Visualization of results

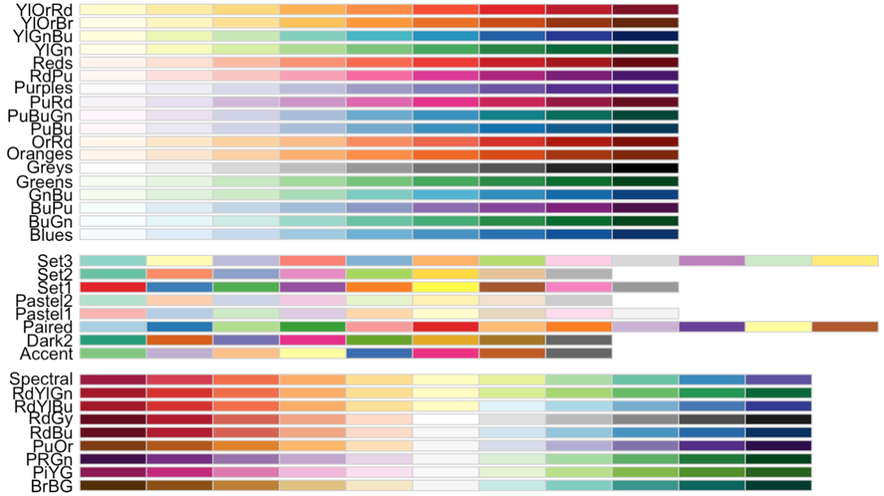
RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

datacamp

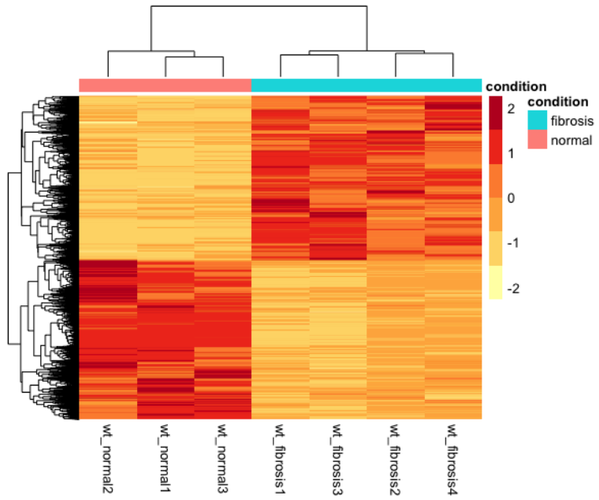
Visualizing results - Expression heatmap

```
# Subset normalized counts to significant genes
sig_norm_counts_wt <- normalized_counts_wt[wt_res_sig$ensgene, ]
# Choose a color palette from RColorBrewer
library(RColorBrewer)
heat_colors <- brewer.pal(6, "YlOrRd")
display.brewer.all()
```



Visualizing results - Expression heatmap

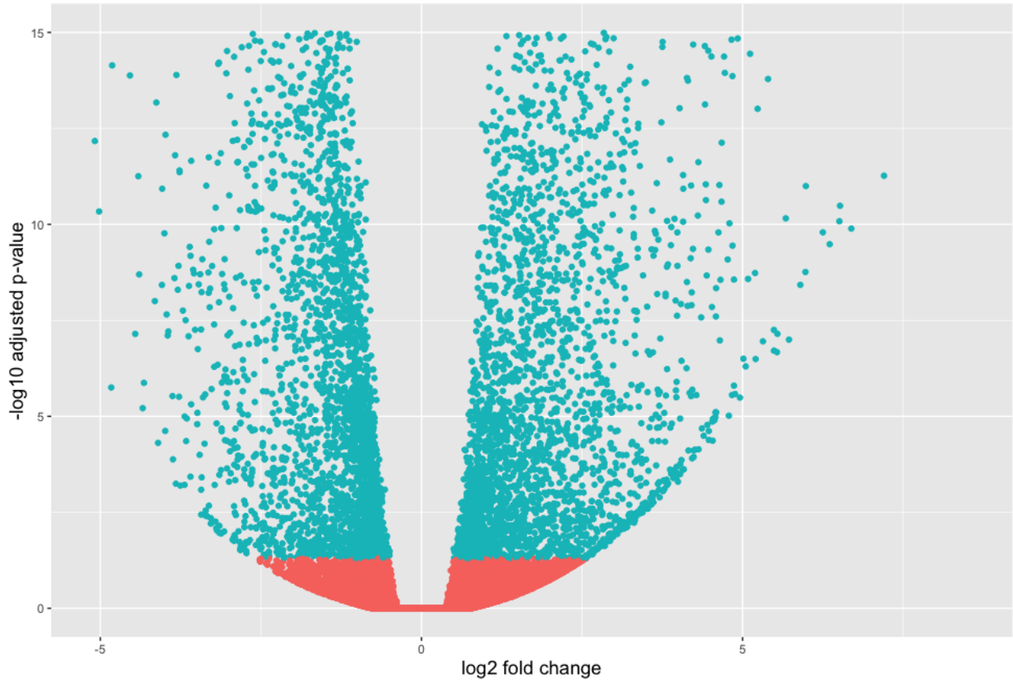
```
# Run pheatmap
pheatmap(sig_norm_counts_wt,
         color = heat_colors,
         cluster_rows = T,
         show_rownames = F,
         annotation = select(wt_metadata, condition),
         scale = "row")
```



Visualizing results - Volcano plot

```
# Obtain logical vector regarding whether padj values are less than 0.05
wt_res_all <- wt_res_all %>%
              rownames_to_column(var = "ensgene") %>%
              mutate(threshold = padj < 0.05)
# Volcano plot
ggplot(wt_res_all) +
        geom_point(aes(x = log2FoldChange, y = -log10(padj),
                       color = threshold)) +
        xlab("log2 fold change") +
        ylab("-log10 adjusted p-value") +
        theme(legend.position = "none",
              plot.title = element_text(size = rel(1.5), hjust = 0.5),
              axis.title = element_text(size = rel(1.25)))
```

```
ggplot(wt_res_all) +
        geom_point(aes(x = log2FoldChange, y = -log10(padj), color = threshold)) +
        xlab("log2 fold change") +
        ylab("-log10 adjusted p-value") +
        ylim=c(0, 15) +
        theme(legend.position = "none",
              plot.title = element_text(size = rel(1.5), hjust = 0.5),
              axis.title = element_text(size = rel(1.25)))
```



Visualizing results - Expression plot

Significant results:

ensgene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	description
ENSMUSG00000053113	1318.1717	4.875042	0.16021506	28.35016	8.330958e-177	1.830145e-172	Socs3	suppressor of cytokine signaling 3 [Source:M...
ENSMUSG00000005087	2943.7403	6.121134	0.20721978	27.89891	2.750356e-171	3.020991e-167	Cd44	CD44 antigen [Source:MGI Symbol;Acc:MGI:8...
ENSMUSG00000036887	3899.5135	3.866162	0.12740248	27.83465	1.652344e-170	1.209957e-166	C1qa	complement component 1, q subcomponent,...
ENSMUSG00000026822	8870.1712	6.466148	0.23782361	25.82294	4.901029e-147	2.691645e-143	Lcn2	lipocalin 2 [Source:MGI Symbol;Acc:MGI:96757]
ENSMUSG00000036905	3237.6046	3.835279	0.13773926	25.52164	1.134018e-143	4.982421e-140	C1qb	complement component 1, q subcomponent,...
ENSMUSG00000027962	9298.5984	5.781446	0.21949603	24.88019	1.219153e-136	4.463724e-133	Vcam1	vascular cell adhesion molecule 1 [Source:MG...

Normalized counts for significant genes:

	wt_normal1	wt_normal2	wt_normal3	wt_fibrosis1	wt_fibrosis2	wt_fibrosis3	wt_fibrosis4
ENSMUSG00000053113	65.703858	71.721818	86.197669	2333.84777	2238.98786	1970.73101	2460.01195
ENSMUSG00000005087	81.034758	70.342553	58.422864	5514.68486	3822.86446	6104.83370	4953.99902
ENSMUSG00000036887	373.416926	433.089442	512.397254	6299.09745	6195.36233	6902.67290	6580.55850
ENSMUSG00000026822	145.643552	199.993532	143.662782	11563.47537	18720.42627	19678.32879	11639.66809
ENSMUSG00000036905	313.188390	346.195700	454.932141	5318.14104	5484.69100	5393.93852	5352.14511
ENSMUSG00000027962	226.678310	300.679931	304.565097	17327.58581	10431.19566	19517.22664	16982.25769

Visualizing results - Expression plot

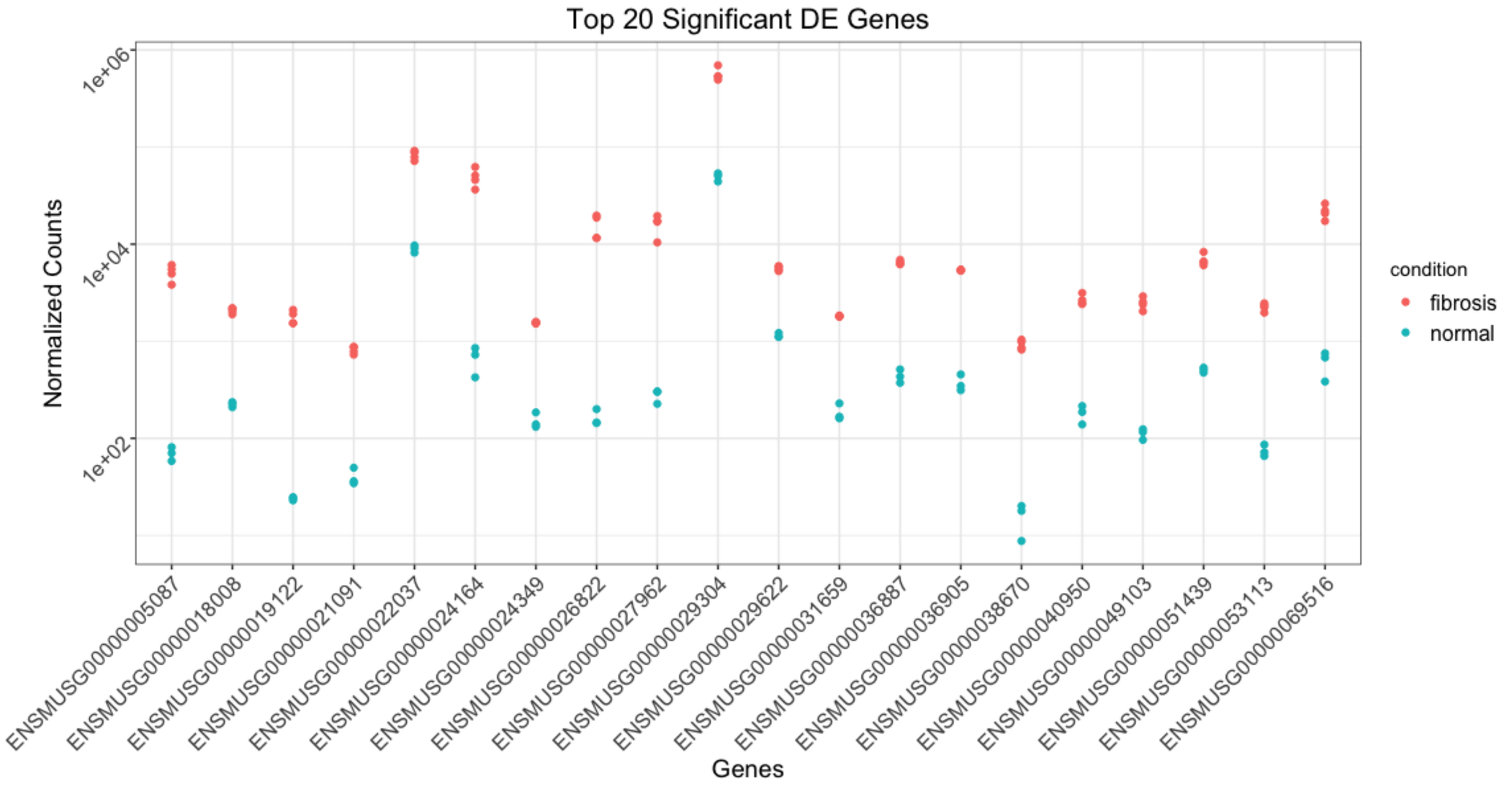
```
top_20 <- data.frame(sig_norm_counts_wt)[1:20, ] %>%
        rownames_to_column(var = "ensgene")
top_20 <- gather(top_20, key = "samplename", value = "normalized_counts", 2:8)
```

	ensgene	samplename	normalized_counts
1	ENSMUSG00000053113	wt_normal1	6.570386e+01
2	ENSMUSG00000005087	wt_normal1	8.103476e+01
3	ENSMUSG00000036887	wt_normal1	3.734169e+02
4	ENSMUSG00000026822	wt_normal1	1.456436e+02
5	ENSMUSG00000036905	wt_normal1	3.131884e+02
6	ENSMUSG00000027962	wt_normal1	2.266783e+02
7	ENSMUSG00000018008	wt_normal1	2.091573e+02
8	ENSMUSG00000051439	wt_normal1	5.343914e+02
9	ENSMUSG00000019122	wt_normal1	2.409141e+01
10	ENSMUSG00000049103	wt_normal1	9.636566e+01
11	ENSMUSG00000024164	wt_normal1	8.508650e+02
12	ENSMUSG00000022037	wt_normal1	8.170275e+03
13	ENSMUSG00000024349	wt_normal1	1.314077e+02
14	ENSMUSG00000029304	wt_normal1	4.423512e+04
15	ENSMUSG00000029622	wt_normal1	1.109300e+03
16	ENSMUSG00000031659	wt_normal1	1.609745e+02
17	ENSMUSG00000069516	wt_normal1	6.800349e+02
18	ENSMUSG00000021091	wt_normal1	3.613712e+01
19	ENSMUSG00000038670	wt_normal1	8.760514e+00
20	ENSMUSG00000040950	wt_normal1	1.390732e+02
21	ENSMUSG00000053113	wt_normal2	7.172182e+01
22	ENSMUSG00000005087	wt_normal2	7.034255e+01

Visualizing results - Expression plot

```
top_20 <- inner_join(top_20,
                     rownames_to_column(wt_metadata, var = "samplename"),
                     by = "samplename")
```

```
ggplot(top_20) +
        geom_point(aes(x = ensgene, y = normalized_counts, color = condition)) +
        scale_y_log10() +
        xlab("Genes") +
        ylab("Normalized Counts") +
        ggtitle("Top 20 Significant DE Genes") +
        theme_bw() +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        theme(plot.title = element_text(hjust = 0.5))
```


Top 20 Significant DE Genes
Normalized Counts
1e+06
1e+04
1e+02
Genes
ENSMUSG00000005087
ENSMUSG00000018008
ENSMUSG00000019122
ENSMUSG00000021091
ENSMUSG00000022037
ENSMUSG00000024164
ENSMUSG00000024349
ENSMUSG00000026822
ENSMUSG00000027962
ENSMUSG00000029304
ENSMUSG00000029622
ENSMUSG00000031659
ENSMUSG00000036887
ENSMUSG00000036905
ENSMUSG00000038670
ENSMUSG00000040950
ENSMUSG00000049103
ENSMUSG00000051439
ENSMUSG00000053113
ENSMUSG00000069516
condition
fibrosis
normal

Let's practice!

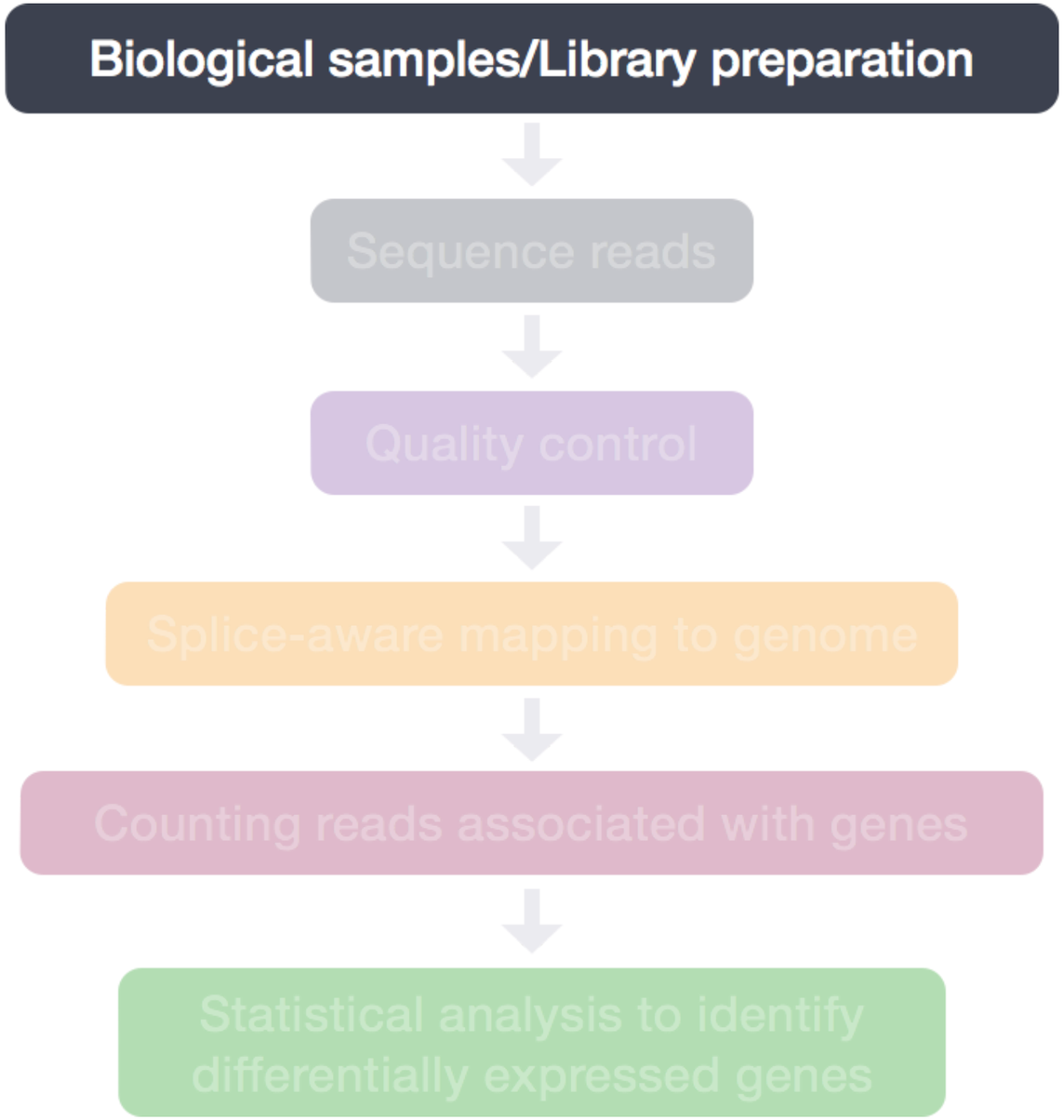
RNA-SEQ WITH BIOCONDUCTOR IN R

RNA-Seq DE analysis summary

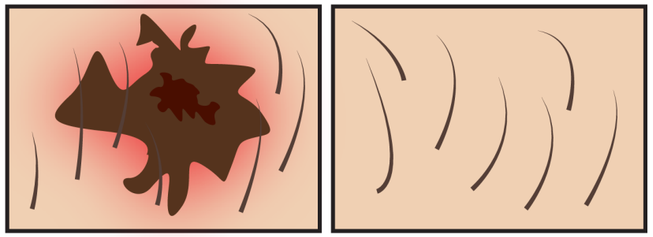
RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

datacamp


Biological samples/Library preparation
Sequence reads
Quality control
Splice-aware mapping to genome
Counting reads associated with genes
Statistical analysis to identify differentially expressed genes

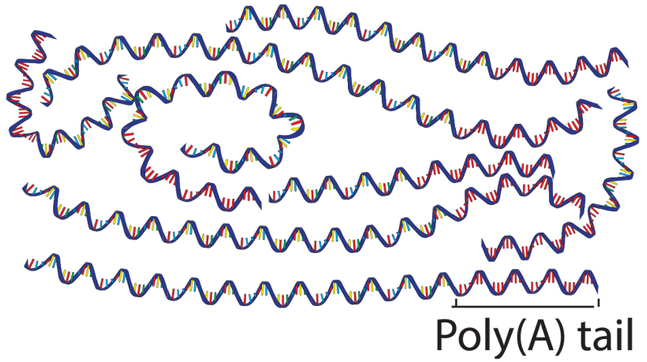
Samples of interest



Condition 1
(e.g. tumor)

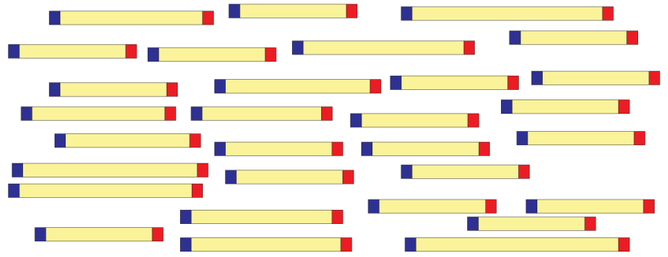
Condition 2
(e.g. normal)

Isolate RNAs

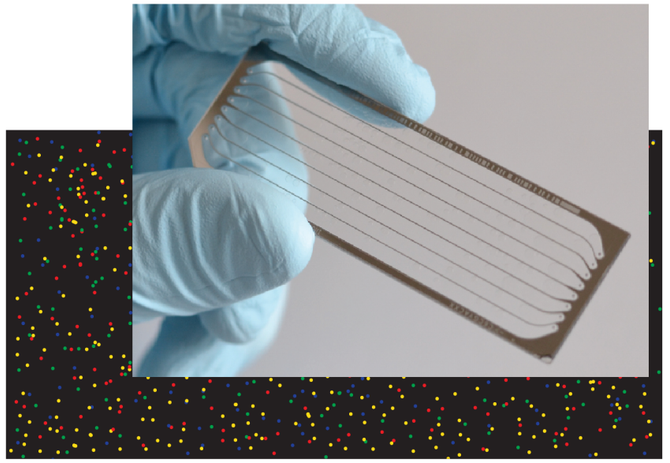


Poly(A) tail

Generate cDNA, fragment, size select, add linkers



Sequence ends

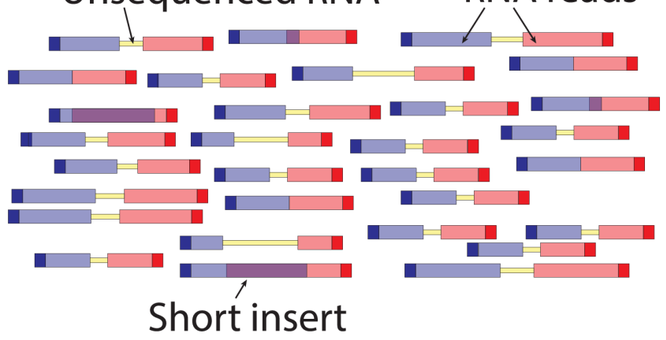


100s of millions of paired reads
10s of billions bases of sequence

FASTQ sequence files

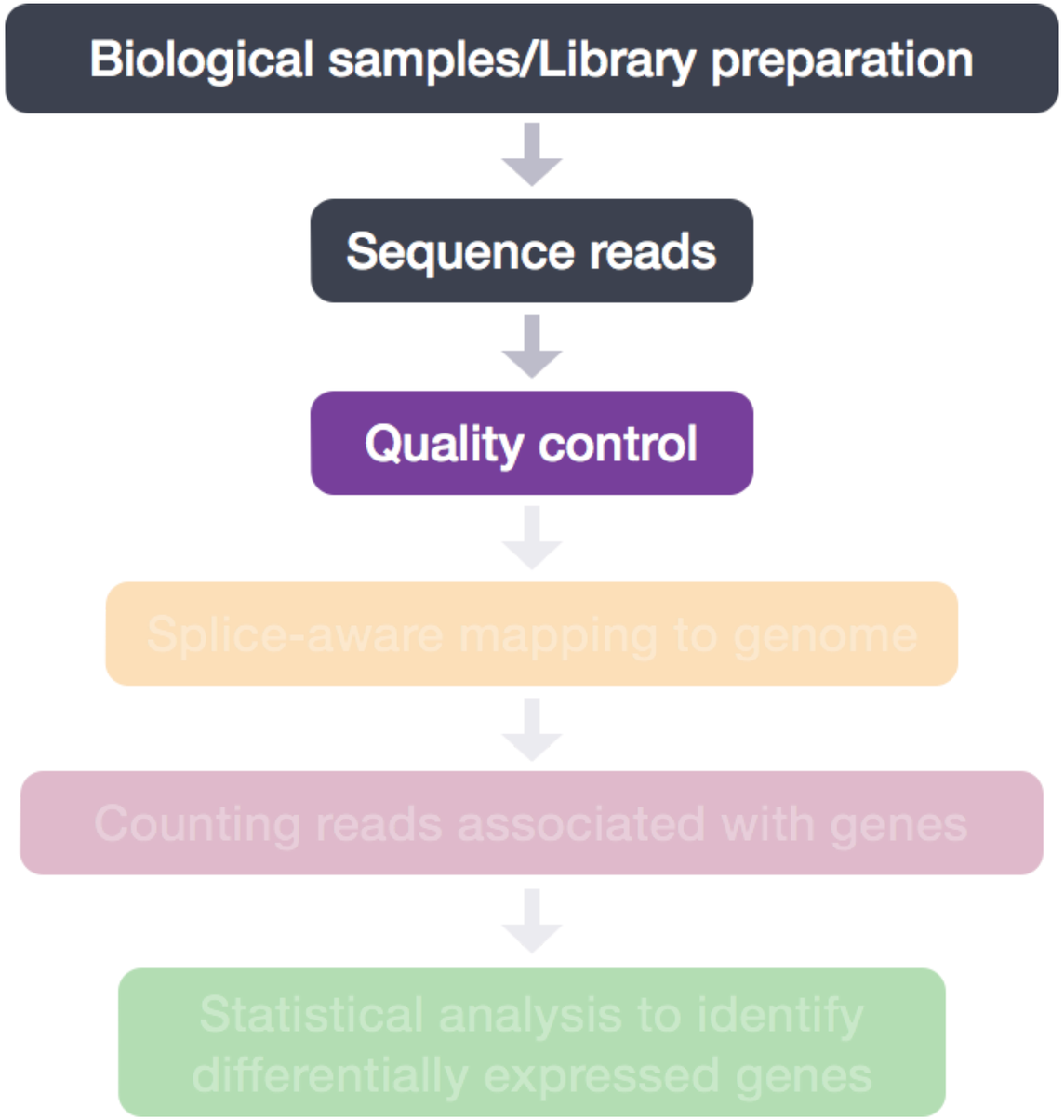
```
@HWI-ST330:304:H045HADXX:1:1101:1162:2055
NAGAACTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTAT
+
#1=DDFFFHHHHGHIJJJJIJJJGEGGAFGBHHEHGFBFFDEDECDDA==CB@BDDDDD?;B-<CBDDD>BBBBDDB5<@DDDC
@HWI-ST330:304:H045HADXX:2:2111:20110:84312
GTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGG
+
@@<FFFFDFFH>DEGFEGIJGJIJD9;CFCG;@;9?DDCD8AHGEF@84ADB?CD>3@CAACBBBDD@@@??90))5055(22
@HWI-ST330:304:H045HADXX:1:1214:9417:35291
CTCCAGACTCCGATCGTACAGCTTGAACTTCACATCTGAGGGCAGCAACGAGACCCCACGGGAGGCCACAGGAAAAAGCATGG
-bash-4.2$ head -n 100 Mov10_oe_1.subset.fq
@HWI-ST330:304:H045HADXX:1:1101:1162:2055
NAGAACTTGGCGGCGAATGGGCTGACCGCTTCCTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTAT
+
#1=DDFFFHHHHGHIJJJJIJJJGEGGAFGBHHEHGFBFFDEDECDDA==CB@BDDDDD?;B-<CBDDD>BBBBDDB5<@DDDC
@HWI-ST330:304:H045HADXX:2:2111:20110:84312
GTCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGG
+
@@<FFFFDFFH>DEGFEGIJGJIJD9;CFCG;@;9?DDCD8AHGEF@84ADB?CD>3@CAACBBBDD@@@??90))5055(22
@HWI-ST330:304:H045HADXX:1:1214:9417:35291
CTCCAGACTCCGATCGTACAGCTTGAACTTCACATCTGAGGGCAGCAACGAGACCCCACGGGAGGCCACAGGAAAAAGCATGG
+
BCCDFFDDHHHHAHHIJIHIJJJGA;9CDFBDGGIHGGBGHIB67;C;CH7@';@CA?B><B/;;??AB?9A<??<A089ACD
@HWI-ST330:304:H045HADXX:2:1212:4967:77898
AAAGCATGGGCCATAGCACCCAGCGCCCCGTCAACTCTAGGGCCCGGTGCTGGACCCGAAGGGGCTGGCGGTTGGAGGGAAAG
+
CCCFFFDFHHHHHJJJJJJJJJIJJJJJJFHIHJGIGIHEHIJJBE:=DFFECEEDDDDDDDDDBDDDDD09-&4(8&)<BB
@HWI-ST330:304:H045HADXX:2:2111:20388:84387
CTCATTTGCTCGGATCAGGTTAGCCAGATTGATATAAACATTTAGGTGGTTAGGGGCAATTCTGGCTGCATATTTTTTACCAG
+
=BBDDFFFHHGHFCEEEGGEHHIIBFE?BB??FC?FI?F>FEBHGI006.7?BCBFGG=D==CDHEA=C;?);@CEE?2>@C@
```

Unsequenced RNA / RNA reads



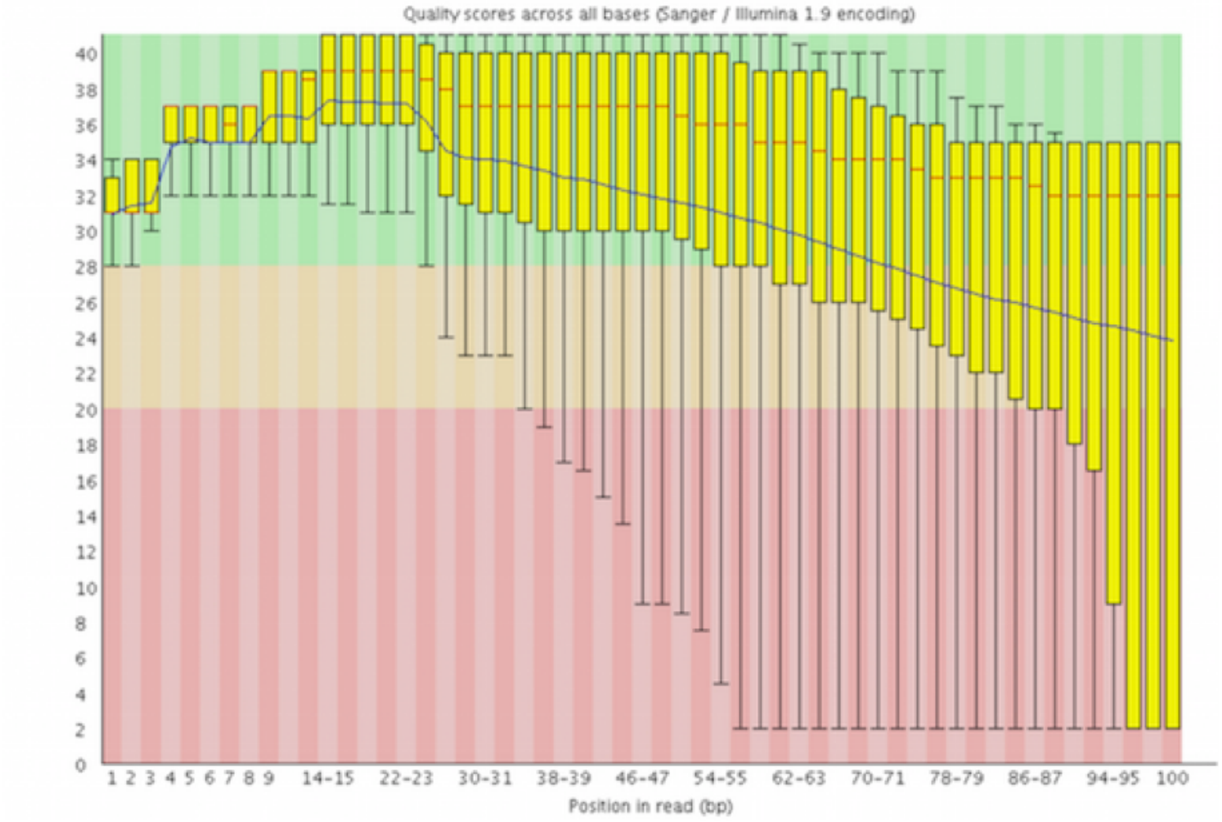
Short insert

Modified from https://doi.org/10.1371/journal.pcbi.1004393.g002


Biological samples/Library preparation
Sequence reads
Quality control
Splice-aware mapping to genome
Counting reads associated with genes
Statistical analysis to identify differentially expressed genes

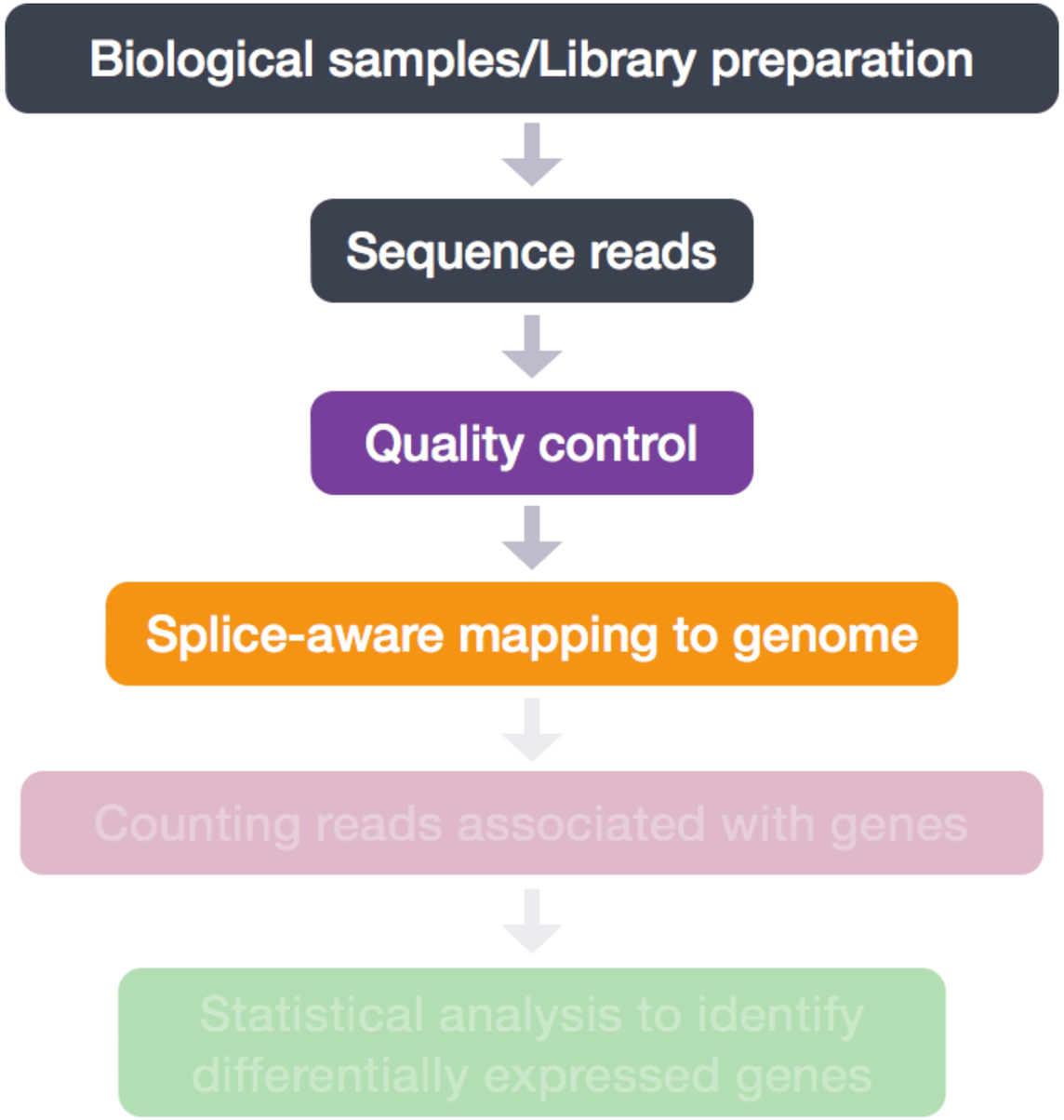
RNA-Seq Workflow: Raw data quality control

Per base sequence quality



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGCTATGGCCACCAGACTCTCAGGCTCCATGCAGTGGCCAGCCTCATCG	2554	0.8349133703824779	No Hit
CAGCGGTCTAGTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAG	2463	0.8051650866296176	No Hit
GTTTGAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATC	1920	0.6276560967636483	No Hit
CCACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATG	1219	0.39849624060150374	No Hit
GAAGAACCTGACCCGAGTCTTGGTGACGAAGGCCAGATTTGCGATCTTCA	1186	0.3877084014383786	No Hit
GGCAGGTGGACCCGGAGCCGCTGACAGAGGAGGTCAGCCCCTGAGTTGGA	1111	0.3631905851585486	No Hit
CACAGGGTCCCAGGTCATGGGTACCGAGTCCAGGTCATAGTGCCGGATGT	1079	0.35272965021248776	No Hit
GTCCCTGCTGCGGGGCCAGGAGACGGTAGATGGAGCTGGGCAGGTGGACCC	1036	0.3386727688787185	No Hit


Biological samples/Library preparation
Sequence reads
Quality control
Splice-aware mapping to genome
Counting reads associated with genes
Statistical analysis to identify differentially expressed genes

RNA-Seq Workflow: Alignment

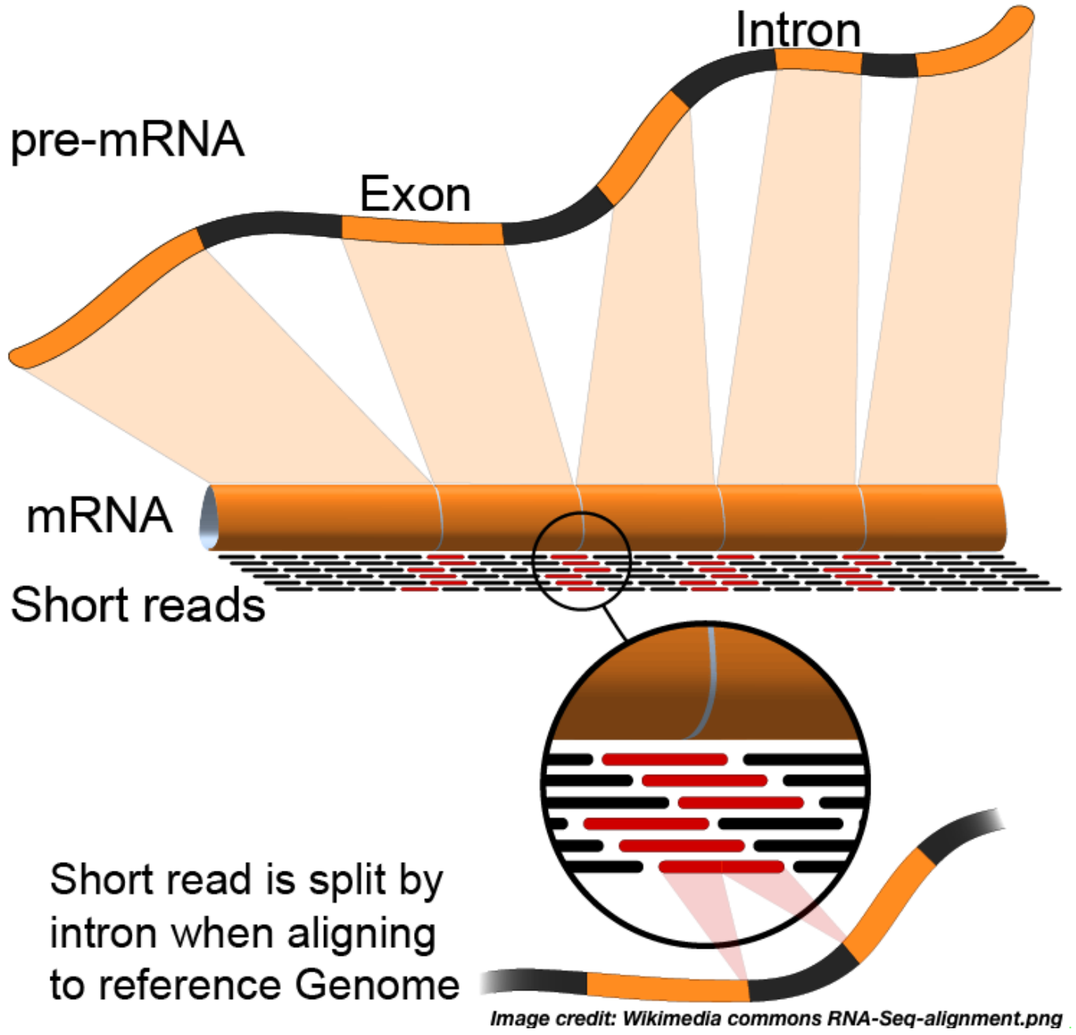
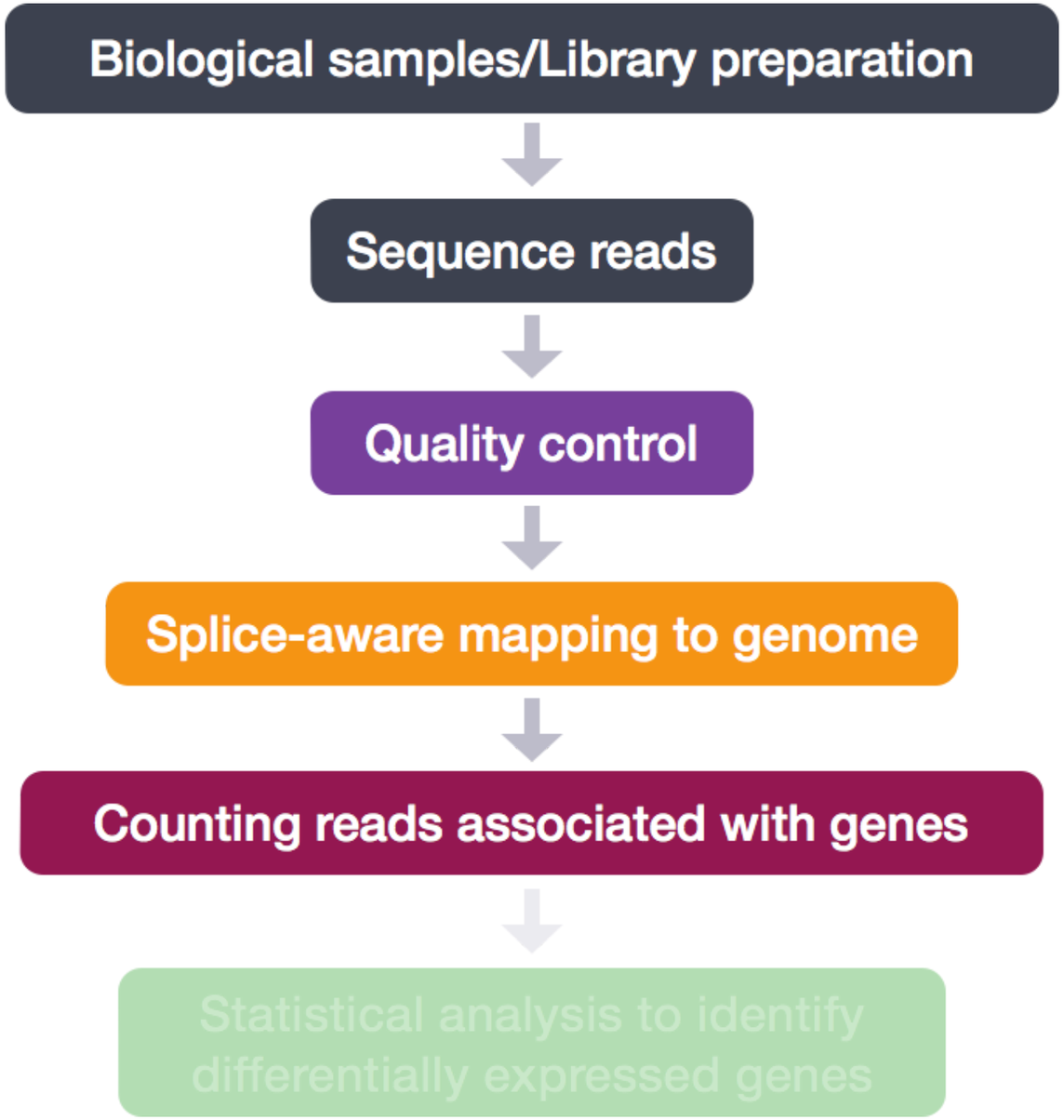
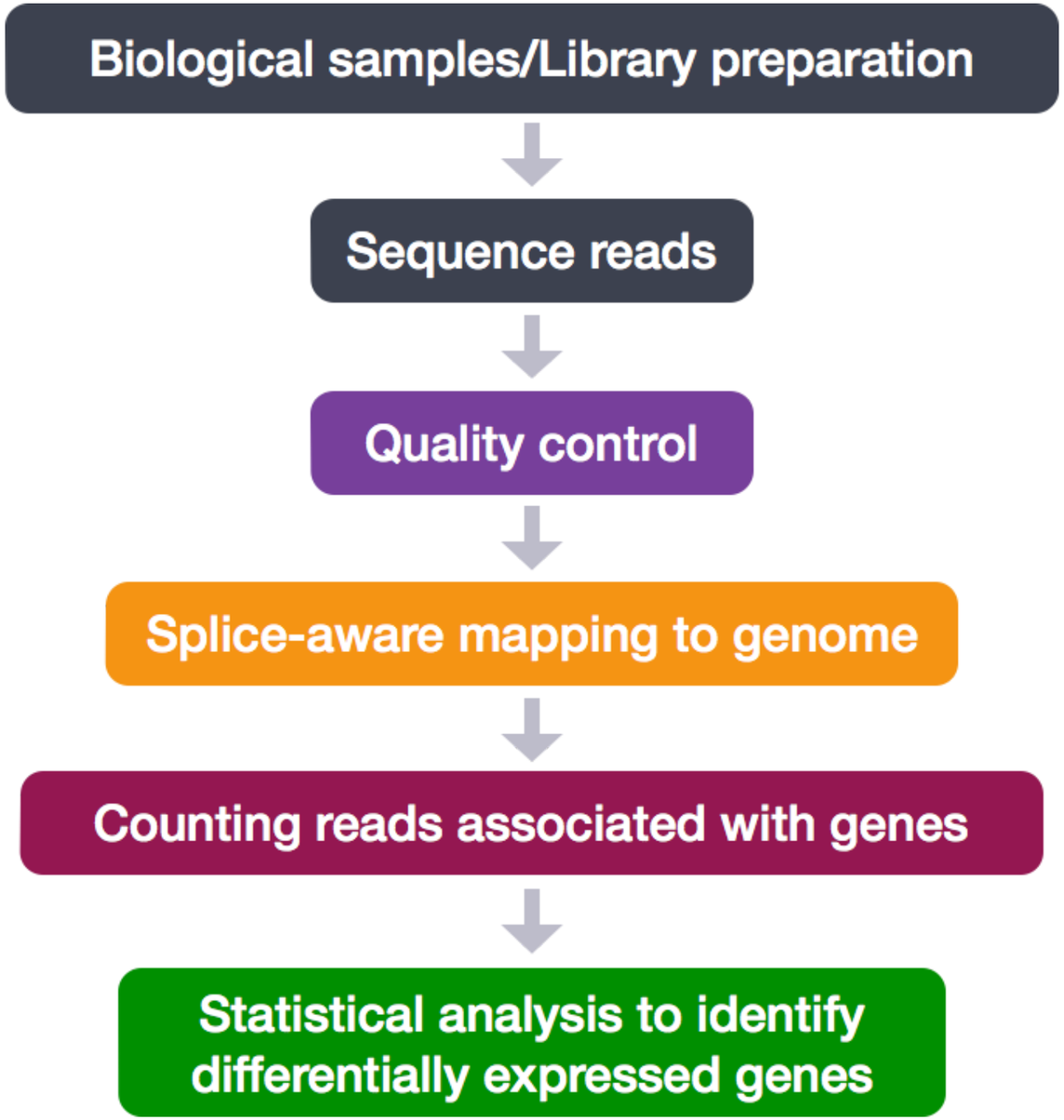


Image credit: Wikimedia commons RNA-Seq-alignment.png


Biological samples/Library preparation
Sequence reads
Quality control
Splice-aware mapping to genome
Counting reads associated with genes
Statistical analysis to identify differentially expressed genes

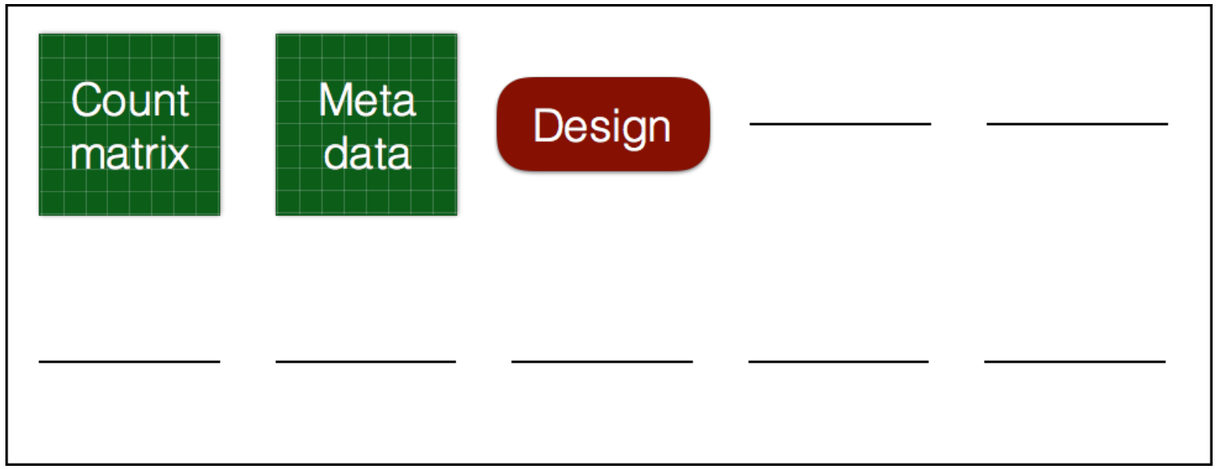
RNA-Seq Workflow: Quantitation

	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


Biological samples/Library preparation
Sequence reads
Quality control
Splice-aware mapping to genome
Counting reads associated with genes
Statistical analysis to identify differentially expressed genes

Preparation for differential expression analysis: DESeq2 object

```
dds <- DESeqDataSetFromMatrix(countData = rawcounts,
                              colData = metadata,
                              design = ~ condition)
```



Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R

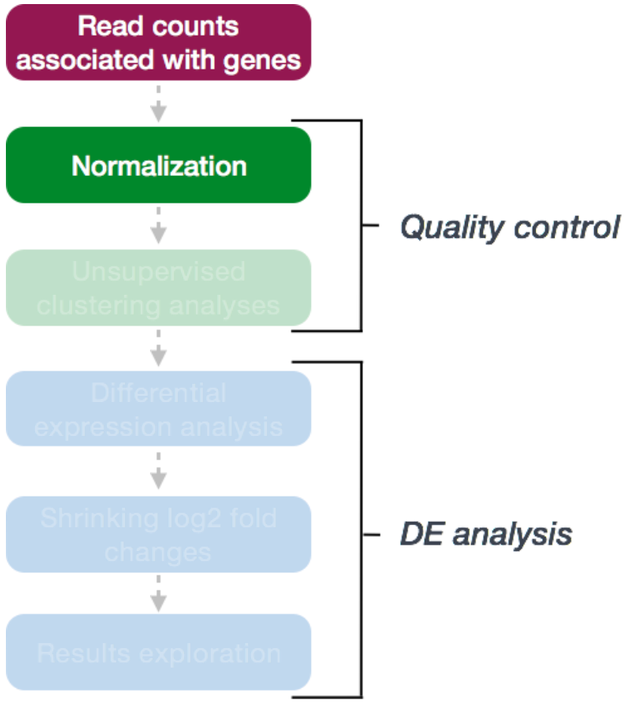
RNA-Seq DE analysis summary 2

RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

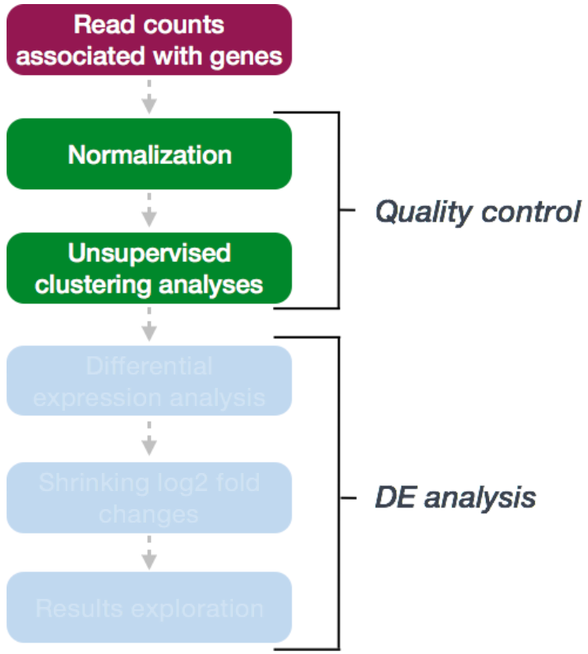
datacamp

DESeq workflow - normalization



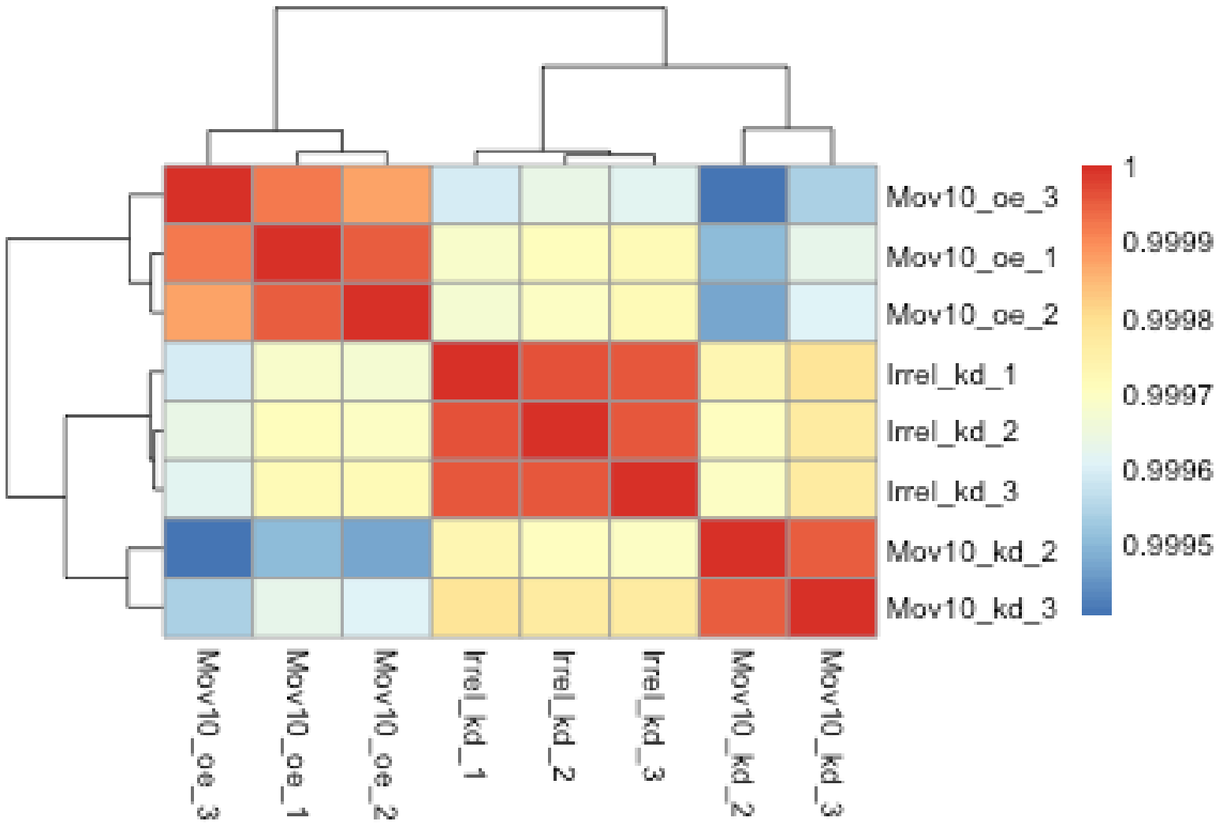

```
dds <- estimateSizeFactors(dds)
normalized_counts <- counts(dds, normalized=TRUE)
```

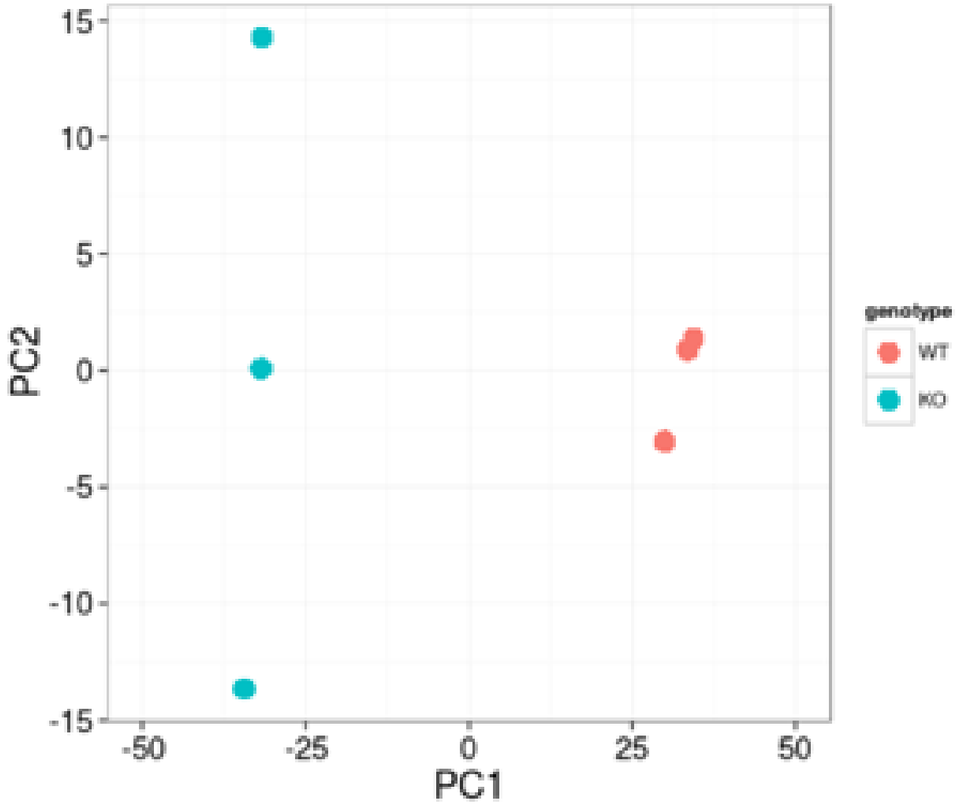
Unsupervised clustering analyses: log transformation




```
# Log transformation of normalized counts
vsd <- vst(dds, blind=TRUE)
```

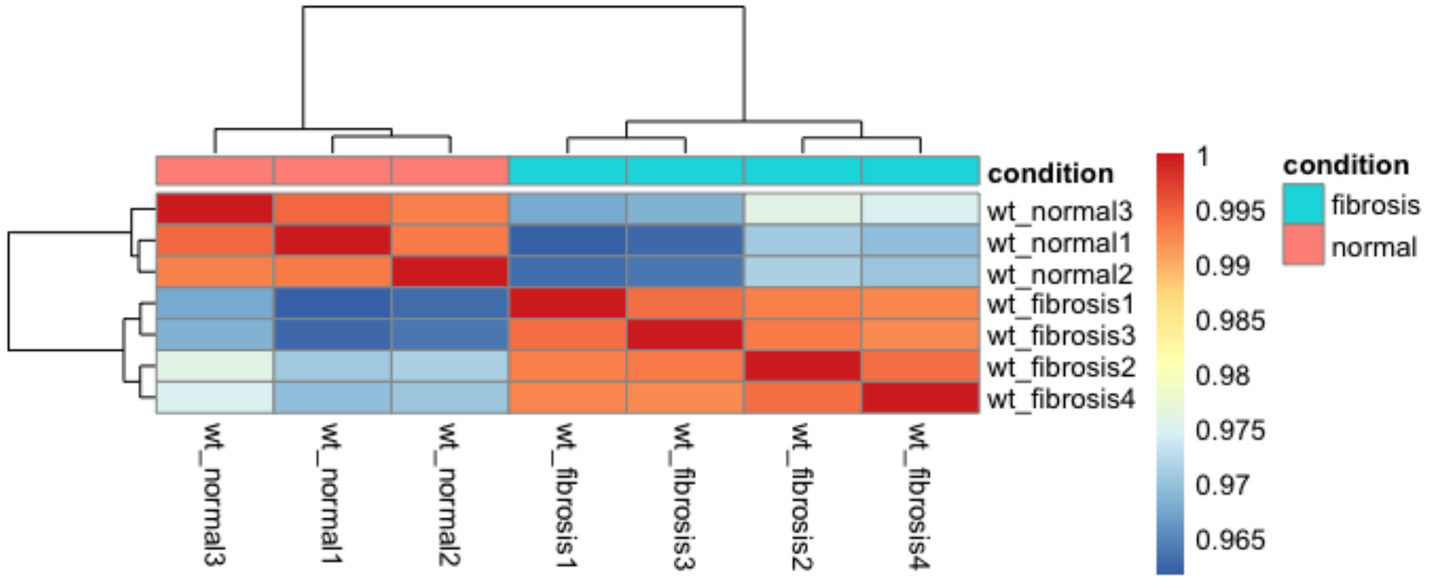
Unsupervised clustering analyses





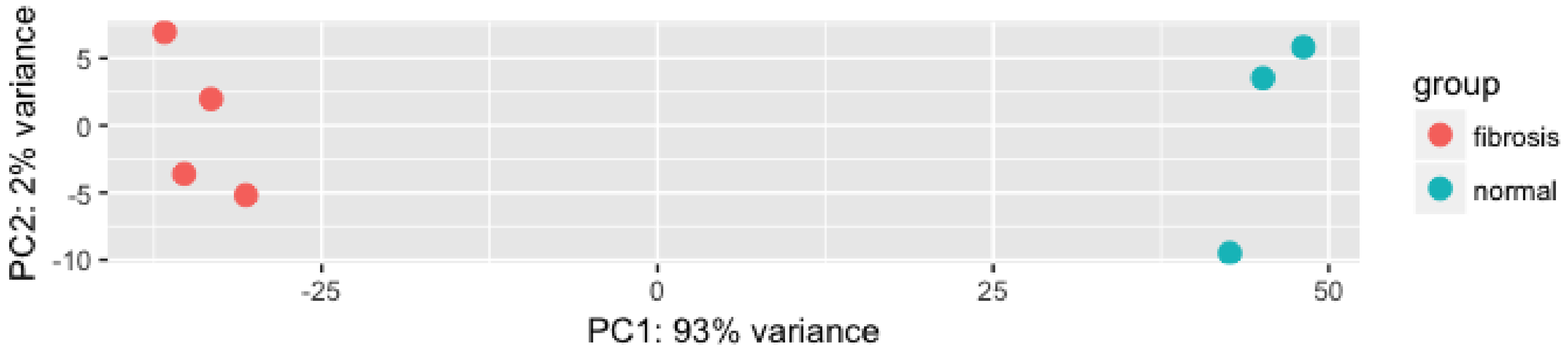
Unsupervised clustering analyses - heatmap

```
vsd %>%
    assay() %>% # Extract the vst matrix from the object
    cor() %>%   # Compute pairwise correlation values
    pheatmap(annotation = metadata[ , c("column_name1", "column_name2])
```

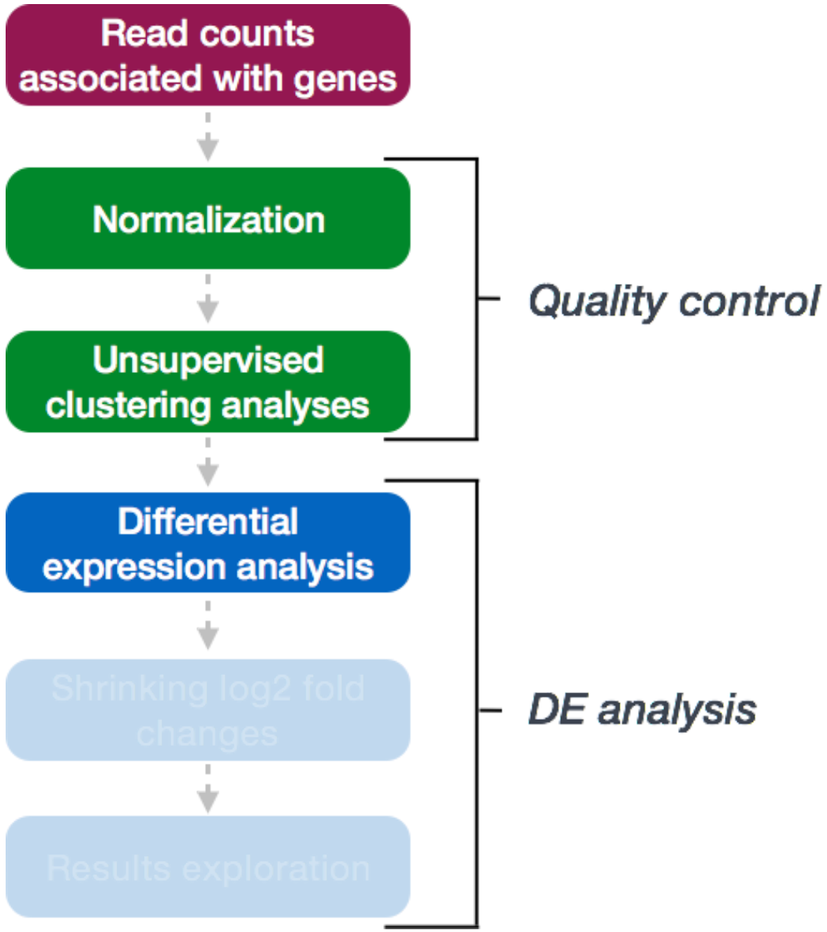


Unsupervised clustering analyses - pca

```
# PCA
plotPCA(vsd, intgroup="condition")
```



Running the DE analysis



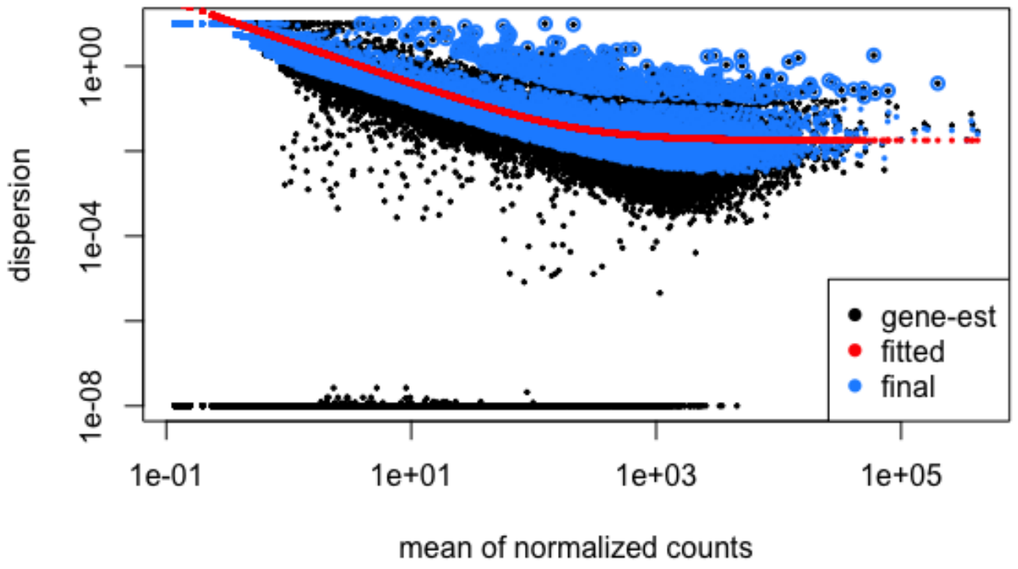
Running the DE analysis

```
# Create DESeq object
dds <- DESeqDataSetFromMatrix(countData = rawcounts,
                              colData = metadata,
                              design = ~ source_of_variation + condition)
```

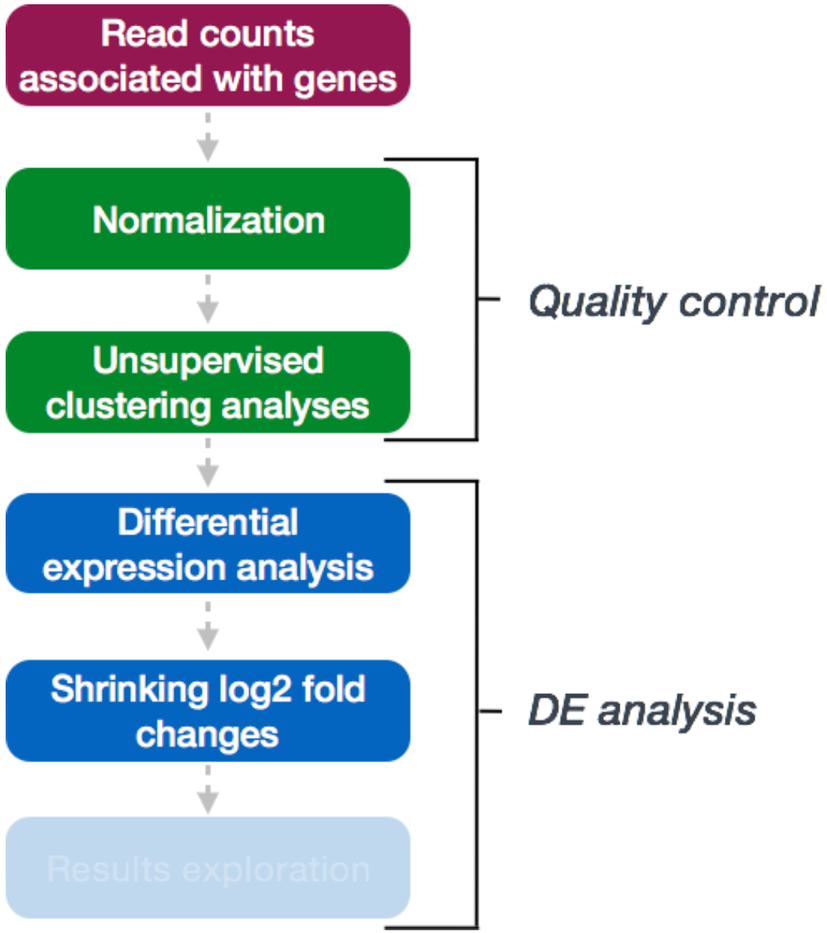
```
# Run analysis
dds <- DESeq(dds)
```

DESeq2 workflow - model

```
# Plot dispersion estimates
plotDispEsts(dds)
```



DESeq2 workflow - contrasts and LFC shrinkage



DESeq2 workflow - contrasts and LFC shrinkage

```
# Extract results for comparison of interest
res <- results(dds,
               contrast = c("condition_factor", "level_to_compare",
               "base_level"),
               alpha = 0.05)
```

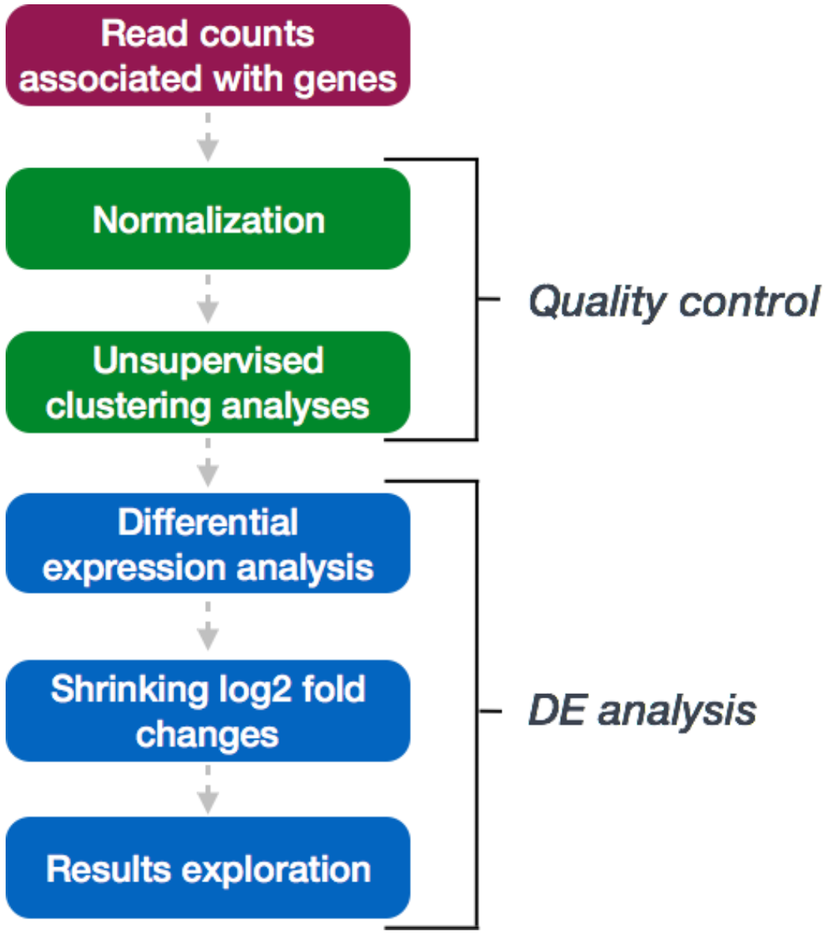
```
# Shrink the log2 foldchanges
res <- lfcShrink(dds,
               contrast = c("condition_factor", "level_to_compare",
               "base_level"),
               res = res)
```

DESeq2 workflow - LFC shrinkage

```
# Extract all results as a data frame
res_all <- data.frame(res) %>%
        rownames_to_column(var = "ensgene")
# Add gene annotations
res_all <- left_join(x = res_all,
                     y = grcm38[, c("ensgene", "symbol", "description")],
                     by = "ensgene")
res_all <- arrange(res_all, padj)
```

ensgene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	description
ENSMUSG00000053113	1318.1717	4.875042	0.16021506	28.35016	8.330958e-177	1.830145e-172	Socs3	suppressor of cytokine signaling 3 [Source:M...
ENSMUSG00000005087	2943.7403	6.121134	0.20721978	27.89891	2.750356e-171	3.020991e-167	Cd44	CD44 antigen [Source:MGI Symbol;Acc:MGI:8...
ENSMUSG00000036887	3899.5135	3.866162	0.12740248	27.83465	1.652344e-170	1.209957e-166	C1qa	complement component 1, q subcomponent,...
ENSMUSG00000026822	8870.1712	6.466148	0.23782361	25.82294	4.901029e-147	2.691645e-143	Lcn2	lipocalin 2 [Source:MGI Symbol;Acc:MGI:96757]
ENSMUSG00000036905	3237.6046	3.835279	0.13773926	25.52164	1.134018e-143	4.982421e-140	C1qb	complement component 1, q subcomponent,...
ENSMUSG00000027962	9298.5984	5.781446	0.21949603	24.88019	1.219153e-136	4.463724e-133	Vcam1	vascular cell adhesion molecule 1 [Source:MG...

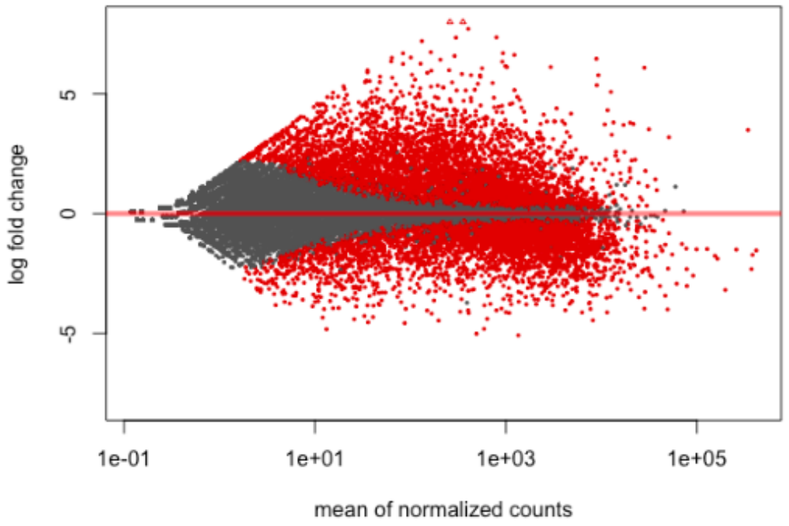
DESeq2 workflow - results exploration

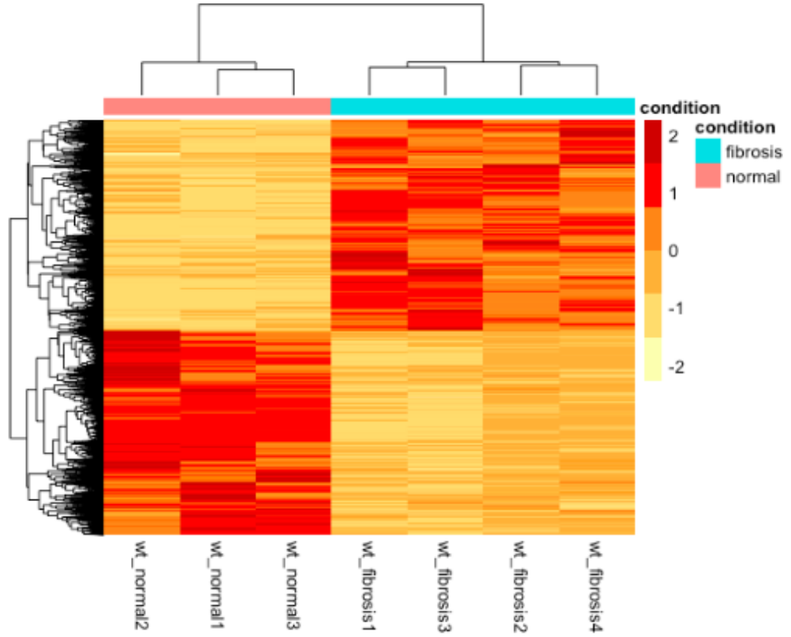


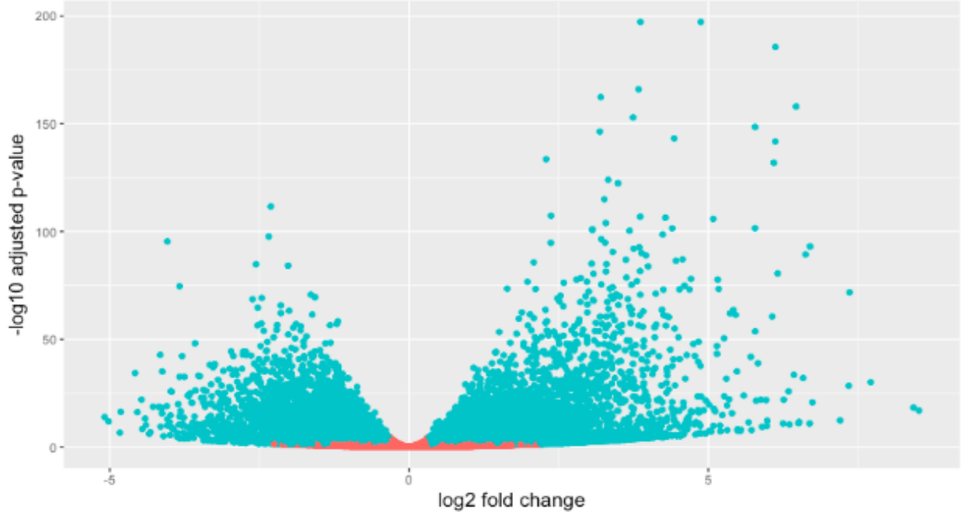
DESeq2 workflow - results exploration

```
# Identify significant genes
res_sig <- subset(res_all, padj < 0.05)
```

	ensgene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	description
1	ENSMUSG00000053113	1318.172	4.875042	0.1602151	28.35016	8.330958e-177	1.830145e-172	Socs3	suppressor of cytokine signaling 3 [Source:MGI Symbol;Acc:MGI:1201791]
2	ENSMUSG00000005087	2943.740	6.121134	0.2072198	27.89891	2.750356e-171	3.020991e-167	Cd44	CD44 antigen [Source:MGI Symbol;Acc:MGI:88338]
3	ENSMUSG00000036887	3899.514	3.866162	0.1274025	27.83465	1.652344e-170	1.209957e-166	C1qa	complement component 1, q subcomponent, alpha polypeptide [Source:MGI Symbol;Acc:MGI:88223]
4	ENSMUSG00000026822	8870.171	6.466148	0.2378236	25.82294	4.901029e-147	2.691645e-143	Lcn2	lipocalin 2 [Source:MGI Symbol;Acc:MGI:96757]
5	ENSMUSG00000036905	3237.605	3.835279	0.1377393	25.52164	1.134018e-143	4.982421e-140	C1qb	complement component 1, q subcomponent, beta polypeptide [Source:MGI Symbol;Acc:MGI:88224]
6	ENSMUSG00000027962	9298.598	5.781446	0.2194960	24.88019	1.219153e-136	4.463724e-133	Vcam1	vascular cell adhesion molecule 1 [Source:MGI Symbol;Acc:MGI:98926]


log fold change
5
0
-5
1e-01
1e+01
1e+03
1e+05
mean of normalized counts


condition
condition
fibrosis
normal
2
1
0
-1
-2
wt_normal2
wt_normal1
wt_normal3
wt_fibrosis1
wt_fibrosis3
wt_fibrosis2
wt_fibrosis4


200
150
100
50
0
-log10 adjusted p-value
-5
0
5
log2 fold change

Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R

RNA-Seq next steps

RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

datacamp

RNA-Seq next steps

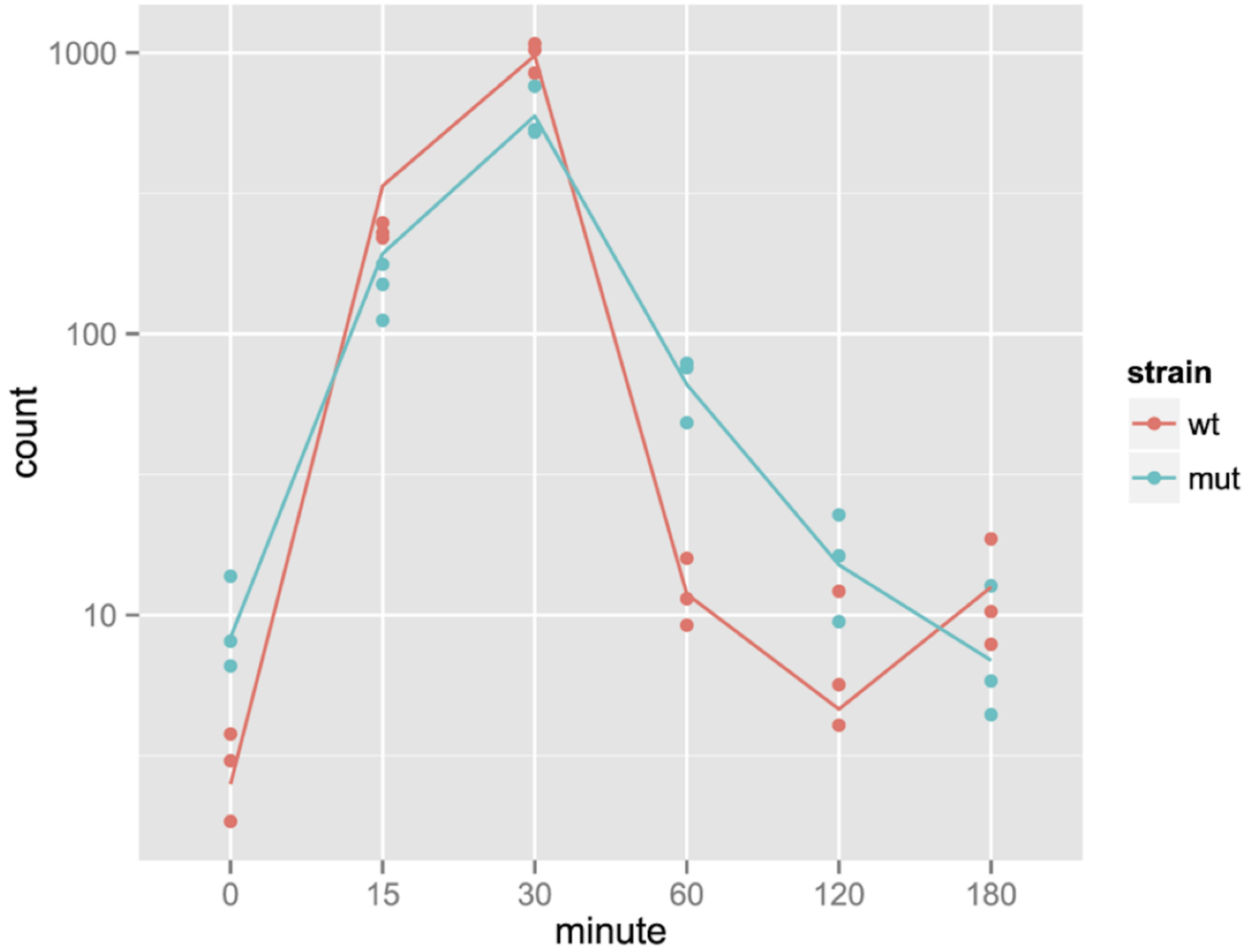
Vignette:

- https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Bioconductor support site:

- https://support.bioconductor.org (tag 'DESeq2')

Normalized counts for a gene with condition-specific changes over time.



Love MI, Anders S, Kim V and Huber W. RNA-Seq workflow: gene-level exploratory analysis and differential expression [version 2]. F1000Research 2016, 4:1070 (doi:

F1000Research

Overview of goals

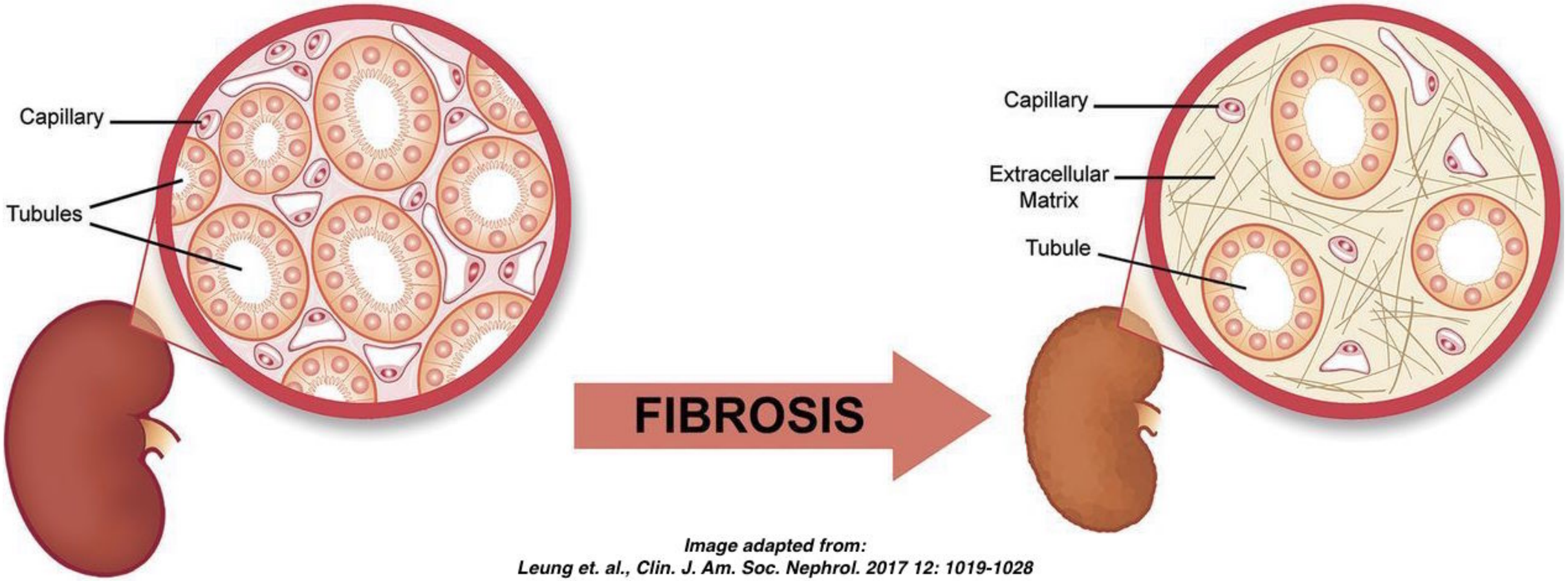


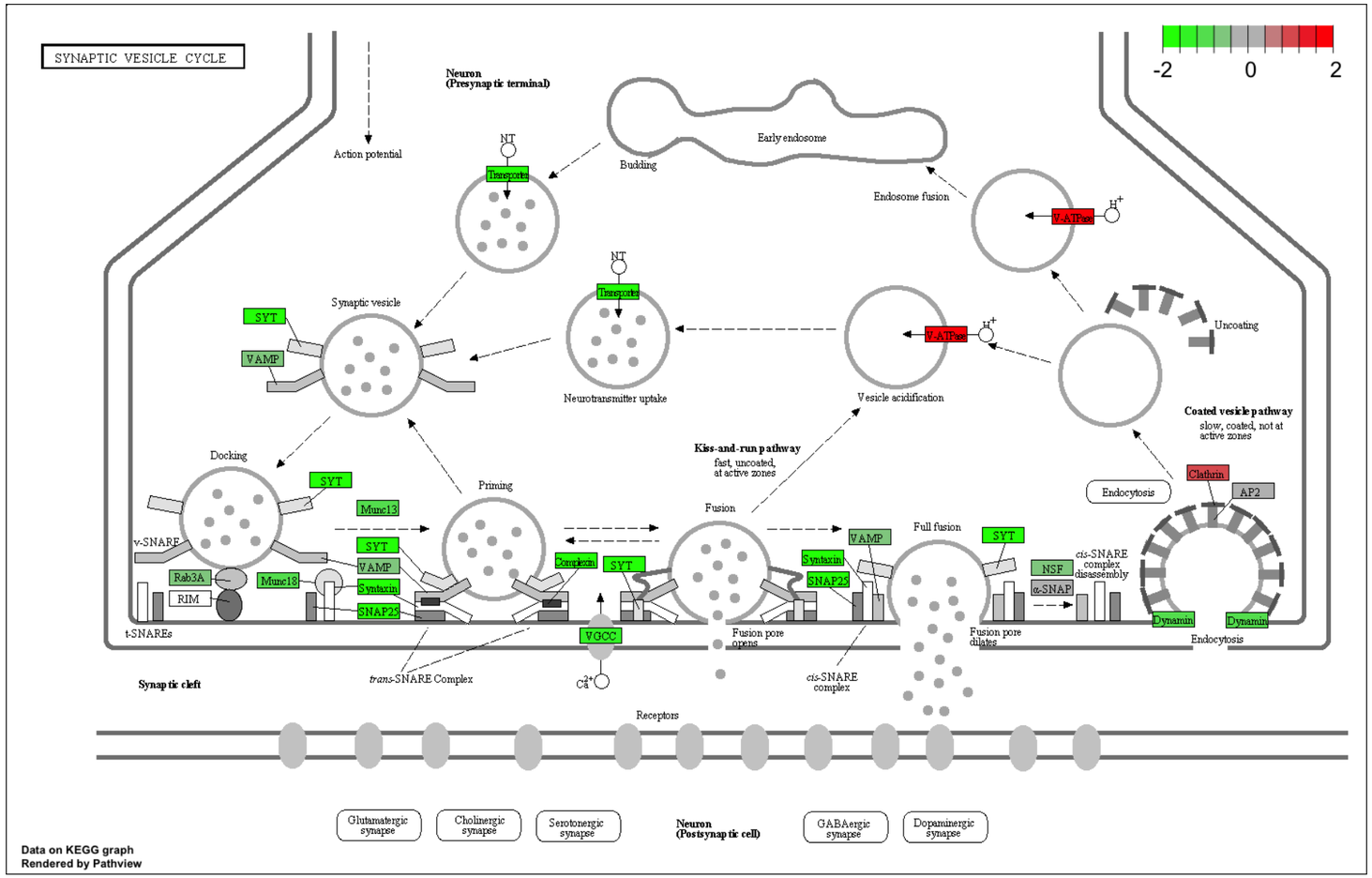
Image adapted from:
Leung et. al., Clin. J. Am. Soc. Nephrol. 2017 12: 1019-1028

Significant genes interpretation

ensgene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	description
ENSMUSG00000053113	1318.1717	4.875042	0.16021506	28.35016	8.330958e-177	1.830145e-172	Socs3	suppressor of cytokine signaling 3 [Source:M...
ENSMUSG00000005087	2943.7403	6.121134	0.20721978	27.89891	2.750356e-171	3.020991e-167	Cd44	CD44 antigen [Source:MGI Symbol;Acc:MGI:8...
ENSMUSG00000036887	3899.5135	3.866162	0.12740248	27.83465	1.652344e-170	1.209957e-166	C1qa	complement component 1, q subcomponent,...
ENSMUSG00000026822	8870.1712	6.466148	0.23782361	25.82294	4.901029e-147	2.691645e-143	Lcn2	lipocalin 2 [Source:MGI Symbol;Acc:MGI:96757]
ENSMUSG00000036905	3237.6046	3.835279	0.13773926	25.52164	1.134018e-143	4.982421e-140	C1qb	complement component 1, q subcomponent,...
ENSMUSG00000027962	9298.5984	5.781446	0.21949603	24.88019	1.219153e-136	4.463724e-133	Vcam1	vascular cell adhesion molecule 1 [Source:MG...

Significant genes interpretation

Significant genes interpretation



Conclusion

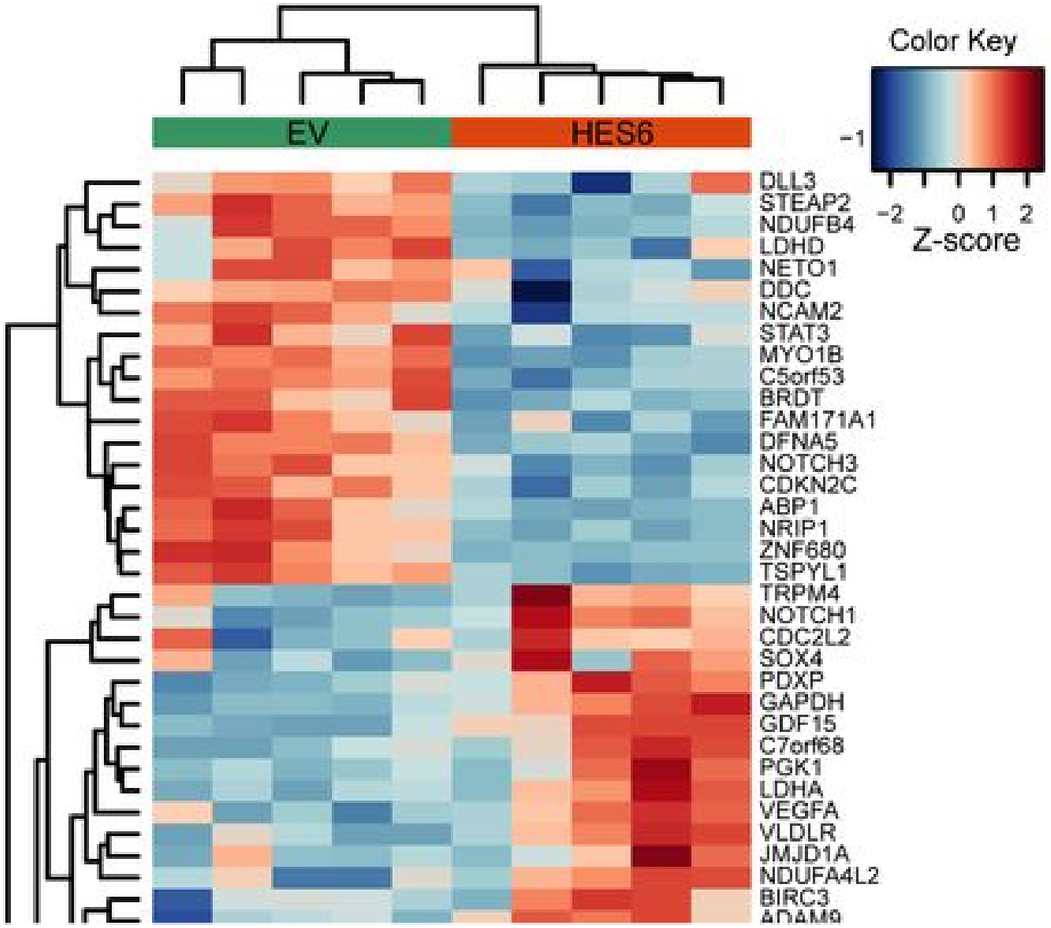


Image adapted from: Ramos-Montoya, A., et. al., EMBO Mol Med (2014) 6: 651–661.

Congratulations!

RNA-SEQ WITH BIOCONDUCTOR IN R