

# TRANSCRIPTOMICS

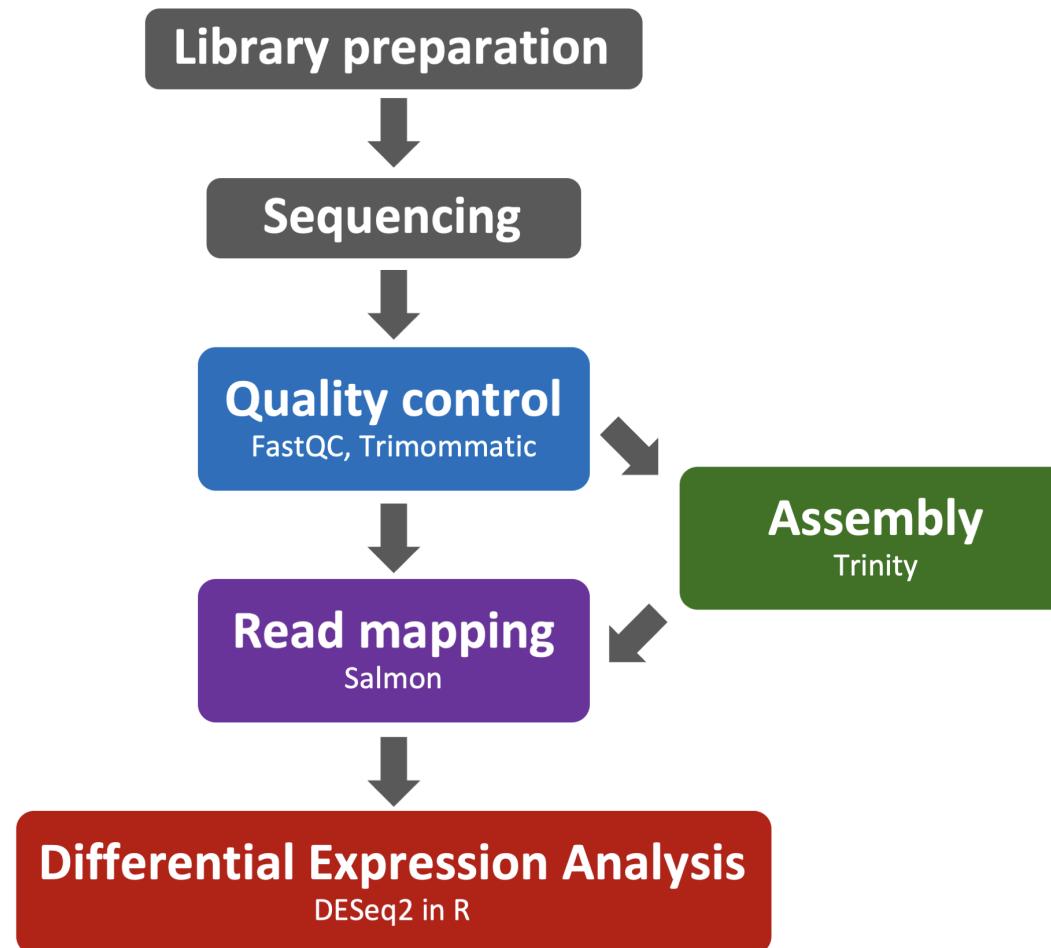
## Differential Gene Expression

Day 06

<https://totorres.github.io/transcriptomics/>

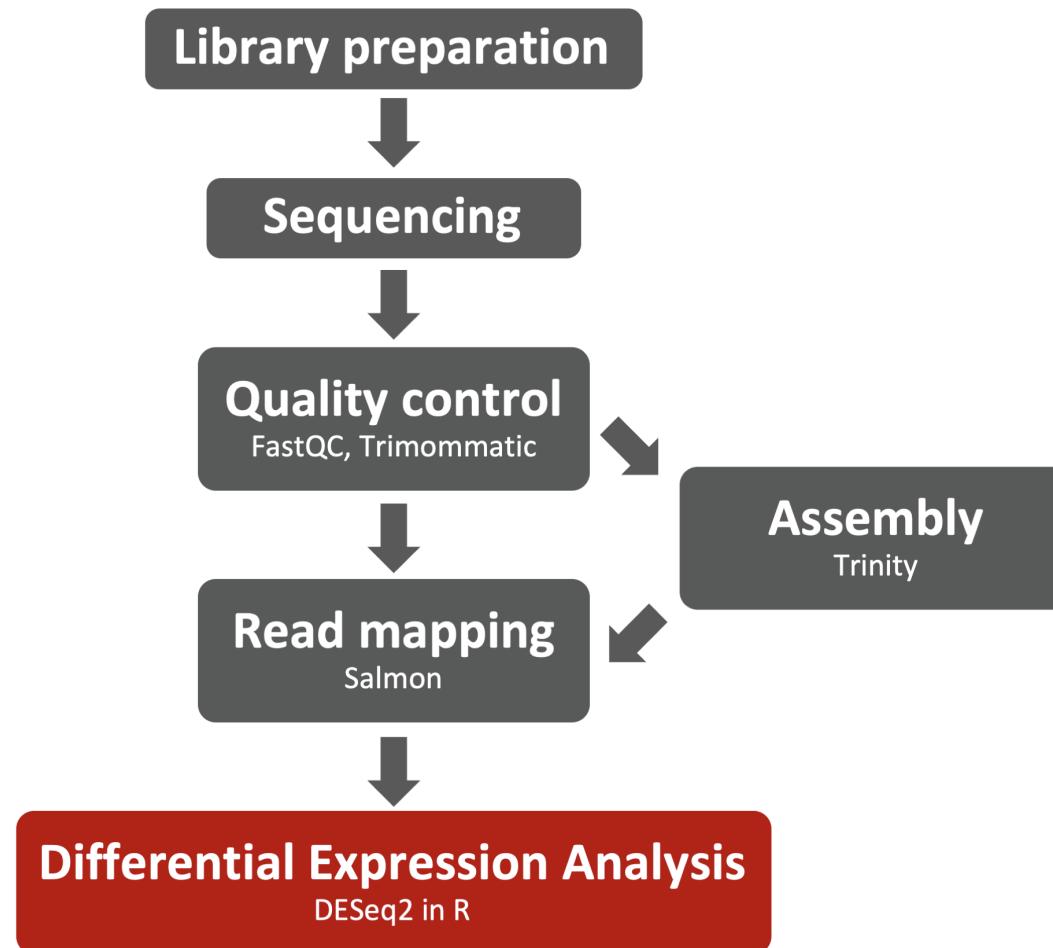
# RNA-seq workflow

## Differential Gene Expression



# RNA-seq workflow

## Differential Gene Expression



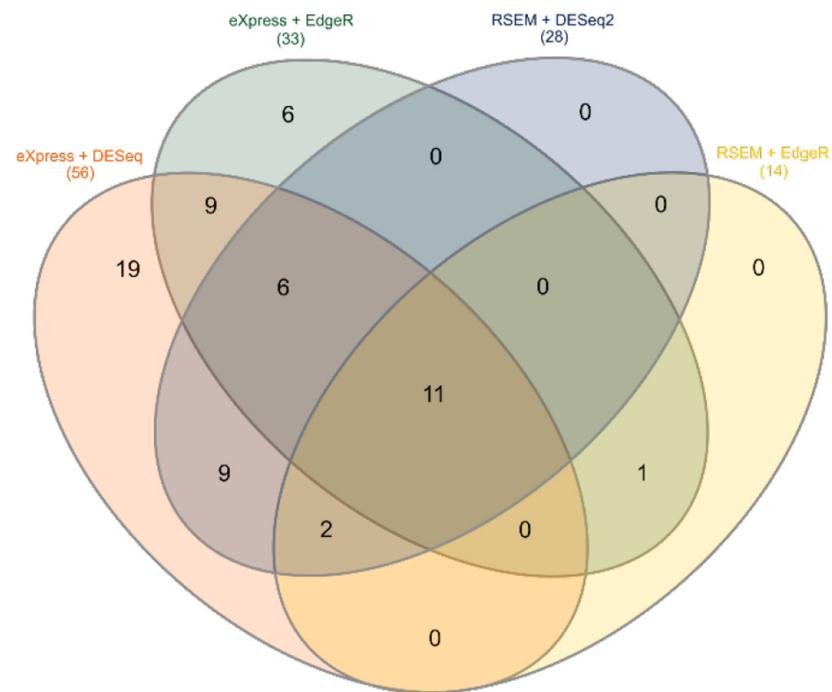
# RNA-Seq: Differential Gene Expression

- What genes/transcripts are being expressed at higher/lower levels in different groups of samples?
- Are differences observed in read counts real gene expression differences? Are these differences 'significant', accounting for variance/noise?
- Tying gene expression back to genotype/phenotype

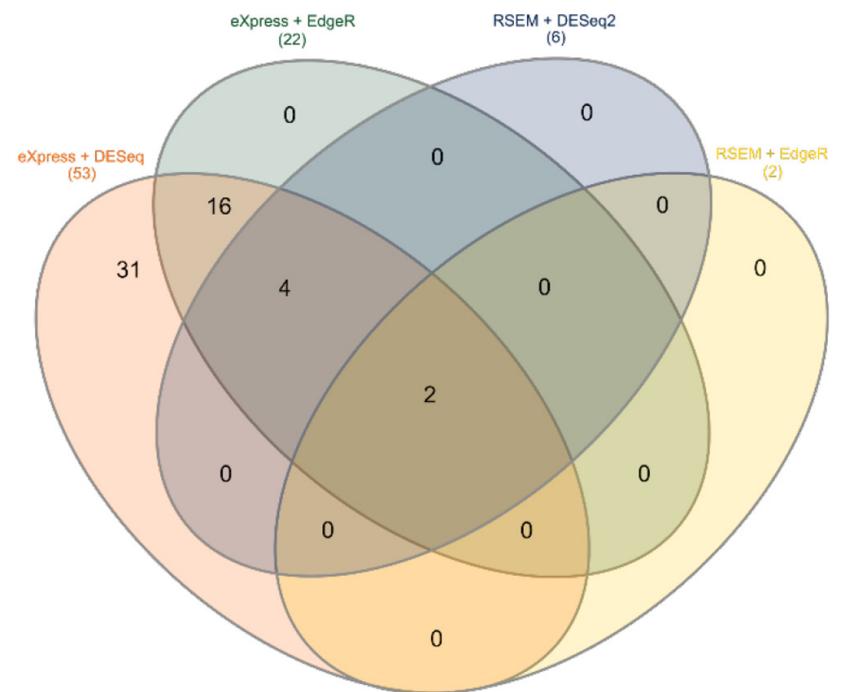
# RNA-Seq: Differential Gene Expression

## Choosing package/model

A

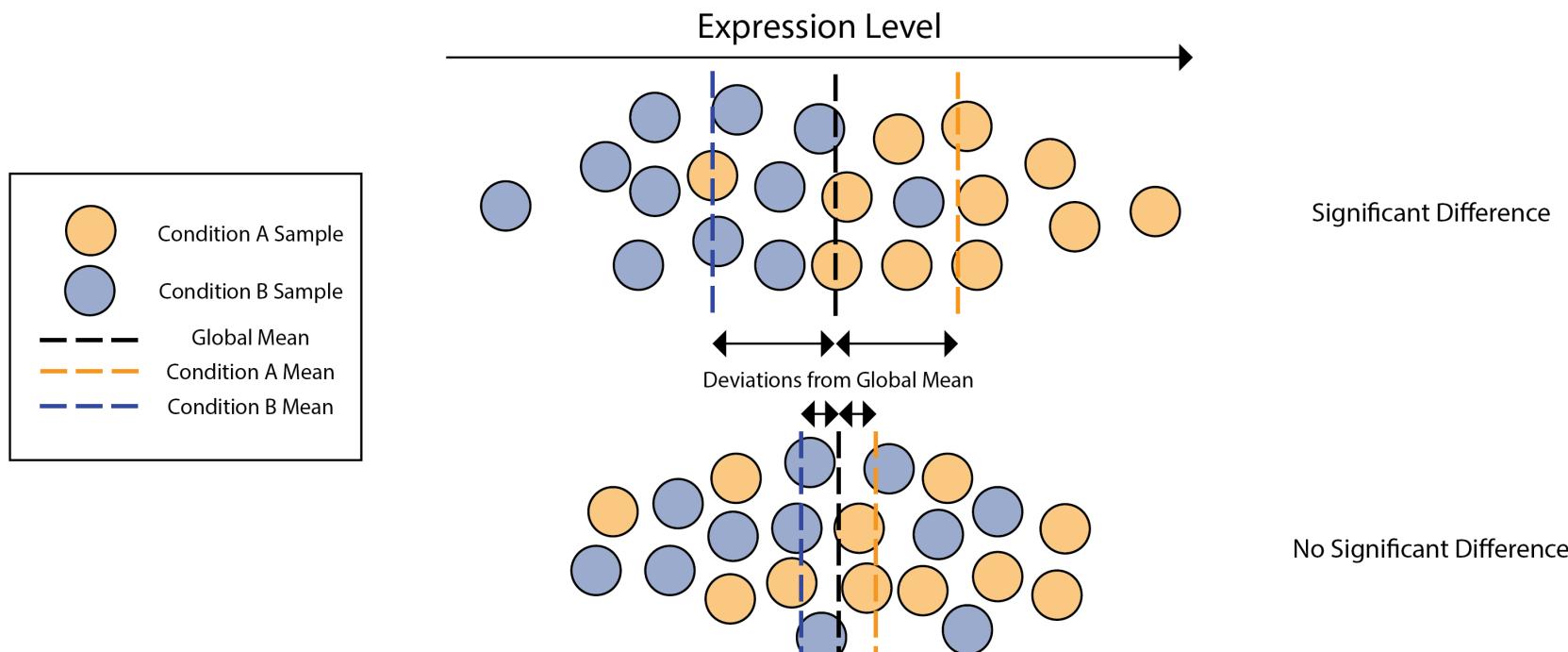


B



# RNA-Seq: Differential Gene Expression

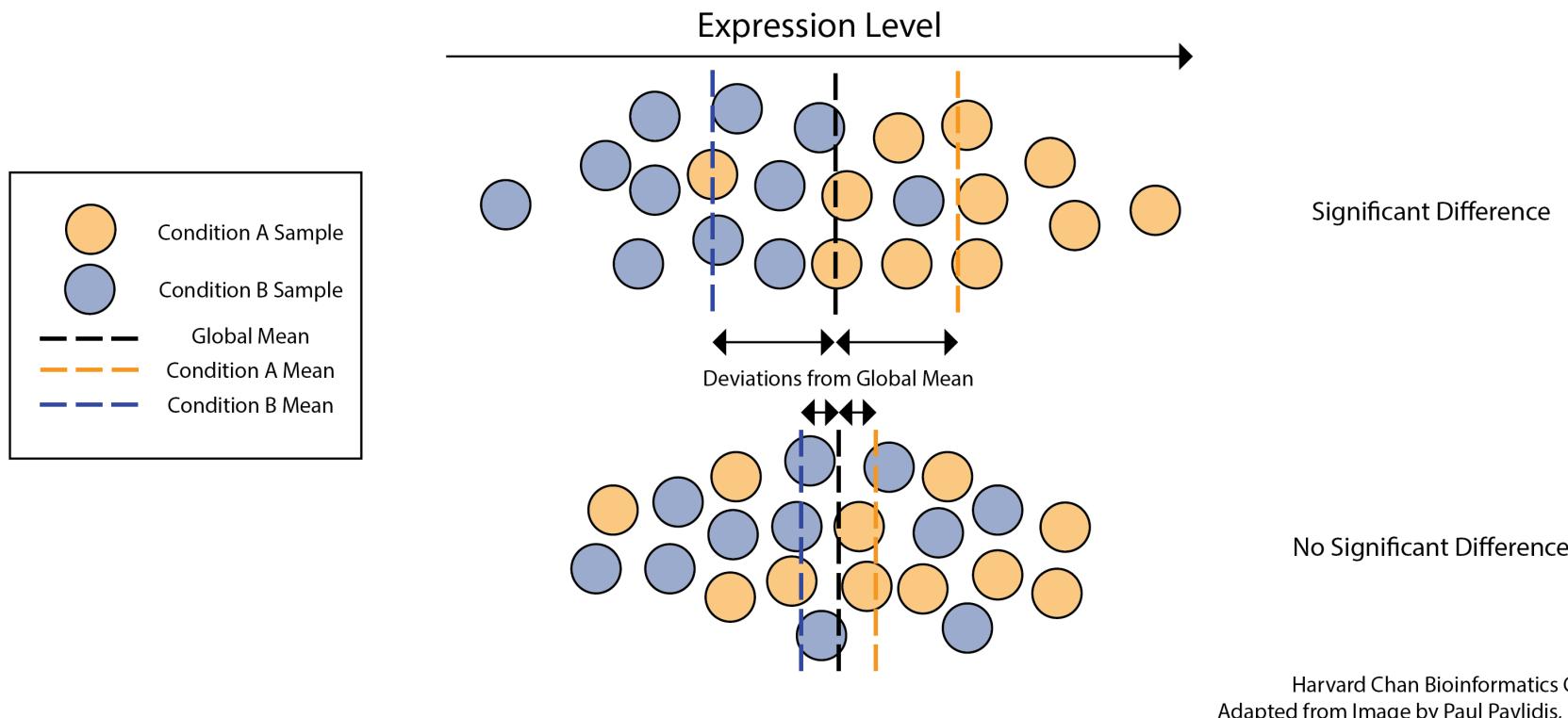
## Differential expression analysis with DESeq2



Harvard Chan Bioinformatics Core  
Adapted from Image by Paul Pavlidis, UBC

# RNA-Seq: Differential Gene Expression

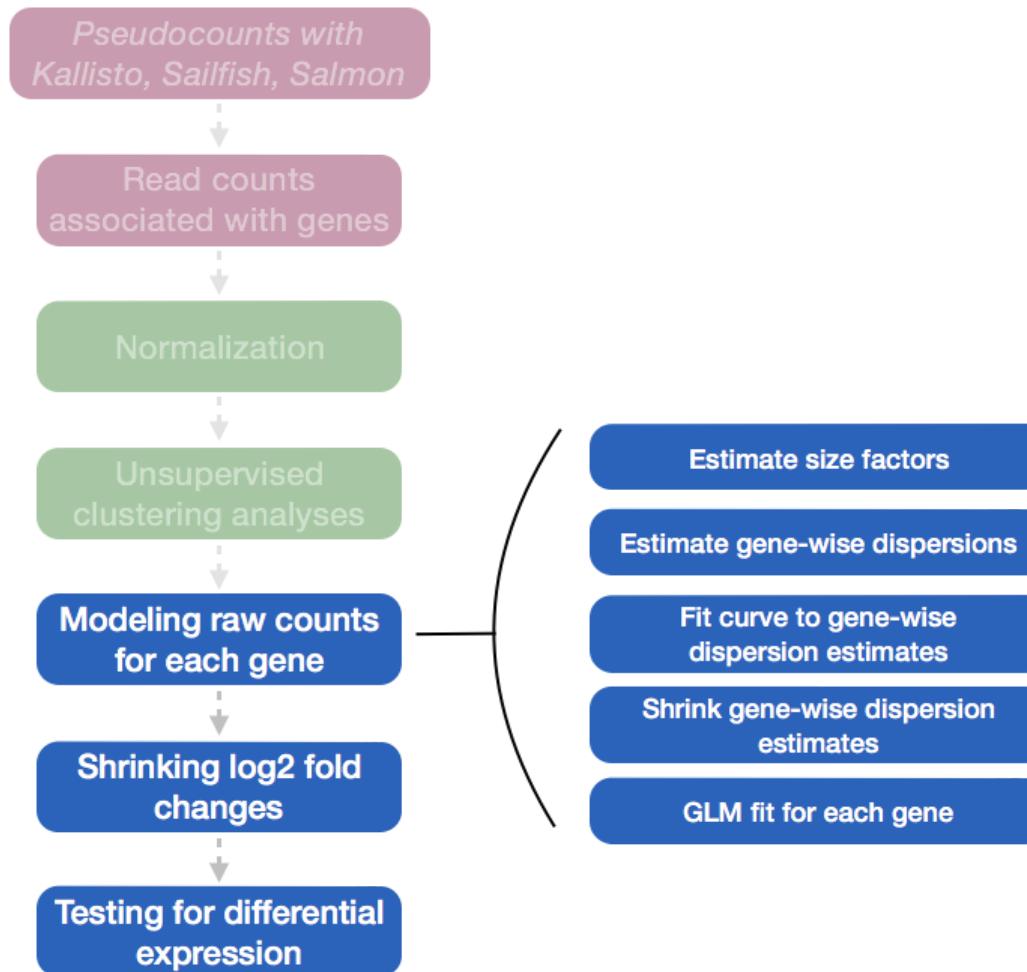
## Differential expression analysis with DESeq2



Fitting the raw counts to the model and performing the statistical test for differentially expressed genes

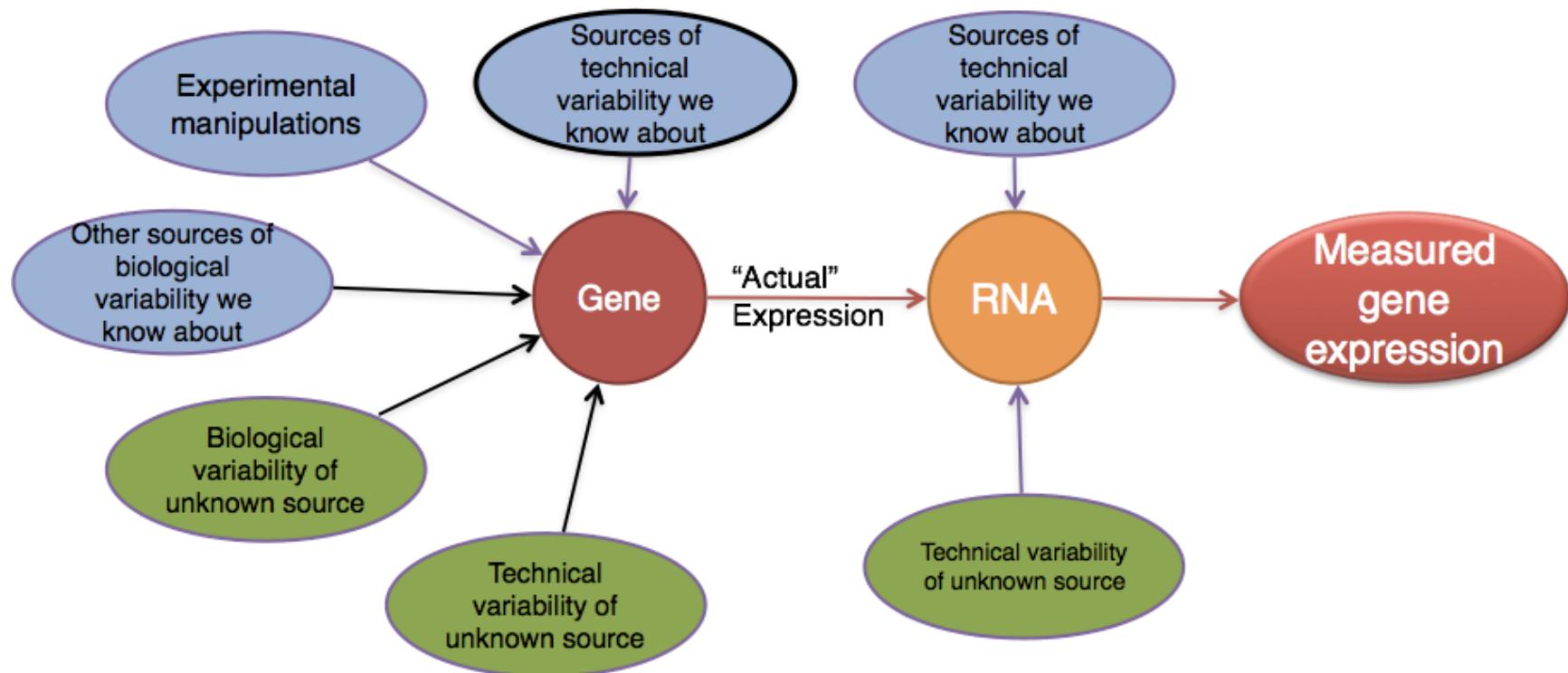
# RNA-Seq: Differential Gene Expression

## Differential expression analysis with DESeq2



# RNA-Seq: Differential Gene Expression

## Review of the dataset



Courtesy of Paul Pavlidis, UBC

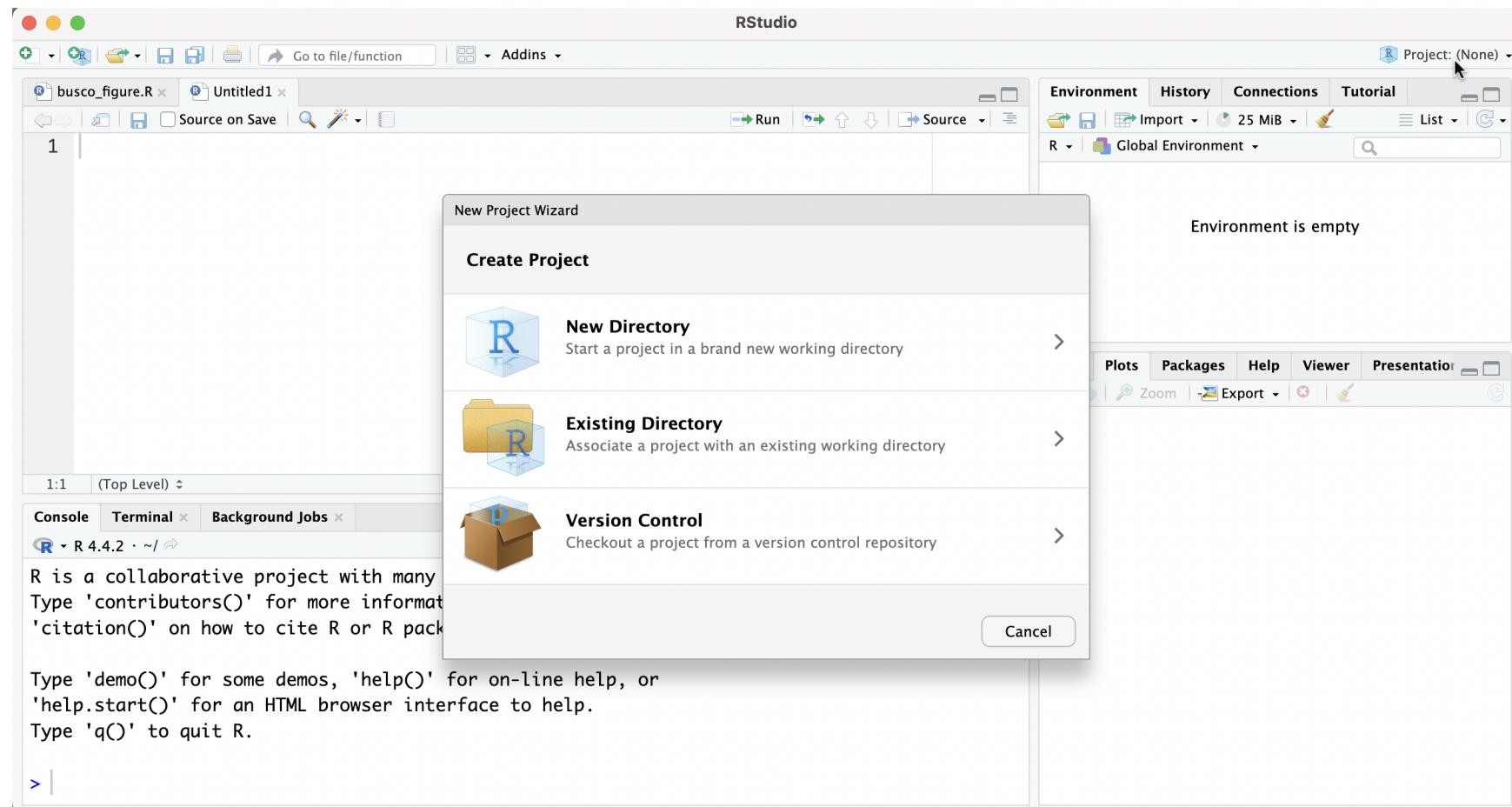
# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

1. Opening up RStudio.
2. Set up a new project for this analysis.
  - Go to the `File` menu and select `New Project`.
  - In the `New Project` window, choose `New Directory`.
  - Choose `New Project`.

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio



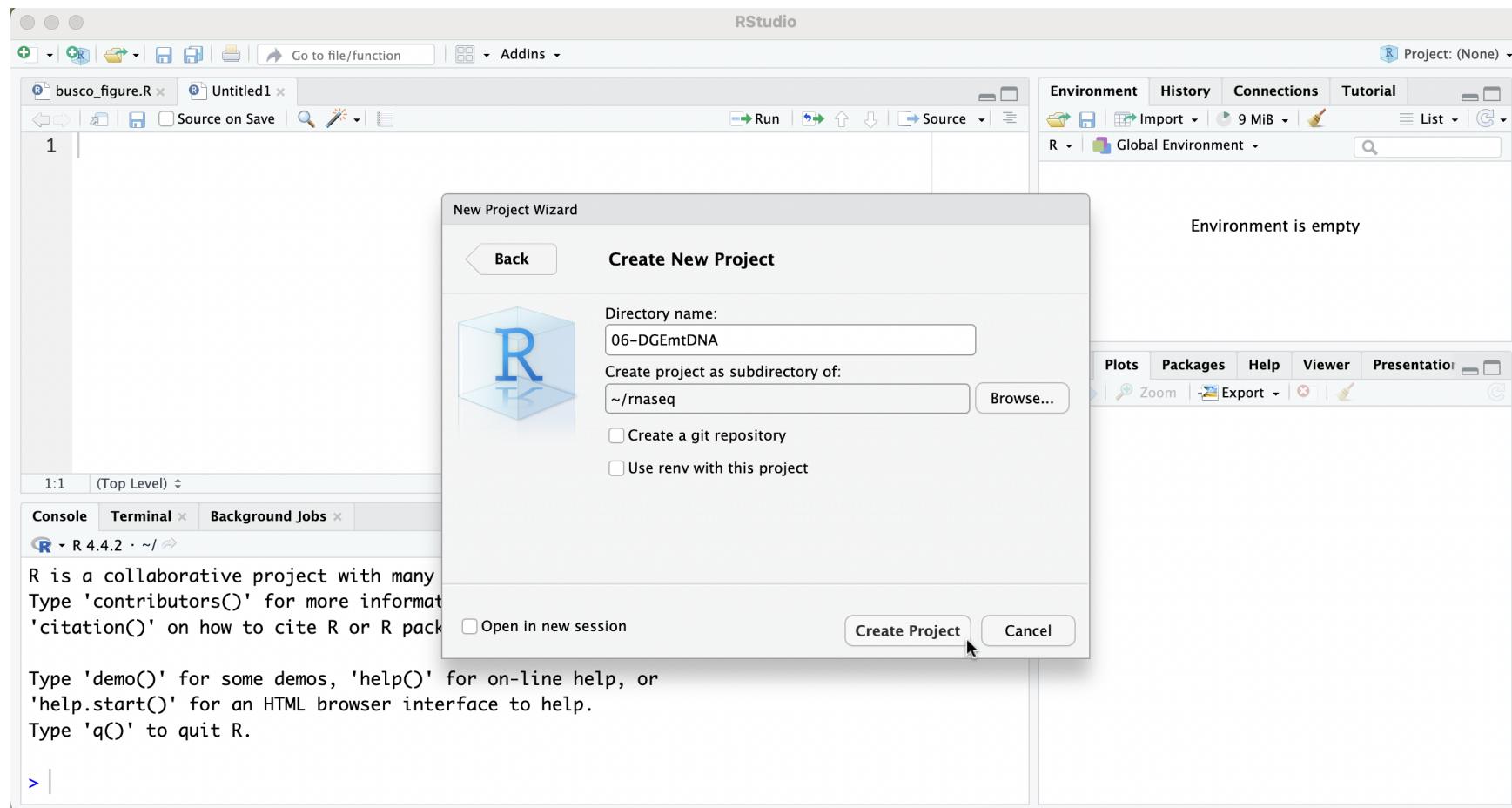
# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

1. Open RStudio.
2. Set up a new project for this analysis.
  - Go to the `File` menu and select `New Project`.
  - In the `New Project` window, choose `New Directory`.
  - Choose `New Project`.
  - Name your new directory `06-DGEmtDNA`
  - Create the project as subdirectory of `rnaseq`
  - The new project should automatically open in RStudio.

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio



# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

- A new R session is started and some files are created:
  - `.Rprofile`
  - `.RData`
  - `.Rhistory`

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

3. Check (in R console) if you are in the correct working directory

```
getwd()
```

4. Download the DGEmtDNA.R script from the course website (save it in the project folder)
5. On RStudio go to `File > Open file` to open the DGEmtDNA.R script

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

The screenshot shows the RStudio interface with the following components:

- Script Editor:** Displays the script `DGEmtDNA.R` containing R code for setting up Bioconductor and CRAN libraries.
- Console:** Shows the R environment setup, including locale information, contributor details, and help messages.
- File Browser:** Shows the directory structure under `rnaseq/06-DGEmtDNA`, listing files like `06-DGEmtDNA.Rproj` and `DGEmtDNA.R`.
- Environment:** Shows the Global Environment pane which is currently empty.

```
## Setup----  
### Loading Bioconductor and CRAN libraries  
library(DESeq2)  
library(tidyverse)  
library(RColorBrewer)  
library(pheatmap)  
library(DEGreport)  
library(tximport)  
library(ggplot2)  
library(ggrepel)
```

Natural language support but running in an English locale  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

```
> getwd()  
[1] "/Users/tatiana/rnaseq/06-DGEmtDNA"  
> library(DESeq2)  
Error in library(DESeq2) : there is no package called 'DESeq2'  
>
```

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

```
## Setup----  
### Loading Bioconductor and CRAN libraries  
library(DESeq2)  
library(tidyverse)  
library(RColorBrewer)  
library(pheatmap)  
library(DEGreport)  
library(tximport)  
library(ggplot2)  
library(ggrepel)
```

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

Loading libraries

```
library(DESeq2)
Error in library(DESeq2) : there is no package called
'DESeq2'
```

Installing libraries

```
install.packages(DESeq2)
Error in install.packages : object 'DESeq2' not found
```

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

What repository R is searching for packages?

```
setRepositories()
```

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

What repository R is searching for packages?

```
setRepositories()
--- Please select repositories for use in this session ---

1: + CRAN
2: BioC software
3: BioC annotation
4: BioC experiment
5: CRAN (extras)
6: R-Forge
7: rforge.net
```

Enter one or more numbers separated by spaces and then ENTER

1:

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

### 6. Set repositories

```
setRepositories()  
--- Please select repositories for use in this session ---  
  
1: + CRAN  
2: BioC software  
3: BioC annotation  
4: BioC experiment  
5: CRAN (extras)  
6: R-Forge  
7: rforge.net  
  
Enter one or more numbers separated by spaces and then EN  
1: 1 2 3 4 5 6 7
```

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

### 7. Load packages

```
library(DESeq2)
library(tidyverse)
library(RColorBrewer)
library(pheatmap)
library(DEGreport)
library(tximport)
library(ggplot2)
library(ggrepel)
```

# RNA-Seq: Differential Gene Expression

## Loading data

### Salmon output: quantification File

Name	Length	EffectiveLength	TPM	NumReads
ND2	1017	840.387	19495.097280	94155.999
COX1	1539	1362.387	191863.560563	1502229.185
COX2	688	511.403	188051.013249	552690.368
ATP8	165	27.164	1651.052638	257.753
ATP6	678	501.404	81711.503320	235458.584
COX3	789	612.391	264354.551853	930375.771
ND3	357	183.851	52965.972446	55963.632
ND5	1720	1543.387	7352.180412	65213.001

# RNA-Seq: Differential Gene Expression

# Loading data

## Package tximport to import quant.sf files

8. Download count data from the course shared drive
  9. Create a variable with the paths to all quant.sf files.

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

The screenshot shows the RStudio interface with the following components:

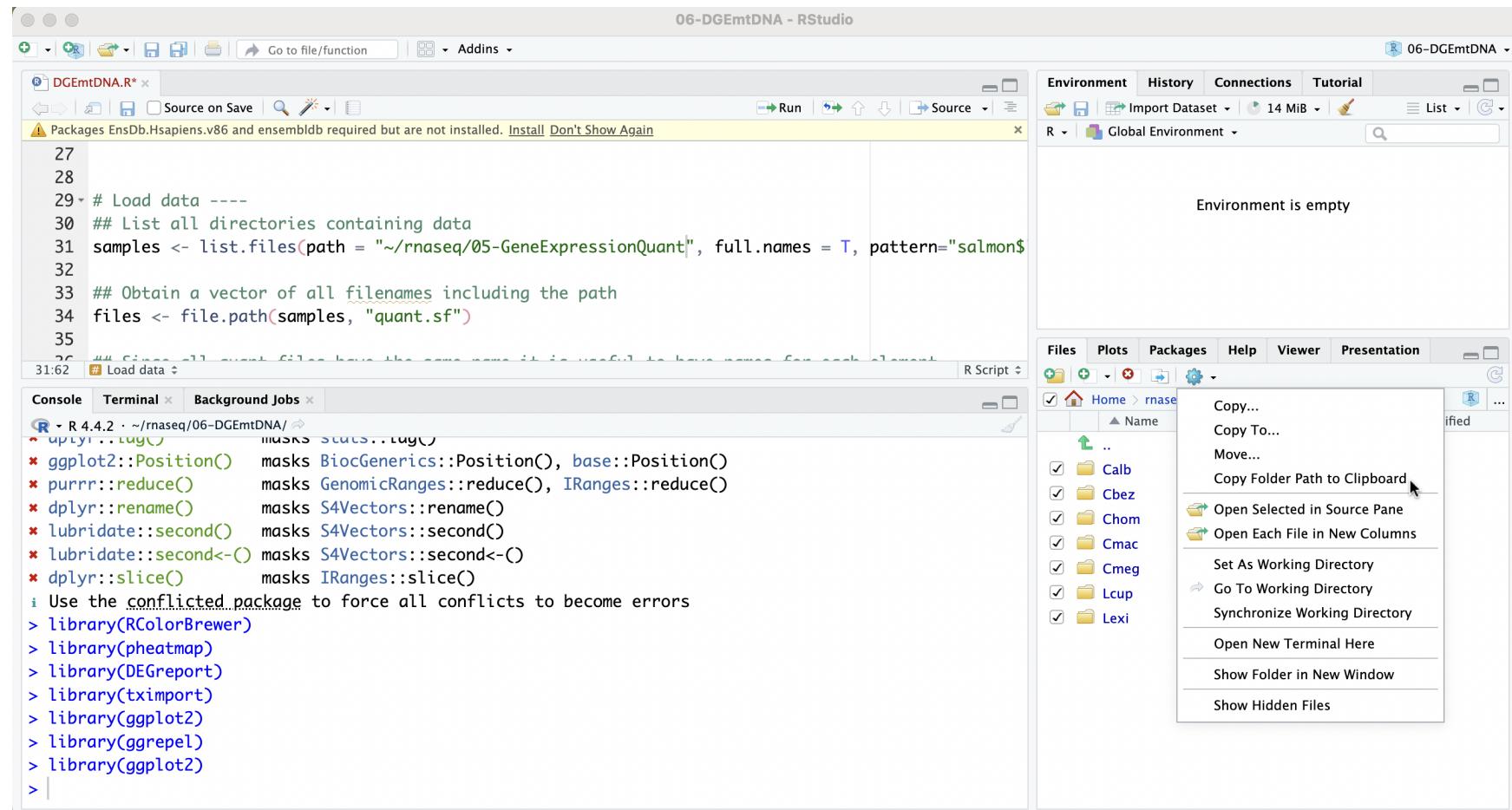
- Script Editor (Top Left):** Displays the R script `DGEmtDNA.R*`. The code loads data from a directory containing salmon files and creates a vector of filenames.
- Environment Browser (Top Right):** Shows the Global Environment tab with the message "Environment is empty".
- File Browser (Bottom Right):** Shows a file tree under "rnaseq > 05-GeneExpressionQuant".
- Console (Bottom Left):** Displays R session output, including package conflicts and library imports.

```
27
28
29 - # Load data ----
30 ## List all directories containing data
31 samples <- list.files(path = "~/rnaseq/05-GeneExpressionQuant", full.names = T, pattern="salmon$"
32
33 ## Obtain a vector of all filenames including the path
34 files <- file.path(samples, "quant.sf")
35
36 ## Since all input files have the same name it is useful to have names for each element
37
38 # Load data
39
```

```
R 4.4.2 · ~/rnaseq/06-DGEmtDNA/ ↵
  upyr...log()  masks stats..log()
  * ggplot2::Position()  masks BiocGenerics::Position(), base::Position()
  * purrr::reduce()  masks GenomicRanges::reduce(), IRanges::reduce()
  * dplyr::rename()  masks S4Vectors::rename()
  * lubridate::second()  masks S4Vectors::second()
  * lubridate::second<-()  masks S4Vectors::second<-()
  * dplyr::slice()  masks IRanges::slice()
  i Use the conflicted package to force all conflicts to become errors
> library(RColorBrewer)
> library(pheatmap)
> library(DEGreport)
> library(tximport)
> library(ggplot2)
> library(ggrepel)
> library(ggplot2)
>
```

# RNA-Seq: Differential Gene Expression

# Setting up on RStudio



# RNA-Seq: Differential Gene Expression

## Loading data

10. Download count data from the course website (unzip it in the project folder)
11. Obtain a vector of all filenames including the path

```
files <- file.path(samples, "quant.sf")
```

12. Give unique name for each element

```
names(files) <- str_replace(samples, "./allCountData/", "")  
str_replace("-salmon", "") %>% str_replace("/", "")
```

13. Load the annotation table for the mtDNA genes

```
tx2gene <- read.delim("~/rnaseq/00-Databases/mtDNA.txt")
```

# RNA-Seq: Differential Gene Expression

## Loading data

Package `tximport` to import `quant.sf` files

14. See `tximport help`

```
?tximport
```

# RNA-Seq: Differential Gene Expression

## Setting up on RStudio

The screenshot shows the RStudio interface with the following components:

- Script Editor (Top Left):** Displays the script `DGEmtDNA.R*` containing R code for reading RNA-seq data. A warning message at the top indicates required packages `EnsDb.Hsapiens.v86` and `ensemblDb` are not installed.
- Environment Browser (Top Right):** Shows the global environment with variables `files` and `samples` defined.
- Help Documentation (Bottom Right):** Provides details for the `tximport` package, including its description and usage.
- Console (Bottom Left):** Shows the R console output, starting with a blank session and then the command `?tximport`.

```
34 files <- file.path(samples, "quant.sf")
35
36 ## Since all quant files have the same name it is useful to have names for each element
37 names(files) <- str_replace(samples, "./allCountData/", "") %>%
38   str_replace("-salmon", "")
39
40
41 #there are LOTS of ways to read data into R, but the readr package (from tidyverse) is one of the
42 targets <- read_tsv("studydesign.txt")
43
44 # Load data
```

R: Import transcript-level abundances and counts for transcript- and gene-level analysis packages

**Description**

`tximport` imports transcript-level estimates from various external software and optionally summarizes abundances, counts, and transcript lengths to the gene-level (default) or outputs transcript-level matrices (see `txOut` argument).

**Usage**

# RNA-Seq: Differential Gene Expression

## Loading data

Package `tximport` to import `quant.sf` files

14. See `tximport help`

```
?tximport
```

`tximport` imports transcript-level estimates from various external software (Salmon, Kallisto, Sailfish, etc) and optionally summarizes abundances, counts, and transcript lengths to the gene-level (default) or outputs transcript-level matrices.

# RNA-Seq: Differential Gene Expression

## Loading data

Package `tximport` to import `quant.sf` files

Some of `tximport` arguments:

- ▶ `files`: vector of filenames for the transcript-level abundances
- ▶ `type`: software used to generate the abundance. Options:
  - salmon,
  - sailfish,
  - kallisto,
  - rsem,
  - others, or
  - none

# RNA-Seq: Differential Gene Expression

## Loading data

Package `tximport` to import `quant.sf` files

Some of `tximport` arguments:

→ `countsFromAbundance`: whether to generate estimated counts using abundance estimates

- no (default)
- scaledTPM: scaled up to library size
- lengthScaledTPM: scaled using the average transcript length over samples and then the library size
- dtuScaledTPM: scaled using the median transcript length among isoforms of a gene, and then the library size.

# RNA-Seq: Differential Gene Expression

## Loading data

Package `tximport` to import `quant.sf` files

Some of `tximport` arguments:

- `countsFromAbundance`: whether to generate estimated counts using abundance estimates

$$\text{TPM} = \frac{\text{reads per Kb}}{\text{total RPK in sample}} * 10^6$$

# RNA-Seq: Differential Gene Expression

## Loading data

Package `tximport` to import `quant.sf` files

Some of `tximport` arguments:

→ `tx2gene`: a two-column data.frame linking transcript id (column 1) to gene id (column 2). The column names are not relevant, but this column order must be used.

This argument is required for gene-level summarization. It is important when reads are mapped against transcripts or contigs.

# RNA-Seq: Differential Gene Expression

## Loading data

Package `tximport` to import `quant.sf` files

15. Run `tximport`

```
txi <- tximport(files, type="salmon",
                 countsFromAbundance="lengthScaledTPM",
                 tx0ut="TRUE" )
```

- ➔ `txOut`: logical, whether the function should just output transcript-level (default FALSE). Our references are already at the gene-level.

# RNA-Seq: Differential Gene Expression

## Viewing data

15. View how `txi` object is organized

```
class(txi)
```

# RNA-Seq: Differential Gene Expression

## Viewing data

15. View how `txi` object is organized

```
class(txi)
[1] "list"
```

# RNA-Seq: Differential Gene Expression

## Viewing data

15. View how `txi` object is organized

```
class(txi)
[1] "list"
attributes(txi)
```

# RNA-Seq: Differential Gene Expression

## Viewing data

15. View how `txi` object is organized

```
class(txi)
[1] "list"
attributes(txi)
$names
[1] "abundance" "counts"   "length"   "countsFromAbundance"
```

# RNA-Seq: Differential Gene Expression

## Viewing data

15. View the counts

```
txi$counts %>% View()
```

# RNA-Seq: Differential Gene Expression

## Viewing data

The screenshot shows the RStudio interface for an R session titled "06-DGEmtDNA - RStudio".

**Data View:** A grid showing differential gene expression data across various genes and samples. The columns include CalbF\_1\_mtDNA, CalbF\_4\_mtDNA, CalbL\_1\_mtDNA, CalbL\_4\_mtDNA, CbezF\_1\_mtDNA, CbezL\_1\_mtDNA, and ChomF\_1.

	CalbF_1_mtDNA	CalbF_4_mtDNA	CalbL_1_mtDNA	CalbL_4_mtDNA	CbezF_1_mtDNA	CbezL_1_mtDNA	ChomF_1
NAD2	13338.000	9209.882	4277.000	1992.000	52750.995	83333.902	
COX1	550546.339	633534.659	340000.948	140846.664	980982.446	3266084.093	
COX2	153949.241	135999.086	81661.996	44389.479	208500.952	1081209.747	
ATP8	384.319	420.091	98.212	93.753	4848.458	6188.946	
ATP6	154399.212	136590.126	47777.419	42332.396	225728.867	507499.188	
COX3	350302.849	299529.932	158793.378	76511.253	503434.356	1122978.899	
NAD3	8552.196	14303.001	3914.000	4963.454	16160.398	42497.005	
NADS	52583.744	50085.784	16228.000	9744.000	130641.297	227652.322	

**Environment View:** Shows the global environment with objects like txi and lists of files and samples.

**Console View:** Displays the R session history, showing commands related to reading salmon files and creating a tximport object.

```
R 4.4.2 · ~/rnaseq/06-DGEmtDNA/ ▶
[1] "./allCountData//Lcupl_1_mtDNA-salmon" "./allCountData//Lcupl_3_mtDNA-salmon"
[21] "./allCountData//Lcupl_1_mtDNA-salmon" "./allCountData//Lcupl_3_mtDNA-salmon"
[23] "./allCountData//Lexif_1_mtDNA-salmon" "./allCountData//Lexif_2_mtDNA-salmon"
[25] "./allCountData//Lexil_1_mtDNA-salmon" "./allCountData//Lexil_2_mtDNA-salmon"
> names(files) <- str_replace(samples, "./allCountData/", "") %>%
+   str_replace("-salmon", "") %>% str_replace("/", "")
>
> ## Run tximport
> txi <- tximport(files, type="salmon",
+                  # countsFromAbundance="lengthScaledTPM",
+                  txOut=TRUE)
reading in files with read_tsv
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
> beep(sound = 2)
> txi$counts %>% View()
>
```

**File View:** Shows the file tree structure under the current working directory.

- CalbF\_1\_mtDNA-salmon
- CalbF\_4\_mtDNA-salmon
- CalbL\_1\_mtDNA-salmon
- CalbL\_4\_mtDNA-salmon
- CbezF\_1\_mtDNA-salmon
- CbezL\_1\_mtDNA-salmon
- ChomF\_1\_mtDNA-salmon
- ChomF\_3\_mtDNA-salmon
- ChomL\_1\_mtDNA-salmon
- ChomL\_3\_mtDNA-salmon
- CmacF\_1\_mtDNA-salmon
- CmacF\_3\_mtDNA-salmon
- CmacL\_1\_mtDNA-salmon
- CmacL\_3\_mtDNA-salmon

# RNA-Seq: Differential Gene Expression

## Viewing data

15. View the counts

```
txi$counts %>% View()
```

16. Create an object with the counts

```
data <- txi$counts %>%
  round() %>%
  data.frame()
```

# RNA-Seq: Differential Gene Expression

## Viewing data

17. Create sample table with metadata

```
# Extract sample names  
sampleNames <- colnames(data)
```

# RNA-Seq: Differential Gene Expression

## Viewing data

### 17. Create sample table with metadata

```
# Extract sample names
sampleNames <- colnames(data)
sampleNames
[1] "CalbF_1" "CalbF_2" "CalbF_4" "CalbL_1" "CalbL_2"
[6] "CalbL_4" "CbezF_1" "CbezL_1" "ChomF_1" "ChomF_2"
[11] "ChomF_3" "ChomL_1" "ChomL_2" "ChomL_3" "CmacF_1"
[16] "CmacF_2" "CmacF_3" "CmacL_1" "CmacL_2" "CmacL_3"
[21] "CmegF_1" "CmegF_2" "CmegF_4" "CmegL_1" "CmegL_2"
[26] "CmegL_4" "LexiF_1" "LexiF_2" "LexiF_3" "LexiL_1"
[31] "LexiL_2" "LexiL_3"
```

# RNA-Seq: Differential Gene Expression

## Viewing data

17. Create sample table with metadata

```
# samplename = First five characters
samplename <- substr(sampleNames, 1, 5)
```

# RNA-Seq: Differential Gene Expression

## Viewing data

### 17. Create sample table with metadata

```
# samplename = First five characters
samplename <- substr(sampleNames, 1, 5)
samplename
[1] "CalbF" "CalbF" "CalbF" "CalbL" "CalbL" "CalbL"
[7] "CbezF" "CbezL" "ChomF" "ChomF" "ChomF" "ChomL"
[13] "ChomL" "ChomL" "CmacF" "CmacF" "CmacF" "CmacL"
[19] "CmacL" "CmacL" "CmegF" "CmegF" "CmegF" "CmegL"
[25] "CmegL" "CmegL" "LexiF" "LexiF" "LexiF" "LexiL"
[31] "LexiL" "LexiL"
```

# RNA-Seq: Differential Gene Expression

## Viewing data

17. Create sample table with metadata

```
# Species = First four characters  
species <- substr(sampleNames, 1, 4)
```

# RNA-Seq: Differential Gene Expression

## Viewing data

17. Create sample table with metadata

```
# Species = First four characters
species <- substr(sampleNames, 1, 4)
species
[1] "Calb" "Calb" "Calb" "Calb" "Calb" "Calb" "Calb" "Cbez"
[8] "Cbez" "Chom" "Chom" "Chom" "Chom" "Chom" "Chom" "Chom"
[15] "Cmac" "Cmac" "Cmac" "Cmac" "Cmac" "Cmac" "Cmac" "Cmeg"
[22] "Cmeg" "Cmeg" "Cmeg" "Cmeg" "Cmeg" "Lexi" "Lexi" "Lexi"
[29] "Lexi" "Lexi" "Lexi" "Lexi"
```

# RNA-Seq: Differential Gene Expression

## Viewing data

### 17. Create sample table with metadata

```
# Species = First four characters
species <- substr(sampleNames, 1, 4)
species
[1] "Calb" "Calb" "Calb" "Calb" "Calb" "Calb" "Calb" "Cbez"
[8] "Cbez" "Chom" "Chom" "Chom" "Chom" "Chom" "Chom" "Chom"
[15] "Cmac" "Cmac" "Cmac" "Cmac" "Cmac" "Cmac" "Cmac" "Cmeg"
[22] "Cmeg" "Cmeg" "Cmeg" "Cmeg" "Cmeg" "Lexi" "Lexi" "Lexi"
[29] "Lexi" "Lexi" "Lexi" "Lexi"

# Stage = Fifth characters (F or L)
stage <- substr(sampleNames, 5, 5)
```

# RNA-Seq: Differential Gene Expression

## Viewing data

### 17. Create sample table with metadata

```
# Species = First four characters
species <- substr(sampleNames, 1, 4)
species
[1] "Calb" "Calb" "Calb" "Calb" "Calb" "Calb" "Calb" "Cbez"
[8] "Cbez" "Chom" "Chom" "Chom" "Chom" "Chom" "Chom" "Chom"
[15] "Cmac" "Cmac" "Cmac" "Cmac" "Cmac" "Cmac" "Cmac" "Cmeg"
[22] "Cmeg" "Cmeg" "Cmeg" "Cmeg" "Cmeg" "Lexi" "Lexi" "Lexi"
[29] "Lexi" "Lexi" "Lexi" "Lexi"

# Stage = Fifth characters (F or L)
stage <- substr(sampleNames, 5, 5)
stage
[1] "F" "F" "F" "L" "L" "L" "F" "L" "F" "F" "F" "L" "L"
[14] "L" "F" "F" "F" "L" "L" "L" "F" "F" "F" "F" "L" "L"
[27] "F" "F" "F" "L" "L" "L"
```

# RNA-Seq: Differential Gene Expression

## Viewing data

17. Create sample table with metadata

```
# Replacing F with "female" and L with "larva"  
stage <- ifelse(stage == "F", "female", "larva")
```

# RNA-Seq: Differential Gene Expression

## Viewing data

### 17. Create sample table with metadata

```
# Replacing F with "female" and L with "larva"
stage <- ifelse(stage == "F", "female", "larva")
stage
[1] "female" "female" "female" "larva"  "larva"  "larva"
[7] "female" "larva"  "female" "female" "female" "larva"
[13] "larva"  "larva"  "female" "female" "female" "larva"
[19] "larva"  "larva"  "female" "female" "female" "larva"
[25] "larva"  "larva"  "female" "female" "female" "larva"
[31] "larva"  "larva"
```

# RNA-Seq: Differential Gene Expression

## Viewing data

### 17. Create sample table with metadata

```
#Create datframe
meta <- data.frame(sampletype, species, stage,
                     row.names = sampleNames)
meta
  sampletype species stage
CalbF_1      CalbF    Calb female
CalbF_2      CalbF    Calb female
CalbF_4      CalbF    Calb female
CalbL_1      CalbL    Calb larva
CalbL_2      CalbL    Calb larva
CalbL_4      CalbL    Calb larva
CbezF_1      CbezF    Cbez female
CbezL_1      CbezL    Cbez larva
...
...
```



Step 1 complete!

# RNA-Seq: Differential Gene Expression

## Exploring RNA-seq count data

### Count matrix

Gene	CalbF_1	CalbF_4	CalbL_1	CalbL_4	CbezF_1	CbezL_1
NAD2	13338	9210	4277	1992	52751	83334
COX1	550546	633535	340001	140847	980982	3266084
COX2	153949	135999	81662	44389	208501	1081210
ATP8	384	420	98	94	4848	6189
ATP6	154399	136590	47777	42332	225729	507499
COX3	350303	299530	158793	76511	503434	1122979
NAD3	8552	14303	3914	4963	16160	42497
NAD5	52584	50086	16228	9744	130641	227652
NAD4	52447	65090	18445	12634	226526	337825
NAD4L	786	597	116	96	4002	5829

# RNA-Seq: Differential Gene Expression

## Exploring RNA-seq count data

### Distribution of RNA-Seq counts

```
ggplot(data) +  
  geom_histogram(aes(x = CalbF_1_mtDNA),  
                 stat = "bin", bins = 10) +  
  xlab("Raw expression counts") +  
  ylab("Number of genes")
```

# RNA-Seq: Differential Gene Expression

## Exploring RNA-seq count data

The screenshot shows the RStudio interface for an R session titled "06-DGEmtDNA - RStudio".

**Data View:** A data frame titled "DGEmtDNA.R" is displayed, showing counts for various genes across different samples. The columns include NAD2, COX1, COX2, ATP8, ATP6, COX3, NAD3, and NADS. The rows show sample identifiers like CalbF\_1\_mtDNA, CalbF\_4\_mtDNA, CalbL\_1\_mtDNA, etc.

	CalbF_1_mtDNA	CalbF_4_mtDNA	CalbL_1_mtDNA	CalbL_4_mtDNA	CbezF_1_mtDNA	CbezL_1_mtDNA	ChomF_1
NAD2	13338.000	9209.882	4277.000	1992.000	52750.995	83333.902	
COX1	550546.339	633534.659	340000.948	140846.664	980982.446	3266084.093	
COX2	153949.241	135999.086	81661.996	44389.479	208500.952	1081209.747	
ATP8	384.319	420.091	98.212	93.753	4848.458	6188.946	
ATP6	154399.212	136590.126	47777.419	42332.396	225728.867	507499.188	
COX3	350302.849	299529.932	158793.378	76511.253	503434.356	1122978.899	
NAD3	8552.196	14303.001	3914.000	4963.454	16160.398	42497.005	
NADS	52583.744	50085.784	16228.000	9744.000	130641.297	227652.322	

**Environment View:** Shows the global environment with objects "txi" and "files".

**Console View:** Displays R code for reading Salmon count data and creating a tximport object.

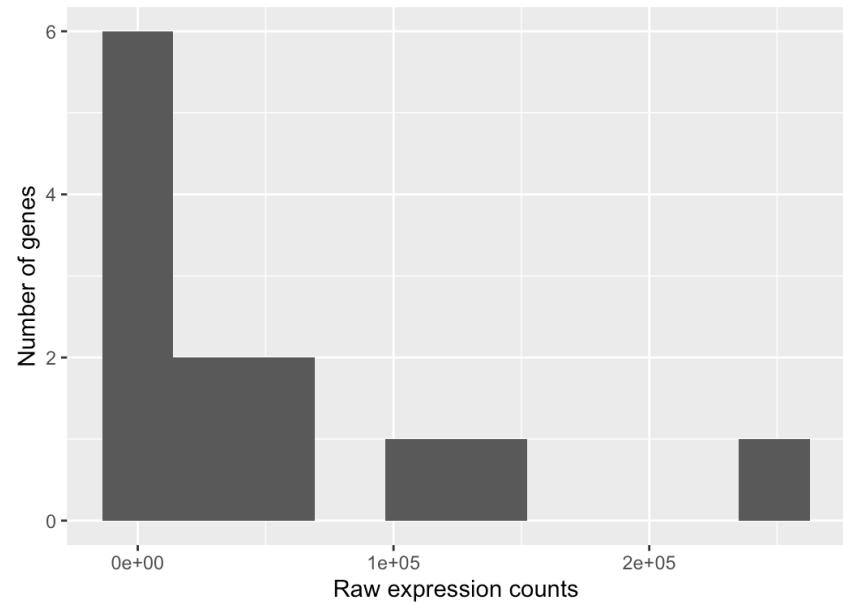
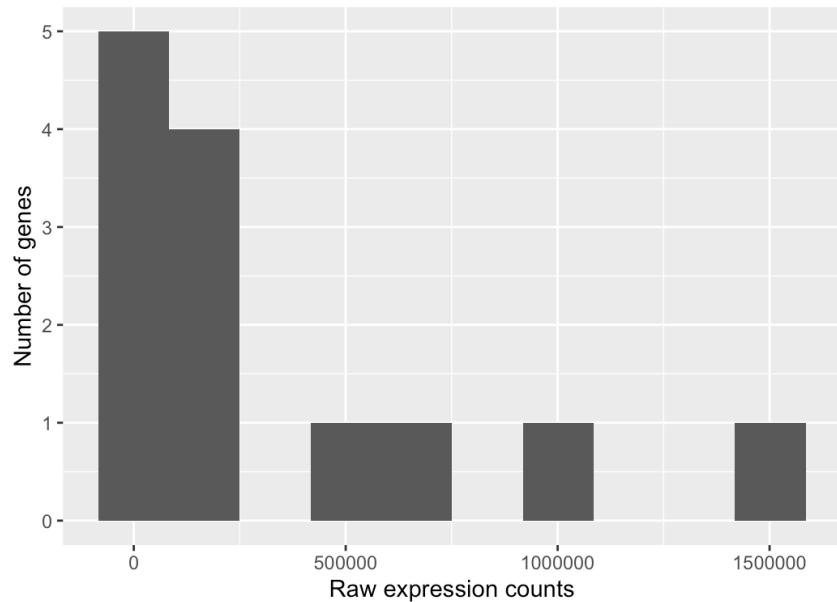
```
R 4.4.2 · ~/rnaseq/06-DGEmtDNA/ ▶
[1] "./allCountData//LcupL_1_mtDNA-salmon" "./allCountData//LcupL_3_mtDNA-salmon"
[21] "./allCountData//LcupL_1_mtDNA-salmon" "./allCountData//LcupL_3_mtDNA-salmon"
[23] "./allCountData//LexiF_1_mtDNA-salmon" "./allCountData//LexiF_2_mtDNA-salmon"
[25] "./allCountData//LexiL_1_mtDNA-salmon" "./allCountData//LexiL_2_mtDNA-salmon"
> names(files) <- str_replace(samples, "./allCountData/", "") %>%
+   str_replace("-salmon", "") %>% str_replace("/", "") 
>
> ## Run tximport
> txi <- tximport(files, type="salmon",
+                  # countsFromAbundance="lengthScaledTPM",
+                  txOut=TRUE)
reading in files with read_tsv
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
> beep(sound = 2)
> txi$counts %>% View()
>
```

**File View:** Shows the file tree structure under the current working directory.

- ..
- CalbF\_1\_mtDNA-salmon
- CalbF\_4\_mtDNA-salmon
- CalbL\_1\_mtDNA-salmon
- CalbL\_4\_mtDNA-salmon
- CbezF\_1\_mtDNA-salmon
- CbezL\_1\_mtDNA-salmon
- ChomF\_1\_mtDNA-salmon
- ChomF\_3\_mtDNA-salmon
- ChomL\_1\_mtDNA-salmon
- ChomL\_3\_mtDNA-salmon
- CmacF\_1\_mtDNA-salmon
- CmacF\_3\_mtDNA-salmon
- CmacL\_1\_mtDNA-salmon
- CmacL\_3\_mtDNA-salmon

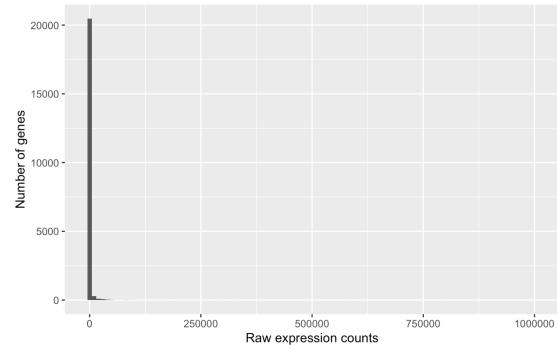
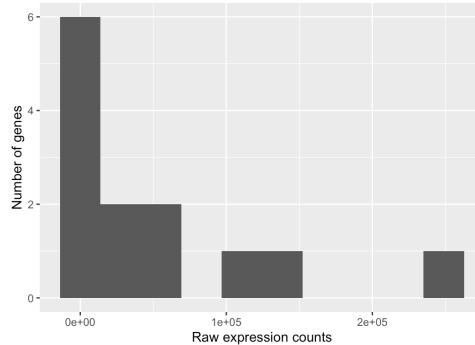
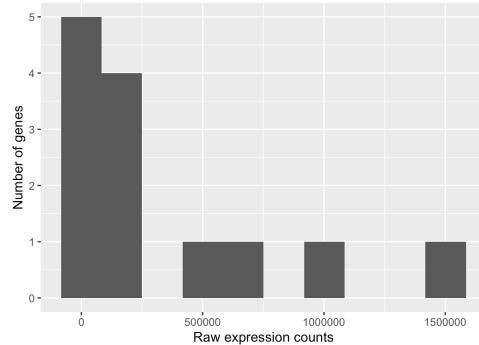
# RNA-Seq: Differential Gene Expression

## Exploring RNA-seq count data



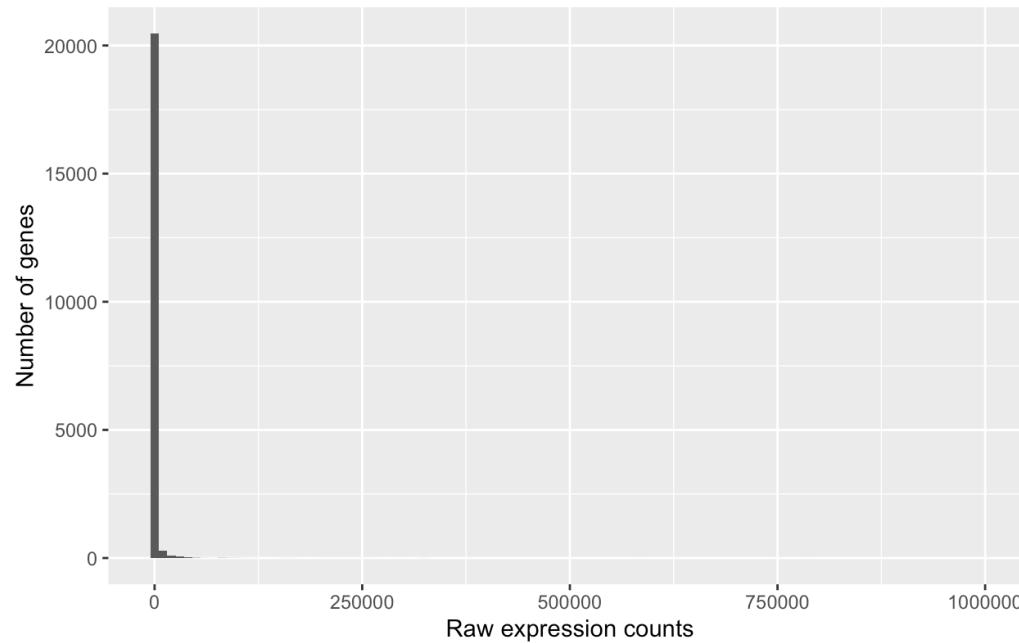
# RNA-Seq: Differential Gene Expression

## Exploring RNA-seq count data



# RNA-Seq: Differential Gene Expression

## Exploring RNA-seq count data



# RNA-Seq: Differential Gene Expression

## Exploring RNA-seq count data

### Common features of RNA-seq count data:

- ☞ a low number of counts associated with a large proportion of genes;
- ☞ a long right tail due to the lack of any upper limit for expression;
- ☞ large dynamic range;
- ☞ not normally distributed.

# RNA-Seq: Differential Gene Expression

## Modeling count data

What model should we use for RNA-Seq data?

Binomial Distribution

Poisson Distribution

# RNA-Seq: Differential Gene Expression

## Modeling count data

### Binomial Distribution

The binomial distribution is used when you're counting the number of successes in a fixed number of independent experiments, where each experiment has only two possible outcomes: success or failure.

#### Key Features:

- Fixed number of trials ( $n$ ).
- Each trial has the same probability of success ( $p$ ).
- Trials are independent.

# RNA-Seq: Differential Gene Expression

## Modeling count data

### Poisson Distribution

The Poisson distribution is used when you're counting how many events happen in a fixed interval of time or space, assuming the events happen independently and at a constant average rate.

#### Key Features:

- The average rate of occurrence ( $\lambda$ , lambda) is known.
- The events are random and independent.
- No fixed number of trials; you count occurrences.

# RNA-Seq: Differential Gene Expression

## Modeling count data

### Comparing the Two:

- Binomial: Used when you have a fixed number of trials and want to count successes (e.g., flipping a coin 10 times).
- Poisson: Used when you count occurrences over time or space with no fixed number of trials and the likelihood of any particular event is low. (e.g., gaining the lottery).

Binomial: "How many people in a group of 100 will like chocolate if each person has a 70% chance of liking it?"

Poisson: "How many people will show up to my party if, on average, 10 people show up every hour?"

# RNA-Seq: Differential Gene Expression

## Modeling count data

What model should we use for RNA-Seq data?

Binomial Distribution

Poisson Distribution

# RNA-Seq: Differential Gene Expression

## Modeling count data

What model should we use for RNA-Seq data?

Binomial Distribution X

Poisson Distribution ?

# RNA-Seq: Differential Gene Expression

## Modeling count data

What model should we use for RNA-Seq data?

Binomial Distribution X

Poisson Distribution ?

It depends on the mean and variance!

# RNA-Seq: Differential Gene Expression

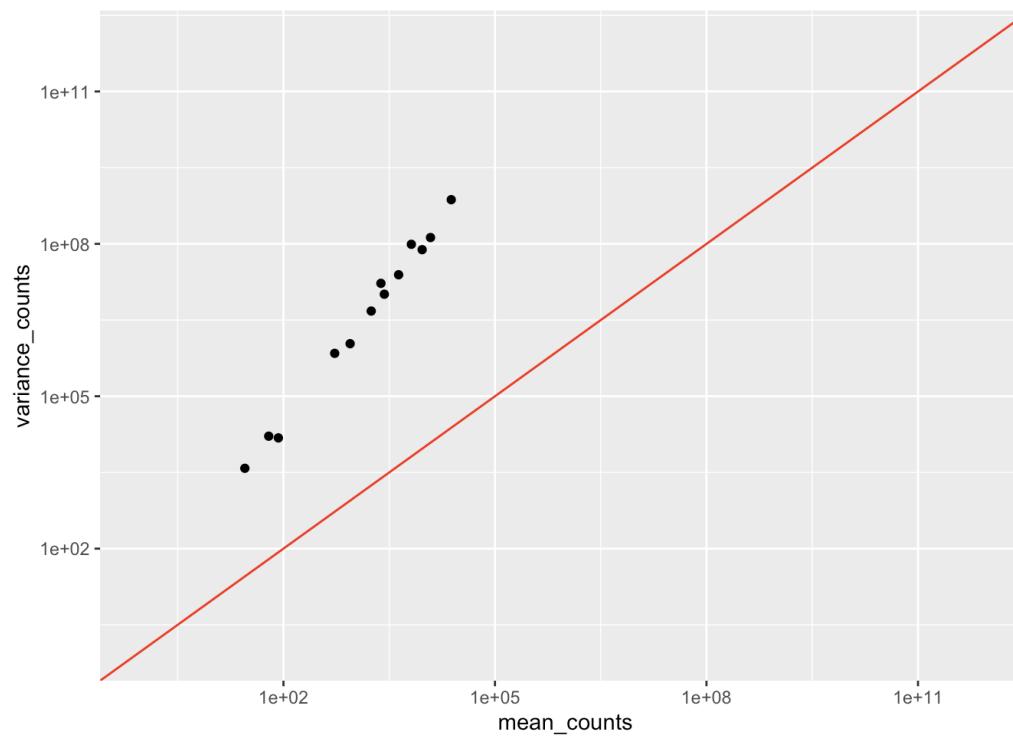
## Modeling count data

```
mean_counts <- apply(data[,1:26], 1, mean)
variance_counts <- apply(data[,1:26], 1, var)
df <- data.frame(mean_counts, variance_counts)

ggplot(df) +
  geom_point(aes(x=mean_counts, y=variance_counts)) +
  scale_y_log10(limits = c(1,1e12)) +
  scale_x_log10(limits = c(1,1e12)) +
  geom_abline(intercept = 0, slope = 1, color="red")
```

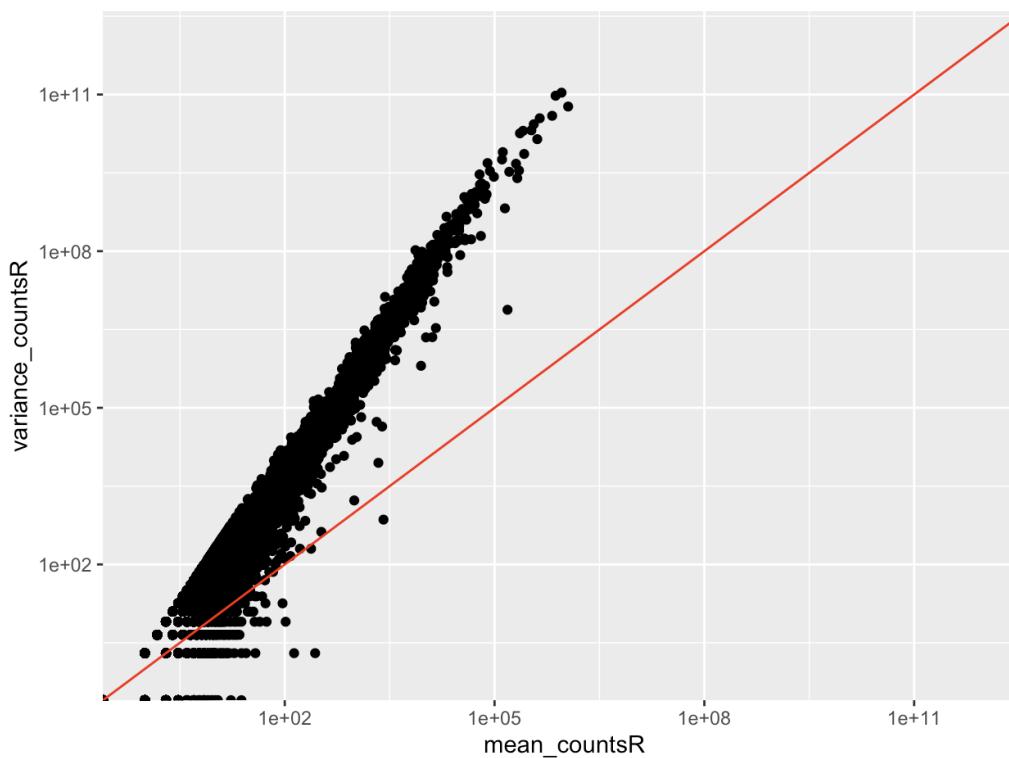
# RNA-Seq: Differential Gene Expression

## Modeling count data



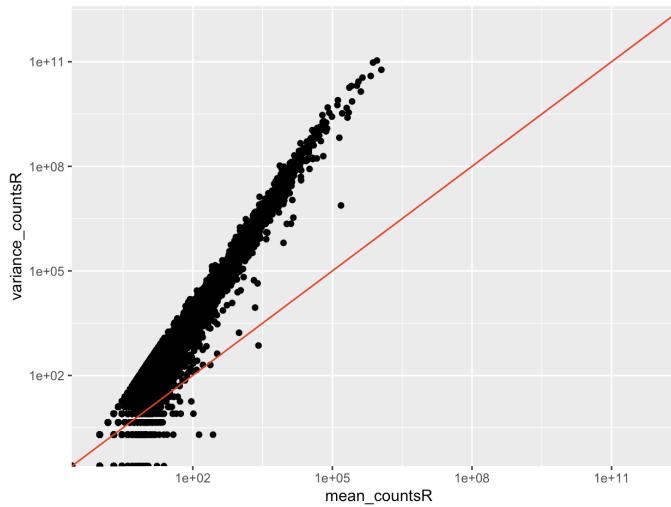
# RNA-Seq: Differential Gene Expression

## Modeling count data



# RNA-Seq: Differential Gene Expression

## Modeling count data



- 💡 The mean is not equal to the variance
- 💡 For the genes with high mean expression, the variance across replicates tends to be greater than the mean
- 💡 For the genes with low mean expression, the variability or spread in the data is high

# RNA-Seq: Differential Gene Expression

## Modeling count data

What model should we use for RNA-Seq data?

Binomial Distribution

Poisson Distribution

# RNA-Seq: Differential Gene Expression

## Modeling count data

What model should we use for RNA-Seq data?

Binomial Distribution

Poisson Distribution

Negative Binomial Distribution

# RNA-Seq: Differential Gene Expression

## Modeling count data

### Negative Binomial Distribution

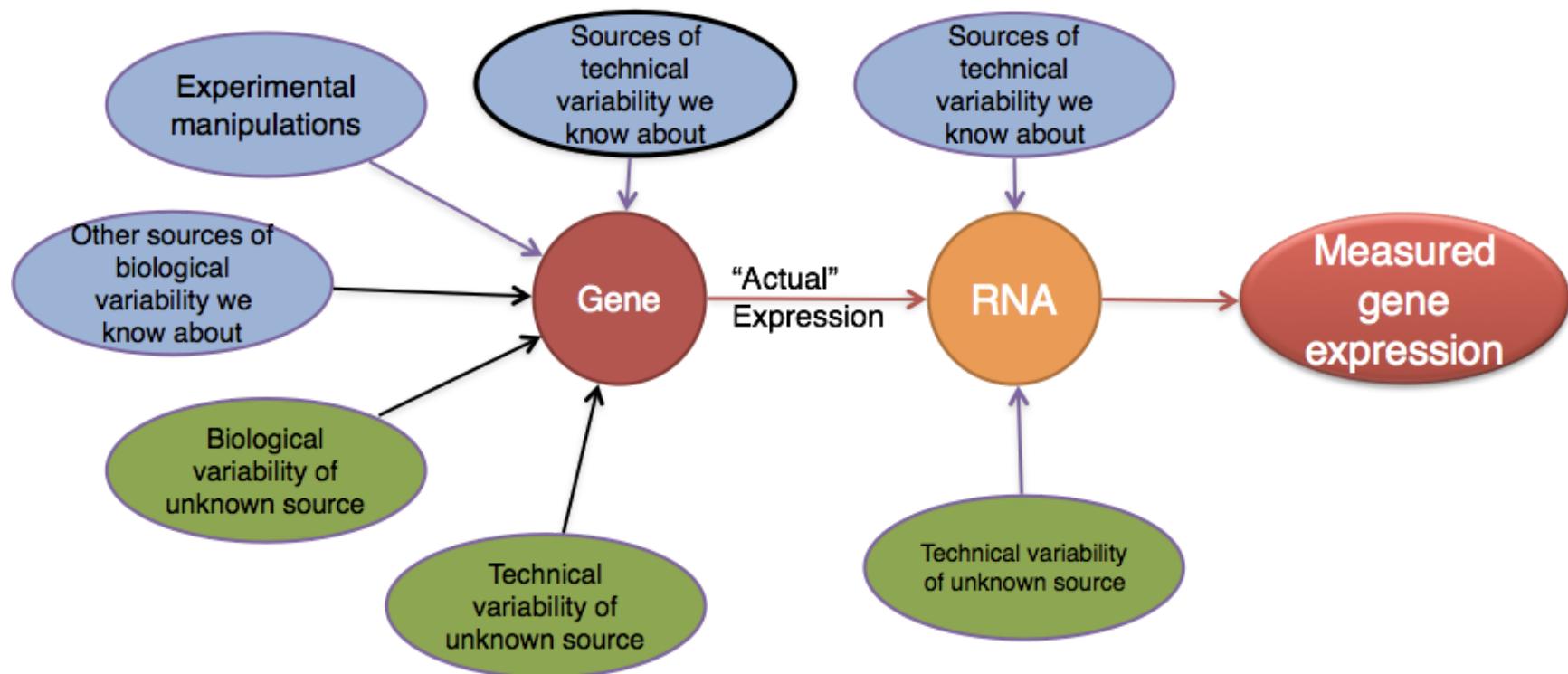
The negative binomial distribution is used when you're counting the number of trials required to achieve a fixed number of successes, where each trial has only two possible outcomes: success or failure.

#### Key Features:

- 💡 Fixed number of successes ( $r$ ).
- 💡 Each trial has the same probability of success ( $p$ ).
- 💡 Trials are independent.
- 💡 The number of trials required is variable.

# RNA-Seq: Differential Gene Expression

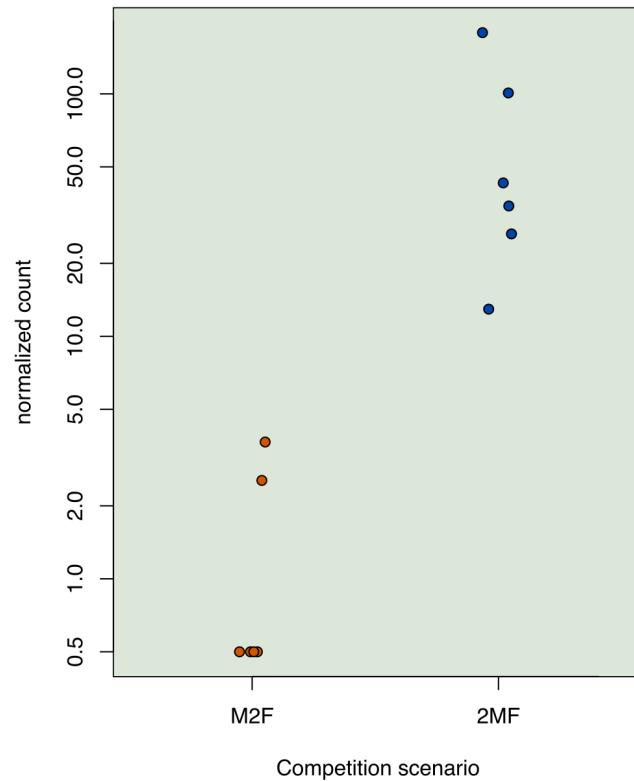
## Replicates and variability



Courtesy of Paul Pavlidis, UBC

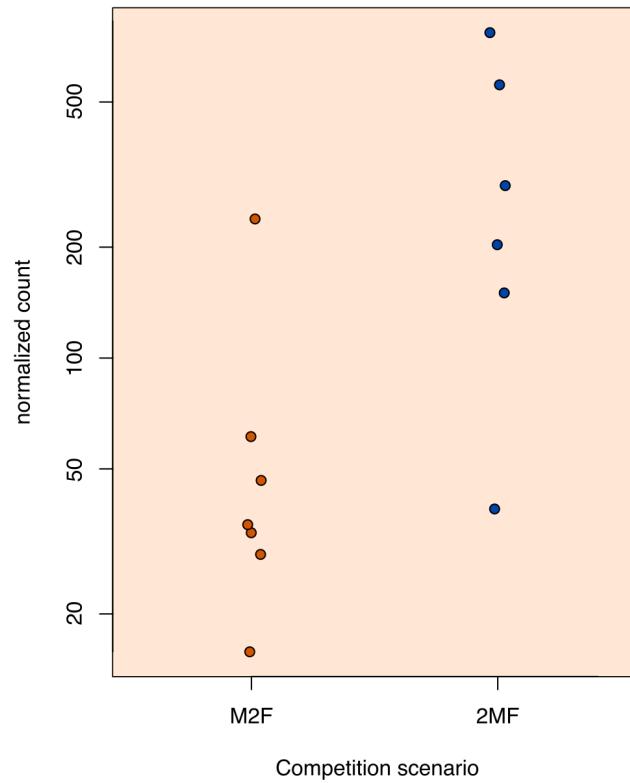
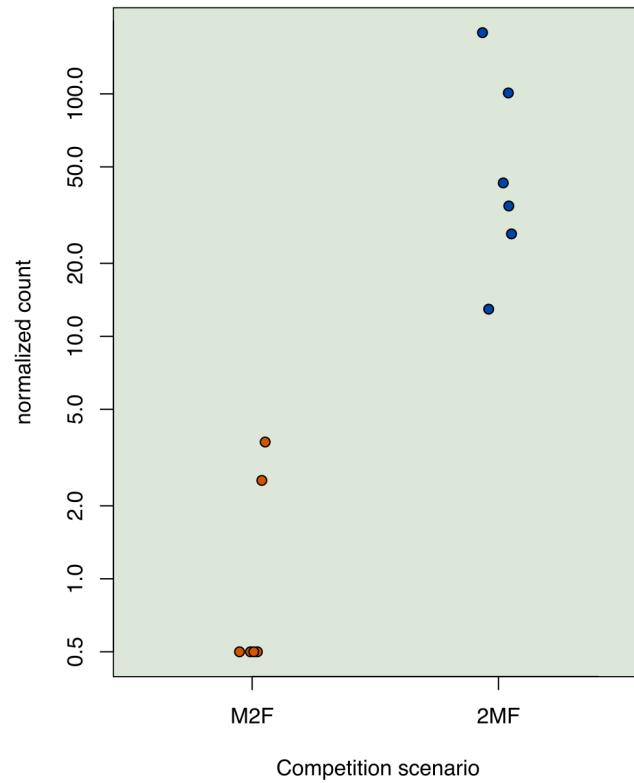
# RNA-Seq: Differential Gene Expression

## Replicates and variability



# RNA-Seq: Differential Gene Expression

## Replicates and variability



# RNA-Seq: Differential Gene Expression

## How many replicates?

- estimate variation for each gene
- randomize out unknown covariates
- spot outliers
- improve precision of expression and fold-change estimates

