEQ WITH BIOCONDUCTOR IN R

Unsupervised clustering analyses

RNA-SEQ WITH BIOCONDUCTOR IN R

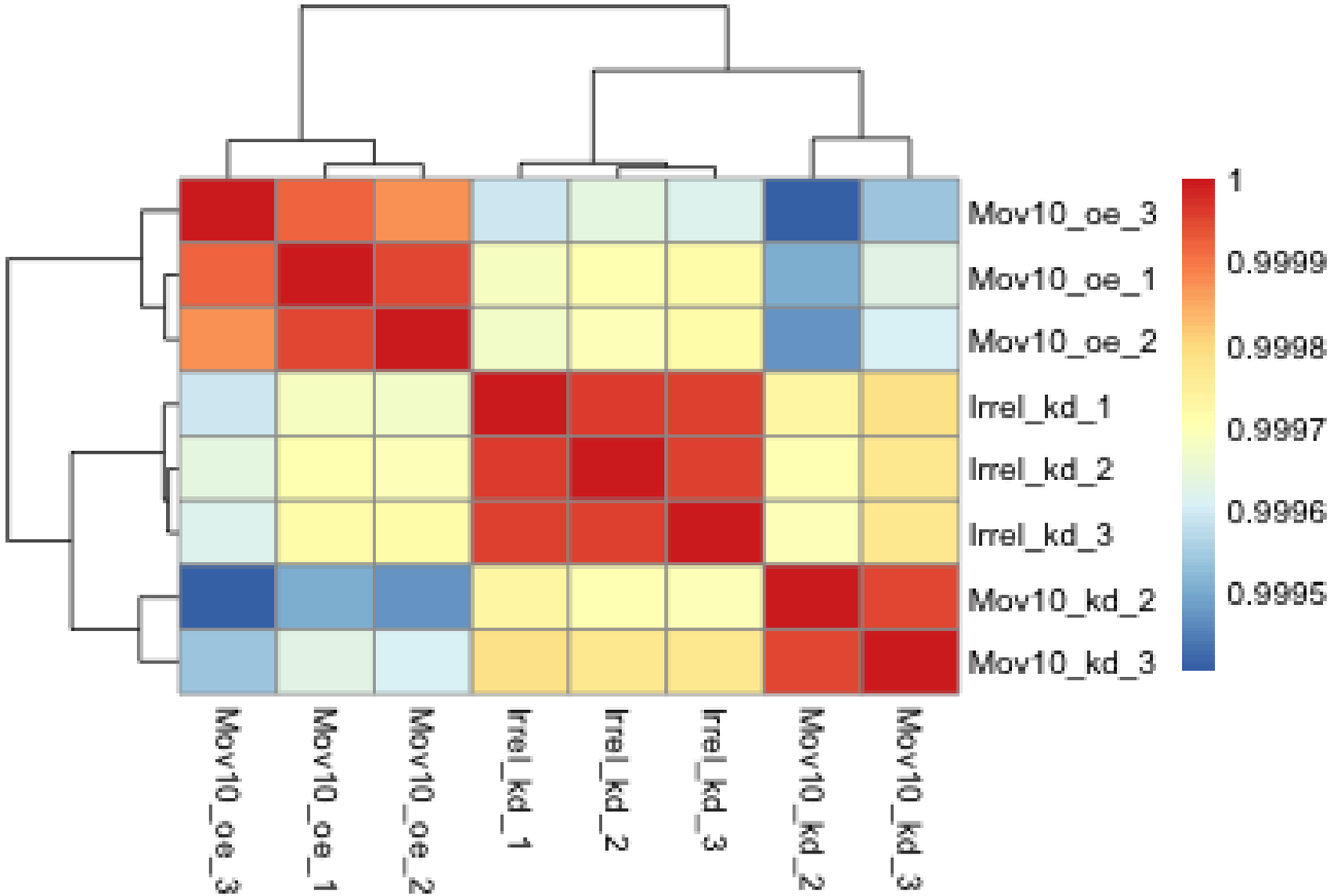
Mary Piper
Instructor

datacamp


Read counts associated with genes
Normalization
Unsupervised clustering analyses
Differential expression analysis
Shrinking log2 fold changes
Results exploration
Quality control
DE analysis

Unsupervised clustering analyses: log transformation

```
vsd_wt <- vst(dds_wt, blind=TRUE)
```


Mov10_oe_3
Mov10_oe_1
Mov10_oe_2
Irrel_kd_1
Irrel_kd_2
Irrel_kd_3
Mov10_kd_2
Mov10_kd_3
1
0.9999
0.9998
0.9997
0.9996
0.9995

Hierarchical clustering with correlation heatmaps

```
# Extract the vst matrix from the object
vsd_mat_wt <- assay(vsd_wt)
# Compute pairwise correlation values
vsd_cor_wt <- cor(vsd_mat_wt)
View(vsd_cor_wt)
```

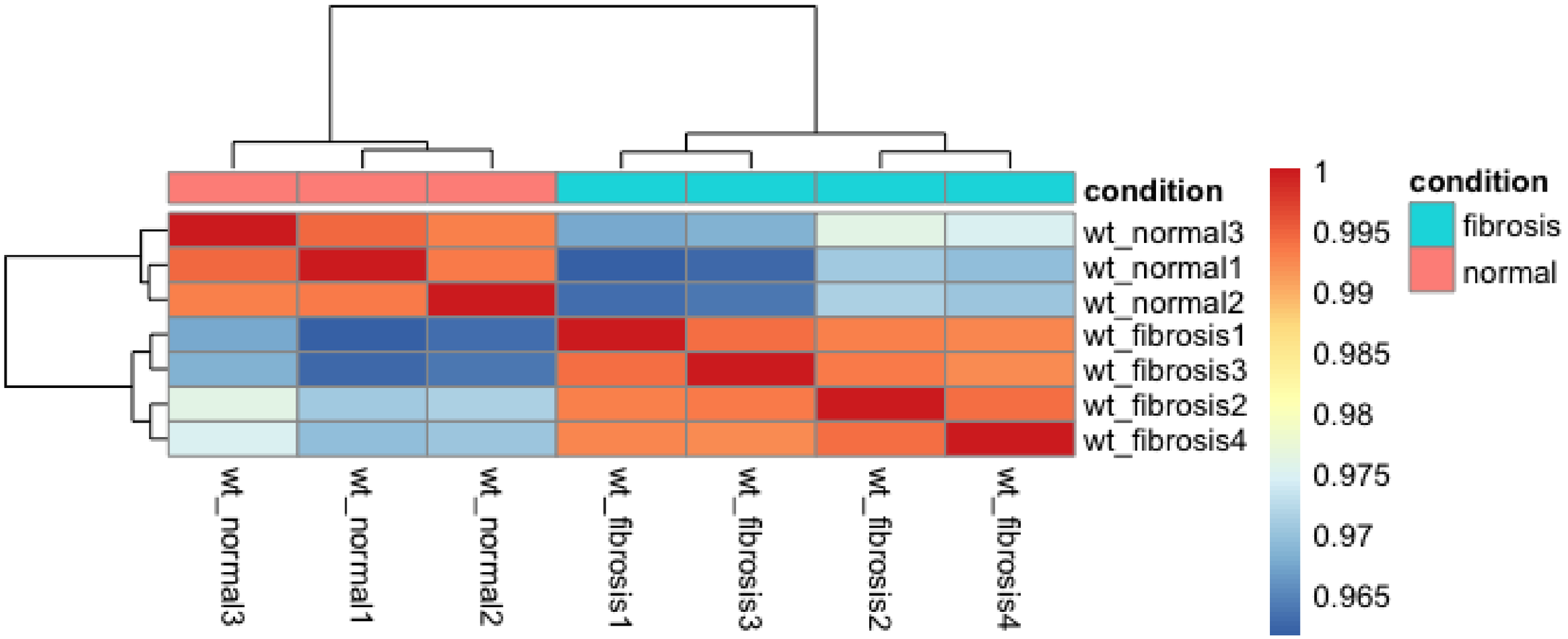
	wt_normal1	wt_normal2	wt_normal3	wt_fibrosis1	wt_fibrosis2	wt_fibrosis3	wt_fibrosis4
wt_normal1	1.0000000	0.9934287	0.9945298	0.9616998	0.9708459	0.9626185	0.9696097
wt_normal2	0.9934287	1.0000000	0.9930148	0.9629644	0.9713154	0.9639685	0.9704541
wt_normal3	0.9945298	0.9930148	1.0000000	0.9678018	0.9758950	0.9683519	0.9750891
wt_fibrosis1	0.9616998	0.9629644	0.9678018	1.0000000	0.9930090	0.9939055	0.9926560
wt_fibrosis2	0.9708459	0.9713154	0.9758950	0.9930090	1.0000000	0.9931793	0.9939010
wt_fibrosis3	0.9626185	0.9639685	0.9683519	0.9939055	0.9931793	1.0000000	0.9922991
wt_fibrosis4	0.9696097	0.9704541	0.9750891	0.9926560	0.9939010	0.9922991	1.0000000

Hierarchical clustering with correlation heatmaps

```
# Load pheatmap libraries
library(pheatmap)


# Plot heatmap
pheatmap(vsd_cor_wt, annotation = select(wt_metadata, condition))
```

Hierarchical clustering with correlation heatmaps



Let's practice!

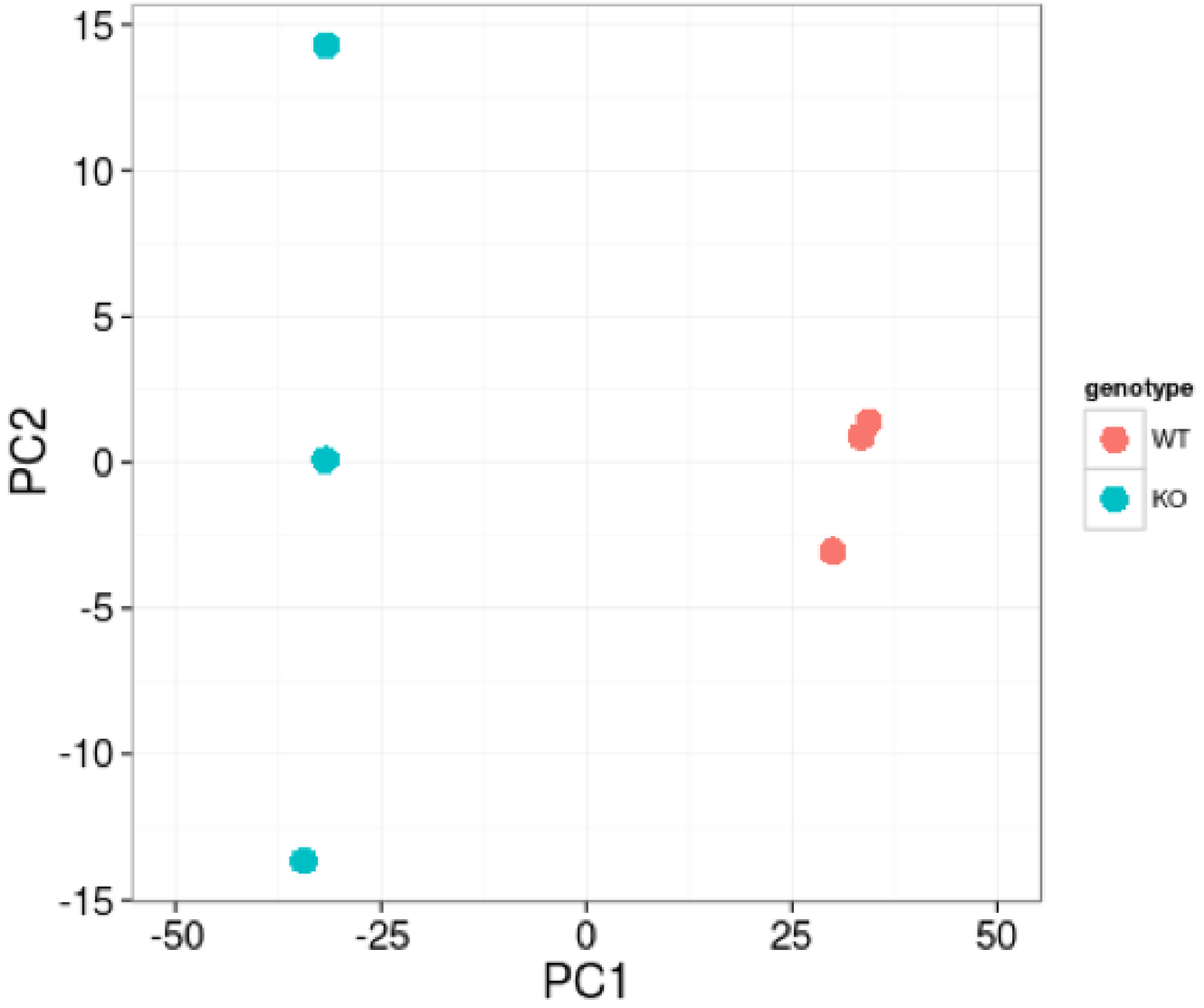
RNA-SEQ WITH BIOCONDUCTOR IN R

Principal component analysis

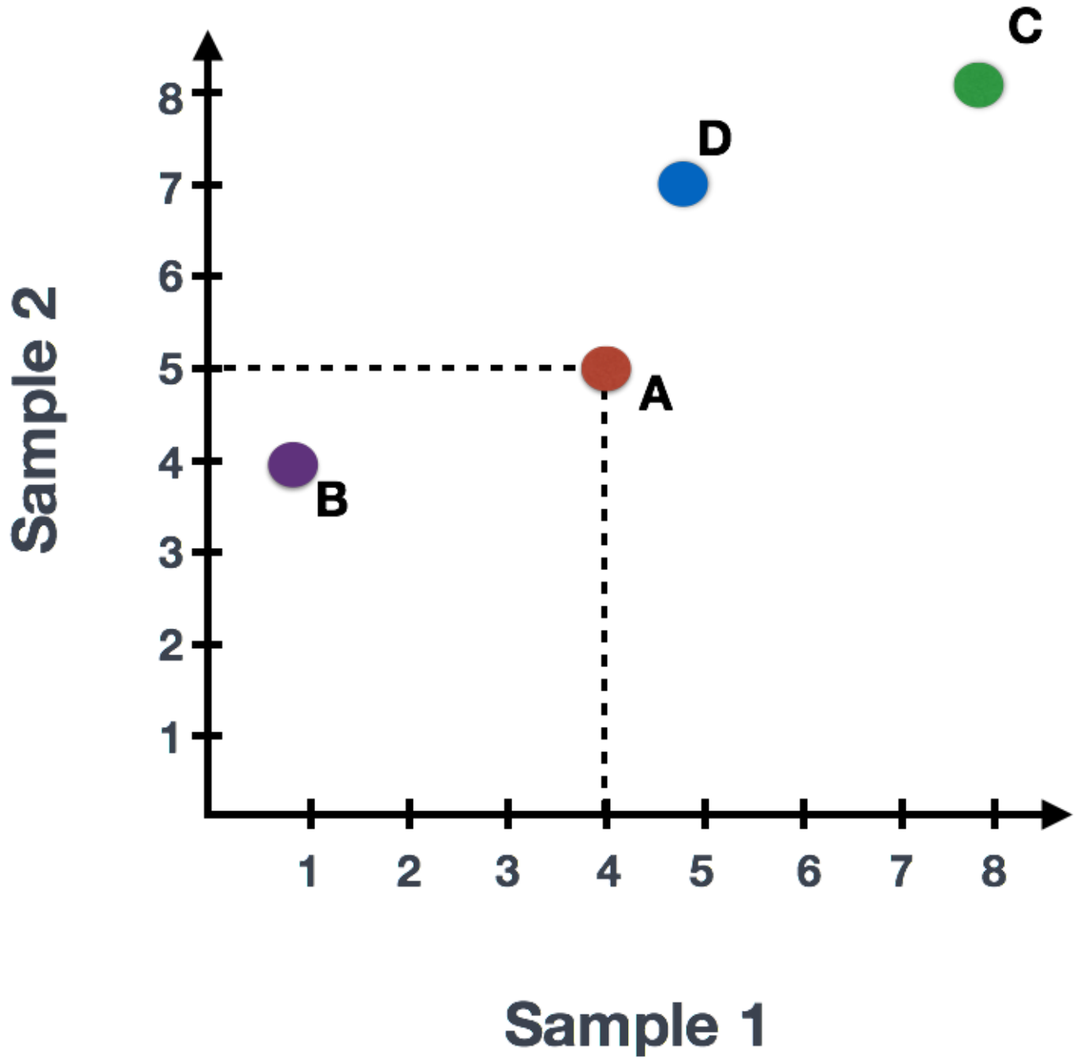
RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

datacamp

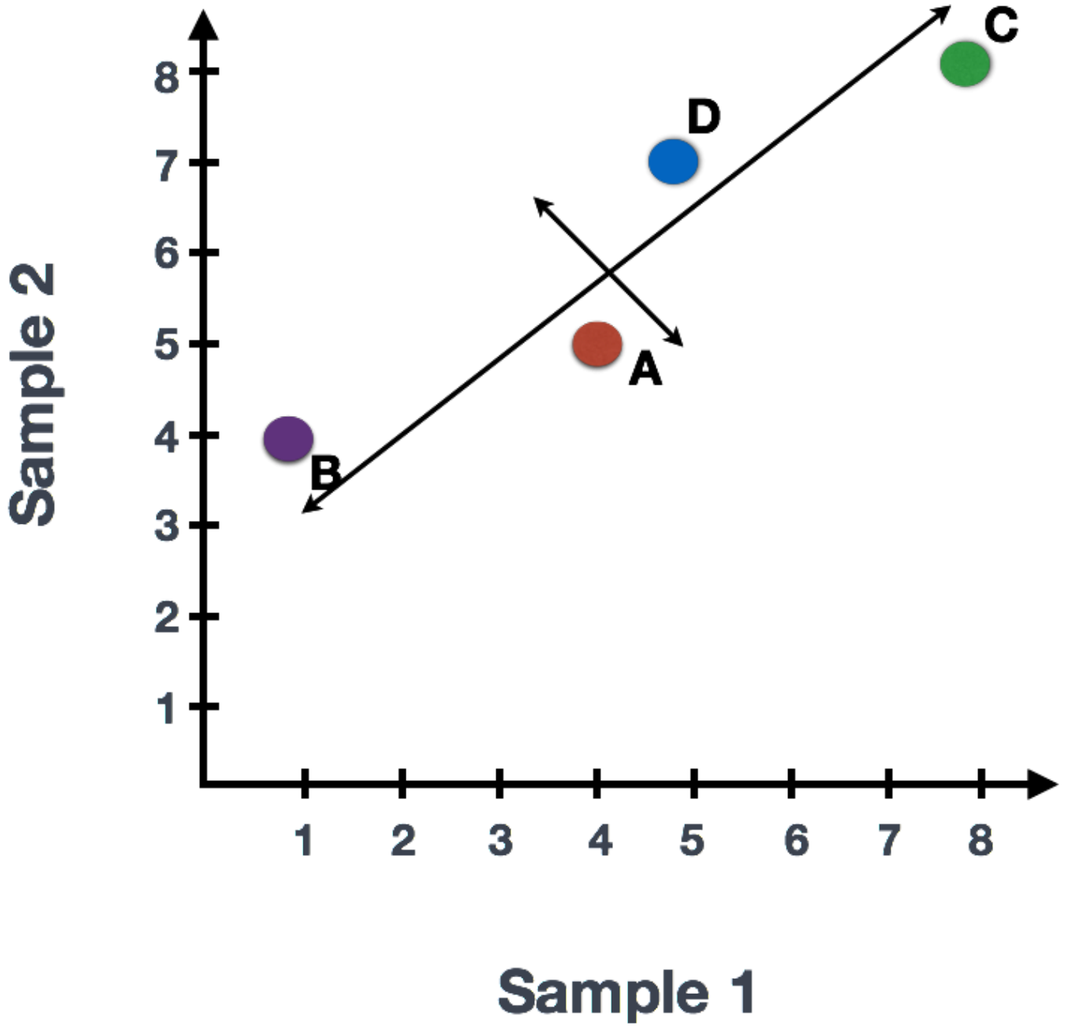

15
10
5
0
-5
-10
-15
PC2
-50
-25
0
25
50
PC1
genotype
WT
KO

Principal component analysis (PCA): Theory

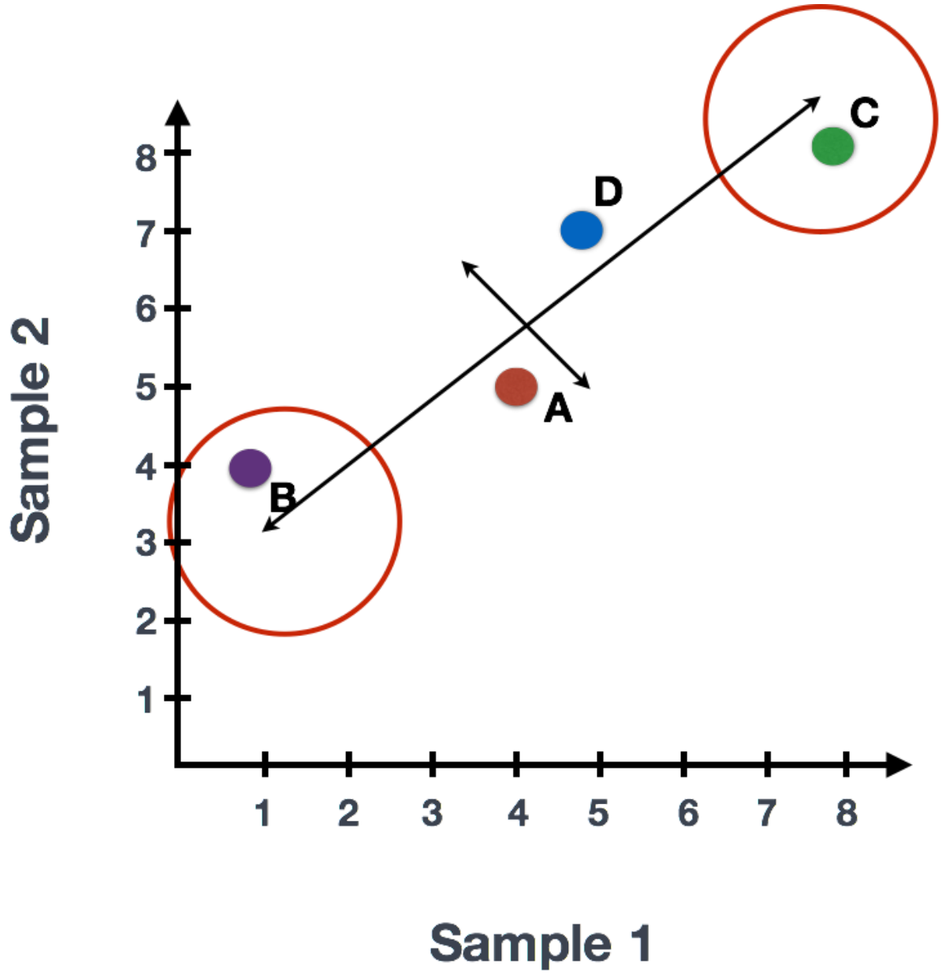


	Sample 1	Sample 2
Gene A	4	5
Gene B	1	4
Gene C	8	8
Gene D	5	7

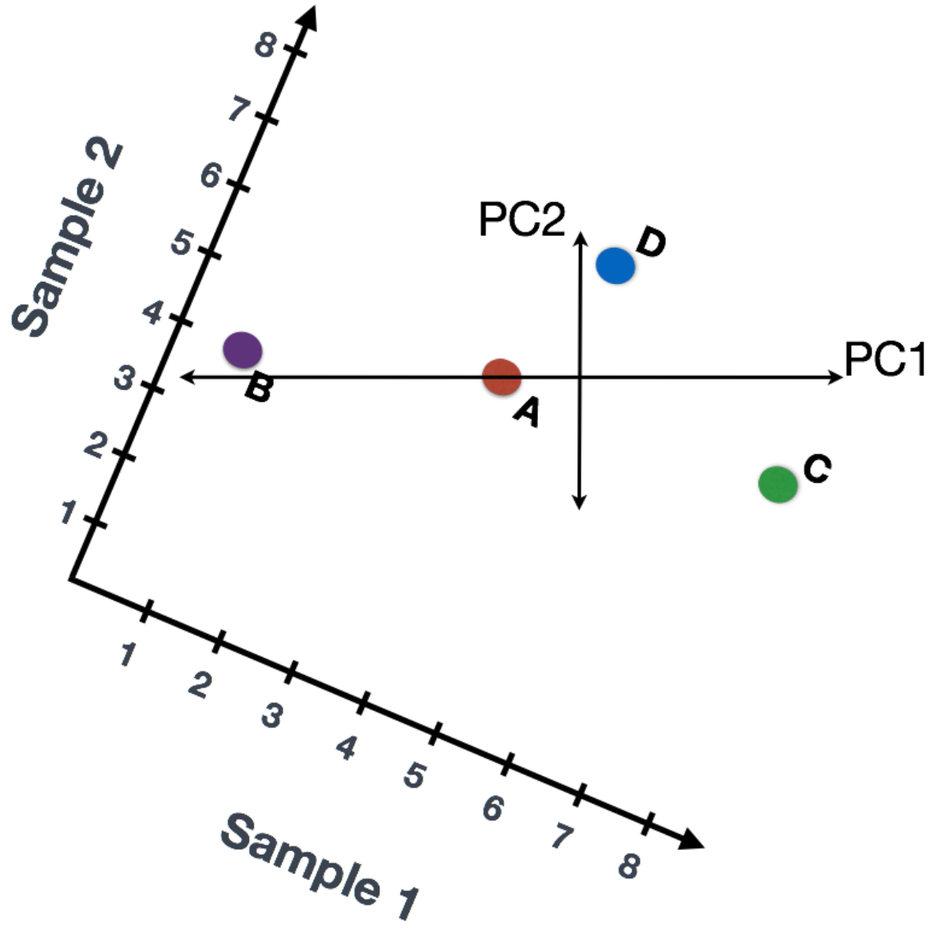
Principal component analysis (PCA): Theory



Principal component analysis (PCA): Theory

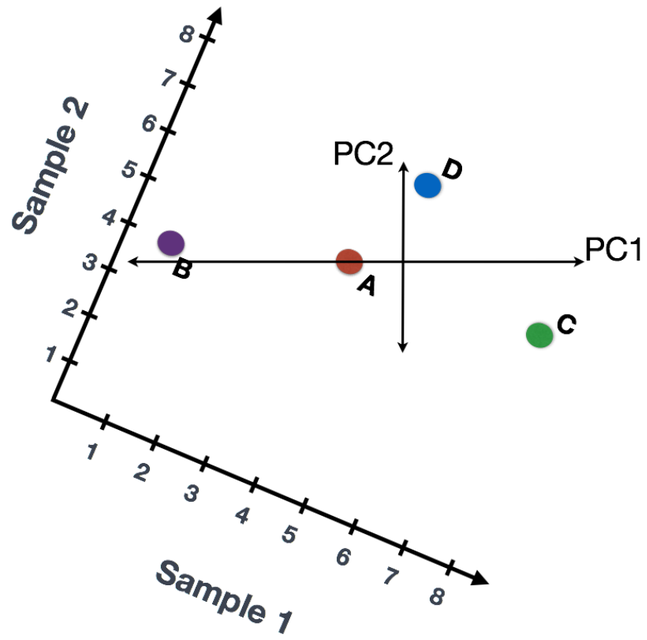


Principal component analysis (PCA): Theory



	Sample 1	Sample 2	Influence on PC1	Influence on PC2
Gene A	4	5	-2	0.5
Gene B	1	4	-10	1
Gene C	8	8	8	-5
Gene D	5	7	1	6

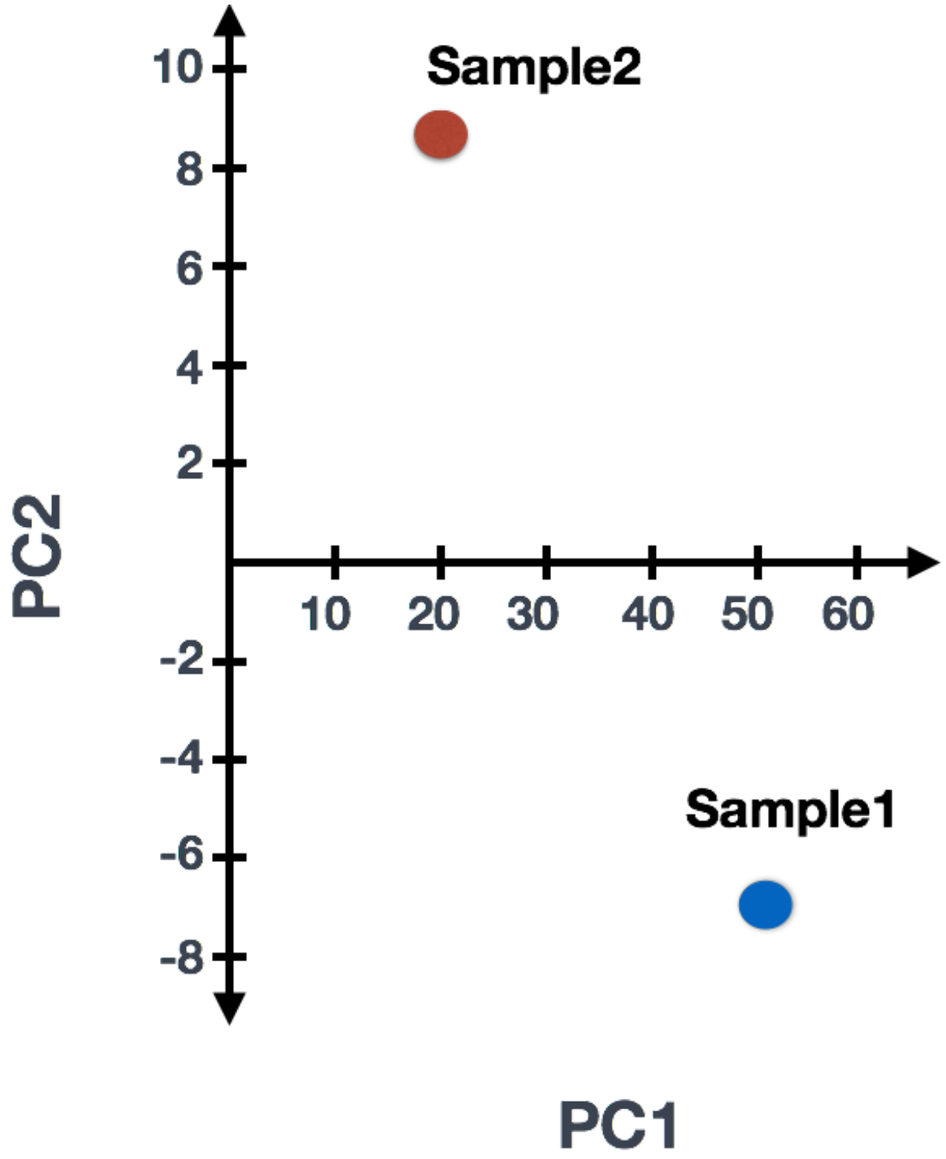
Principal component analysis (PCA): Theory



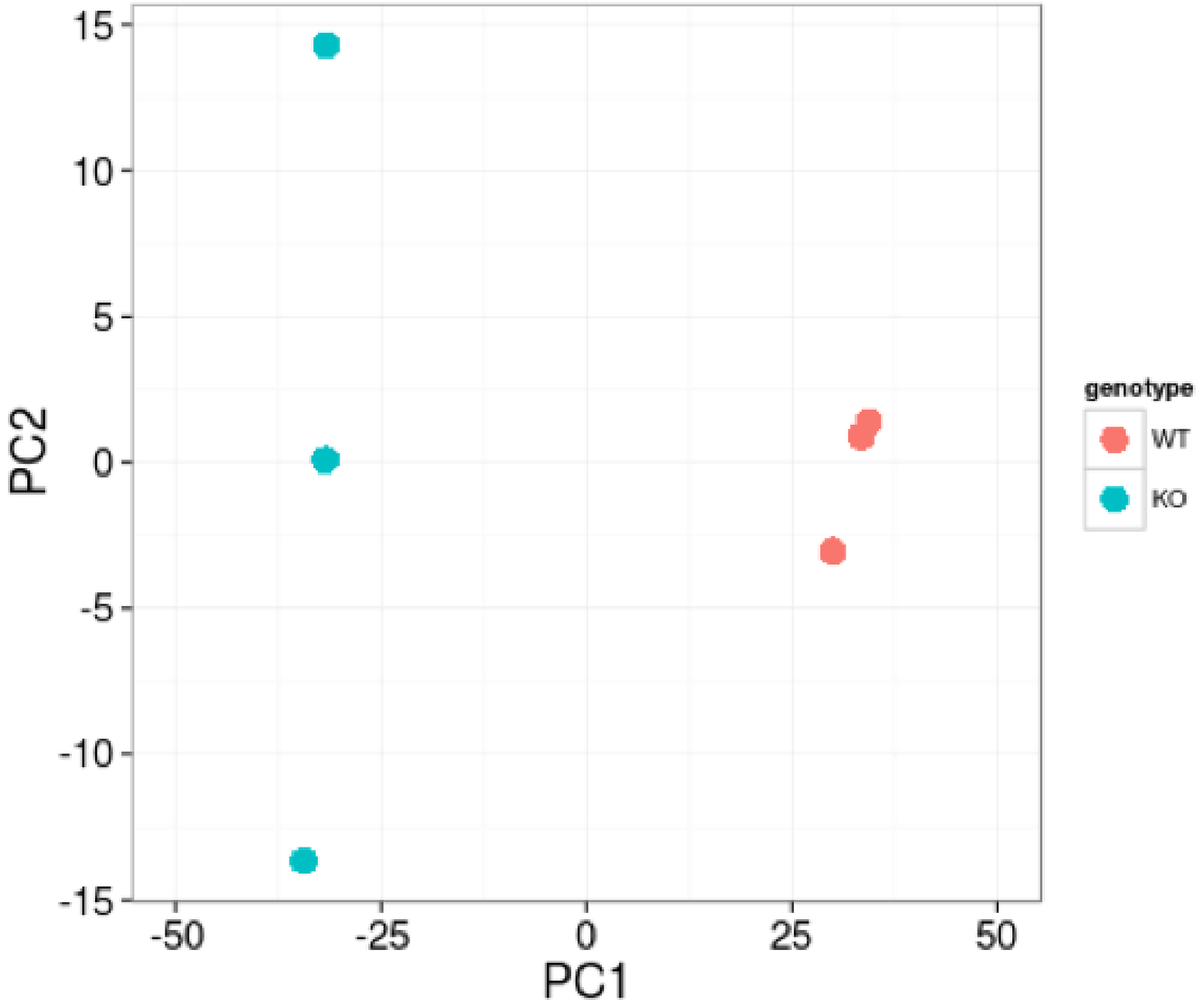
	Sample 1	Sample 2	Influence on PC1	Influence on PC2
Gene A	4	5	-2	0.5
Gene B	1	4	-10	1
Gene C	8	8	8	-5
Gene D	5	7	1	6

```
Sample1 PC1 score = (4 * -2) + (1 * -10) + (8 * 8) + (5 * 1) = 51
Sample1 PC2 score = (4 * 0.5) + (1 * 1) + (8 * -5) + (5 * 6) = -7

Sample2 PC1 score = (5 * -2) + (4 * -10) + (8 * 8) + (7 * 1) = 21
Sample2 PC2 score = (5 * 0.5) + (4 * 1) + (8 * -5) + (7 * 6) = 8.5
```

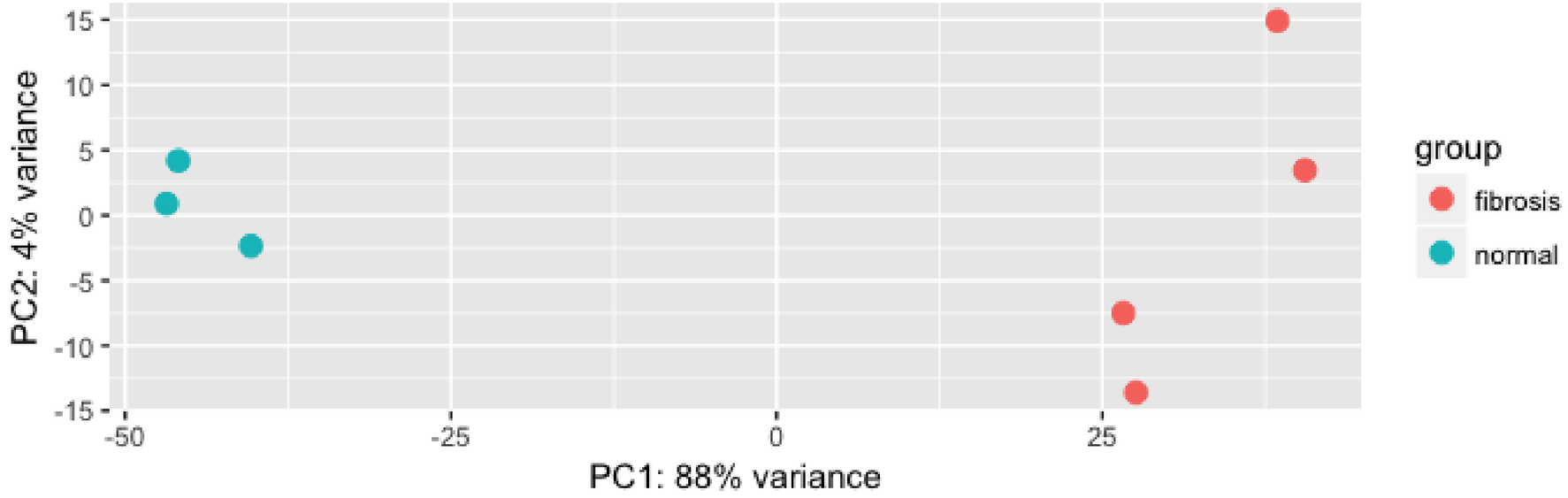


	PC1	PC2
Sample1	51	-7
Sample2	21	8.5


15
10
5
0
-5
-10
-15
PC2
-50
-25
0
25
50
PC1
genotype
WT
KO

Principal component analysis (PCA): Theory

```
# Plot PCA
plotPCA(vsd_wt, intgroup="condition")
```



Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R