Overview of the DE analysis
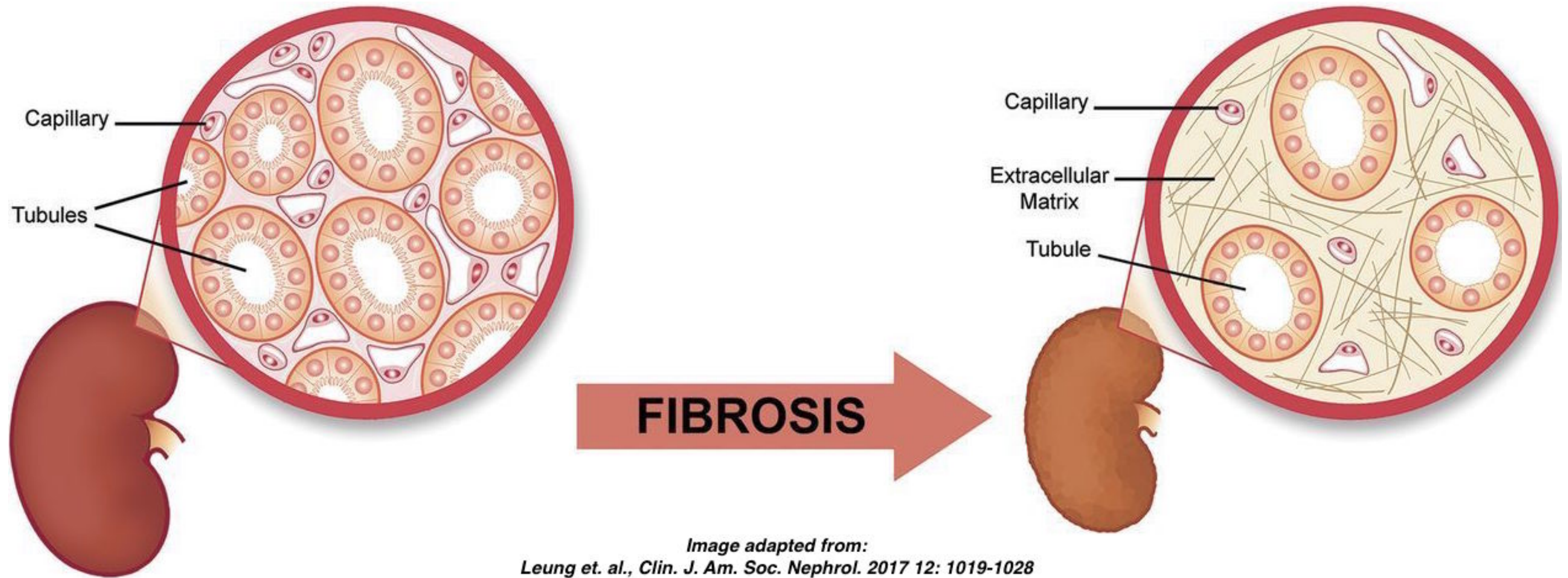
RNA-SEQ WITH BIOCONDUCTOR IN R
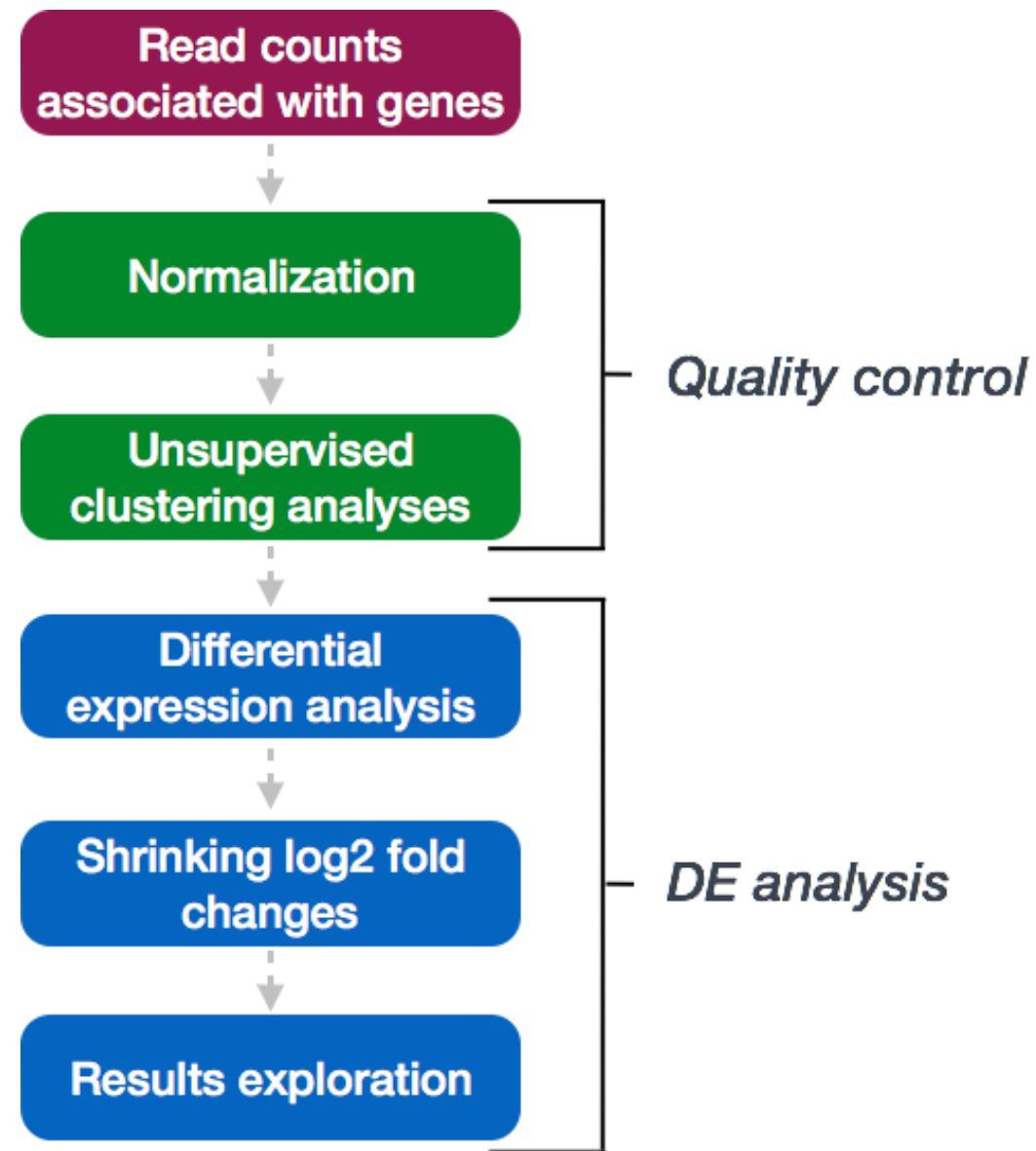
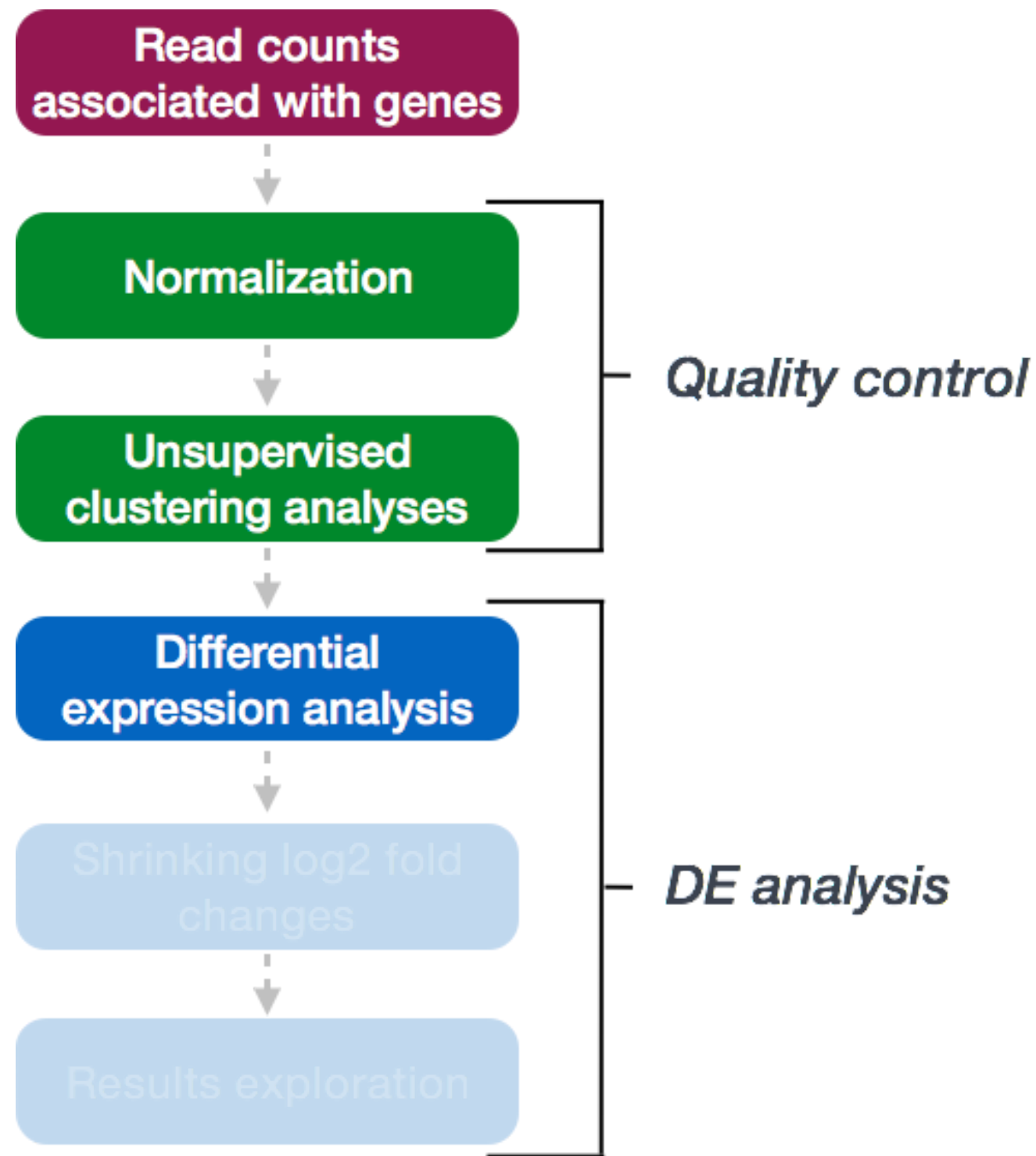**Mary Piper**
Bioinformatics Consultant and Trainer

Review the dataset/question

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

Overview of the DE analysis
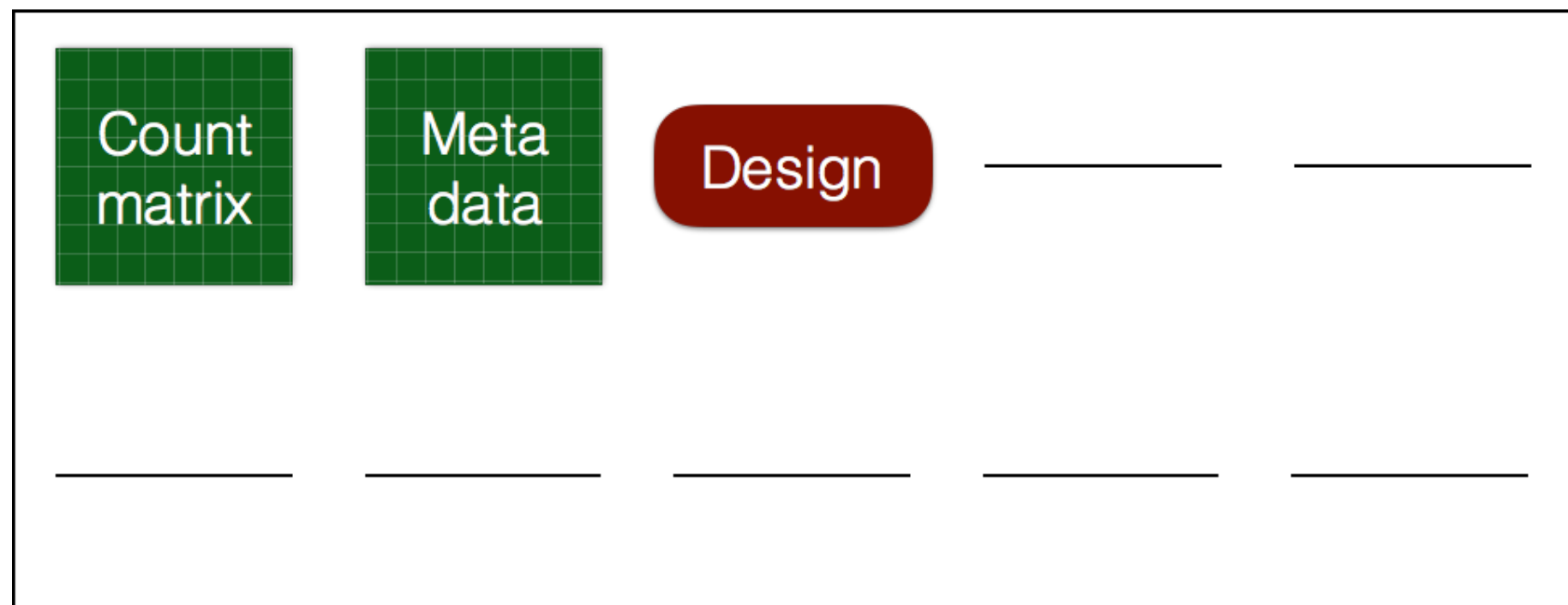
DSeq2 workflow: Model

```
# Create DESeq object

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

DSeq2 workflow: Design formula

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

```
# Design formula
~ strain + sex + treatment
```

DSeq2 workflow: Design formula

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

```
# Design formula
~ strain + sex + treatment + sex:treatment
```
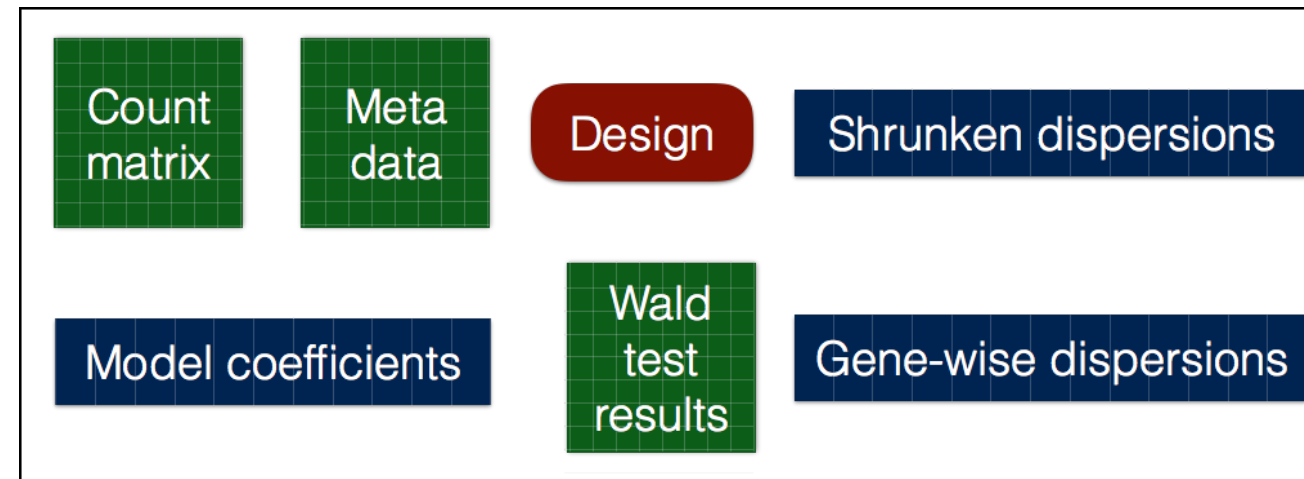
DSeq2 workflow: Running

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

```
using pre-existing size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
```

Let's practice!
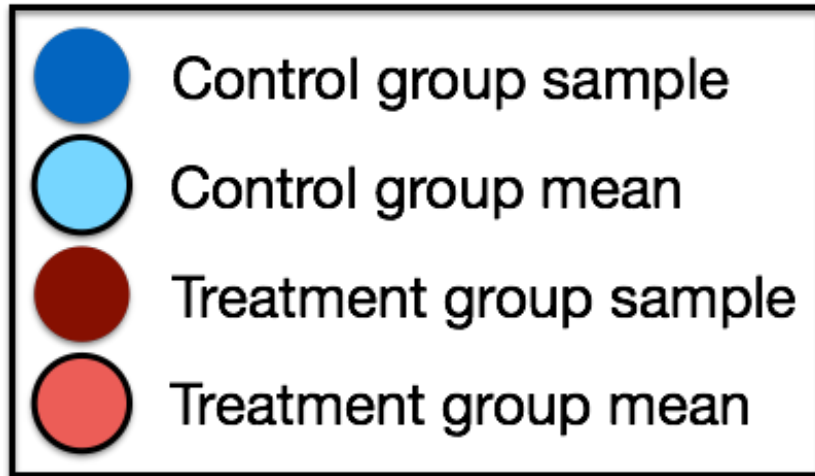
RNA-SEQ WITH BIOCONDUCTOR IN R

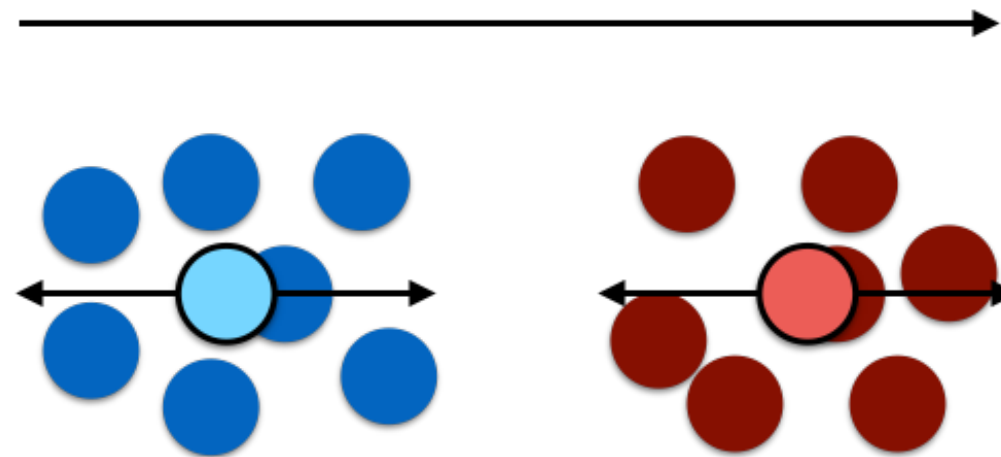# DESeq2 model

RNA-SEQ WITH BIOCONDUCTOR IN R

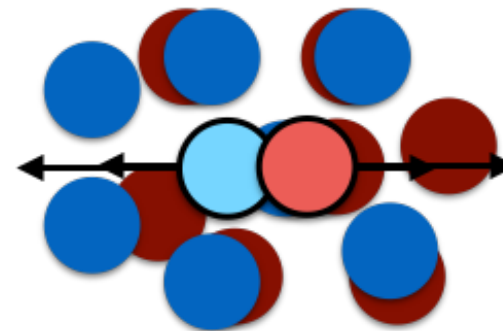**Mary Piper**
Bioinformatics Consultant and Trainer

DESeq2 model

Expression level for Gene A

$$\text{Log2} \frac{\text{treatment mean}}{\text{control mean}} \neq 0$$

$$\text{Log2} \frac{\text{treatment mean}}{\text{control mean}} \simeq 0$$

Legend:
- Control group sample
- Control group mean
- Treatment group sample
- Treatment group mean

DESeq2 model - mean-variance relationship

```r
# Syntax for apply()
apply(data, rows/columns, function_to_apply)
```

```r
# Calculating mean for each gene (each row)
mean_counts <- apply(wt_rawcounts[, 1:3], 1, mean)
```

```r
# Calculating variance for each gene (each row)
variance_counts <- apply(wt_rawcounts[, 1:3], 1, var)
```
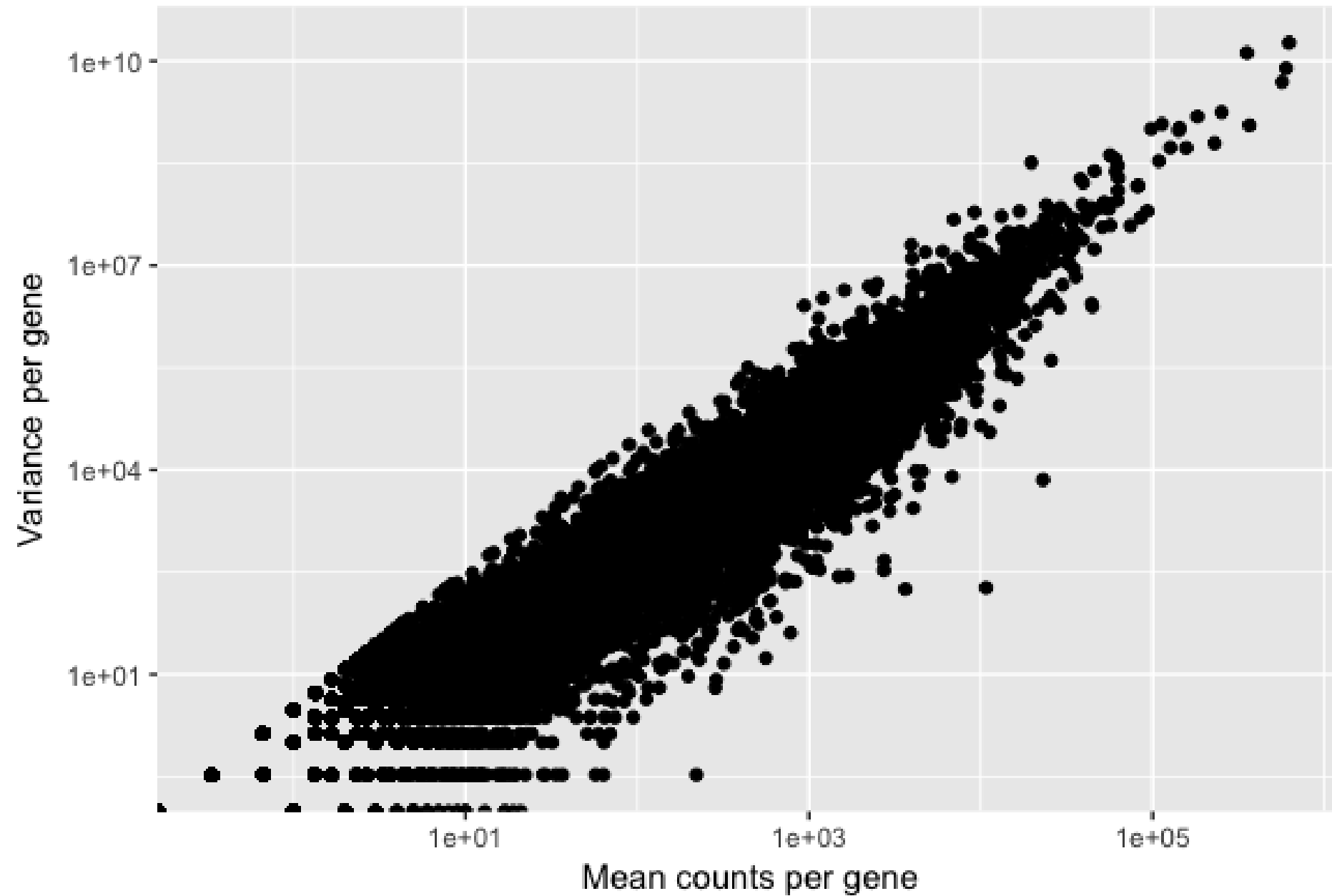
DESeq2 model - dispersion

Plotting relationship between mean and variance:

```r
# Creating data frame with mean and variance for every gene
df <- data.frame(mean_counts, variance_counts)
```

```r
ggplot(df) +
        geom_point(aes(x=mean_counts, y=variance_counts)) +
        scale_y_log10() +
        scale_x_log10() +
        xlab("Mean counts per gene") +
        ylab("Variance per gene")
```

DESeq2 model - dispersion

DESeq2 model - dispersion

Var: variance

μ: mean

α: dispersion

Dispersion formula: $Var = \mu + \alpha * \mu^2$

Relationship between mean, variance and dispersion:

$$\uparrow variance \Rightarrow \uparrow dispersion$$

$$\uparrow mean \Rightarrow \downarrow dispersion$$

DESeq2 model - dispersion

```
# Plot dispersion estimates
plotDispEsts(dds_wt)
```
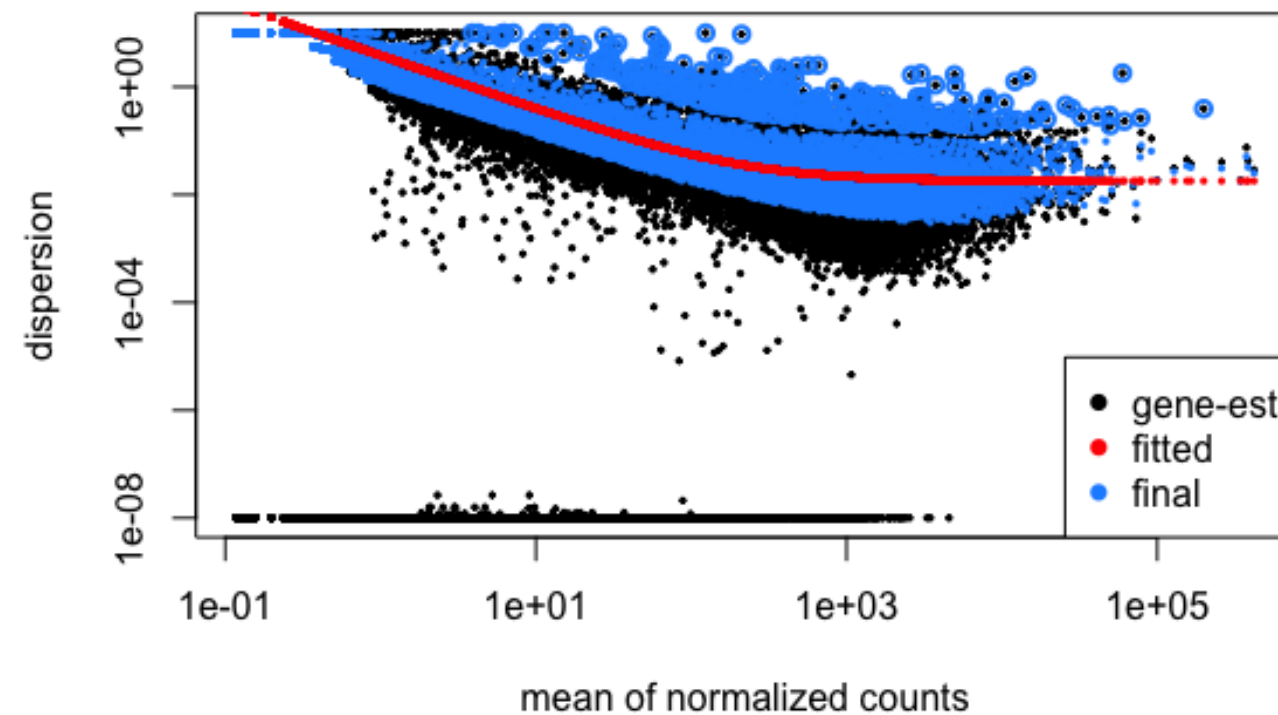
DESeq2 model - dispersion
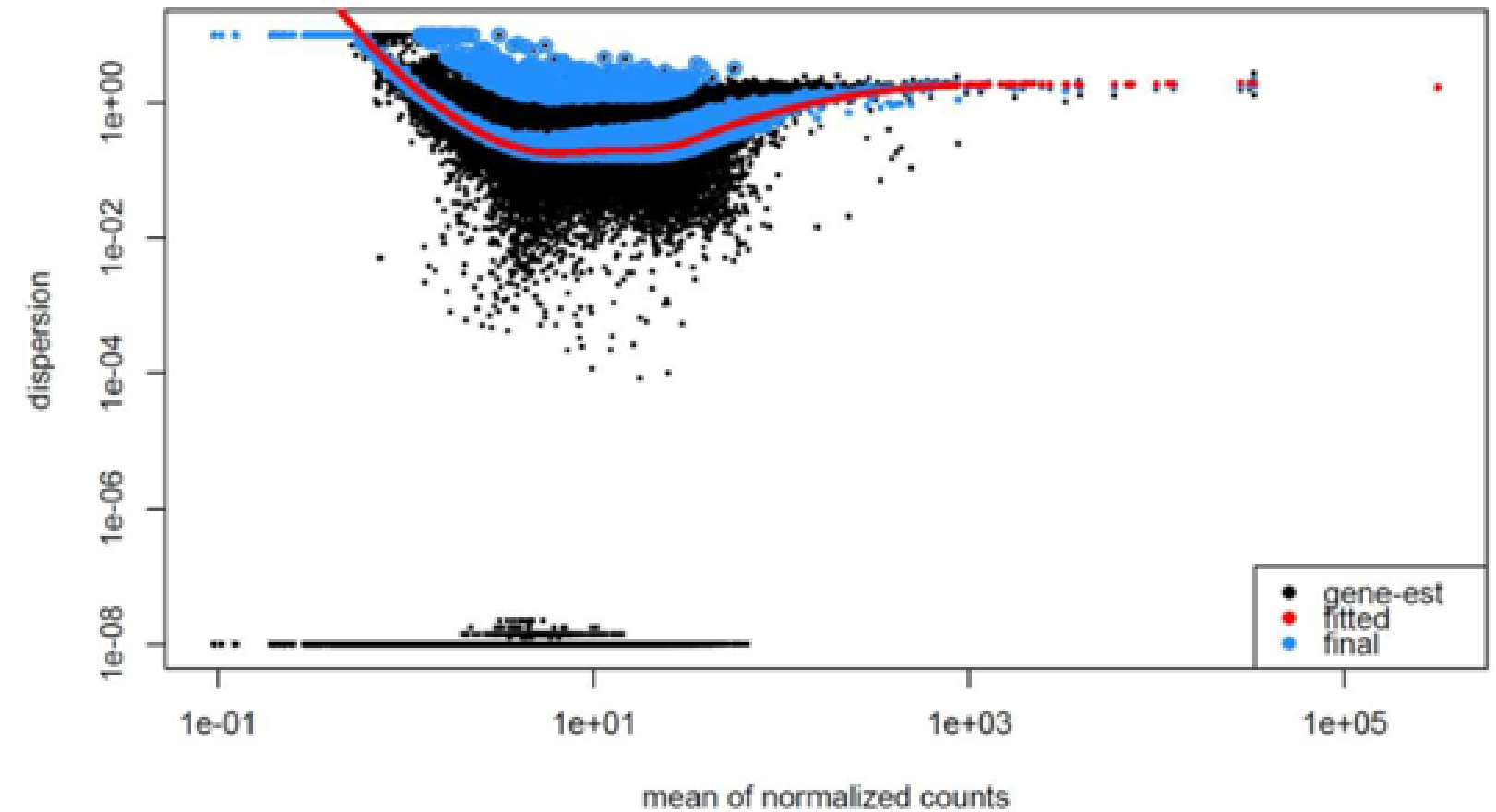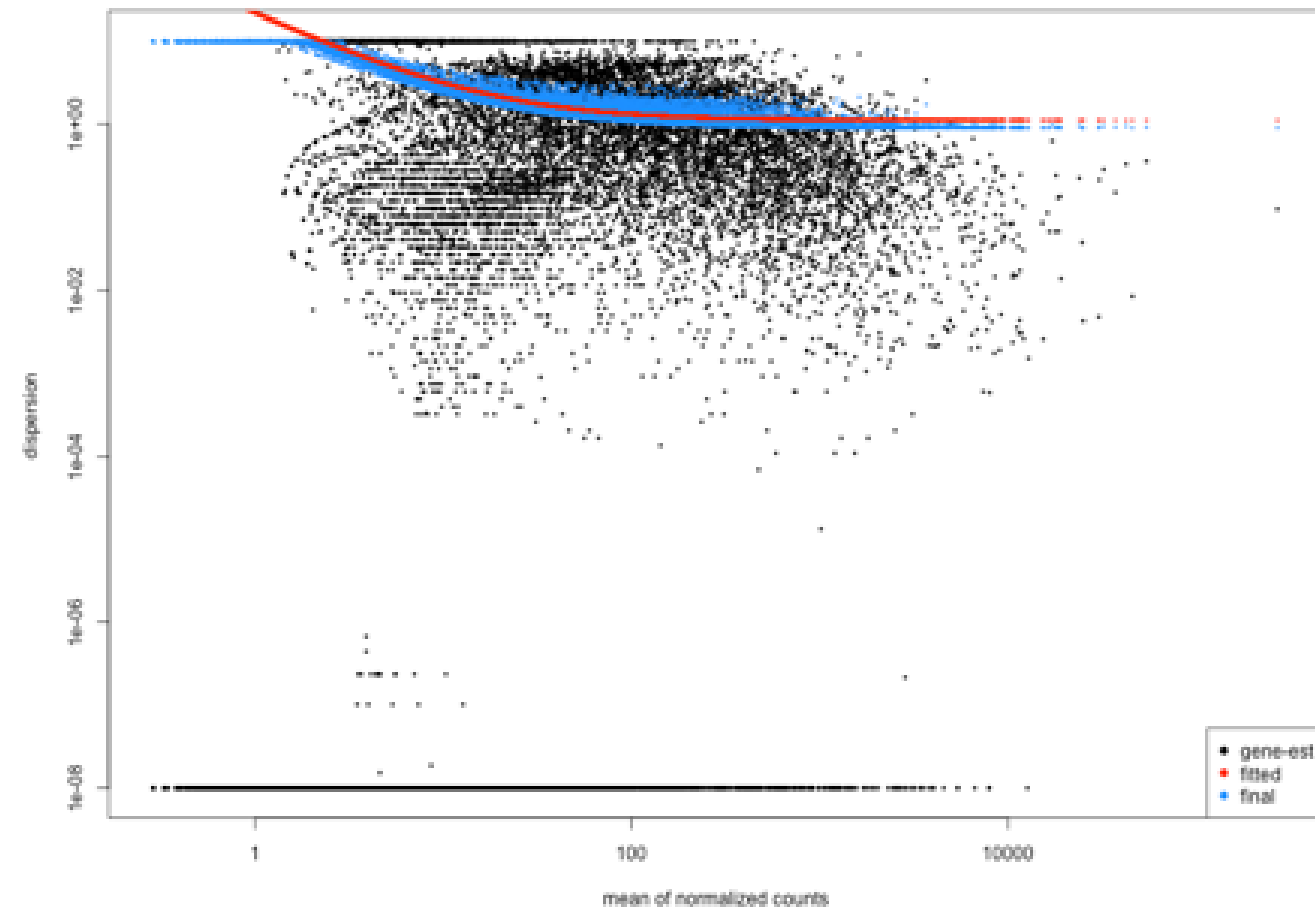
Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R

DESeq2 model - contrasts

RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

DESEq2 workflow
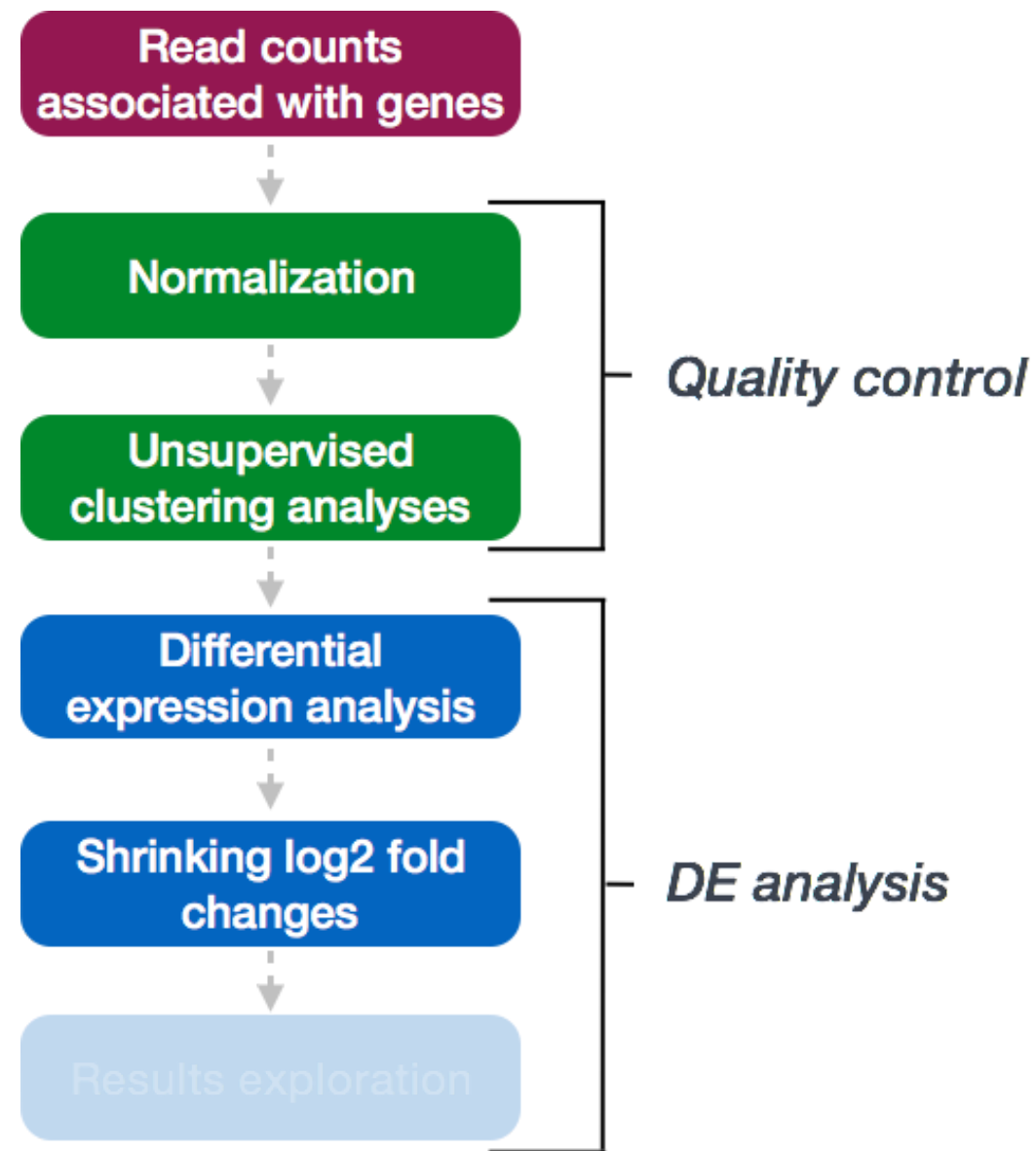
DESeq2 workflow

```r
# Run analysis
dds_wt <- DESeq(dds_wt)
```

```
using pre-existing size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
```

DESeq2 Negative Binomial Model

raw count for gene i, sample j

The mean is taken as "normalized counts" scaled by a normalization factor

one dispersion per gene

$$K_{ij} \sim \mathrm{NB}(s_{ij}q_{ij}, \alpha_i)$$

DESeq2 Negative Binomial Model

normalized counts for gene i, sample j

log2 fold change between conditions

$$\log 2\, q_{ij} = \Sigma_r X_{jr} \beta_{ir}$$

DESeq2 contrasts

```
results(wt_dds, alpha = 0.05)
```

```
log2 fold change (MLE): condition normal vs fibrosis
Wald test p-value: condition normal vs fibrosis
DataFrame with 47729 rows and 6 columns
```

	baseMean	log2FoldChange	lfcSE
	<numeric>	<numeric>	<numeric>
ENSMUSG00000102693	0	NA	NA
ENSMUSG00000064842	0	NA	NA
ENSMUSG00000051951	19.5084656230804	-4.1291703663434	0.815892075046039
ENSMUSG00000102851	0	NA	NA
ENSMUSG00000103377	0	NA	NA

DESeq2 contrasts

The syntax is:

```
results(dds,
        contrast = c("condition_factor", "level_to_compare",
        "base_level"),
        alpha = 0.05)
```

```
wt_res <- results(dds_wt,
        contrast = c("condition", "fibrosis",
        "normal"),
        alpha = 0.05)
```

DESeq2 contrasts

wt_res

```
log2 fold change (MLE): condition fibrosis vs normal
Wald test p-value: condition fibrosis vs normal
DataFrame with 47729 rows and 6 columns
```

	baseMean	log2FoldChange	lfcSE
	<numeric>	<numeric>	<numeric>
ENSMUSG00000102693	0	NA	NA
ENSMUSG00000064842	0	NA	NA
ENSMUSG00000051951	19.5084656230804	4.1291703663434	0.815892075046039
ENSMUSG00000102851	0	NA	NA
ENSMUSG00000103377	0	NA	NA
...	...	...	...

DESeq2 LFC shrinkage

```
plotMA(wt_res, ylim=c(-8,8))
```
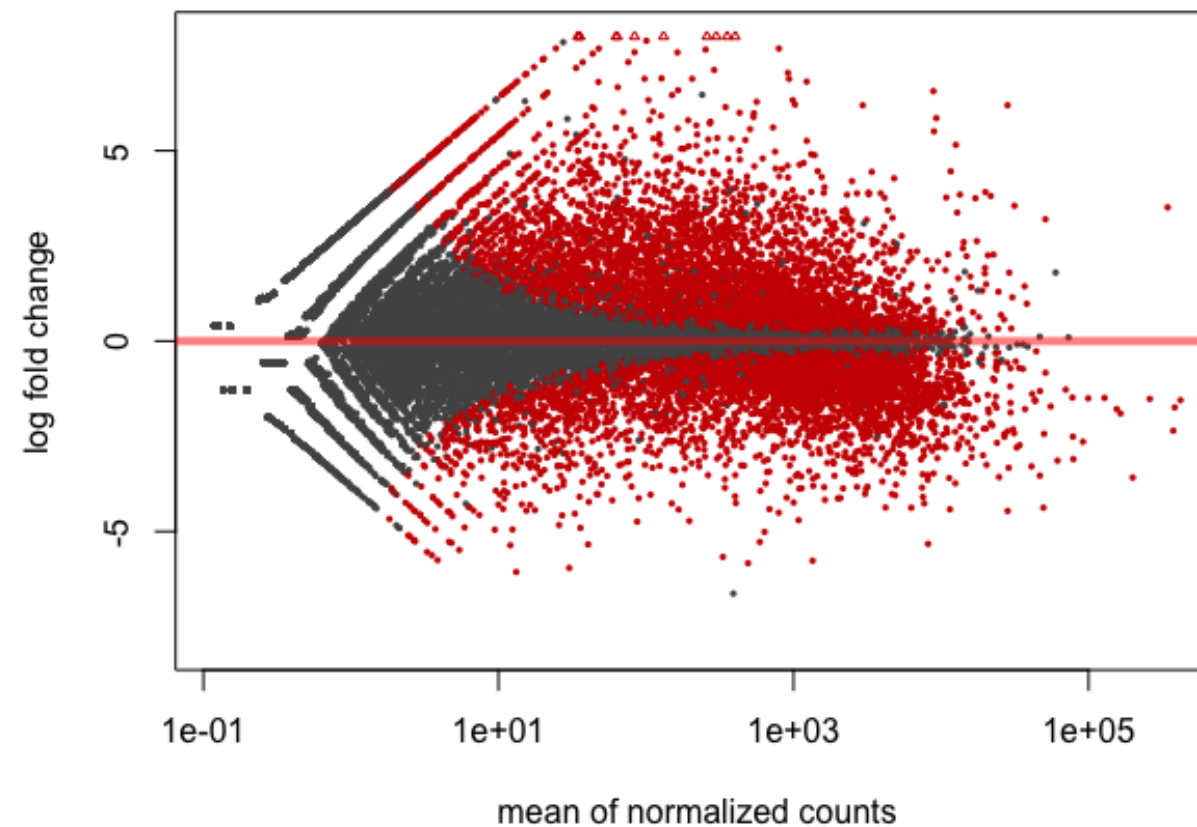
LFC shrinkage

```r
wt_res <- lfcShrink(dds_wt,
          contrast=c("condition", "fibrosis", "normal"),
          res=wt_res)
```

```r
plotMA(wt_res, ylim=c(-8,8))
```

LFC shrinkage
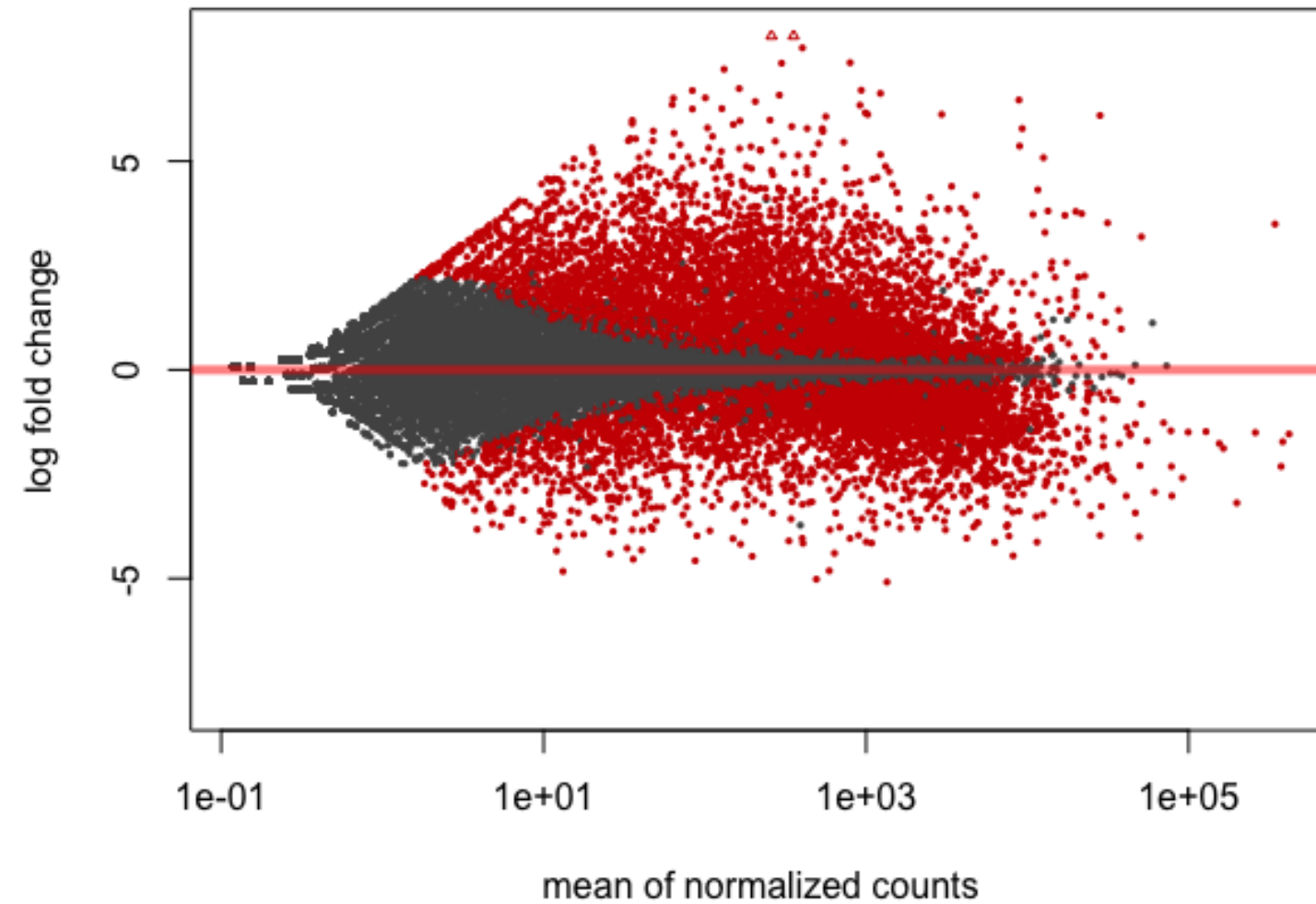
Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R

DESeq2 results

RNA-SEQ WITH BIOCONDUCTOR IN R

Mary Piper
Bioinformatics Consultant and Trainer

DESeq2 results table

```
mcols(wt_res)
```

```
DataFrame with 6 rows and 2 columns
            type                              description
     <character>                              <character>
1 intermediate           mean of normalized counts for all samples
2      results  log2 fold change (MAP): condition fibrosis vs normal
3      results           standard error: condition fibrosis vs normal
4      results            Wald statistic: condition fibrosis vs normal
5      results       Wald test p-value: condition fibrosis vs normal
6      results                              BH adjusted p-values
```

```
head(wt_res, n=10)
```

```
log2 fold change (MAP): condition fibrosis vs normal
Wald test p-value: condition fibrosis vs normal
data frame with 6 rows and 6 columns
                              baseMean    log2FoldChange                lfcSE               stat
                             <numeric>         <numeric>            <numeric>          <numeric>
ENSMUSG00000102693                   0                NA                   NA                 NA
ENSMUSG00000064842                   0                NA                   NA                 NA
ENSMUSG00000051951   19.5084656230804  3.55089043143673    0.648400500074659   4.66871842838828
ENSMUSG00000102851                   0                NA                   NA                 NA
ENSMUSG00000103377                   0                NA                   NA                 NA
ENSMUSG00000104017                   0                NA                   NA                 NA
                                pvalue                   padj
                             <numeric>              <numeric>
ENSMUSG00000102693                  NA                     NA
ENSMUSG00000064842                  NA                     NA
ENSMUSG00000051951  3.03084428526558e-06  1.93776447202312e-05
ENSMUSG00000102851                  NA                     NA
ENSMUSG00000103377                  NA                     NA
ENSMUSG00000104017                  NA                     NA
```

Significant DE genes - summary

```
summary(wt_res)
```

```
out of 29866 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)        : 5413, 18%
LFC < 0 (down)      : 5235, 18%
outliers [1]        : 47, 0.16%
low counts [2]      : 8412, 28%
(mean count < 2)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Significant DE genes - fold-change threshold

```r
wt_res <- results(dds_wt,
                  contrast = c("condition", "fibrosis", "normal"),
                  alpha = 0.05,
                  lfcThreshold = 0.32)
```

```r
wt_res <- lfcShrink(dds_wt,
                    contrast=c("condition", "fibrosis", "normal"),
                    res=wt_res)
```

Significant DE genes - summary

```
summary(wt_res)
```

```
       out of 29866 with nonzero total read count
       adjusted p-value < 0.05
       LFC > 0 (up)         : 3376, 11%
       LFC < 0 (down)       : 3071, 10%
       outliers [1]         : 47, 0.16%
       low counts [2]       : 7851, 26%
       (mean count < 1)
       [1] see 'cooksCutoff' argument of ?results
       [2] see 'independentFiltering' argument of ?results
```

Results - annotate

```
library(annotables)


grcm38
```

```
# A tibble: 53,728 x 9
   ensgene            entrez symbol chr      start       end strand biotype        description
   <chr>               <int> <chr>  <chr>    <int>     <int>  <int> <chr>          <chr>
 1 ENSMUSG00000000001  14679 Gnai3  3    108107280 108146146     -1 protein_coding guanine nucleotide binding protein (G p…
 2 ENSMUSG00000000003  54192 Pbsn   X     77837901  77853623     -1 protein_coding probasin [Source:MGI Symbol;Acc:MGI:186…
 3 ENSMUSG00000000028  12544 Cdc45  16    18780447  18811987     -1 protein_coding cell division cycle 45 [Source:MGI Symb…
 4 ENSMUSG00000000031     NA H19    7    142575529 142578143     -1 lincRNA        H19, imprinted maternally expressed tra…
 5 ENSMUSG00000000037 107815 Scml2  X    161117193 161258213      1 protein_coding sex comb on midleg-like 2 (Drosophila) …
 6 ENSMUSG00000000049  11818 Apoh   11   108343354 108414396      1 protein_coding apolipoprotein H [Source:MGI Symbol;Acc…
 7 ENSMUSG00000000056  67608 Narf   11   121237253 121255856      1 protein_coding nuclear prelamin A recognition factor […
 8 ENSMUSG00000000058  12390 Cav2   6     17281185  17289115      1 protein_coding caveolin 2 [Source:MGI Symbol;Acc:MGI:1…
 9 ENSMUSG00000000078  23849 Klf6   13     5861482   5870394      1 protein_coding Kruppel-like factor 6 [Source:MGI Symbo…
10 ENSMUSG00000000085  29871 Scmh1  4    120405281 120530186      1 protein_coding sex comb on midleg homolog 1 [Source:MG…
# ... with 53,718 more rows
```

Results - extract

```
wt_res_all <- data.frame(wt_res) %>%

        rownames_to_column(var = "ensgene") %>%

        left_join(x = wt_res_all,

                  y = grcm38[, c("ensgene", "symbol", "description")],

                  by = "ensgene")

View(wt_res_all)
```

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...
ENSMUSG00000018008	1278.6520	3.202855	0.11631046	24.77939	1.495690e-135	4.693902e-132	Cyth4	cytohesin 4 [Source:MGI Symbol;Acc:MGI:244...
ENSMUSG00000051439	4144.4097	3.743987	0.14014630	24.43589	7.109040e-132	1.952142e-128	Cd14	CD14 antigen [Source:MGI Symbol;Acc:MGI:8...
ENSMUSG00000019122	1022.6759	6.119309	0.23466958	24.38950	2.210634e-131	5.395911e-128	Ccl9	chemokine (C–C motif) ligand 9 [Source:MGI ...
ENSMUSG00000049103	1459.2660	4.429691	0.17109476	23.99096	3.455388e-127	7.590796e-124	Ccr2	chemokine (C–C motif) receptor 2 [Source:M...
ENSMUSG00000024164	28248.5968	6.095037	0.24563004	23.52525	2.250267e-122	4.493989e-119	C3	complement component 3 [Source:MGI Symb...
ENSMUSG00000022037	50990.1309	3.187649	0.12200514	23.51432	2.911008e-122	5.329085e-119	Clu	clusterin [Source:MGI Symbol;Acc:MGI:88423]
ENSMUSG00000024349	951.8453	3.327356	0.13816314	21.75630	6.021862e-105	1.017602e-101	Tmem173	transmembrane protein 173 [Source:MGI Sym...

Significant DE genes - arrange

```
wt_res_sig <- subset(wt_res_all, padj < 0.05)

wt_res_sig  <- wt_res_sig %>%

        arrange(padj)

View(wt_res_all)
```

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...
ENSMUSG00000018008	1278.6520	3.202855	0.11631046	24.77939	1.495690e-135	4.693902e-132	Cyth4	cytohesin 4 [Source:MGI Symbol;Acc:MGI:244...
ENSMUSG00000051439	4144.4097	3.743987	0.14014630	24.43589	7.109040e-132	1.952142e-128	Cd14	CD14 antigen [Source:MGI Symbol;Acc:MGI:8...
ENSMUSG00000019122	1022.6759	6.119309	0.23466958	24.38950	2.210634e-131	5.395911e-128	Ccl9	chemokine (C-C motif) ligand 9 [Source:MGI ...
ENSMUSG00000049103	1459.2660	4.429691	0.17109476	23.99096	3.455388e-127	7.590796e-124	Ccr2	chemokine (C-C motif) receptor 2 [Source:M...
ENSMUSG00000024164	28248.5968	6.095037	0.24563004	23.52525	2.250267e-122	4.493989e-119	C3	complement component 3 [Source:MGI Symb...
ENSMUSG00000022037	50990.1309	3.187649	0.12200514	23.51432	2.911008e-122	5.329085e-119	Clu	clusterin [Source:MGI Symbol;Acc:MGI:88423]
ENSMUSG00000024349	951.8453	3.327356	0.13816314	21.75630	6.021862e-105	1.017602e-101	Tmem173	transmembrane protein 173 [Source:MGI Sym...

ORIGINAL ARTICLE

Loss of SOCS3 in the liver promotes fibrosis by enhancing STAT3-mediated TGF-β1 production

H Ogata[1,2], T Chinen[1], T Yoshida[1,3], I Kinjyo[1], G Takaesu[1], H Shiraishi[1], M Iida[2], T Kobayashi[1] and A Yoshimura[1]

Pericytes and immune cells contribute to complement activation in tubulointerstitial fibrosis

Sandhya Xavier,[1] Ranjit K. Sahu,[1] Susan G. Landes,[1] Jing Yu,[2] Ronald P. Taylor,[3] Srinivas Ayyadevara,[4] Judit Megyesi,[4] William B. Stallcup,[5] Jeremy S. Duffield,[6] Edimara S. Reis,[7] John D. Lambris,[7] and Didier Portilla[1,8]

CD44 Deficiency Increases Tubular Damage But Reduces Renal Fibrosis in Obstructive Nephropathy

Kasper M.A. Rouschop[*], Miguel E. Sewnath[†], Nike Claessen[*], Joris J.T.H. Roelofs[*], Inge Hoedemaeker[*], Ronald van der Neut[*], Jan Aten[*], Steven T. Pals[*], Jan J. Weening[*] and Sandrine Florquin[*]

Lipocalin 2 is essential for chronic kidney disease progression in mice and humans.

Viau A[1], El Karoui K, Laouari D, Burtin M, Nguyen C, Mori K, Pillebout E, Berger T, Mak TW, Knebelmann B, Friedlander G, Barasch J, Terzi F.

Let's practice!

RNA-SEQ WITH BIOCONDUCTOR IN R