TRANSCRIPTOMICS

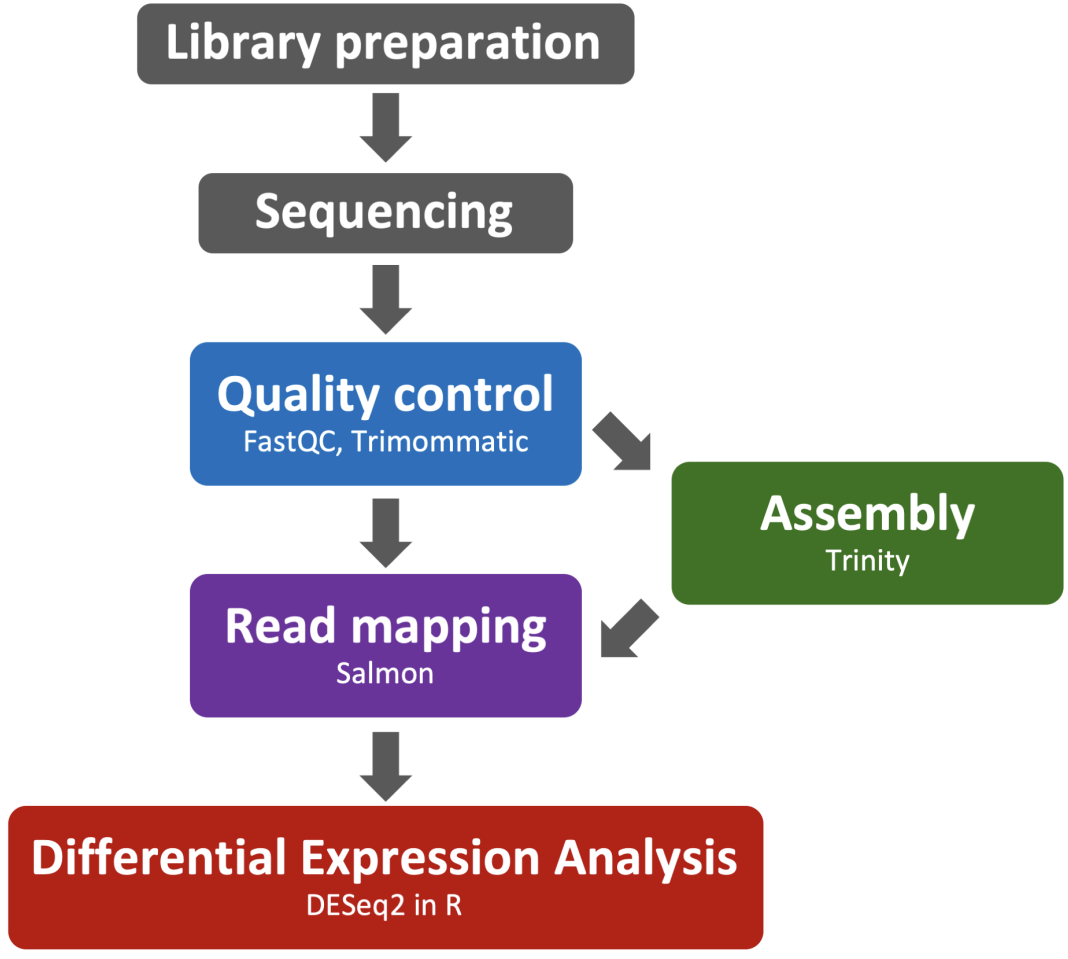
Differential Gene Expression

Day 06

https://tttorres.github.io/transcriptomics/

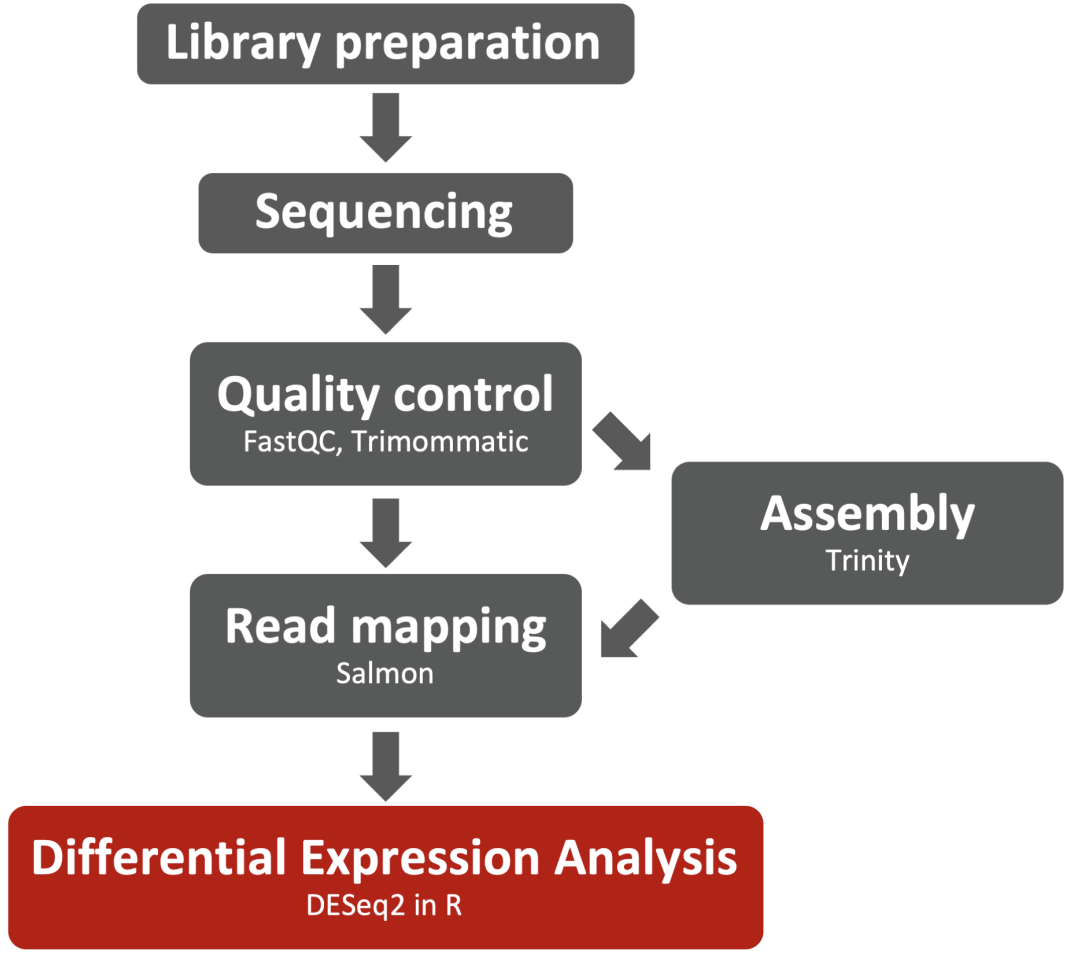
RNA-seq workflow

Differential Gene Expression



RNA-seq workflow

Differential Gene Expression



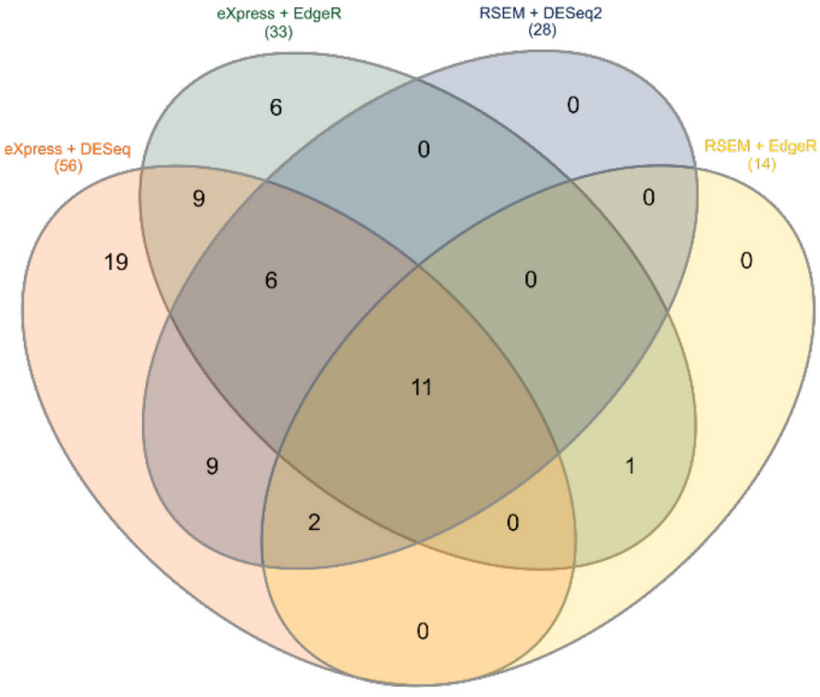
RNA-Seq: Differential Gene Expression

- 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

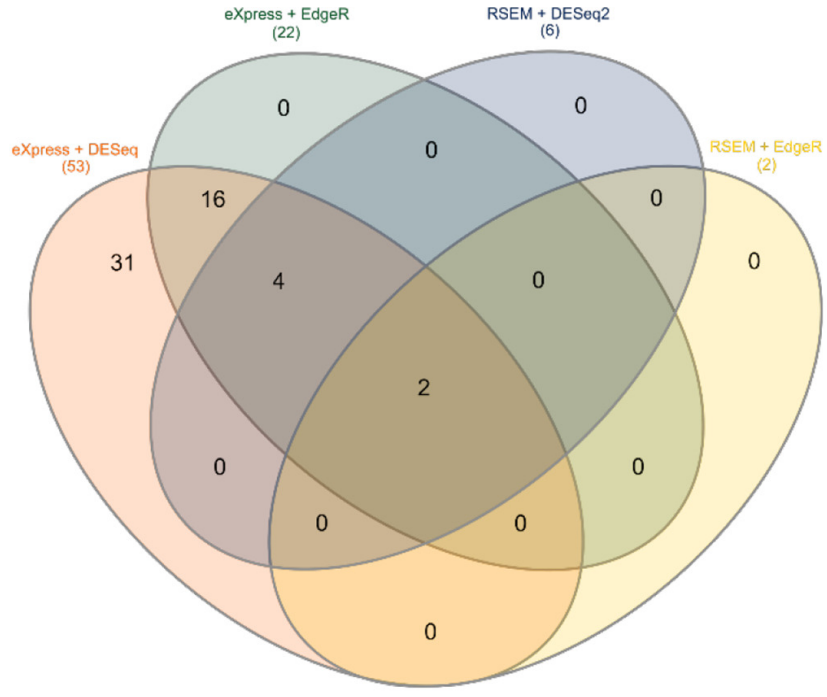
RNA-Seq: Differential Gene Expression

Choosing package/model

A



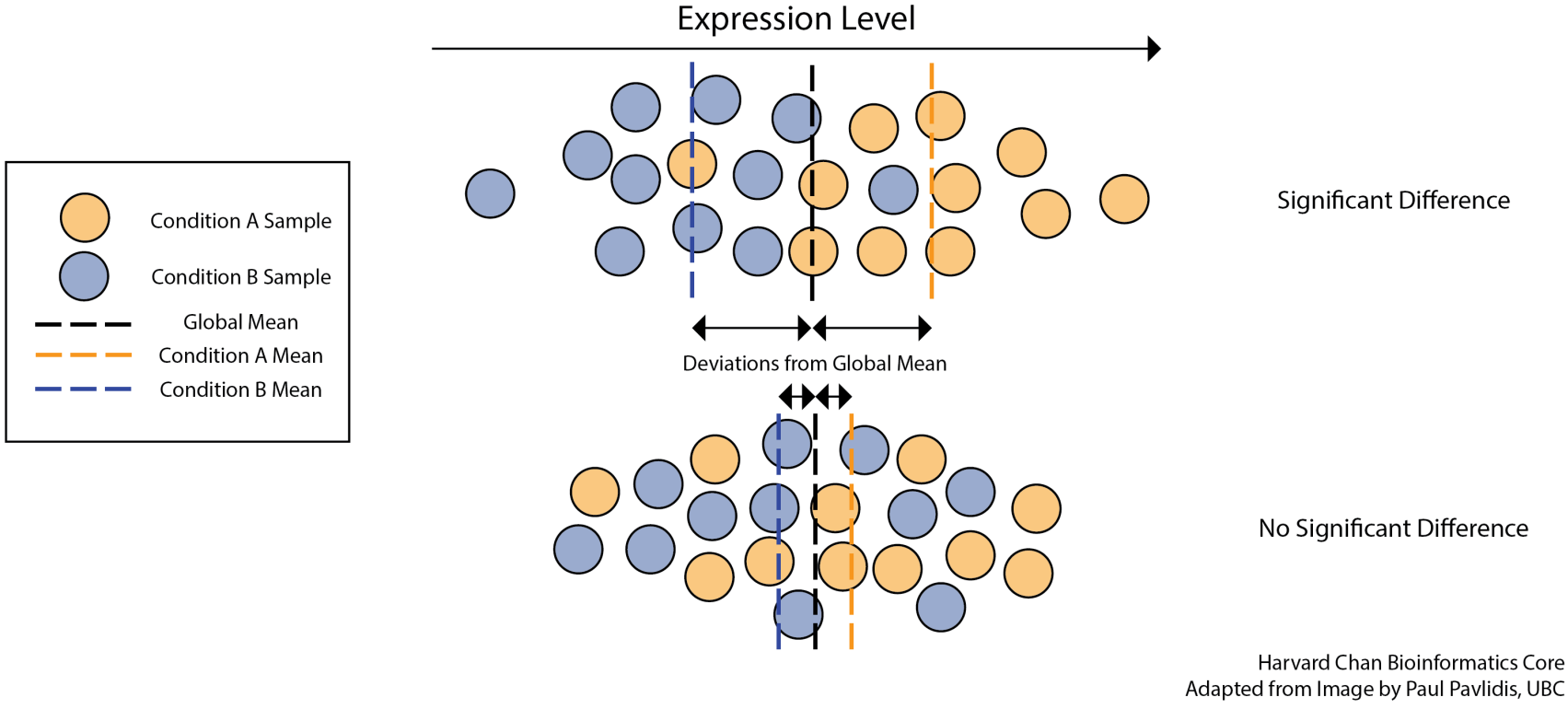
B



Deszo, Cardoso & Torres (2021) Genomics, 113, 699-706

RNA-Seq: Differential Gene Expression

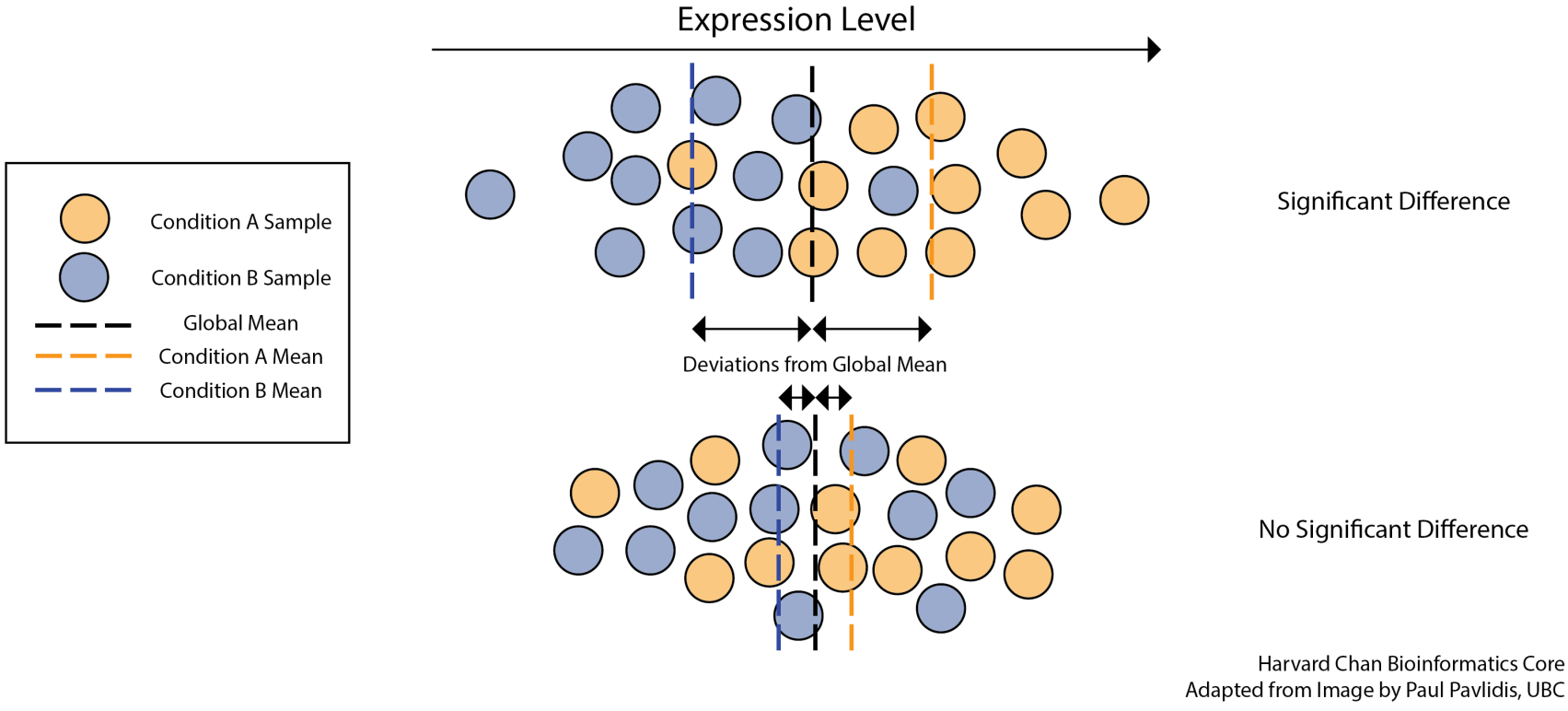
Differential expression analysis with DESeq2



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

RNA-Seq: Differential Gene Expression

Differential expression analysis with DESeq2

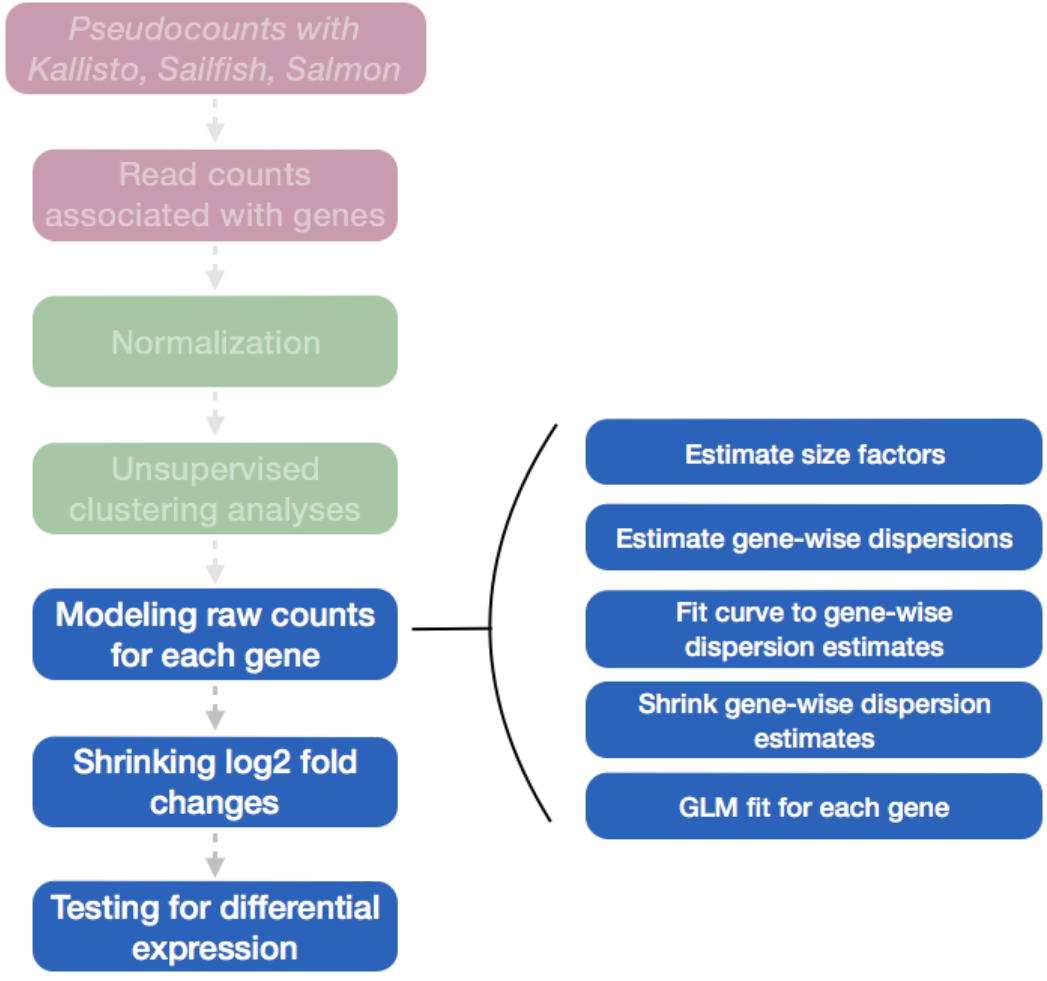


Harvard Chan Bioinformatics Core
Adapted from Image by Paul Pavlidis, UBC

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

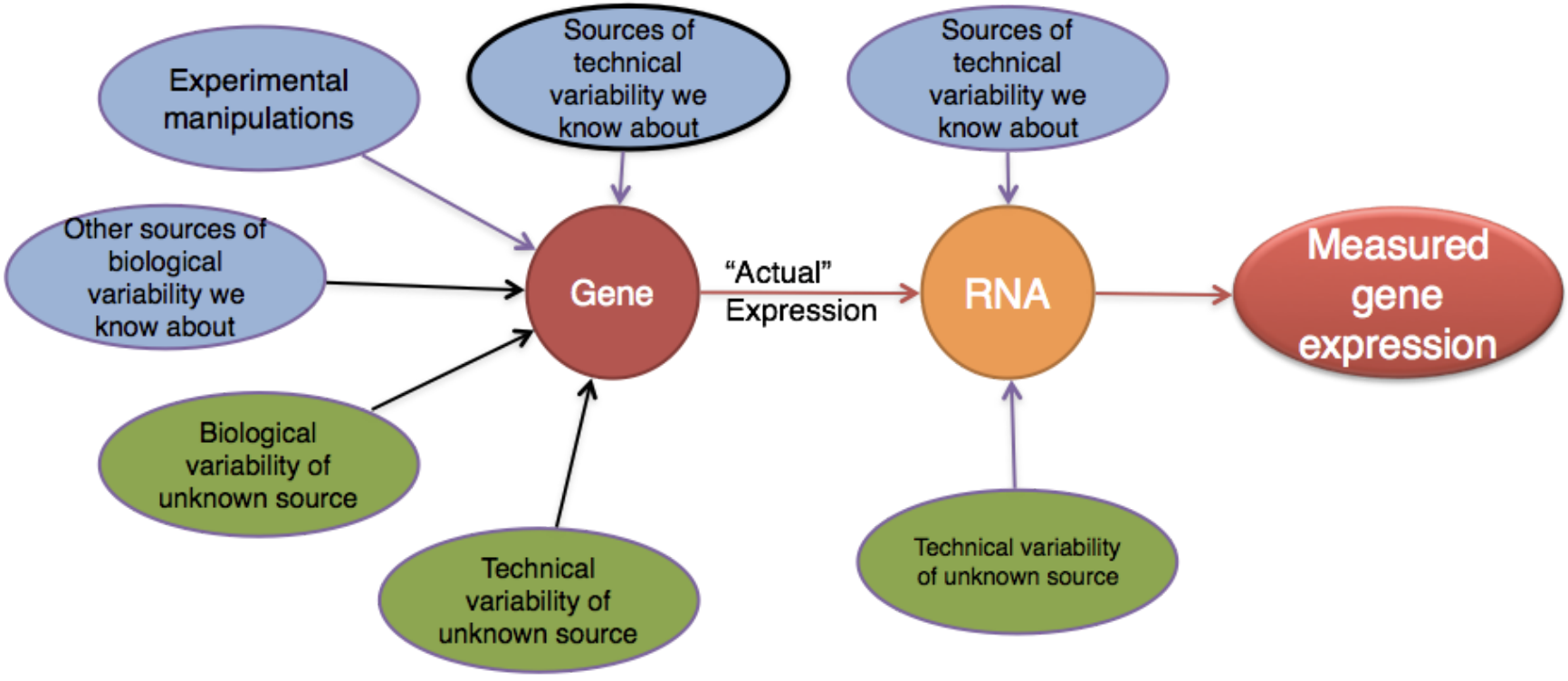
RNA-Seq: Differential Gene Expression

Differential expression analysis with DESeq2



RNA-Seq: Differential Gene Expression

Review of the dataset



Courtesy of Paul Pavlidis, UBC

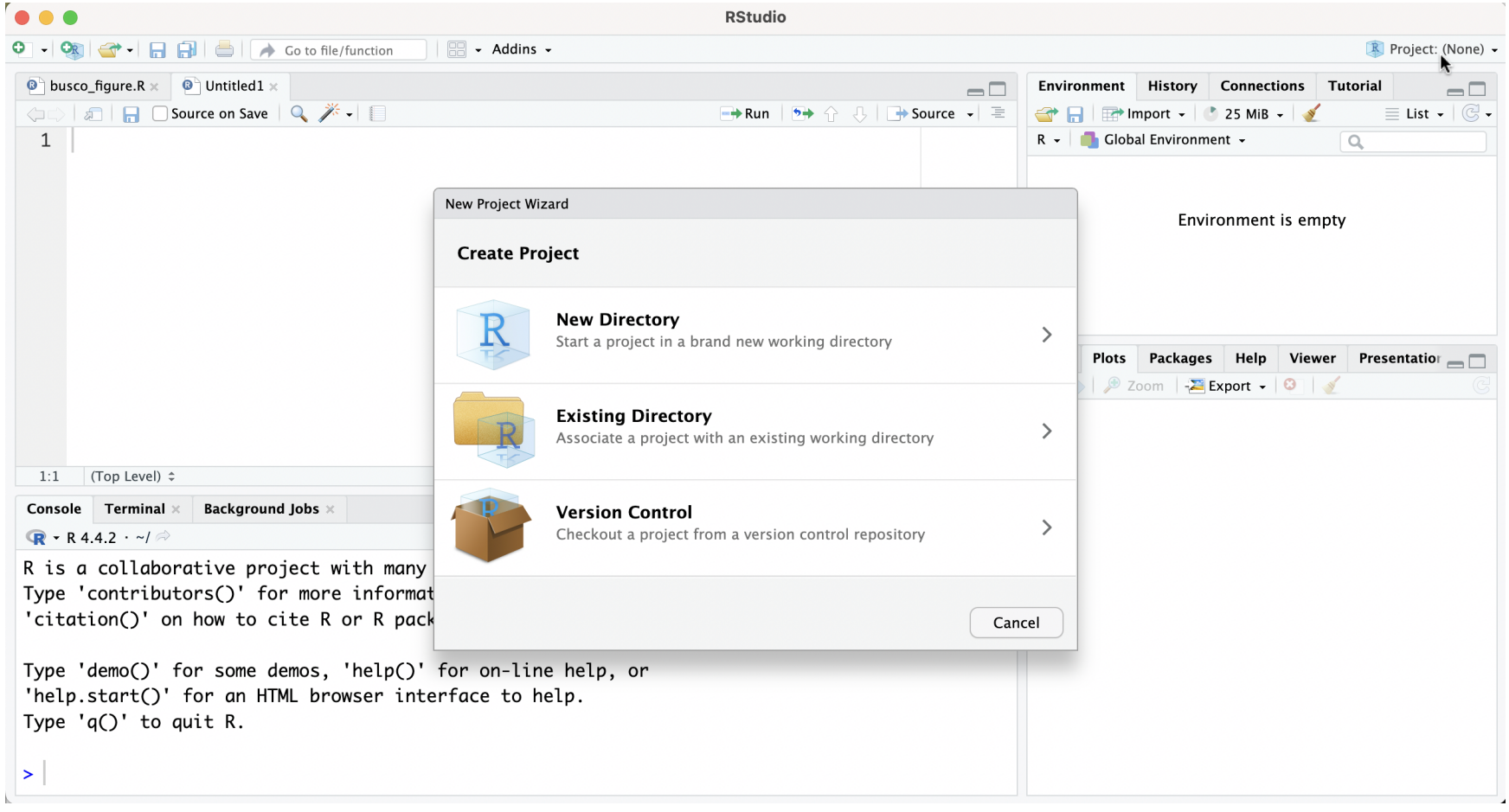
RNA-Seq: Differential Gene Expression

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



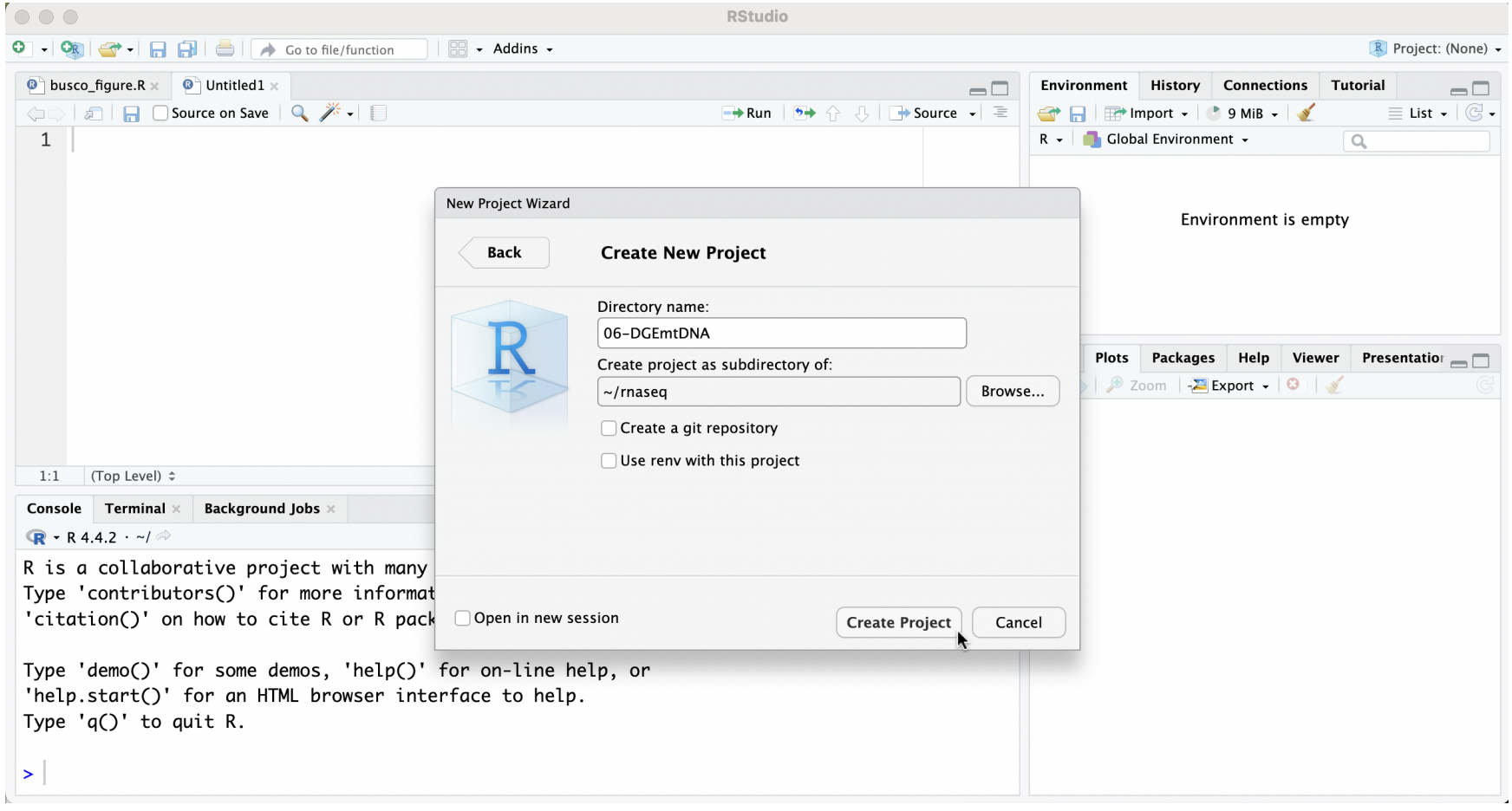
RNA-Seq: Differential Gene Expression

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

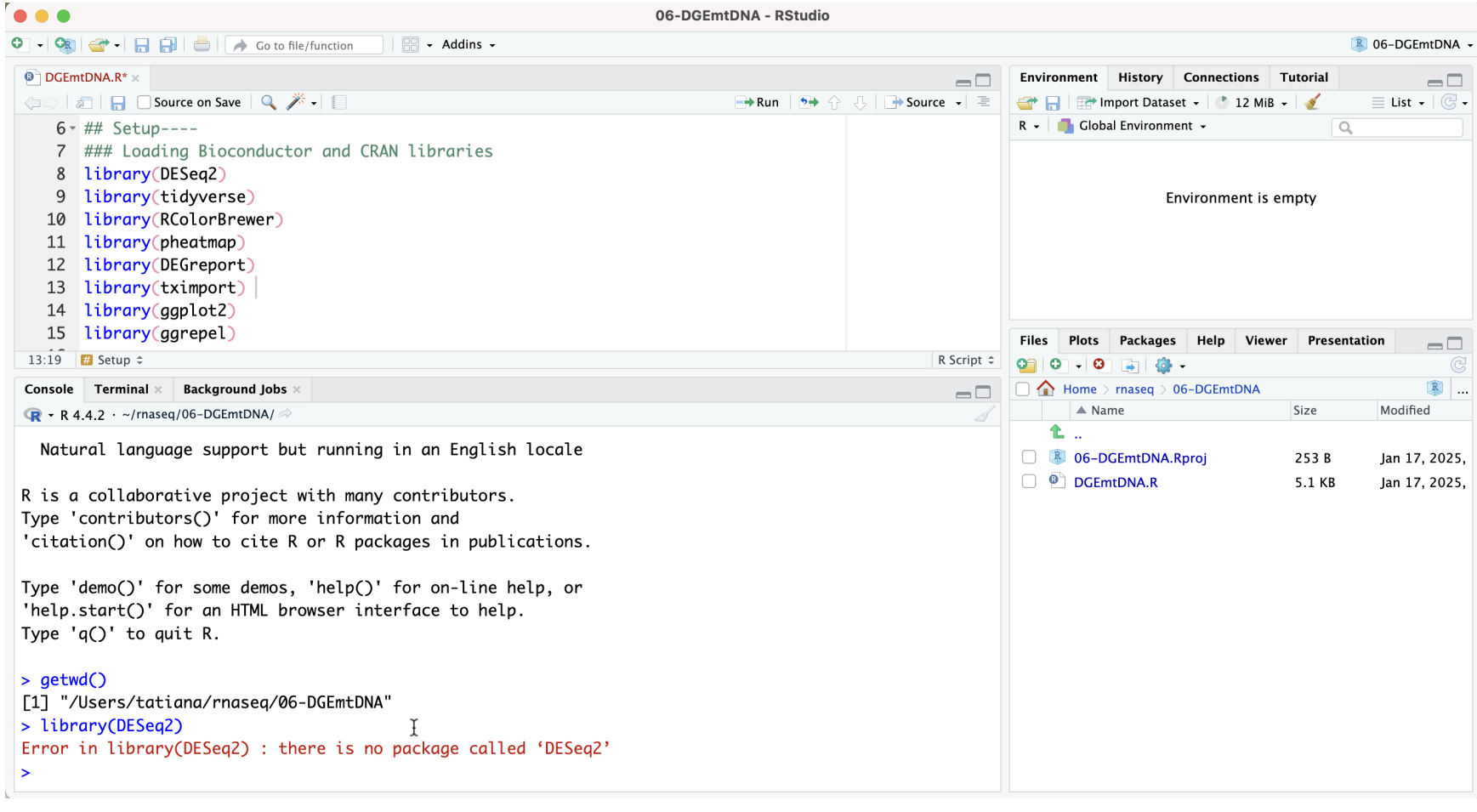
3. Check (in R console) if you are in the correct working directory

```
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



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 EN
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

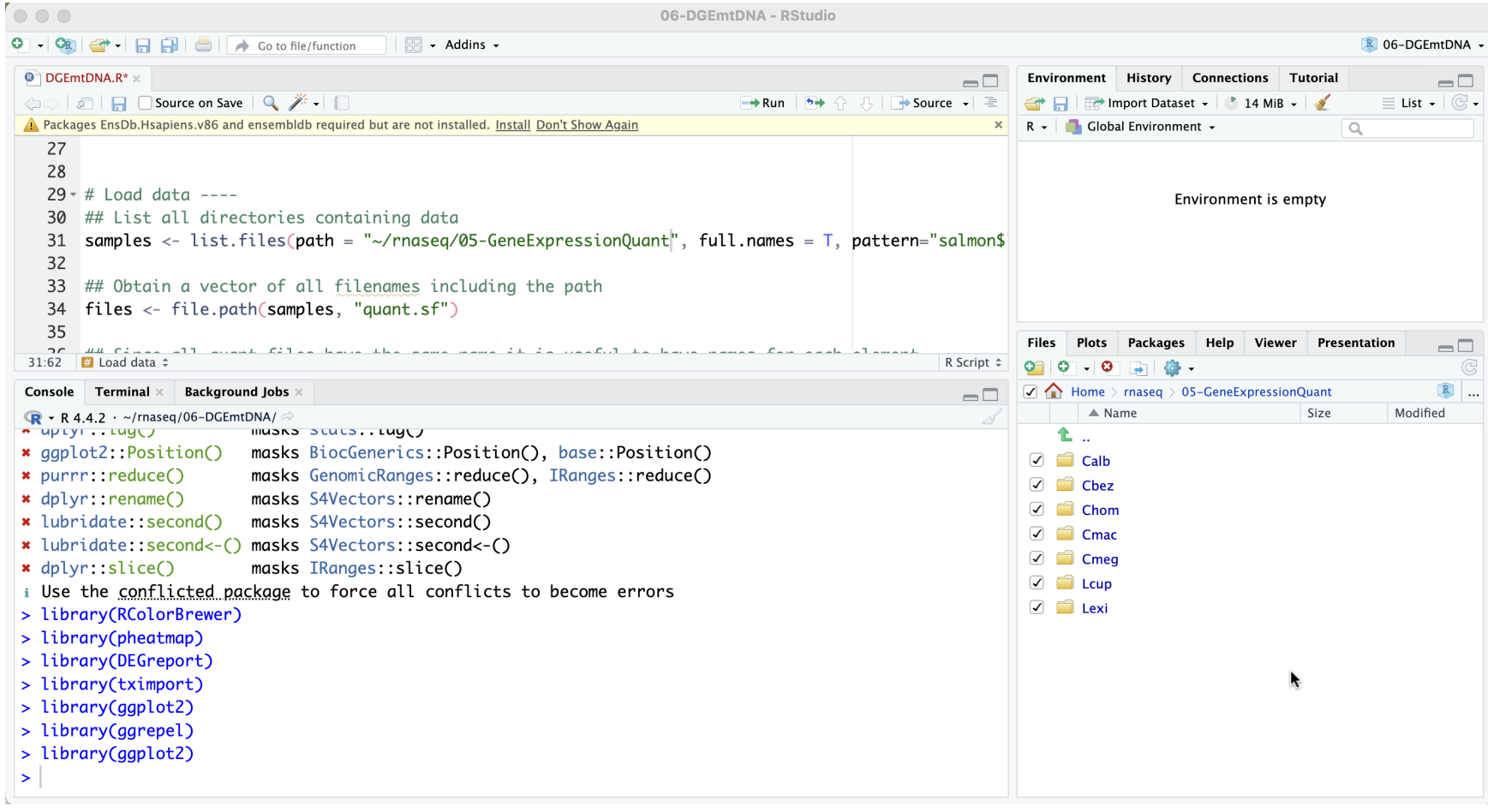
8. Download count data from the course shared drive

9. Create a variable with the paths to all quant.sf files.

```
## List all directories containing data
samples <- list.files(path = "~/rnaseq/06-DGEmtDNA",
                      full.names = T, pattern="salmon$")
```

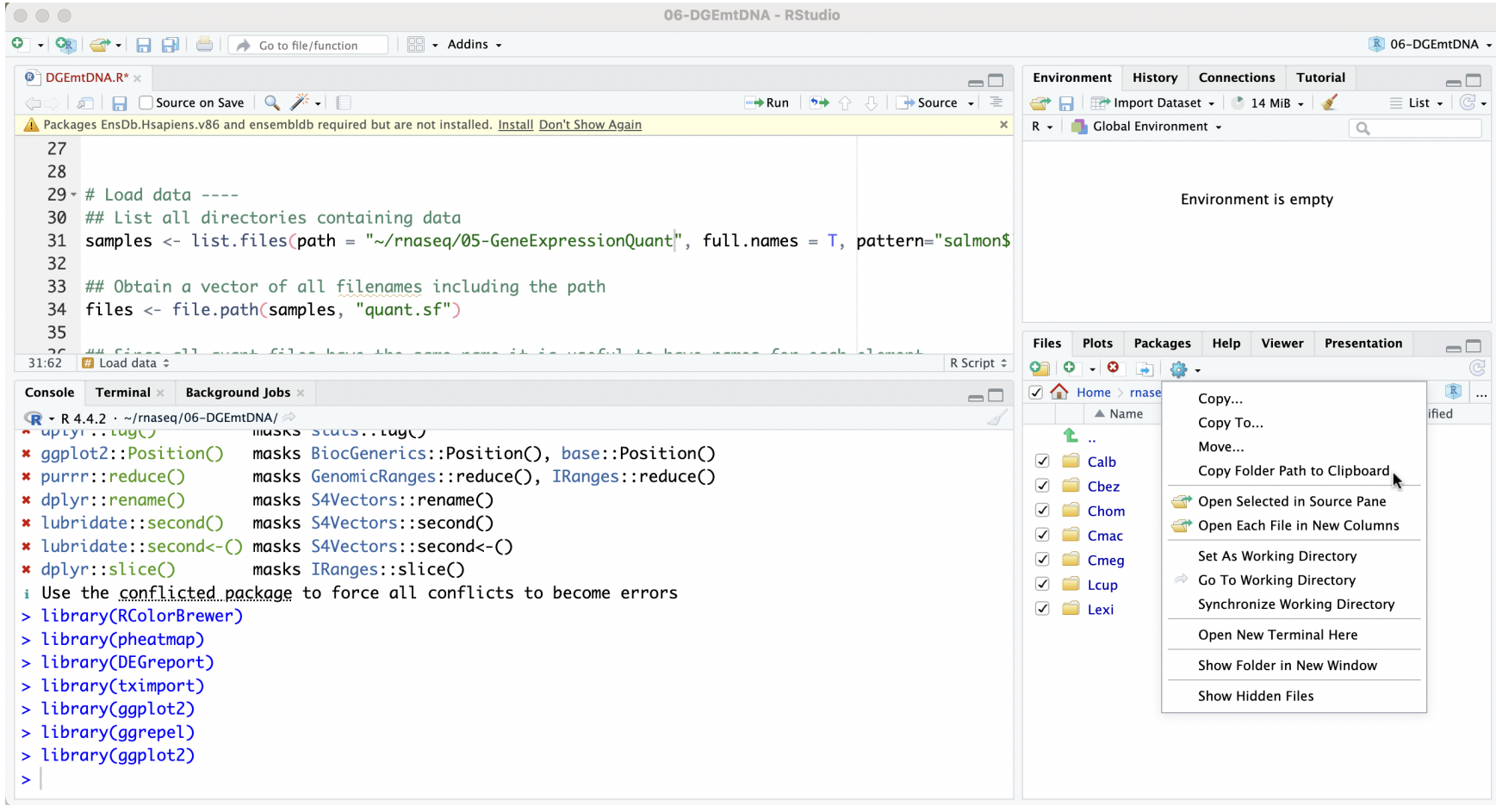
RNA-Seq: Differential Gene Expression

Setting up on RStudio



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

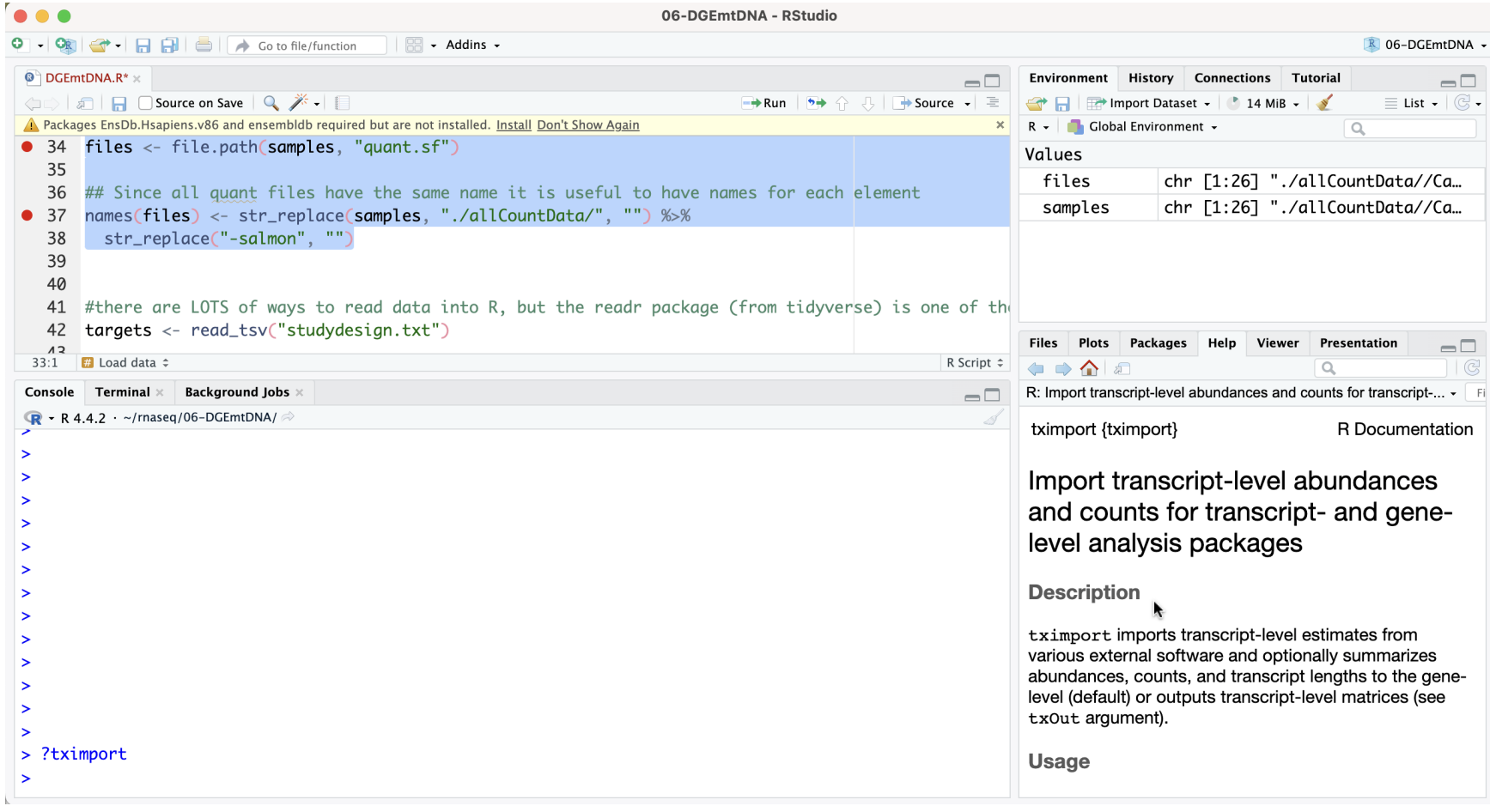
Package `tximport` to import `quant.sf` files

14. See `tximport` help

```
?tximport
```

RNA-Seq: Differential Gene Expression

Setting up on RStudio



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",
                txOut="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

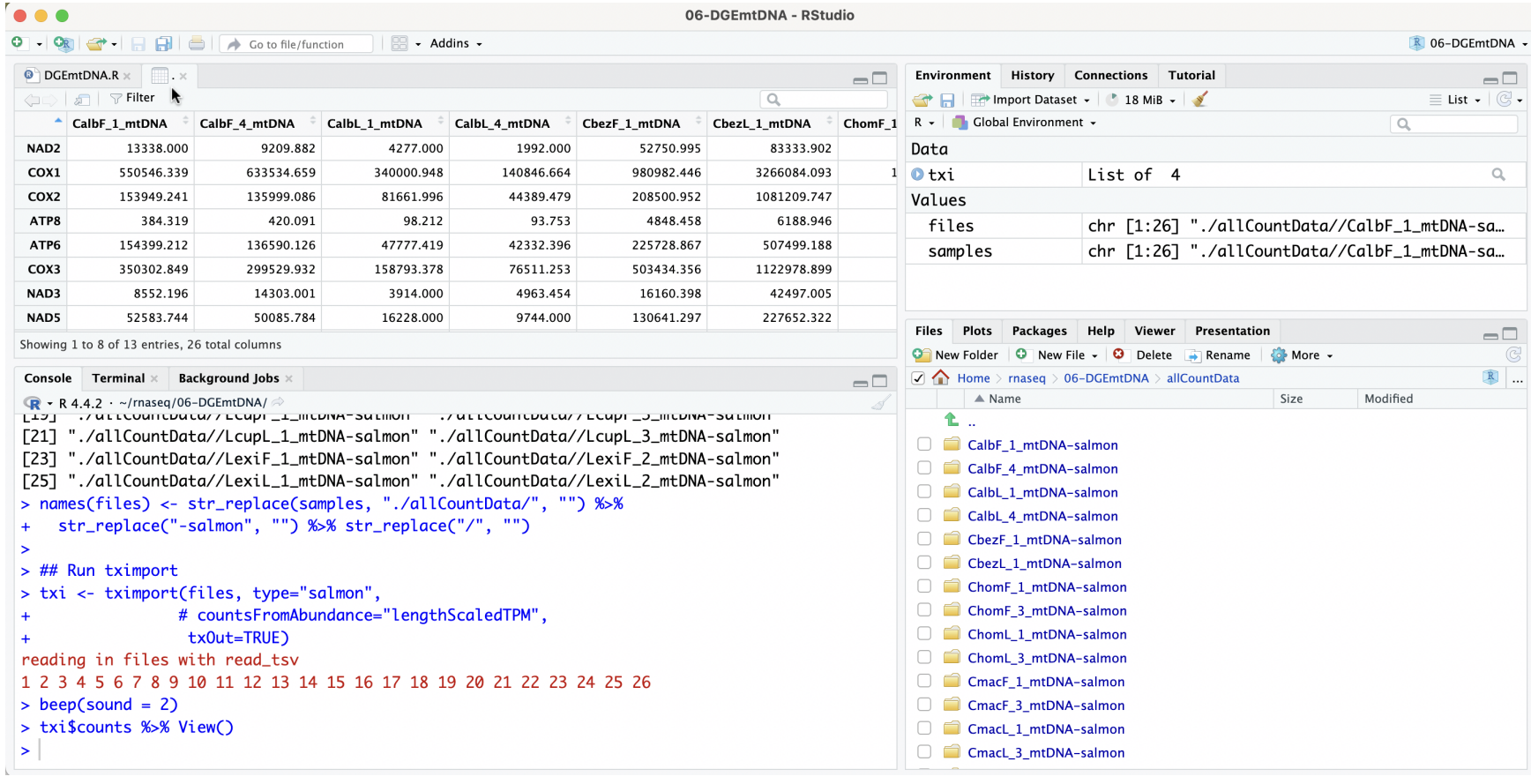
Viewing data

15. View the counts

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

RNA-Seq: Differential Gene Expression

Viewing data



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 metada

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

RNA-Seq: Differential Gene Expression

Viewing data

17. Create sample table with metada

```
# 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 metada

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

RNA-Seq: Differential Gene Expression

Viewing data

17. Create sample table with metada

```
# sampletype = First five characters
sampletype <- substr(sampleNames, 1, 5)
sampletype
 [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 metada

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

RNA-Seq: Differential Gene Expression

Viewing data

17. Create sample table with metada

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

RNA-Seq: Differential Gene Expression

Viewing data

17. Create sample table with metada

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

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

RNA-Seq: Differential Gene Expression

Viewing data

17. Create sample table with metada

```
# 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 metada

```
# 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 metada

```
#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

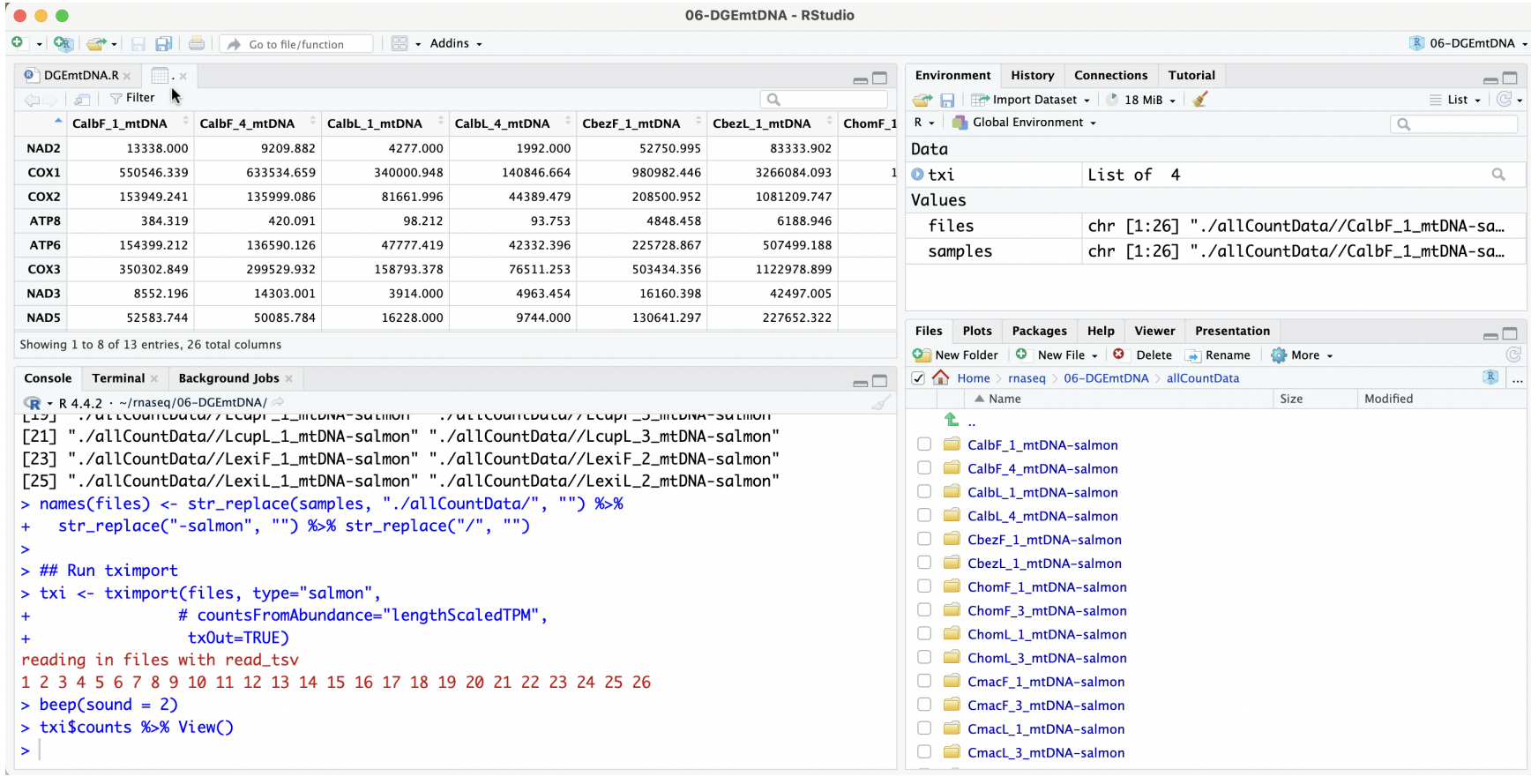
Exploring RNA-seq count data

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")
```

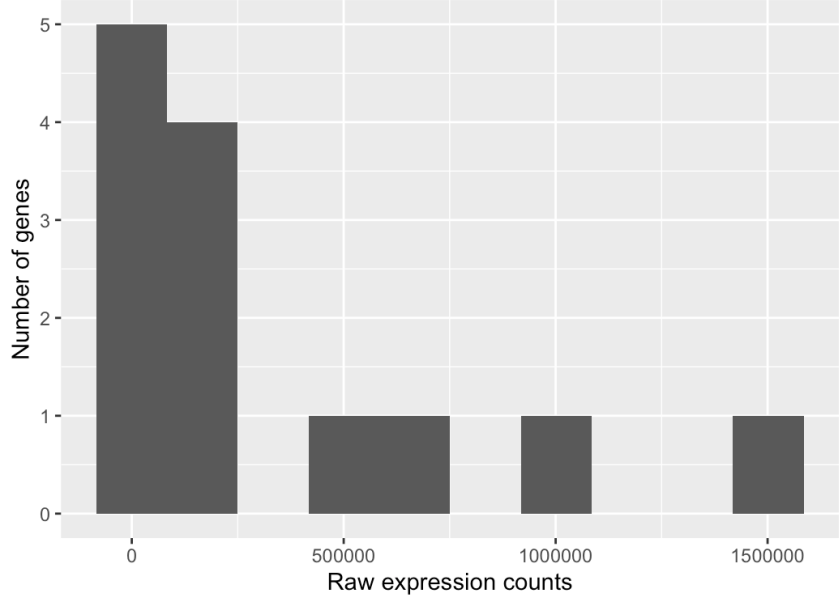
RNA-Seq: Differential Gene Expression

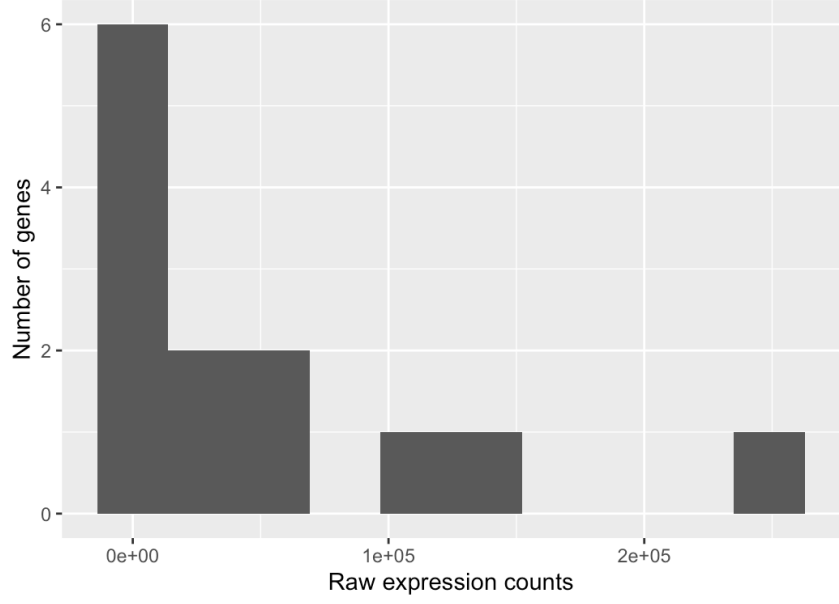
Exploring RNA-seq count data



RNA-Seq: Differential Gene Expression

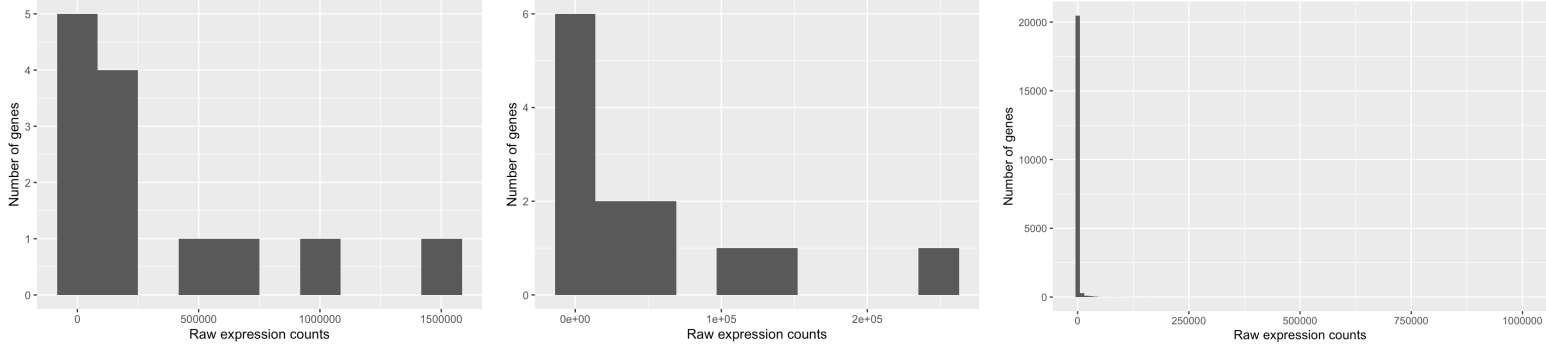
Exploring RNA-seq count data





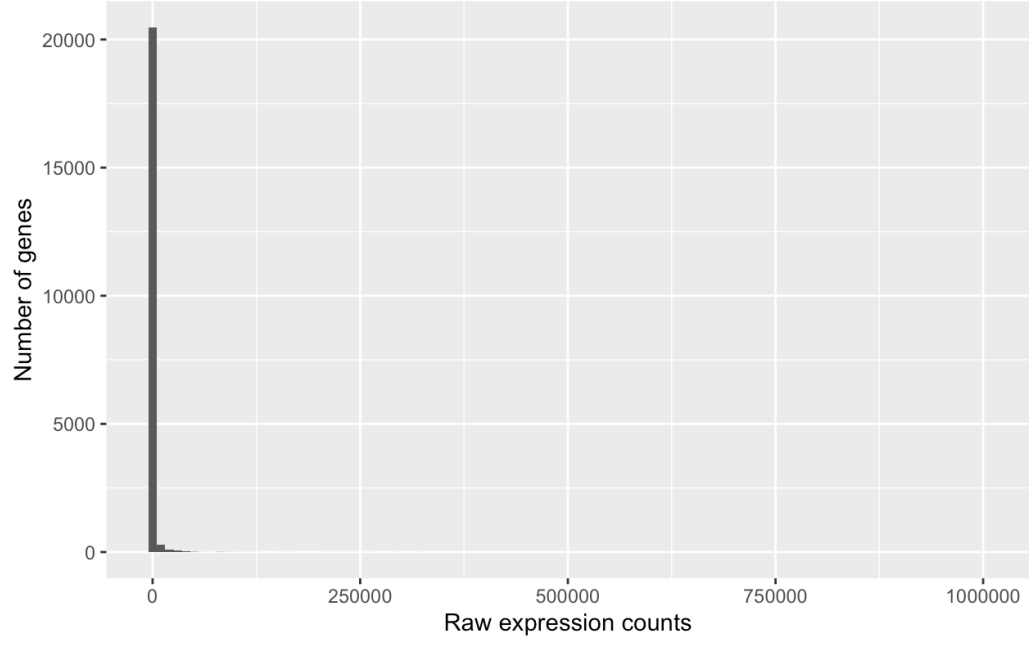
RNA-Seq: Differential Gene Expression

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 (λ, 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 ✗

Poisson Distribution ?

RNA-Seq: Differential Gene Expression

Modeling count data

`What model should we use for RNA-Seq data?`

Binomial Distribution ✗

Poisson Distribution ?

It depends on the mean and variance!

RNA-Seq: Differential Gene Expression

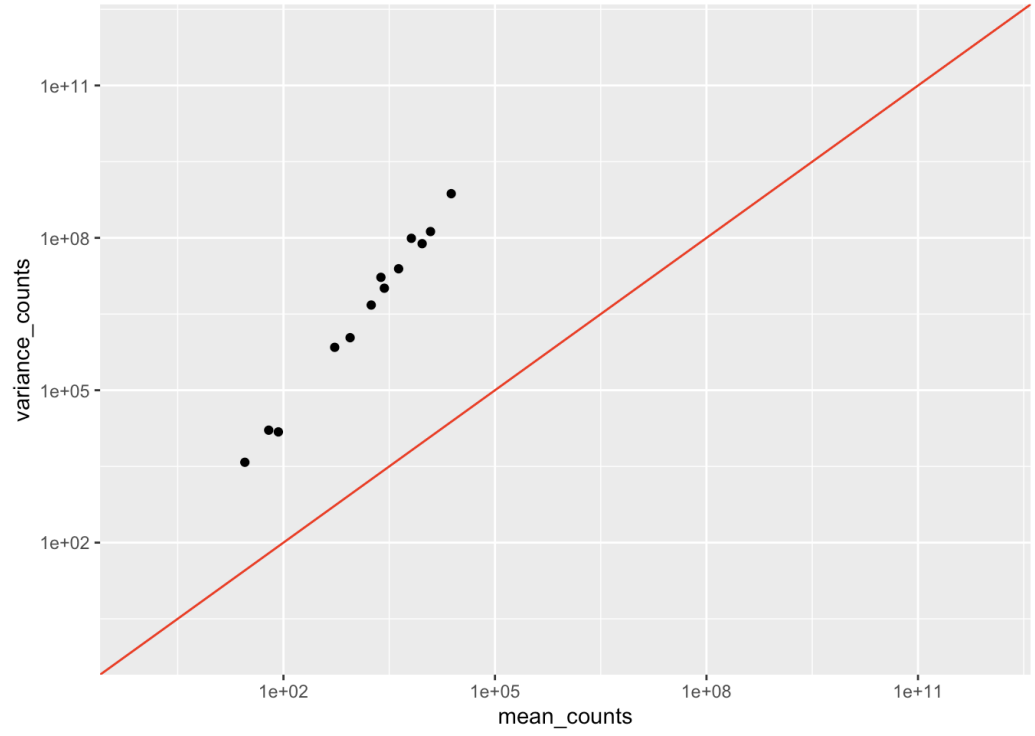
Modeling count data

```
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")
```

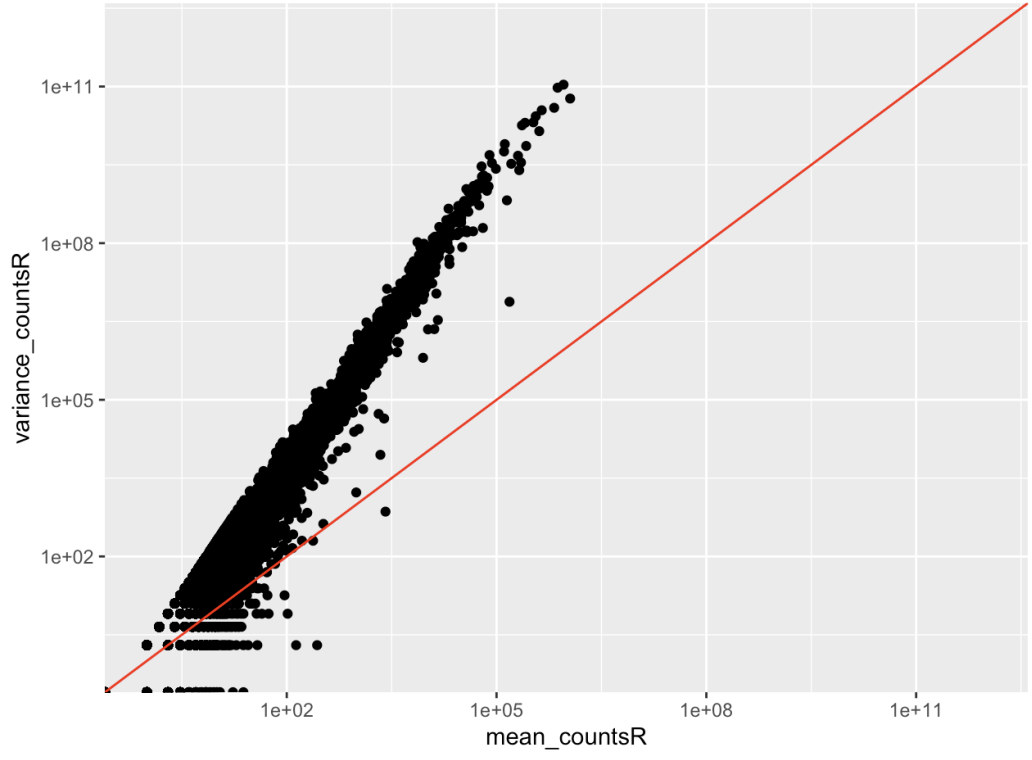
RNA-Seq: Differential Gene Expression

Modeling count data



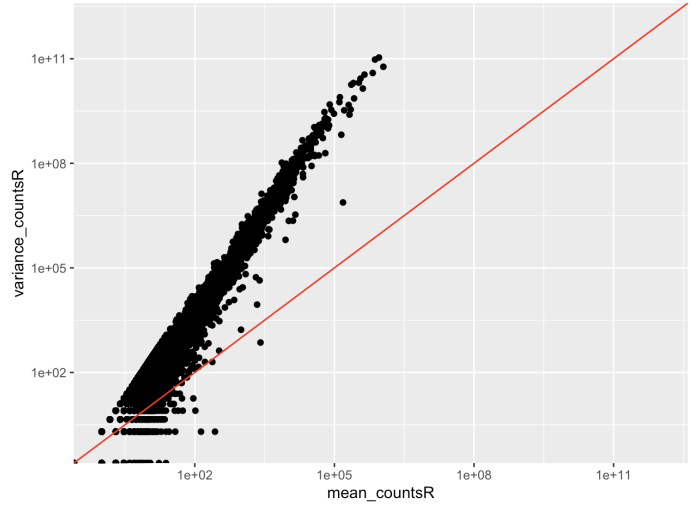
RNA-Seq: Differential Gene Expression

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

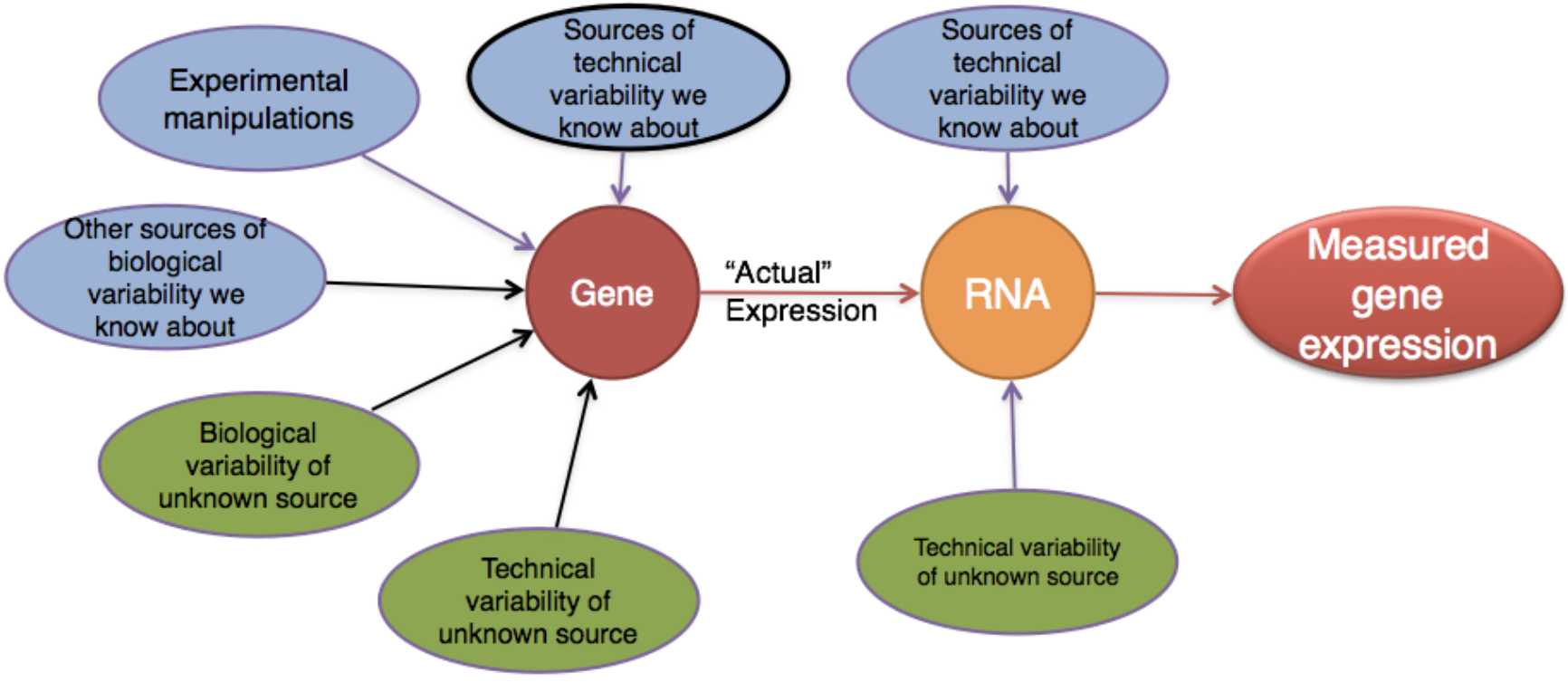
The negative binomial distribution is used when you're counting the number of trials required to achieve a fixed number of successes, where each trial has only two possible outcomes: success or failure.

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

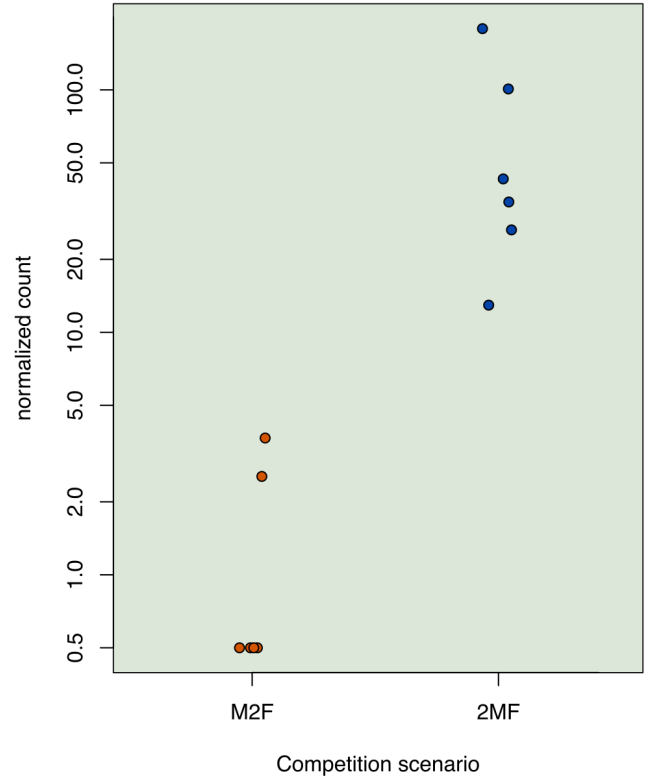
Replicates and variability



Courtesy of Paul Pavlidis, UBC

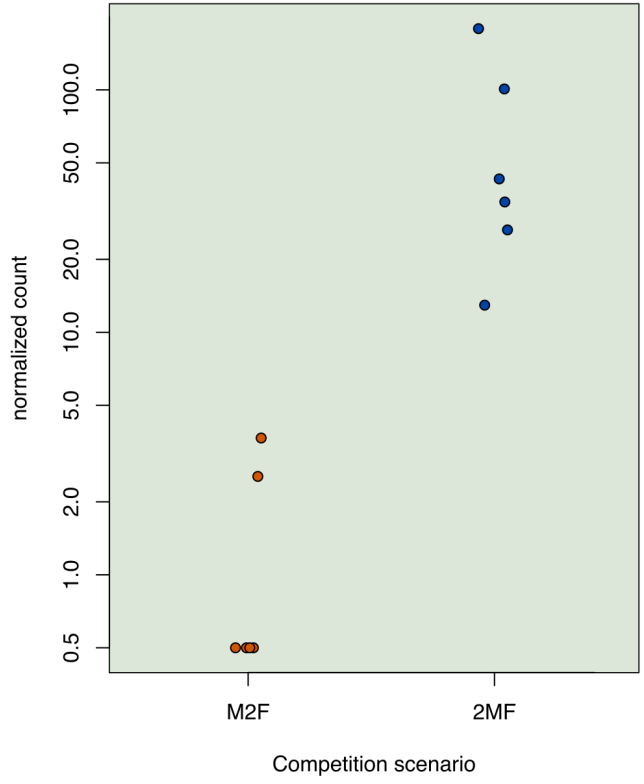
RNA-Seq: Differential Gene Expression

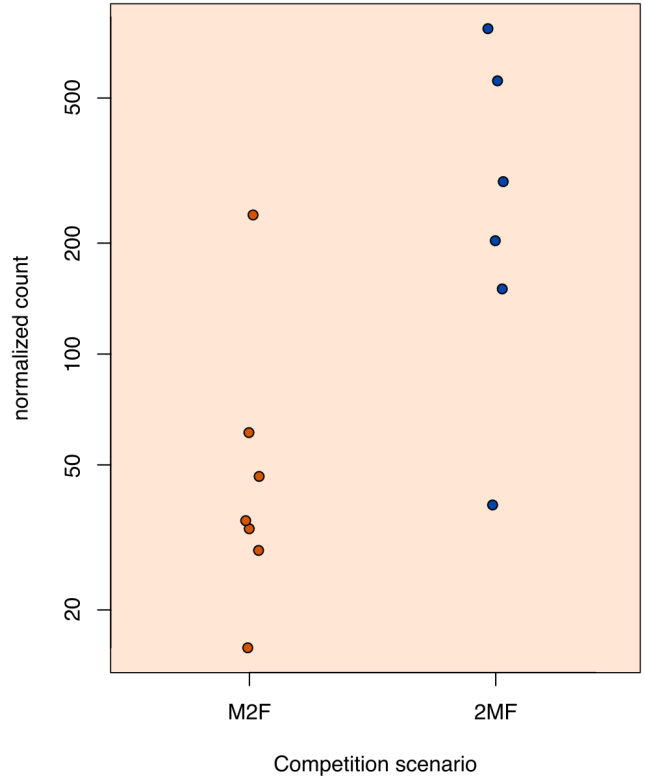
Replicates and variability



RNA-Seq: Differential Gene Expression

Replicates and variability

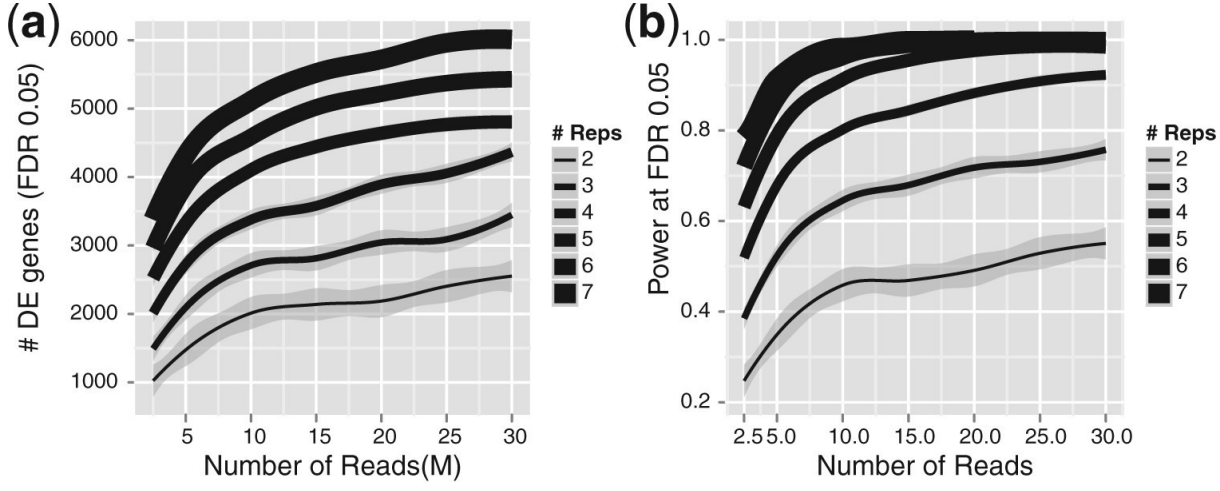




RNA-Seq: Differential Gene Expression

How many replicates?

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



Liu et al.(2014) Bioinformatics, 30:301–4