TRANSCRIPTOMICS

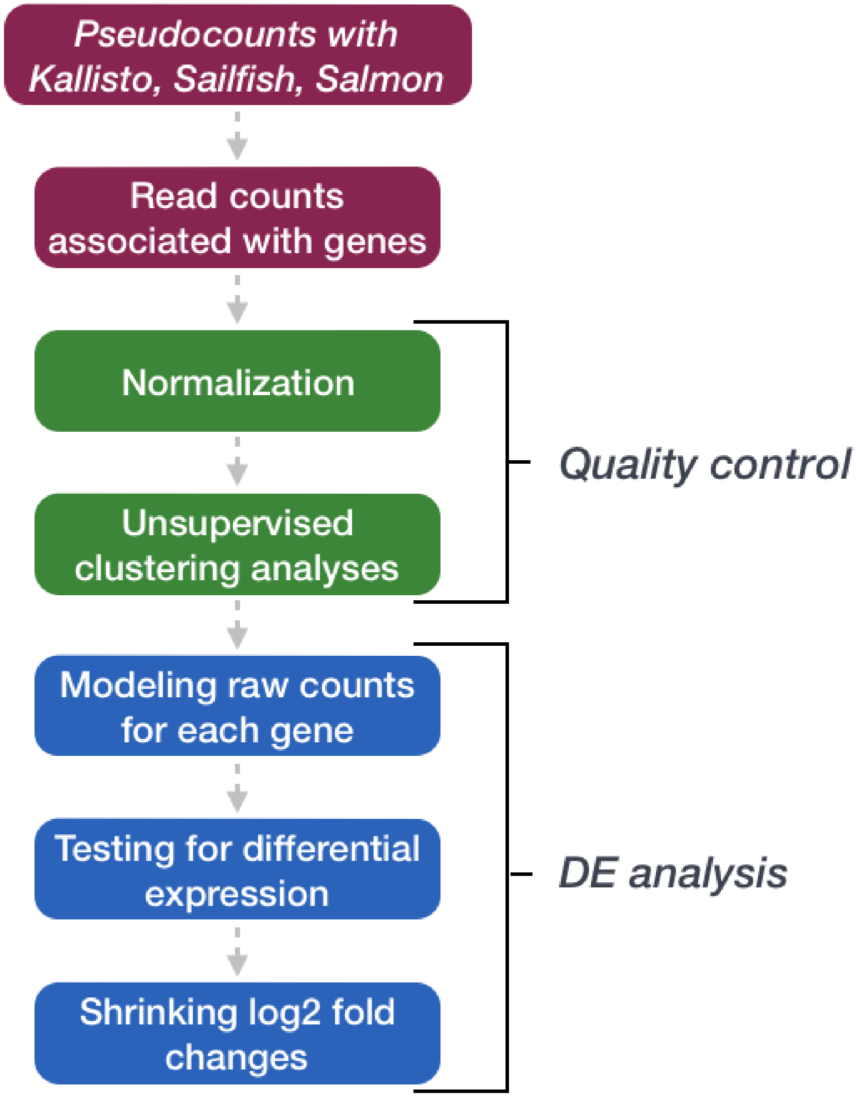
Differential Gene Expression

Day 06

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

RNA-Seq: Differential Gene Expression

Differential Expression with DESeq2


Pseudocounts with Kallisto, Sailfish, Salmon

Read counts associated with genes

Normalization

Unsupervised clustering analyses

Modeling raw counts for each gene

Testing for differential expression

Shrinking log2 fold changes

Quality control

DE analysis

RNA-Seq: Differential Gene Expression

Normalization

> Adjusting raw data to remove biases and technical artifacts, ensuring that observed differences in gene expression levels reflect biological differences rather than extraneous factors.

RNA-Seq: Differential Gene Expression

Normalization

> Adjusting raw data to remove biases and technical artifacts, ensuring that observed differences in gene expression levels reflect biological differences rather than extraneous factors.

RNA-Seq: Differential Gene Expression

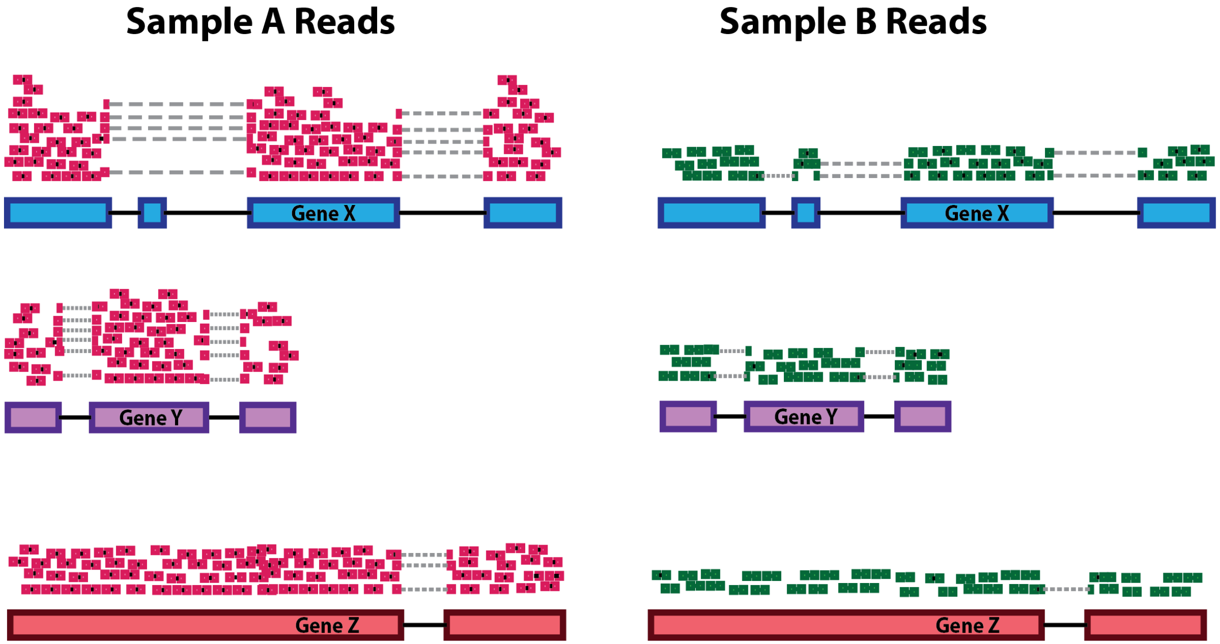
Normalization

1. **Sequencing Depth**: Different samples may have varying numbers of total reads.
2. **Gene Length**: Longer genes naturally produce more reads.
3. **Composition Effects**: A few highly expressed genes in a sample can skew the distribution of expression levels.
4. **Library Preparation or Technical Variability**: Variations in RNA extraction, sequencing efficiency, or sample handling.

RNA-Seq: Differential Gene Expression

Normalization

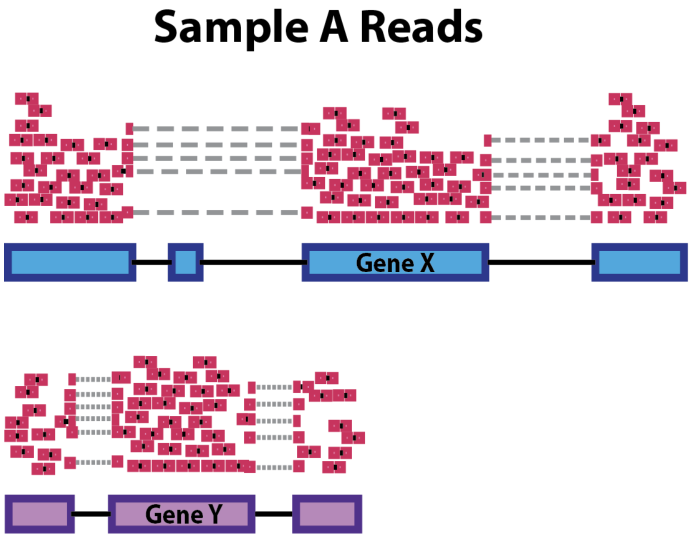
Sequencing Depth



RNA-Seq: Differential Gene Expression

Normalization

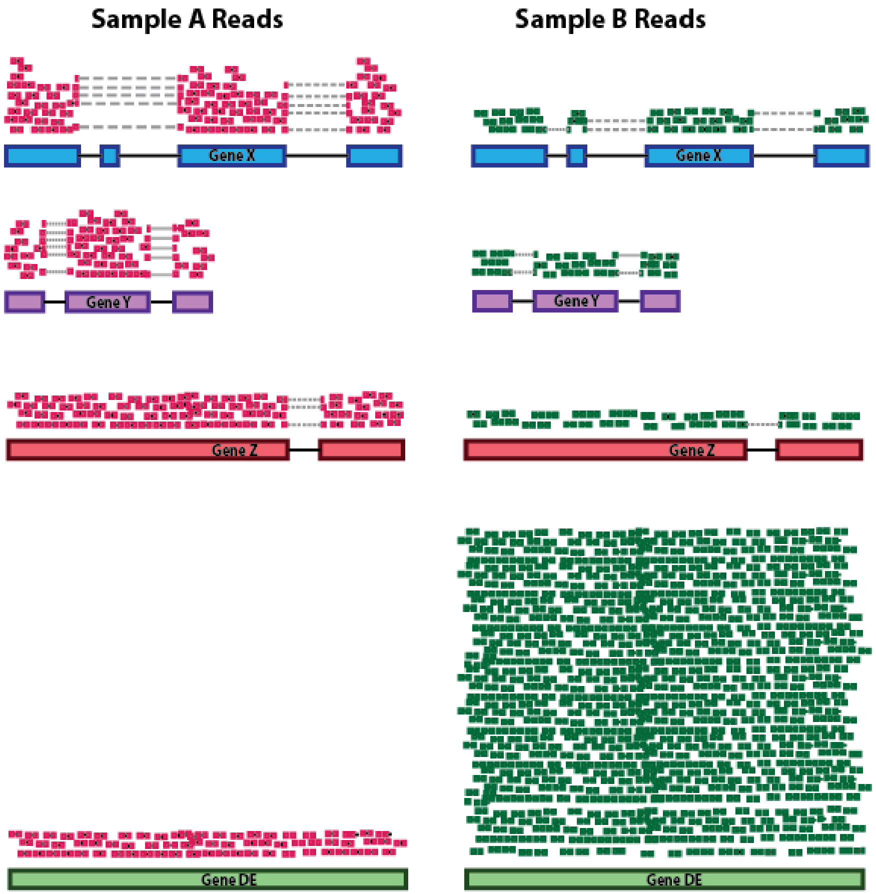
Gene Length



RNA-Seq: Differential Gene Expression

Normalization

Composition Effects



RNA-Seq: Differential Gene Expression

Normalization

Normalization methods

✏️ CPM (counts per million):counts scaled by total number of reads
✏️ RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped): counts per length of transcript (kb) per million reads mapped
✏️ TPM (transcripts per kilobase million): counts per length of transcript (kb) per million reads mapped
✏️ DESeq2's median of ratios: counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene
✏️ EdgeR's trimmed mean of M values (TMM): a weighted trimmed mean of the log expression ratios between samples

RNA-Seq: Differential Gene Expression

Normalization methods

Reads per Kilobase, per Million reads sequenced (RPKM)

$$\text{RPKM} = \frac{\text{\# reads mapped to genomic region}}{\text{(region length in kb)(total \# reads)}} \times 10^{6}$$

$$\text{RPKM} = \frac{1000}{(5)(20000000)} = 10$$

⚠️ RPKM alone, is not sufficient for normalization

RNA-Seq: Differential Gene Expression

Normalization methods

TPM (Transcripts Per Million)

Gene (length)	Gene A (100kb)	Gene B (50kb)	Gene C (25kb)	Gene D (5kb)	Gene E (1kb)	Total RPK
Sample 1	281690	70423	84507	211268	352113	1000000
Sample 2	487	973	973	24331	973236	1000000

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

⚠️ TPM is not sufficient DGE analysis

RNA-Seq: Differential Gene Expression

Normalization methods

DESeq2-normalized counts: Median of ratios method

Step 1: Pseudo-reference sample (row-wise geometric mean)

Gene	Gene A	Gene B	Gene C	Gene D	Gene E
Sample 1	80	10	6	3	1
Sample 2	20	20	10	50	400
Pseudo-reference	sqrt(80*20)	sqrt(10*20)	sqrt(6*10)	sqrt(3*50)	sqrt(1*400)

RNA-Seq: Differential Gene Expression

Normalization methods

DESeq2-normalized counts: Median of ratios method

Step 1: Pseudo-reference sample (row-wise geometric mean)

Gene	Gene A	Gene B	Gene C	Gene D	Gene E
Sample 1	80	10	6	3	1
Sample 2	20	20	10	50	400
Pseudo-reference	40	14.14	7.74	12.25	20

RNA-Seq: Differential Gene Expression

Normalization methods

DESeq2-normalized counts: Median of ratios method

Step 2: Ratio of each sample to the reference

Gene	Gene A	Gene B	Gene C	Gene D	Gene E
Sample 1	80	10	6	3	1
Sample 2	20	20	10	50	400
Pseudo-reference	40	14.14	7.74	12.25	20
Sample 1/Pseudo	80/40	10/14.14	6/7.64	3/12.25	1/20
Sample 2/Pseudo	20/40	20/14.14	10/7.64	50/12.25	400/20

RNA-Seq: Differential Gene Expression

Normalization methods

DESeq2-normalized counts: Median of ratios method

Step 2: Ratio of each sample to the reference

Gene	Gene A	Gene B	Gene C	Gene D	Gene E
Sample 1	80	10	6	3	1
Sample 2	20	20	10	50	400
Pseudo-reference	40	14.14	7.74	12.25	20
Sample 1/Pseudo	2	0,71	0.78	0.24	0.05
Sample 2/Pseudo	0.5	1.41	1.31	0.08	20

RNA-Seq: Differential Gene Expression

Normalization methods

DESeq2-normalized counts: Median of ratios method

Step 3: Normalization factor for each sample (size factor)

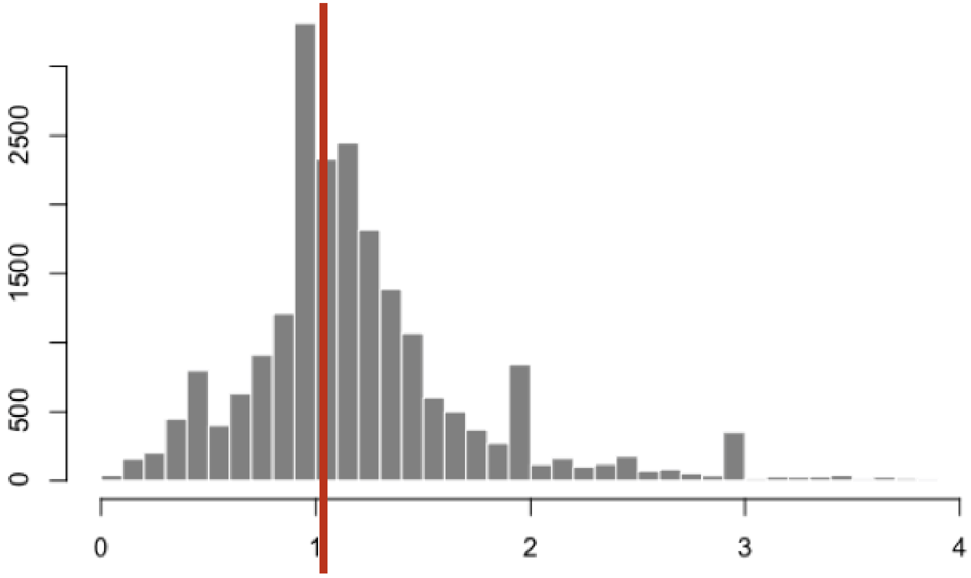
Gene	Gene A	Gene B	Gene C	Gene D	Gene E	Median
Sample 1	80	10	6	3	1	
Sample 2	20	20	10	50	400	
Pseudo-reference	40	14.14	7.74	12.25	20	
Sample 1/Pseudo	2	0.71	0.78	0.24	0.05	0.71
Sample 2/Pseudo	0.5	1.41	1.31	0.08	20	1.31

RNA-Seq: Differential Gene Expression

Normalization methods

DESeq2-normalized counts: Median of ratios method

Step 3: Normalization factor for each sample (size factor)



RNA-Seq: Differential Gene Expression

Normalization methods

DESeq2-normalized counts: Median of ratios method

Step 3: Normalization of raw counts for each gene

Gene	Gene A	Gene B	Gene C	Gene D	Gene E	Median
Sample 1	80	10	6	3	1	
Sample 2	20	20	10	50	400	
Pseudo-reference	40	14.14	7.74	12.25	20	
Sample 1/Pseudo	2	0.71	0.78	0.24	0.05	0.71
Sample 2/Pseudo	0.5	1.41	1.31	0.08	20	1.31
Sample 1 (normal.)	80/0.71	10/0.71	6/0.71	3/0.71	1/ 0.71	
Sample 2 (normal.)	20/1.31	20/1.31	10/1.31	50/1.31	400/1.31	

RNA-Seq: Differential Gene Expression

Normalization methods

DESeq2-normalized counts: Median of ratios method

Step 3: Normalization of raw counts for each gene

Gene	Gene A	Gene B	Gene C	Gene D	Gene E	Median
Sample 1	80	10	6	3	1	
Sample 2	20	20	10	50	400	
Pseudo-reference	40	14.14	7.74	12.25	20	
Sample 1/Pseudo	2	0.71	0.78	0.24	0.05	0.71
Sample 2/Pseudo	0.5	1.41	1.31	0.08	20	1.31
Sample 1 (normal.)	112.68	14.08	8.45	0.86	1.41	
Sample 2 (normal.)	15.27	15.27	7.63	38.17	305.34	

RNA-Seq: Differential Gene Expression

Normalization with DESeq2

1. Check if metadata (meta) and counts (txi) match

```r
all(colnames(txi$counts) %in% rownames(meta))
all(colnames(txi$counts) == rownames(meta))
```

RNA-Seq: Differential Gene Expression

Normalization with DESeq2

1. Check if metadata (meta) and counts (txi) match

```
all(colnames(txi$counts) %in% rownames(meta))
all(colnames(txi$counts) == rownames(meta))
```

2. Create DESEq2 object

```
dds <- DESeqDataSetFromTximport(txi, colData = meta,
                                design = ~ sampletype)
```

RNA-Seq: Differential Gene Expression

Normalization with DESeq2

1. Check if metadata (meta) and counts (txi) match

```
all(colnames(txi$counts) %in% rownames(meta))
all(colnames(txi$counts) == rownames(meta))
```

2. Create DESEq2 object

```
dds <- DESeqDataSetFromTximport(txi, colData = meta,
                                design = ~ sampletype)
```

3. View data

```
View(counts(dds))
```

RNA-Seq: Differential Gene Expression

Normalization with DESeq2

4. Normalize counts

```
dds <- estimateSizeFactors(dds)
```

5. View normalization factors of each sample

```
normalizationFactors(dds)
```

RNA-Seq: Differential Gene Expression

Normalization with DESeq2

4. Normalize counts

```
dds <- estimateSizeFactors(dds)
```

5. View size factors of each sample

```
sizeFactors(ddsR)
  CalbF_1    CalbF_2    CalbF_4    CalbL_1    CalbL_2    CalbL_4
 1.2449508  1.6655969  2.2575536  0.6108855  0.6246485  0.4964350
   CbezF_1    CbezL_1    ChomF_1    ChomF_2    ChomF_3    ChomL_1
 3.3440530 10.1938320  3.4333486  2.3959019  2.9675721  0.8064141
   ChomL_2    ChomL_3    CmacF_1    CmacF_2    CmacF_3    CmacL_1
 0.8004318  0.6583727  1.3194402  1.9110091  3.4938492  0.7244566
   CmacL_2    CmacL_3    CmegF_1    CmegF_2    CmegF_4    CmegL_1
 0.6106440  1.0383915  1.6716498  1.8563761  4.9311681  0.6787449
   CmegL_2    CmegL_4    LexiF_1    LexiF_2    LexiF_3    LexiL_1
 0.7670481  5.2475787  0.1386870  0.2401653  0.3307302  0.2744835
   LexiL_2    LexiL_3
 0.2246053  0.1109496
```

RNA-Seq: Differential Gene Expression

Quality control

Sample-level QC

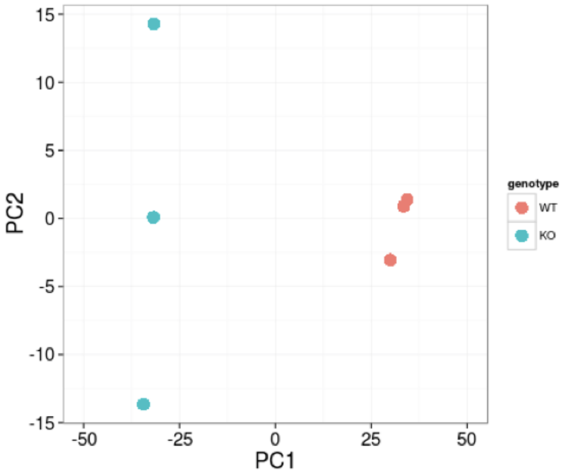
A useful initial step in an RNA-seq analysis is often to assess overall similarity between samples:

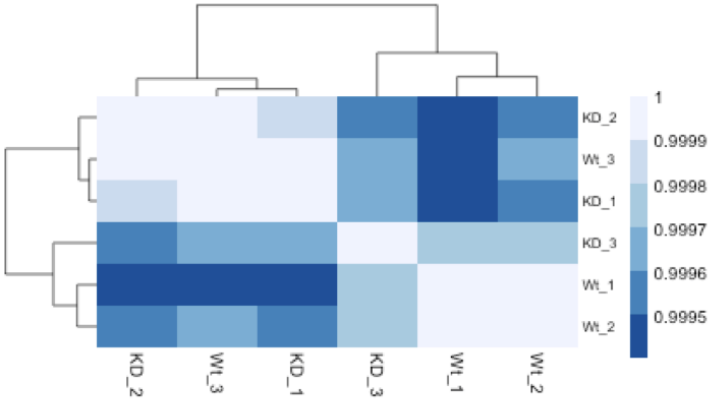
- Which samples are similar to each other, which are different?
- Does this fit to the expectation from the experiment's design?
- What are the major sources of variation in the dataset?

RNA-Seq: Differential Gene Expression

Quality control

Sample-level QC





Differential gene expression workshop using Salmon counts

RNA-Seq: Differential Gene Expression

Quality control

Principal Component Analysis (PCA): Example

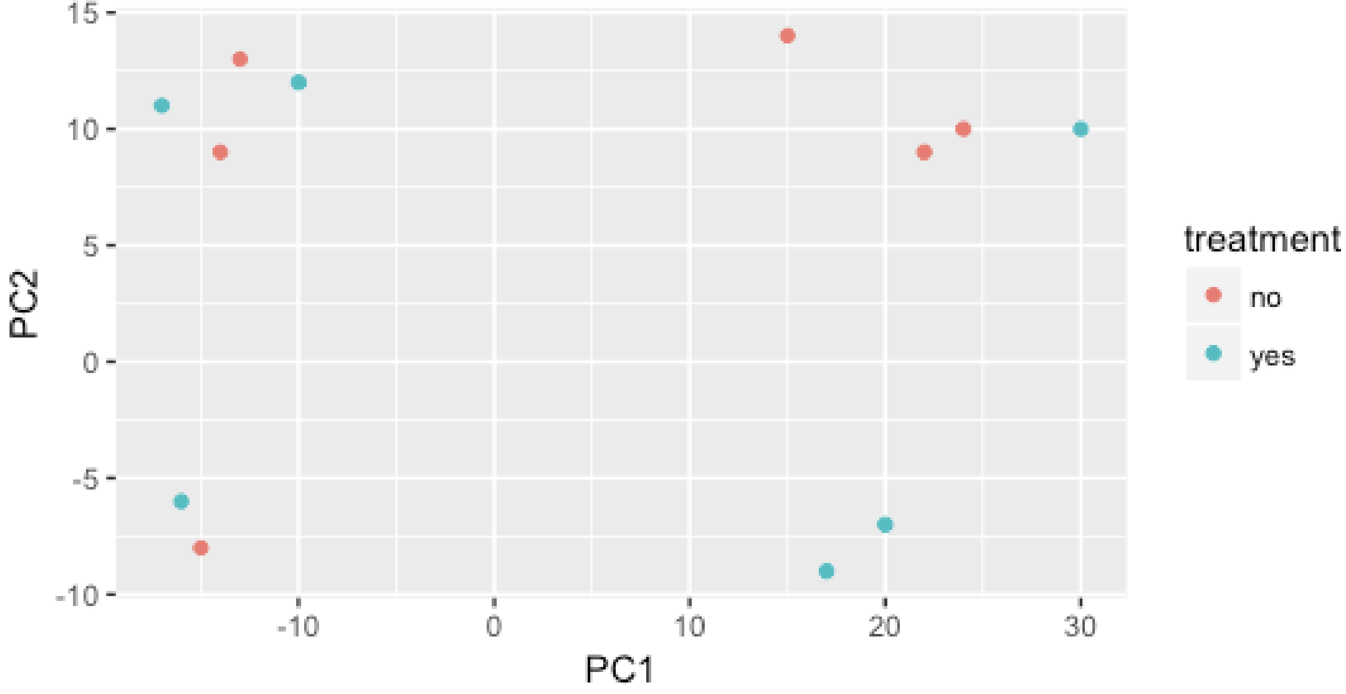
Mouse experiment with two strains, three cages, and one treatment.

RNA-Seq: Differential Gene Expression

Quality control

Principal Component Analysis (PCA): Treatment

Mouse experiment with two strains, three cages, and one treatment.



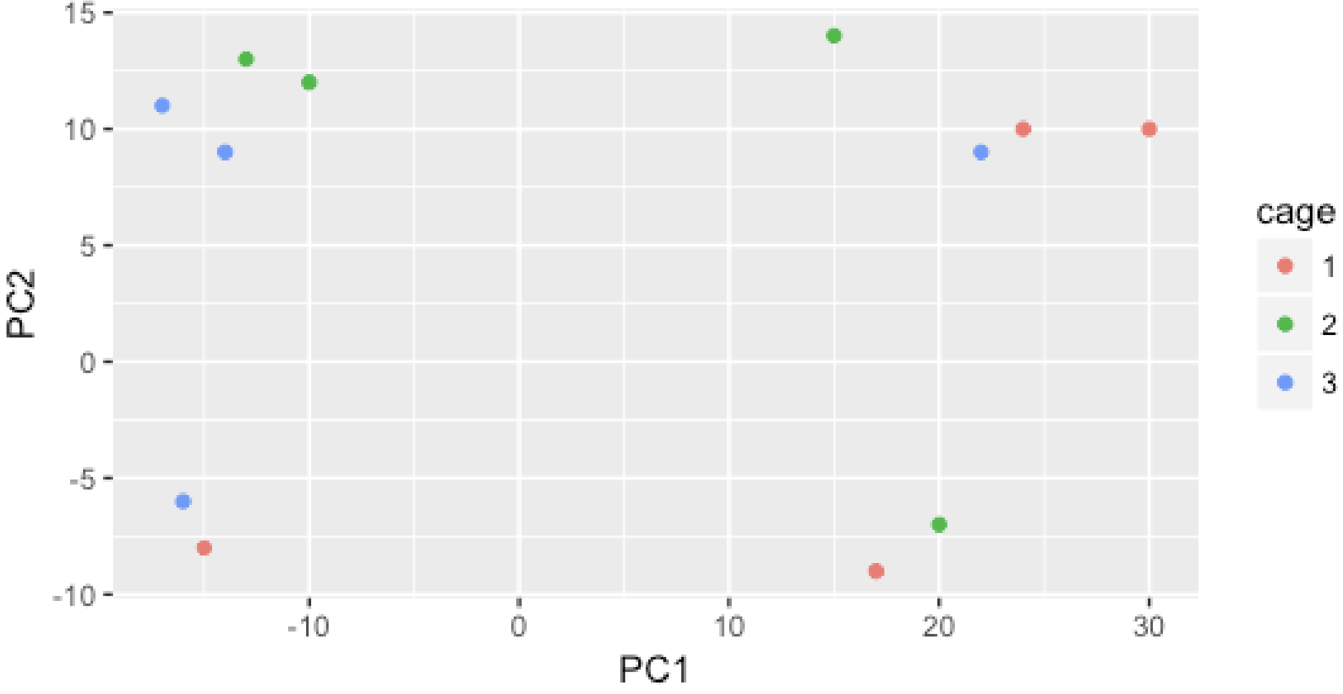
Differential gene expression workshop using Salmon counts

RNA-Seq: Differential Gene Expression

Quality control

Principal Component Analysis (PCA): Cage

Mouse experiment with two strains, three cages, and one treatment.



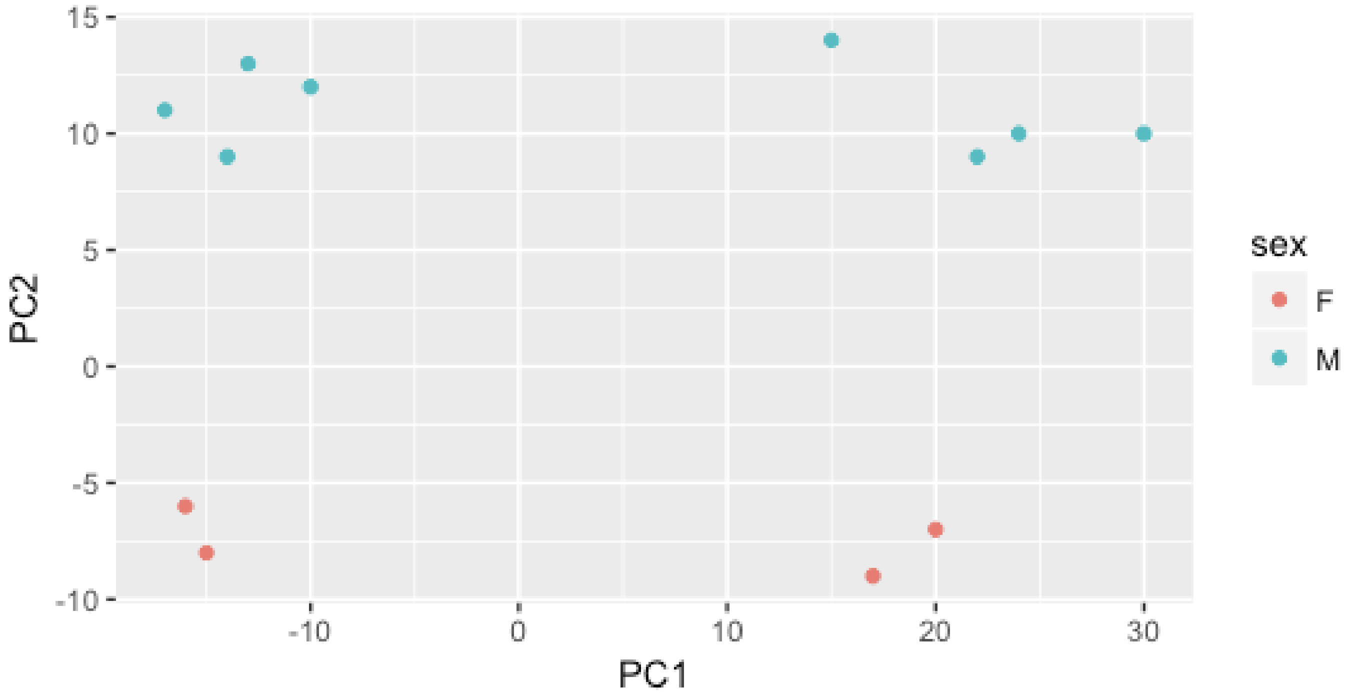
Differential gene expression workshop using Salmon counts

RNA-Seq: Differential Gene Expression

Quality control

Principal Component Analysis (PCA): Sex

Mouse experiment with two strains, three cages, and one treatment.



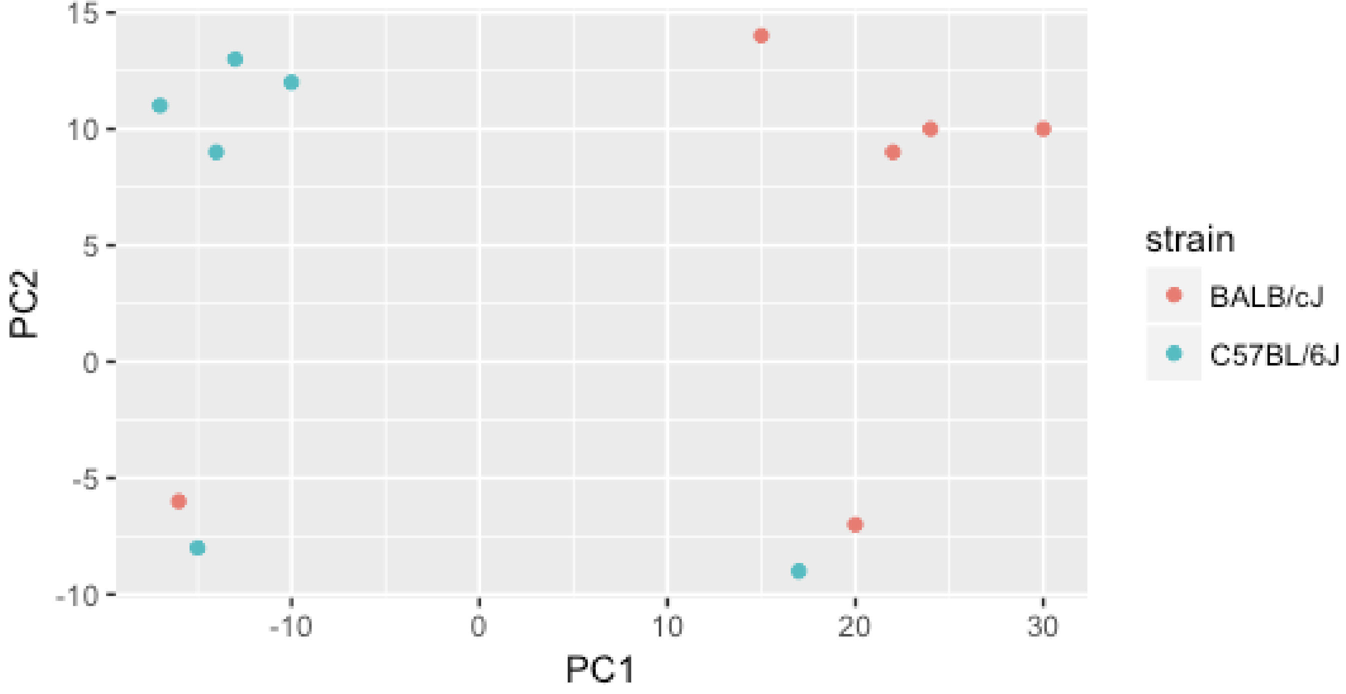
Differential gene expression workshop using Salmon counts

RNA-Seq: Differential Gene Expression

Quality control

Principal Component Analysis (PCA): Strain

Mouse experiment with two strains, three cages, and one treatment.

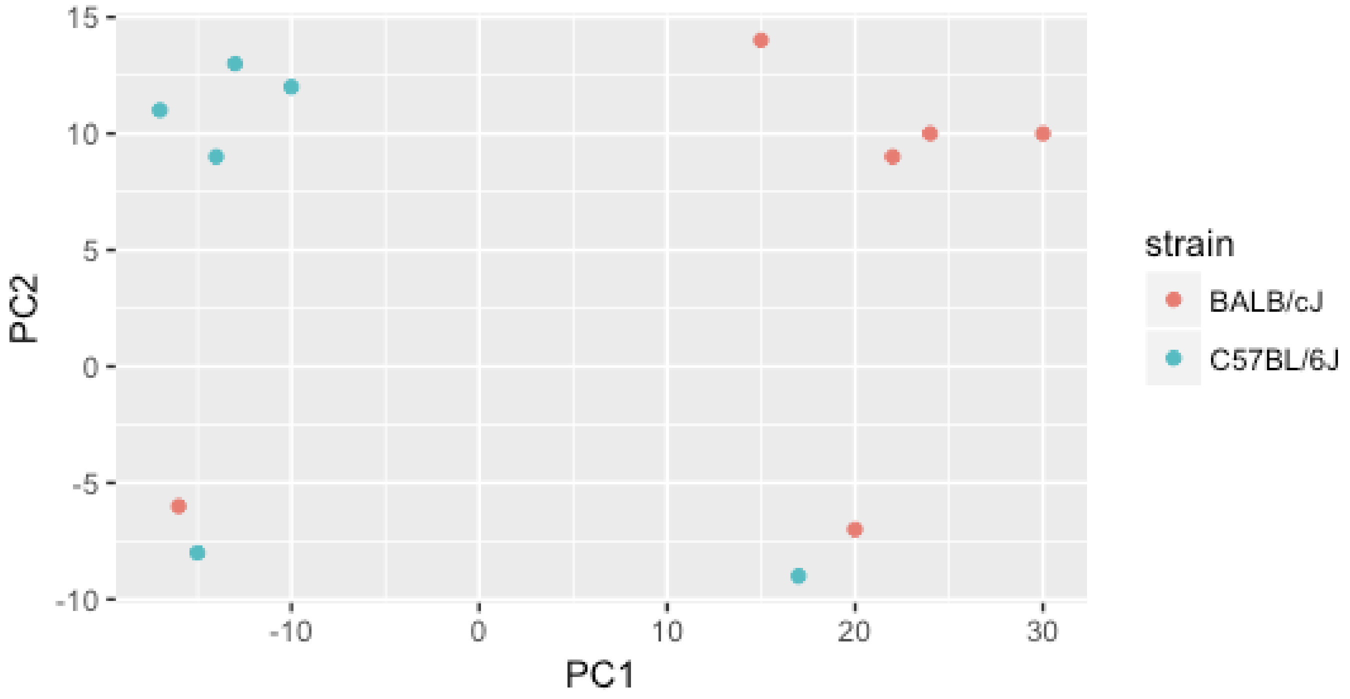


RNA-Seq: Differential Gene Expression

Quality control

Principal Component Analysis (PCA): Strain

⚠️ Two samples do not cluster with the correct strain



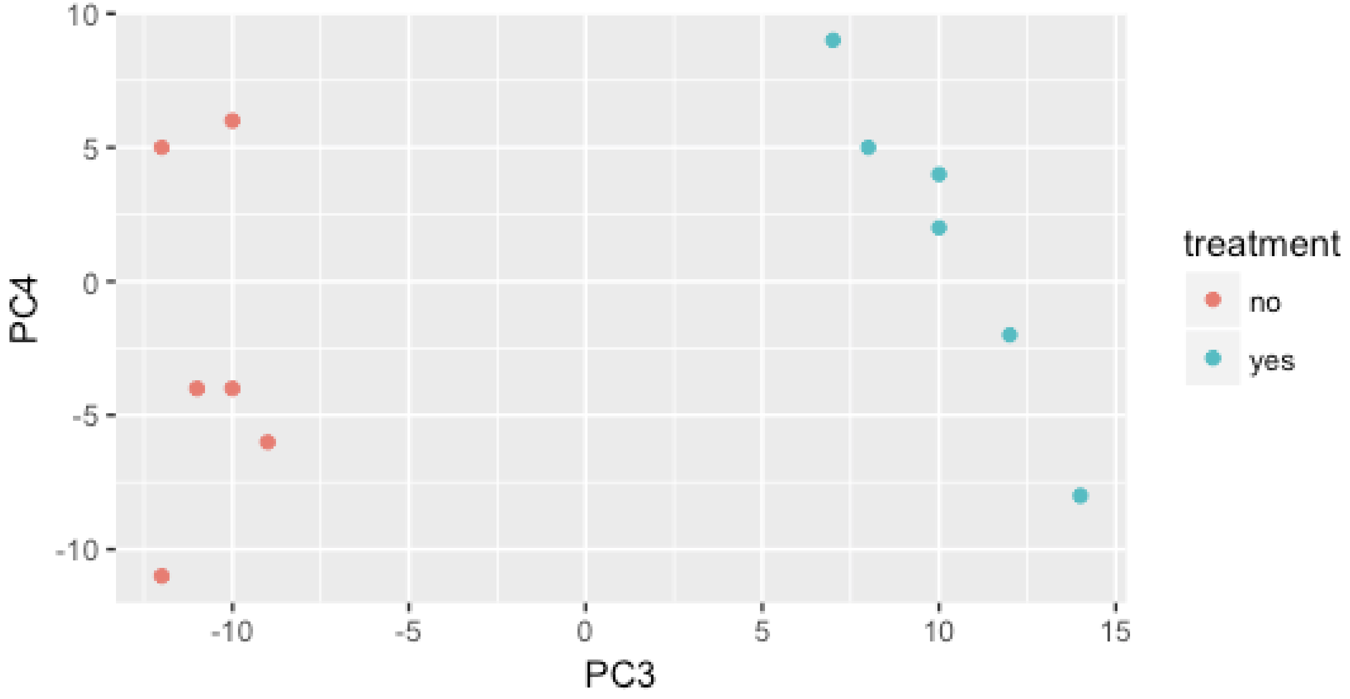
Differential gene expression workshop using Salmon counts

RNA-Seq: Differential Gene Expression

Quality control

Principal Component Analysis (PCA): Treatment

Finding if treatment is a major source of variation.



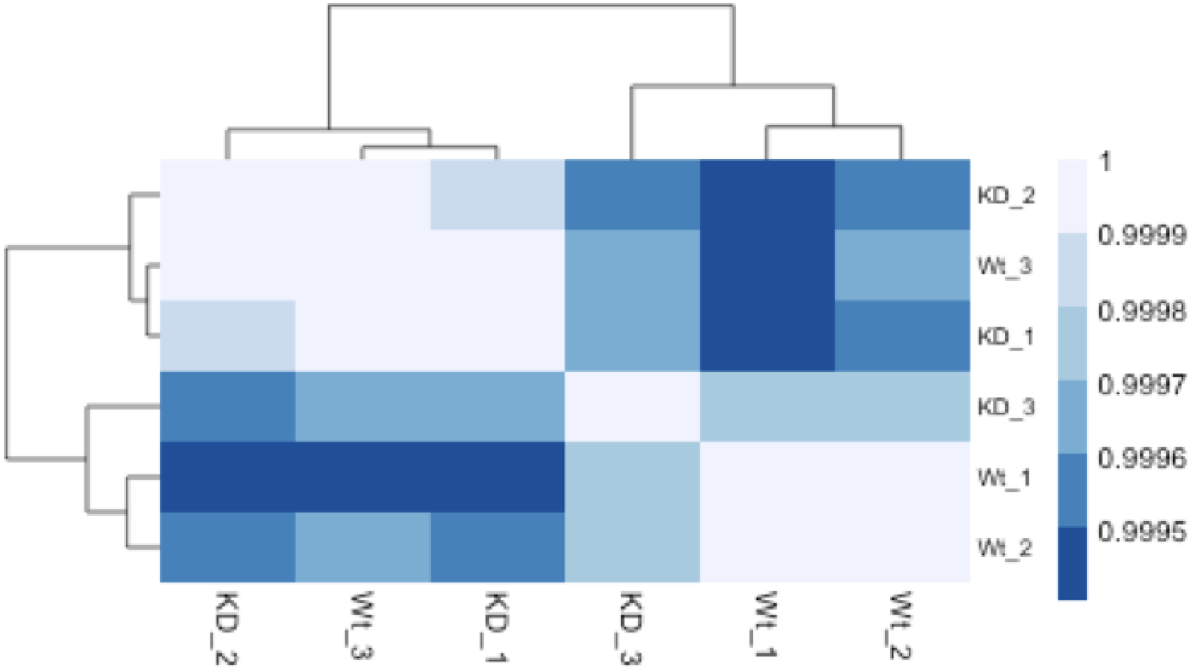
Differential gene expression workshop using Salmon counts

RNA-Seq: Differential Gene Expression

Quality control

Hierarchical Clustering Heatmap

Finding if treatment is a major source of variation.



Differential gene expression workshop using Salmon counts

RNA-Seq: Differential Gene Expression

Exploratory Analysis: PCA

1. Transform counts for data visualization

```
rld <- rlog(dds, blind=TRUE)
```

RNA-Seq: Differential Gene Expression

Exploratory Analysis: PCA

1. Transform counts for data visualization

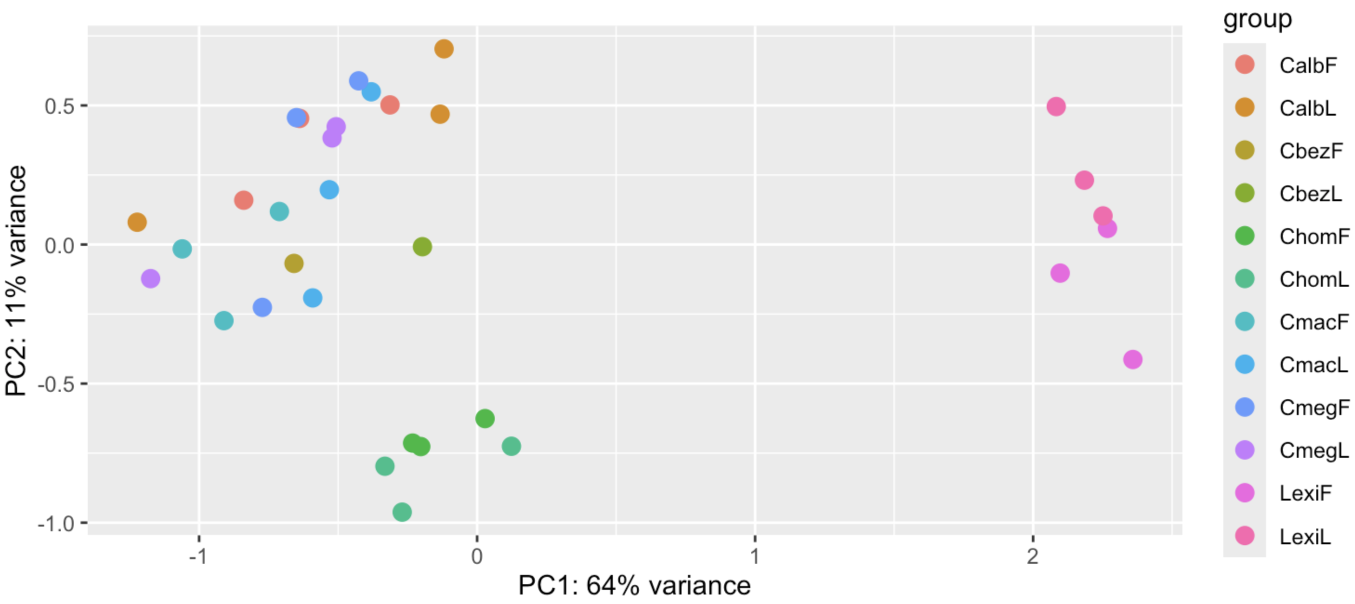
```r
rld <- rlog(dds, blind=TRUE)
```

2. Plot PCA

```r
plotPCA(rld, intgroup="sampletype")
```

RNA-Seq: Differential Gene Expression

Exploratory Analysis: PCA



RNA-Seq: Differential Gene Expression

Exploratory Analysis: PCA

1. Transform counts for data visualization

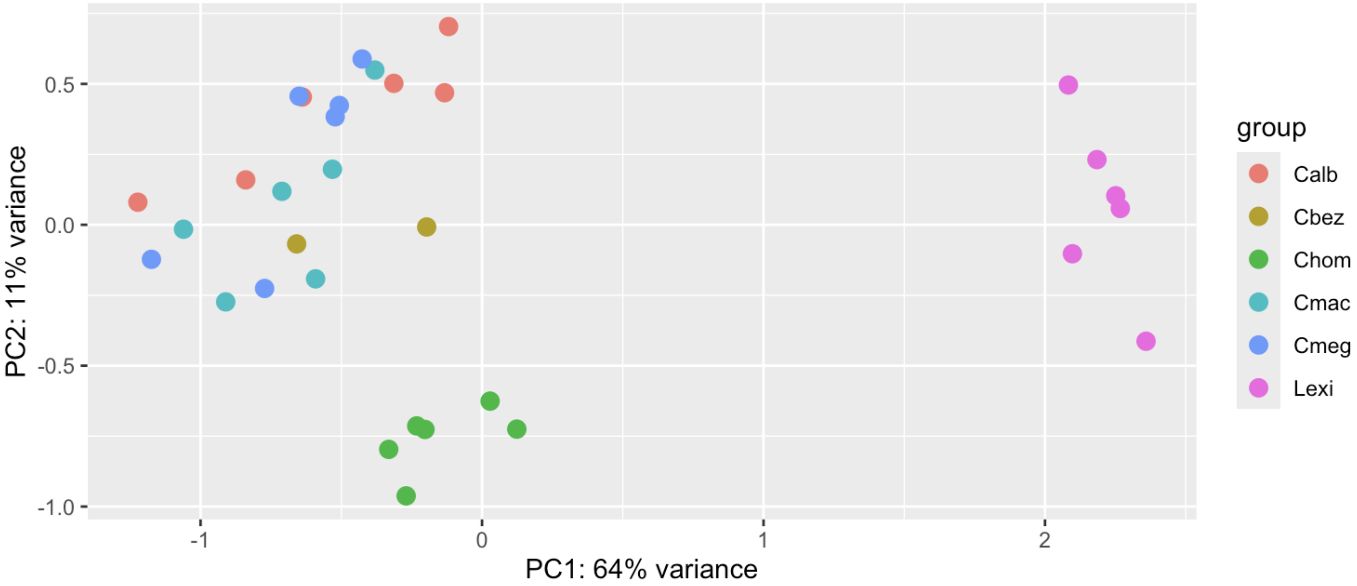
```
rld <- rlog(dds, blind=TRUE)
```

2. Plot PCA

```
plotPCA(rld, intgroup="sampletype")
plotPCA(rld, intgroup="species")
```

RNA-Seq: Differential Gene Expression

Exploratory Analysis: PCA



RNA-Seq: Differential Gene Expression

Exploratory Analysis: PCA

1. Transform counts for data visualization

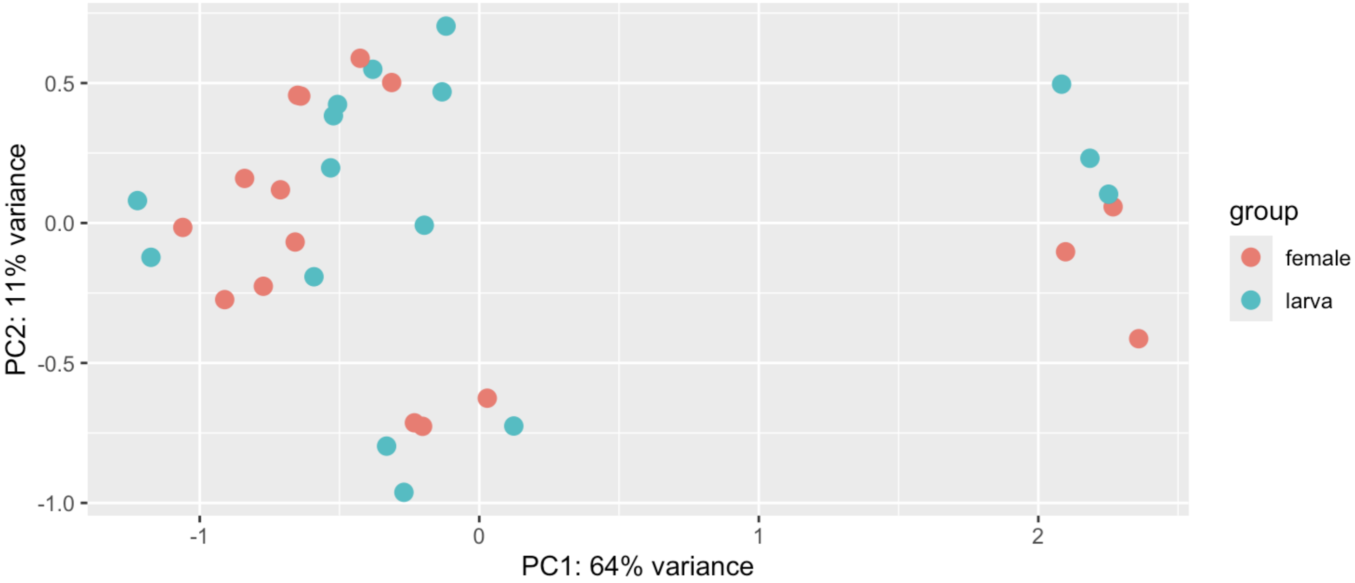
```
rld <- rlog(dds, blind=TRUE)
```

2. Plot PCA

```
plotPCA(rld, intgroup="sampletype")
plotPCA(rld, intgroup="species")
plotPCA(rld, intgroup="stage")
```

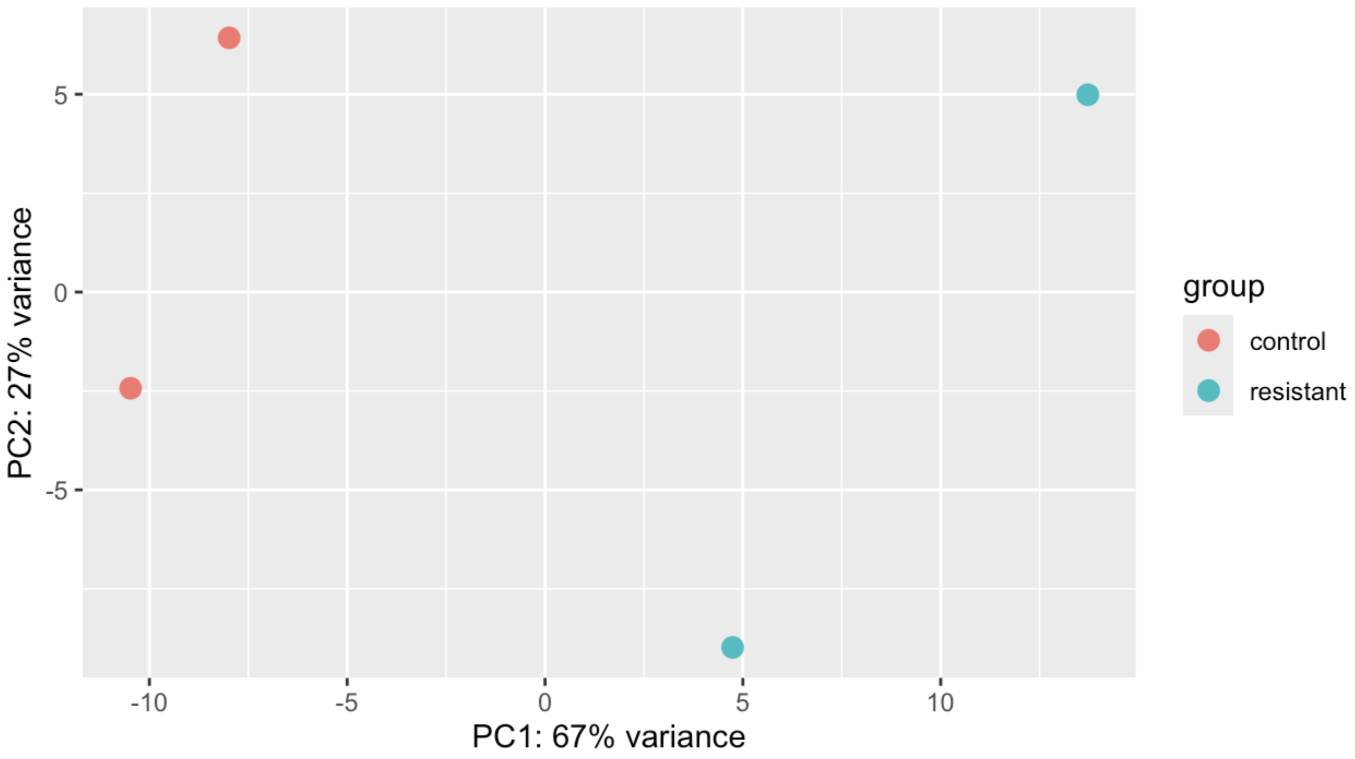
RNA-Seq: Differential Gene Expression

Exploratory Analysis: PCA



RNA-Seq: Differential Gene Expression

Exploratory Analysis: PCA



RNA-Seq: Differential Gene Expression

Exploratory Analysis: Hierarchical Clustering

1. Extract the log matrix from the object

```
rld_mat <- assay(rld)
```

2. Compute pairwise correlation values

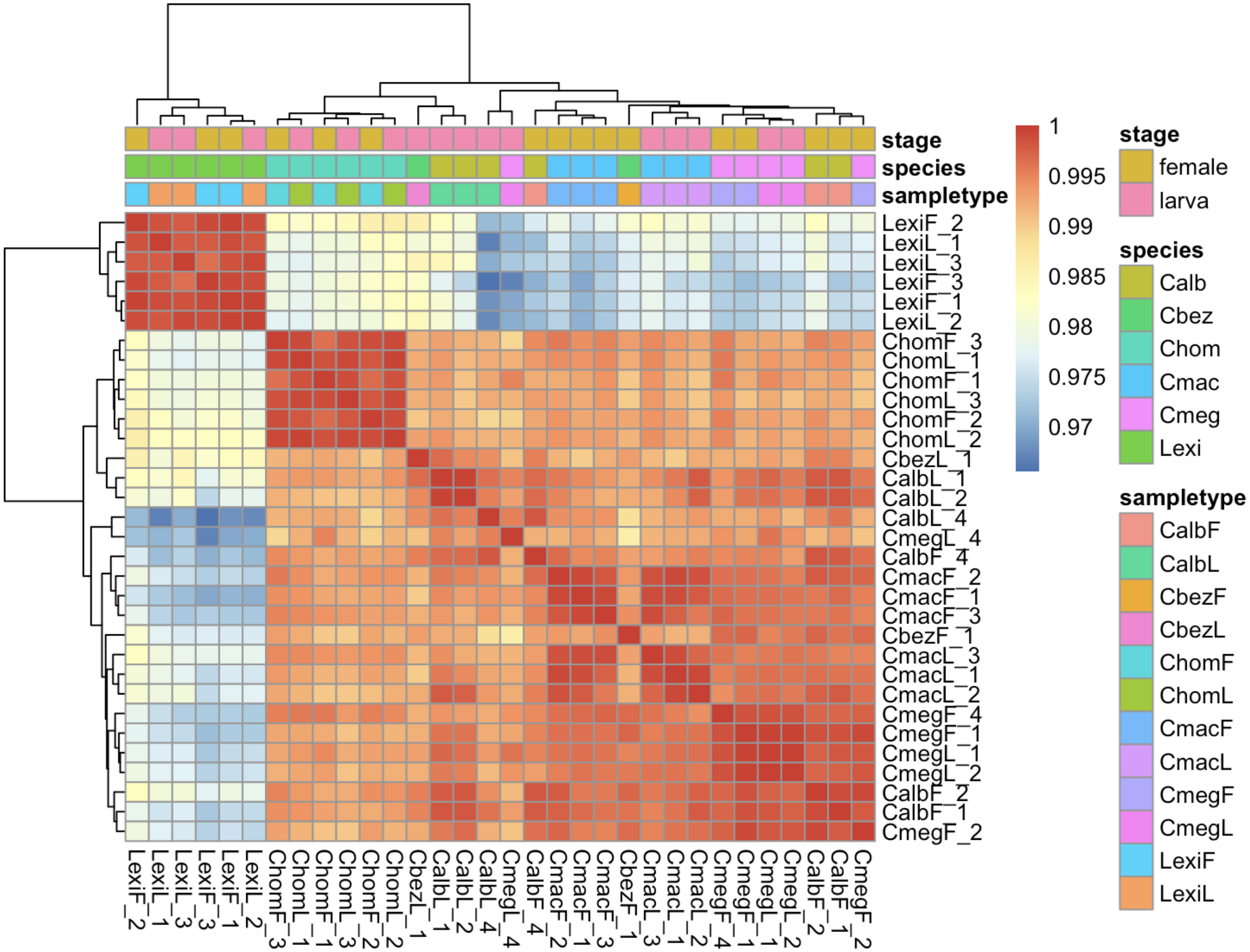
```
rld_cor <- cor(rld_mat)
rld_cor
```

3. Plot the heatmap

```
pheatmap(rld_cor, annotation = meta)
```

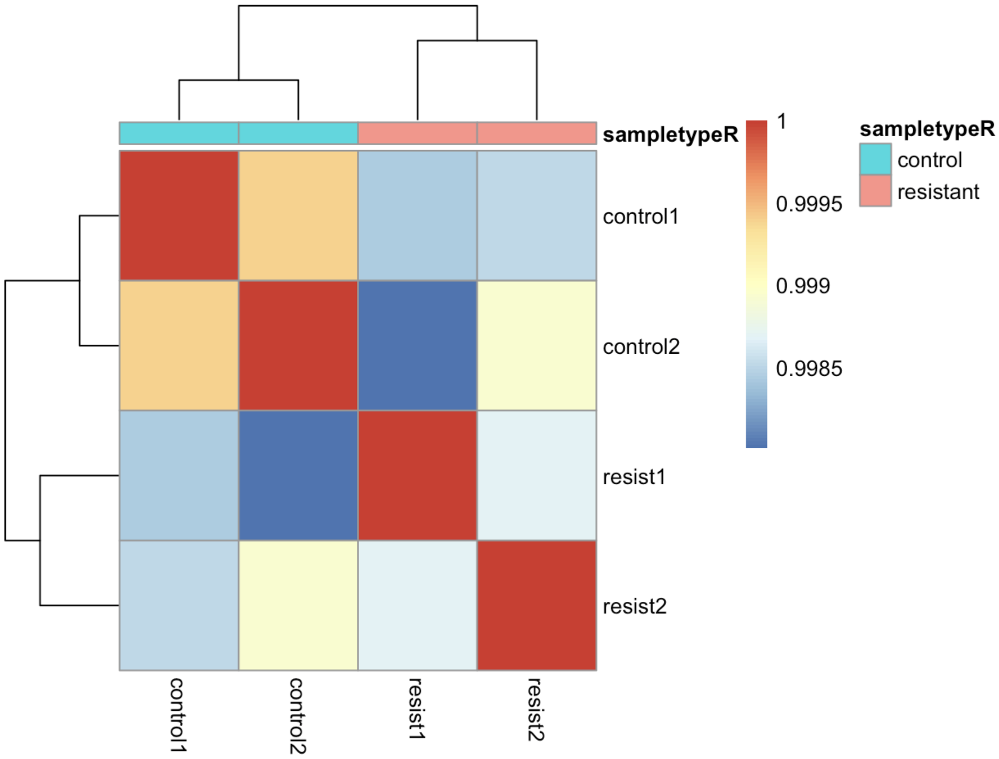
RNA-Seq: Differential Gene Expression

Exploratory Analysis: Hierarchical Clustering



RNA-Seq: Differential Gene Expression

Exploratory Analysis: Hierarchical Clustering



RNA-Seq: Differential Gene Expression

Exploratory Analysis: Hierarchical Clustering

1. Extract the log matrix from the object

```
rld_mat <- assay(rld)
```

2. Compute pairwise correlation values

```
rld_cor <- cor(rld_mat)
rld_cor
```

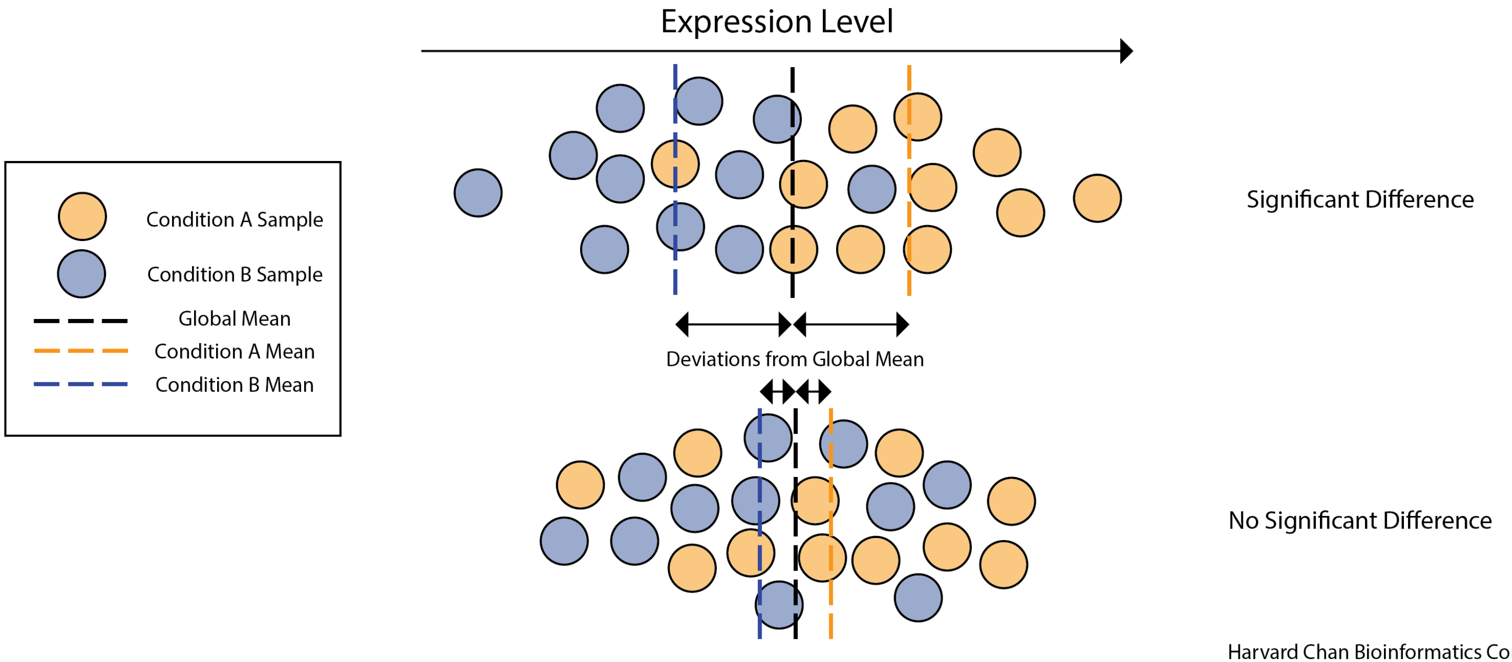
3. Plot the heatmap

```
pheatmap(rld_cor, annotation = meta)
```

💡 `pheatmap` has several options to change the aesthetics of the plot. Explore them with `?pheatmap`.

RNA-Seq: Differential Gene Expression

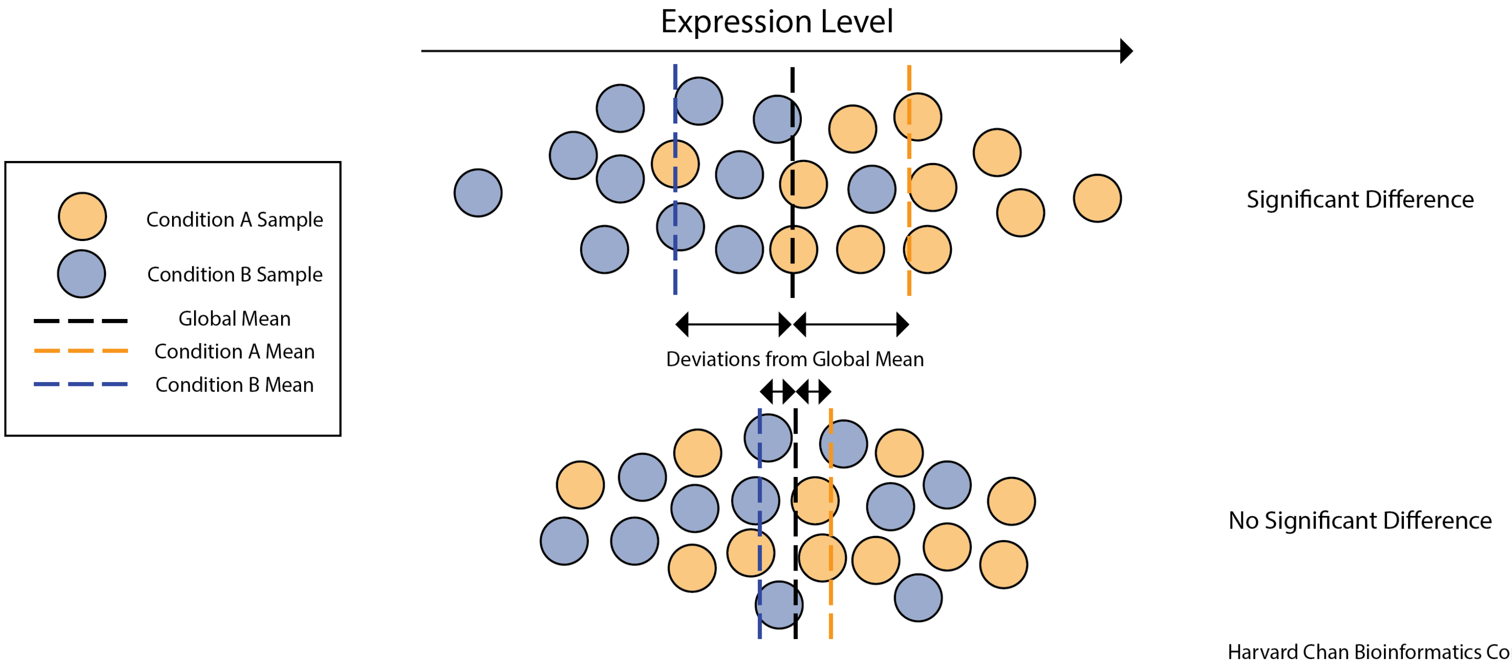
Differential expression analysis with DESeq2



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

RNA-Seq: Differential Gene Expression

Differential expression analysis with DESeq2

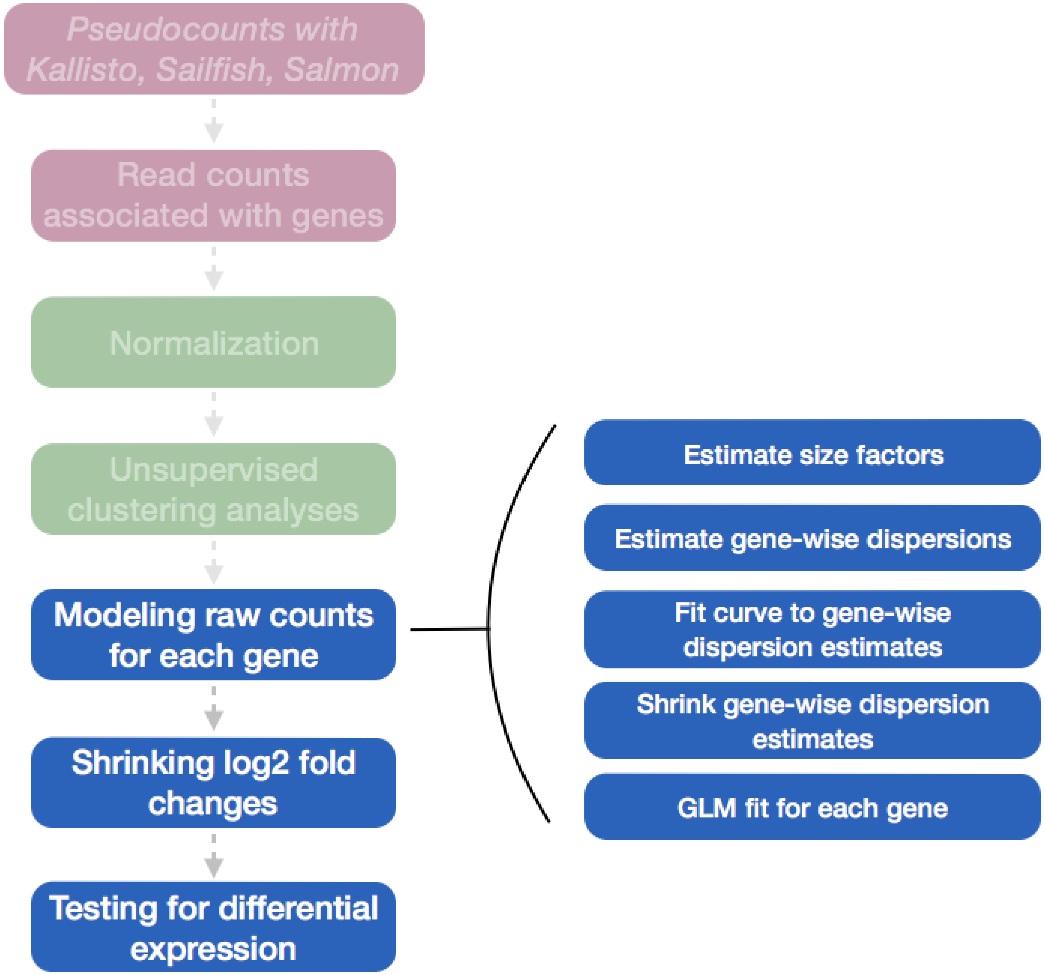


Harvard Chan Bioinformatics Core
Adapted from Image by Paul Pavlidis, UBC

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

RNA-Seq: Differential Gene Expression

Differential expression analysis with DESeq2


Pseudocounts with Kallisto, Sailfish, Salmon
Read counts associated with genes
Normalization
Unsupervised clustering analyses
Modeling raw counts for each gene
Shrinking log2 fold changes
Testing for differential expression
Estimate size factors
Estimate gene-wise dispersions
Fit curve to gene-wise dispersion estimates
Shrink gene-wise dispersion estimates
GLM fit for each gene

RNA-Seq: Differential Gene Expression

Running DESeq2

The Design Formula in RNA-Seq Experiments

- The design formula determines how variation in gene expression is modeled.
- It accounts for biological and technical variables.
- Ensures that comparisons between groups are statistically valid.
- Provides flexibility for complex experimental designs.

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Key concepts

1. **Factors**: variables describing the experimental setup (e.g., treatment, batch).
2. **Levels**: categories within factors (e.g., "control" and "treated").
3. **Interactions**: combined effects of multiple factors (e.g., treatment x time).

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Metadata Table

sampleID	treatment	sex	age	strain
Mouse_1	Control	Male	Young	Strain_A
Mouse_2	Treated	Male	Young	Strain_A
Mouse_3	Control	Female	Young	Strain_A
Mouse_4	Treated	Female	Young	Strain_A
Mouse_5	Control	Male	Old	Strain_B
Mouse_6	Treated	Male	Old	Strain_B
Mouse_7	Control	Female	Old	Strain_B
Mouse_8	Treated	Female	Old	Strain_B

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Metadata Table

> If you want to examine the expression differences between treatments, and you know that major sources of variation include sex and strain, then your design formula would be:

```
design = ~ sex + strain + treatment
```

⚠️ the factors included in the design formula need to match the column names in the metadata.

💡 you can use more complex designs

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

The `+` operator (Main Effects Only)

The `+` operator adds factors to the model, but it does not include interactions between the factors. It only evaluates the main effects, meaning how each factor independently affects the response variable.

Example: `design = ~ strain + treatment`

This means you are testing:

➡️ The main effect of Strain.

➡️ The main effect of Treatment.

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

The `*` operator (Main Effects + Interactions)

The `*` operator includes both the main effects and the interactions between the factors.

Example: `design = ~ strain * treatment`

This means you are testing:

- ➡️ The main effect of Strain.
- ➡️ The main effect of Treatment.
- ➡️ The interaction between Strain and Treatment
  same as `Strain:Treatment`

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Operator `:` (Interaction Only)

The `:` operator models only the interaction between factors, without including their main effects.

Example: `design = ~ strain:treatment`

This means you are testing:

➡️ only the interaction between Strain and Treatment, without considering the independent effects of each.

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Design Formula	Explanation
`~ treatment`	Tests for gene expression differences due to treatment, ignoring other variables.
`~ sex`	Models gene expression differences due to sex, ignoring other variables.
`~ strain + treatment`	Models the independent effects of strain, and treatment.
`~ strain + age + treatment`	Models the independent effects of strain, age, and treatment.
`~ strain * treatment`	Tests for strain-specific treatment effects (interaction between strain and treatment).

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Example: `~ Strain * Age * Treatment`

➡️ **Expands to**:

`~ Strain + Age + Treatment + Strain:Age + Strain:Treatment + Age:Treatment + Strain:Age:Treatment`

➡️ **Tests**:

- Main effects of strain, age, and treatment.
- Interactions:
  - Does strain affect treatment response?
  - Does age modify strain or treatment effects?
  - Is there a combined strain, age, and treatment effect?

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Key Considerations

1. **Main Effects vs. Interactions**:
   - Use interaction terms to study how one factor modifies another.
   - Avoid overly complex models if sample size is small.
2. **Statistical Power**:
   - Ensure sufficient replicates for each group to test interaction terms effectively.

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Our metadata

sampletype	species	stage	replicate
CalbF_1_mtDNA	Calb	female	1
CalbF_4_mtDNA	Calb	female	4
CalbL_1_mtDNA	Calb	larva	1
CalbL_4_mtDNA	Calb	larva	4
CbezF_1_mtDNA	Cbez	female	1
CbezL_1_mtDNA	Cbez	larva	1
ChomF_1_mtDNA	Chom	female	1
ChomF_3_mtDNA	Chom	female	3
ChomL_1_mtDNA	Chom	larva	1
…	…	…	.

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Our metadata

➡️ **Key Factors:**

- species: Different species (e.g., Calb, Cbez, Chom, Cmac).
- stage: Developmental stages (larva, female).

Potential Scientific Questions:

➡️ What is the effect of species on gene expression at each stage (larva and female)?
➡️ What is the effect of stage (larva vs female) on gene expression within each species?
➡️ Are there any interactions between species and stage?

RNA-Seq: Differential Gene Expression

Design Formula in RNA-Seq Experiments

Design Formula:

- To investigate both main effects (species, stage) and their interaction: `design = ~ species * stage`

Explanation:

- species: The main effect of different species.
- stage: The main effect of developmental stage (female vs larva).
- species * stage: The interaction between species and stage, testing if the effect of stage (female vs larva) differs between species.

RNA-Seq: Differential Gene Expression

Running DESeq2

1. Design Formula and Create DESeq2Dataset object

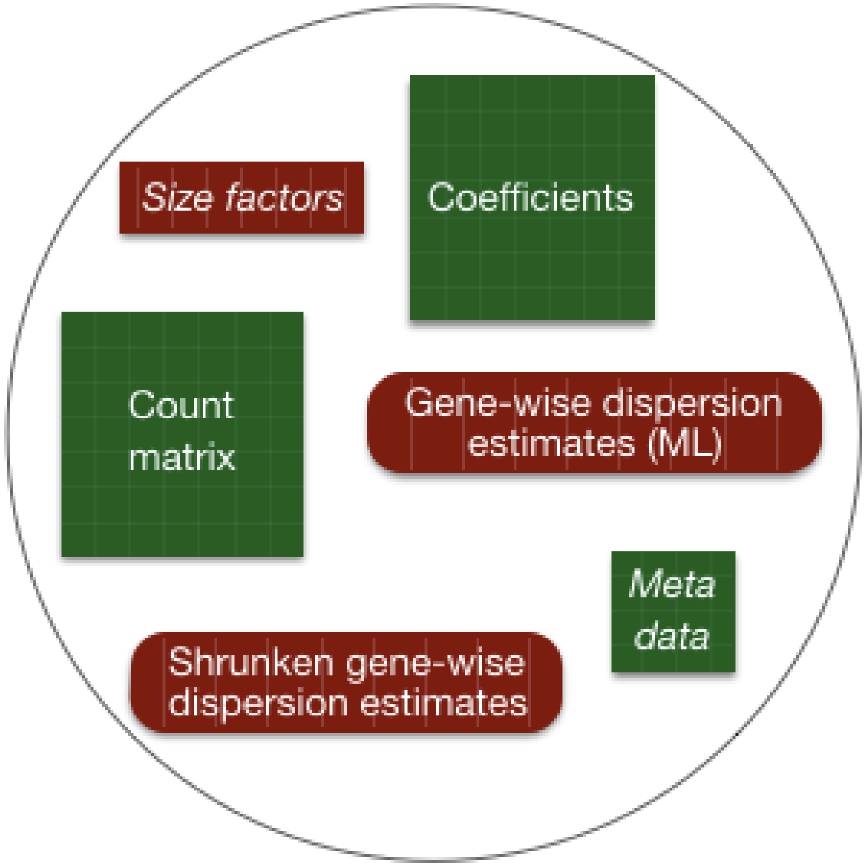
```
## Create DESeq2Dataset object
dds <- DESeqDataSetFromTximport(txi, colData = meta,
                              design = ~ species * stage)
```

2. Run DESeq analysis

```
dds <- DESeq(dds)
```

RNA-Seq: Differential Gene Expression

Running DESeq2


Size factors
Coefficients
Count matrix
Gene-wise dispersion estimates (ML)
Meta data
Shrunken gene-wise dispersion estimates

RNA-Seq: Differential Gene Expression

Running DESeq2

1. Design Formula and Create DESeq2Dataset object

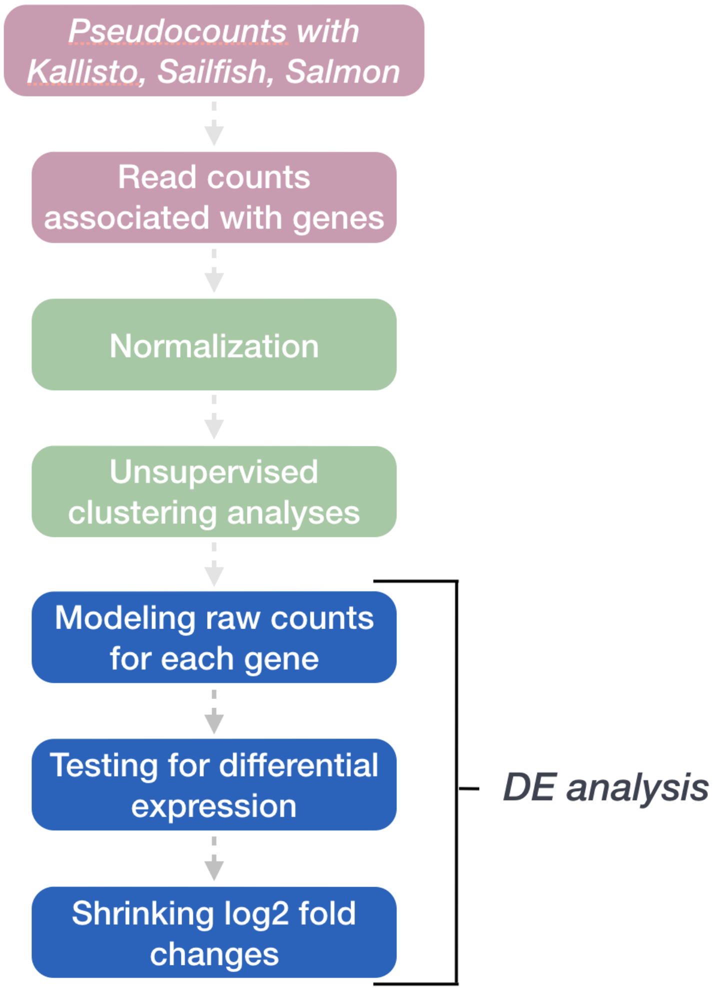
```
## Create DESeq2Dataset object
dds <- DESeqDataSetFromTximport(txi, colData = meta,
                                design = ~ sampletype)
```

2. Run DESeq analysis

```
dds <- DESeq(dds)
estimating size factors
using 'avgTxLength' from assays(dds), correcting for libra
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
```

RNA-Seq: Differential Gene Expression

Model fitting and Hypothesis testing

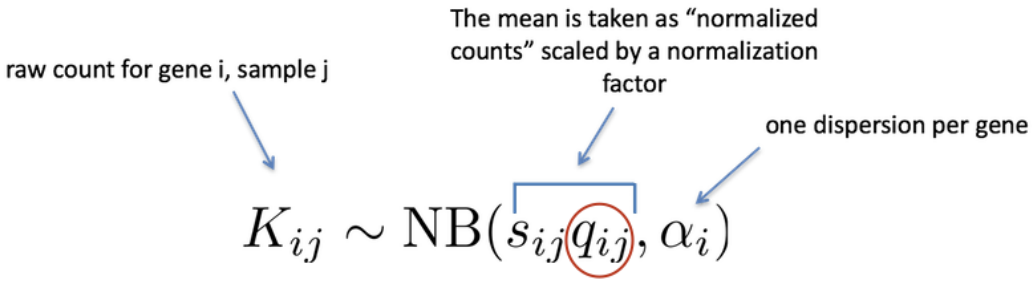

Pseudocounts with Kallisto, Sailfish, Salmon
Read counts associated with genes
Normalization
Unsupervised clustering analyses
Modeling raw counts for each gene
Testing for differential expression
Shrinking log2 fold changes
DE analysis

RNA-Seq: Differential Gene Expression

Model fitting and Hypothesis testing

Generalized Linear Model

Negative binomial distribution to model the RNA-seq counts



raw count for gene i, sample j

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

one dispersion per gene

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

> A GLM is a flexible extension of linear regression that allows modeling data where the response variable has non-normal distributions.

RNA-Seq: Differential Gene Expression

Model fitting and Hypothesis testing

Hypothesis testing

1. **Set up a null hypothesis for each gene**: there is no differential expression across the two sample groups (LFC == 0).
2. Use a statistical test to determine if based on the observed data, the null hypothesis is true.

> In DESeq2, the Wald test is the default used for hypothesis testing when comparing two groups.

RNA-Seq: Differential Gene Expression

Model fitting and Hypothesis testing

DESeq2 implements the Wald test by:

- Taking the LFC and dividing it by its standard error, resulting in a z-statistic
- The z-statistic is compared to a standard normal distribution, and a p-value is computed reporting the probability that a z-statistic at least as extreme as the observed value would be selected at random
- If the p-value is small we reject the null hypothesis and state that there is evidence against the null (i.e. the gene is differentially expressed).

✏️ The model fit and Wald test were already run previously as part of the DESeq() function

RNA-Seq: Differential Gene Expression

Multiple test correction

- As more attributes are compared, differences due solely to chance become more likely!
- Well known from array studies: 10,000s genes/transcripts
- With RNA-seq, more of a problem than ever
- All the complexity of the transcriptome gives huge numbers of potential features
    - Genes, transcripts, exons, junctions, retained introns, microRNAs, IncRNAs, etc

RNA-Seq: Differential Gene Expression

Exploring Results

Specifying contrasts

In our dataset, we have two factors in our design formula:

- species with seven levels
- stage with two levels

There are many possible pairwise comparisons, we will do:

- Chom vs. Cmac
- Cmeg vs. Cbez

RNA-Seq: Differential Gene Expression

Exploring Results

Specifying contrasts

To indicate which two sample classes we are interested in comparing, we need to specify contrasts.

The contrasts are used as input to the DESeq2 `results()` function to extract the desired results.

1. Define contrasts

```
contrast_oe <- c("species", "Chom", "Cmac")
```

RNA-Seq: Differential Gene Expression

Exploring Results

1. Define contrasts

```
contrast_oe <- c("sampletype", "ChomF", "CmacF")
```

2. Extract results for Chom vs Cmac

```
res_tableOE <- results(dds, contrast=contrast_oe,
                        alpha = 0.05)
```

3. View information stored in results

```
res_tableOE %>% data.frame() %>% View()
```

ChomF vs. CmacF

Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NAD2	13516.862	0.2000155	0.5410963	0.3696486	0.71164434	0.7147005
COX1	463957.558	0.2554752	0.5204171	0.4909047	0.62349385	0.7147005
COX2	115717.207	0.6557590	0.4216131	1.5553572	0.11986102	0.2596989
ATP8	1130.872	-2.3780298	1.4837729	-1.6026912	0.10900285	0.2596989
ATP6	76330.244	-0.2564094	0.3326933	-0.7707080	0.44088004	0.6368267
COX3	239220.519	0.2007776	0.5492471	0.3655506	0.71470047	0.7147005
NAD3	7466.317	0.5733087	0.6051762	0.9473418	0.34346462	0.5581300
NAD5	23880.222	-1.0518684	0.4441697	-2.3681678	0.01787642	0.2323935
NAD4	43220.897	0.4951666	0.3024345	1.6372692	0.10157424	0.2596989
NAD4L	629.033	0.8758263	1.6036824	0.5461345	0.58497347	0.7147005
NAD6	1437.998	-2.4738236	1.2267859	-2.0165080	0.04374688	0.2596989
CYTB	143667.904	0.4742206	0.3815155	1.2429918	0.21387085	0.3971887
NAD1	34965.721	0.6931093	0.4080032	1.6987842	0.08935986	0.2596989

ChomL vs. CmacL

Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
NAD2	13516.862	0.26265974	0.5412272	0.4853040	0.62746072	0.83696092
COX1	463957.558	-0.36154357	0.5204227	-0.6947114	0.48723617	0.83696092
COX2	115717.207	0.22436721	0.4216493	0.5321180	0.59464424	0.83696092
ATP8	1130.872	-1.77490767	1.4864819	-1.1940324	0.23246528	0.50367477
ATP6	76330.244	0.12454226	0.3327536	0.3742777	0.70819770	0.83696092
COX3	239220.519	-0.09359385	0.5492584	-0.1704004	0.86469526	0.86469526
NAD3	7466.317	0.72903565	0.6054403	1.2041413	0.22853497	0.50367477
NAD5	23880.222	-1.30998665	0.4443770	-2.9479170	0.00319923	0.04158998
NAD4	43220.897	0.64274453	0.3025368	2.1245166	0.03362697	0.14571688
NAD4L	629.033	0.67842525	1.6052697	0.4226238	0.67256972	0.83696092
NAD6	1437.998	-2.21100504	1.2275464	-1.8011580	0.07167798	0.23295343
CYTB	143667.904	0.09175208	0.3815358	0.2404809	0.80995745	0.86469526
NAD1	34965.721	0.99166437	0.4080991	2.4299594	0.01510051	0.09815333

RNA-Seq: Differential Gene Expression

Exploring Results

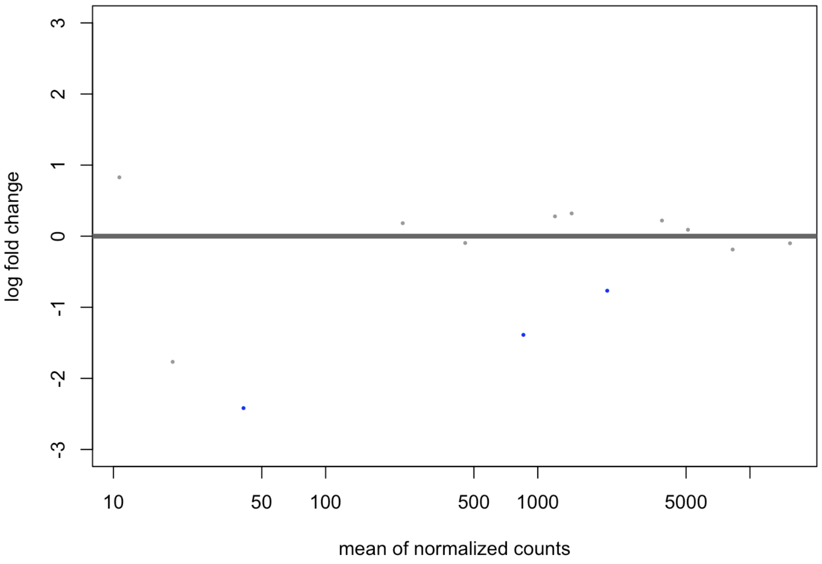
- ✏️ baseMean: mean of normalized counts for all samples
- ✏️ log2FoldChange: log2 fold change
- ✏️ lfcSE: standard error
- ✏️ stat: Wald statistic
- ✏️ pvalue: Wald test p-value
- ✏️ padj: BH adjusted p-values

RNA-Seq: Differential Gene Expression

Exploring Results

1. Plot results

```
plotMA(res_tableOE, ylim=c(-2,2))
```

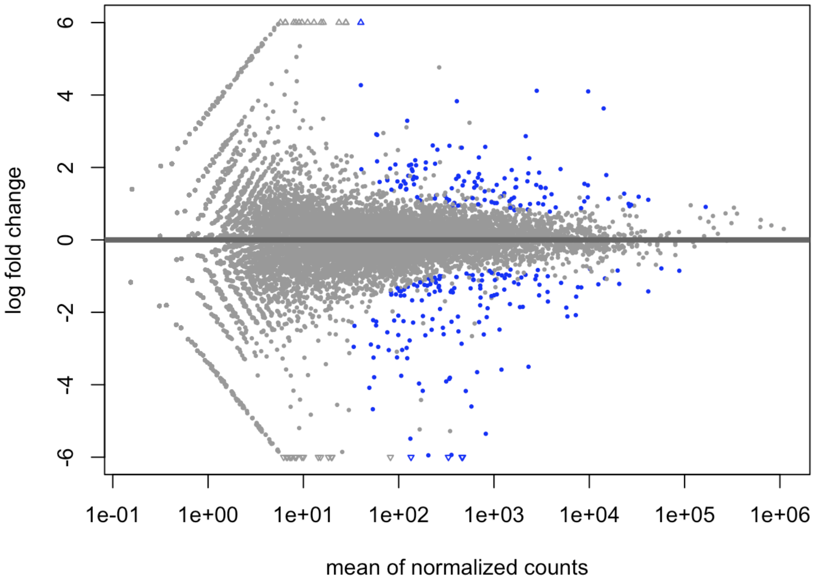


RNA-Seq: Differential Gene Expression

Exploring Results

1. Plot results

```
plotMA(res_tableOE, ylim=c(-2,2))
```



RNA-Seq: Differential Gene Expression

Exploring Results

2. Summarize results

```
summary(res_tableOE, alpha = 0.05)
```

3. Extract significant differentially expressed genes

```
padj.cutoff <- 0.05 #setting threshold

res_tableOE_tb <- res_tableOE %>%
  data.frame() %>%
  rownames_to_column(var="gene") %>%
  as_tibble() ## a tibble is an enhanced version of a dat

sigOE <- res_tableOE_tb %>%
  dplyr::filter(padj < padj.cutoff) # filter the tibble

sigOE
```

RNA-Seq: Differential Gene Expression

Exploring Results

```
  gene    baseMean log2FoldChange lfcSE  stat   pvalue
1 "ATP6 "   2125.          -0.767 0.232 -3.31 9.34e- 4
2 "NAD5 "    855.          -1.39  0.184 -7.55 4.51e-14
3 "NAD6 "     41.0         -2.42  0.733 -3.30 9.72e- 4
```

RNA-Seq: Differential Gene Expression

Exploring Results

4. Plot the expression of a single gene

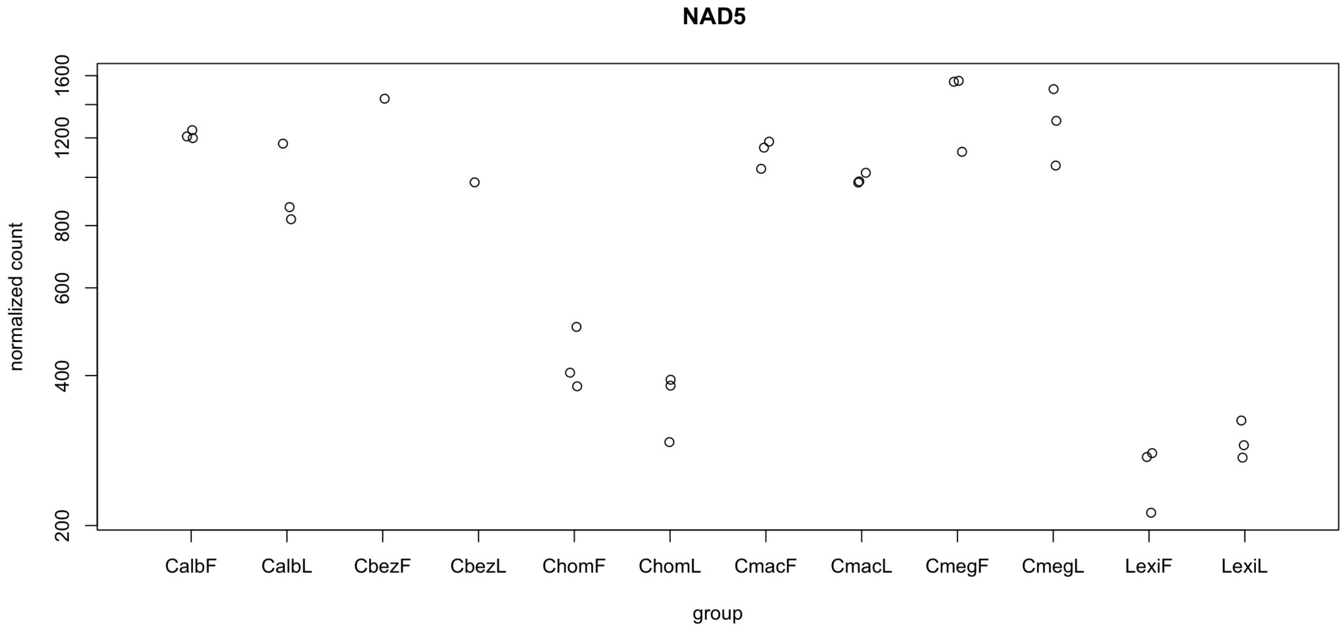
```
plotCounts(dds, gene="NAD5 ", intgroup="sampletype")
plotCounts(dds, gene="NAD6 ", intgroup="sampletype")
plotCounts(dds, gene="ATP8 ", intgroup="sampletype")
```

RNA-Seq: Differential Gene Expression

Exploring Results

4. Plot the expression of a single gene

```
plotCounts(dds, gene="NAD5 ", intgroup="sampletype")
plotCounts(dds, gene="NAD6 ", intgroup="sampletype")
plotCounts(dds, gene="ATP8 ", intgroup="sampletype")
```



RNA-Seq: Differential Gene Expression

Exploring Results

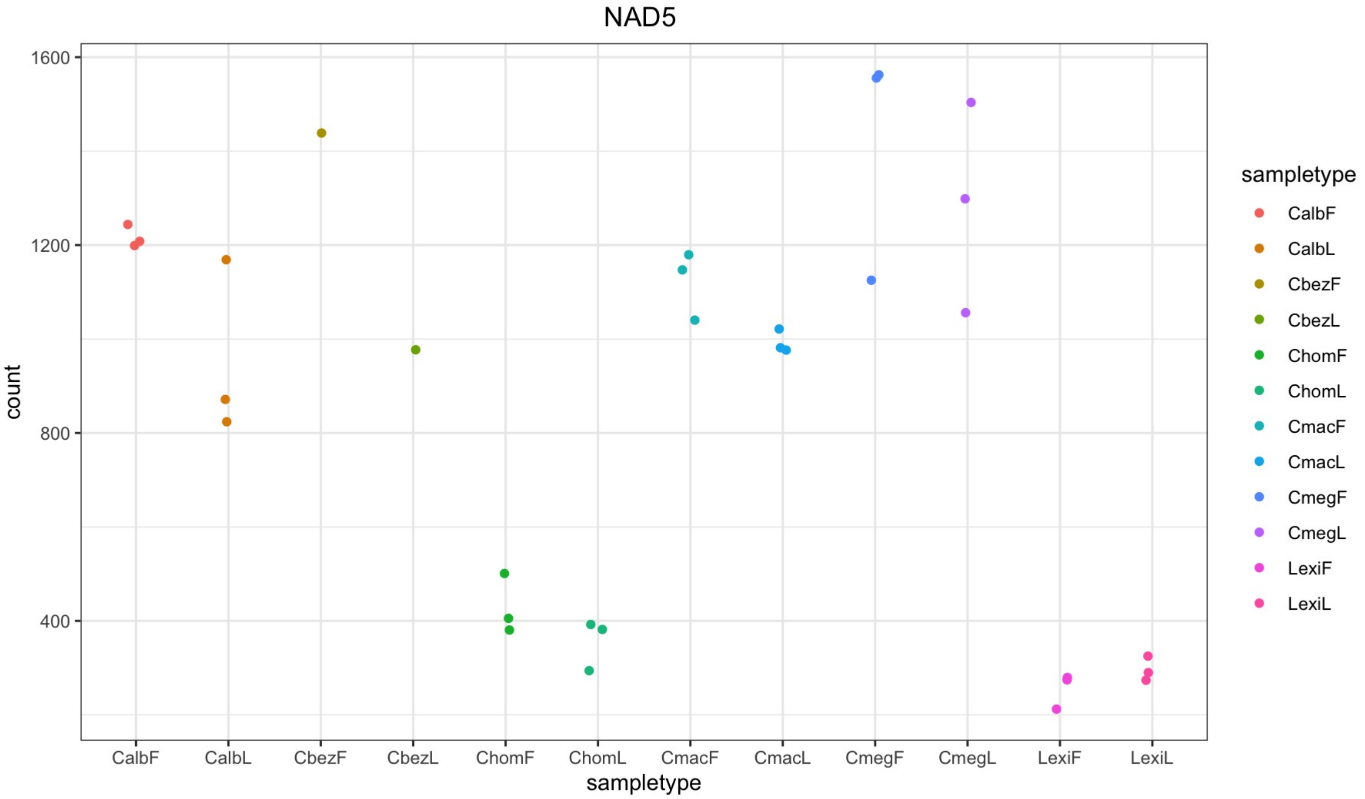
5. Using ggplot2 for the same purpose

```
d <- plotCounts(dds, gene="NAD5 ", intgroup="sampletype",
d %>% View() # View the output of plotCounts()


ggplot(d, aes(x = sampletype, y = count, color = samplety
  geom_point(position=position_jitter(w = 0.1,h = 0)) +
  theme_bw() +
  ggtitle("NAD5 ") +
  theme(plot.title = element_text(hjust = 0.5))
```

RNA-Seq: Differential Gene Expression

Exploring Results



RNA-Seq: Differential Gene Expression

Exploring Results

6. Volcano plot

```
ggplot(res_tableOE_tb) +
  geom_point(aes(x = log2FoldChange, y = -log10(padj), co
  ggtitle("mtDNA expression") +
  xlab("log2 fold change") +
  ylab("-log10 adjusted p-value") +
  #scale_y_continuous(limits = c(0,50)) +
  theme(legend.position = "none",
        plot.title = element_text(size = rel(1.5), hjust
        axis.title = element_text(size = rel(1.25)))
```

RNA-Seq: Differential Gene Expression

Exploring Results

- O `log2FoldChange` (log2FC) representa a mudança na expressão de um gene entre duas condições em escala logarítmica de base 2.

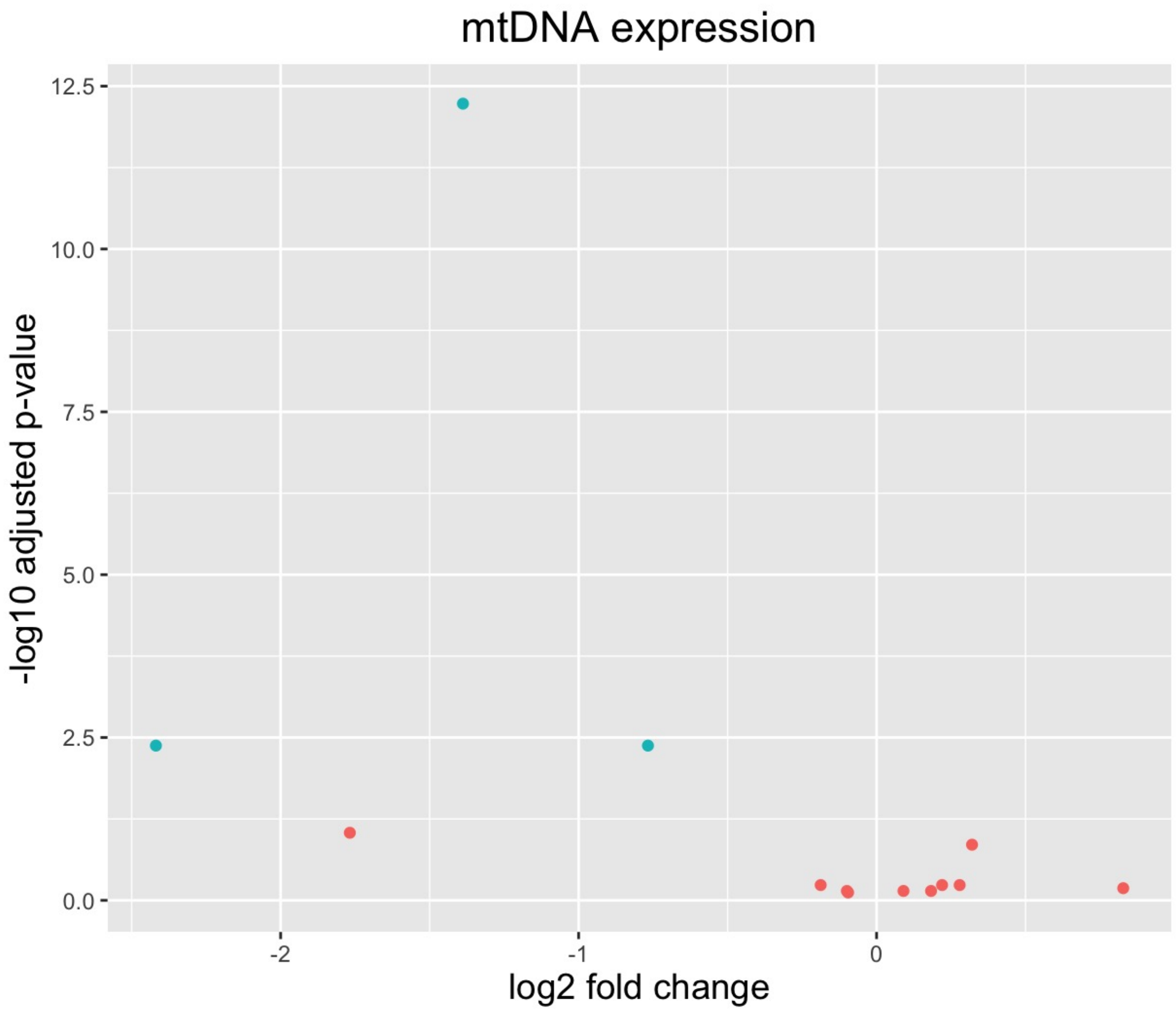
Se **log2FC = 0.58**, então no espaço linear:

$2^{0.58}$ = approx 1.5 x

Isso significa que os genes selecionados apresentam uma **alteração de pelo menos 1.5 vezes** na expressão (50% de aumento ou redução).

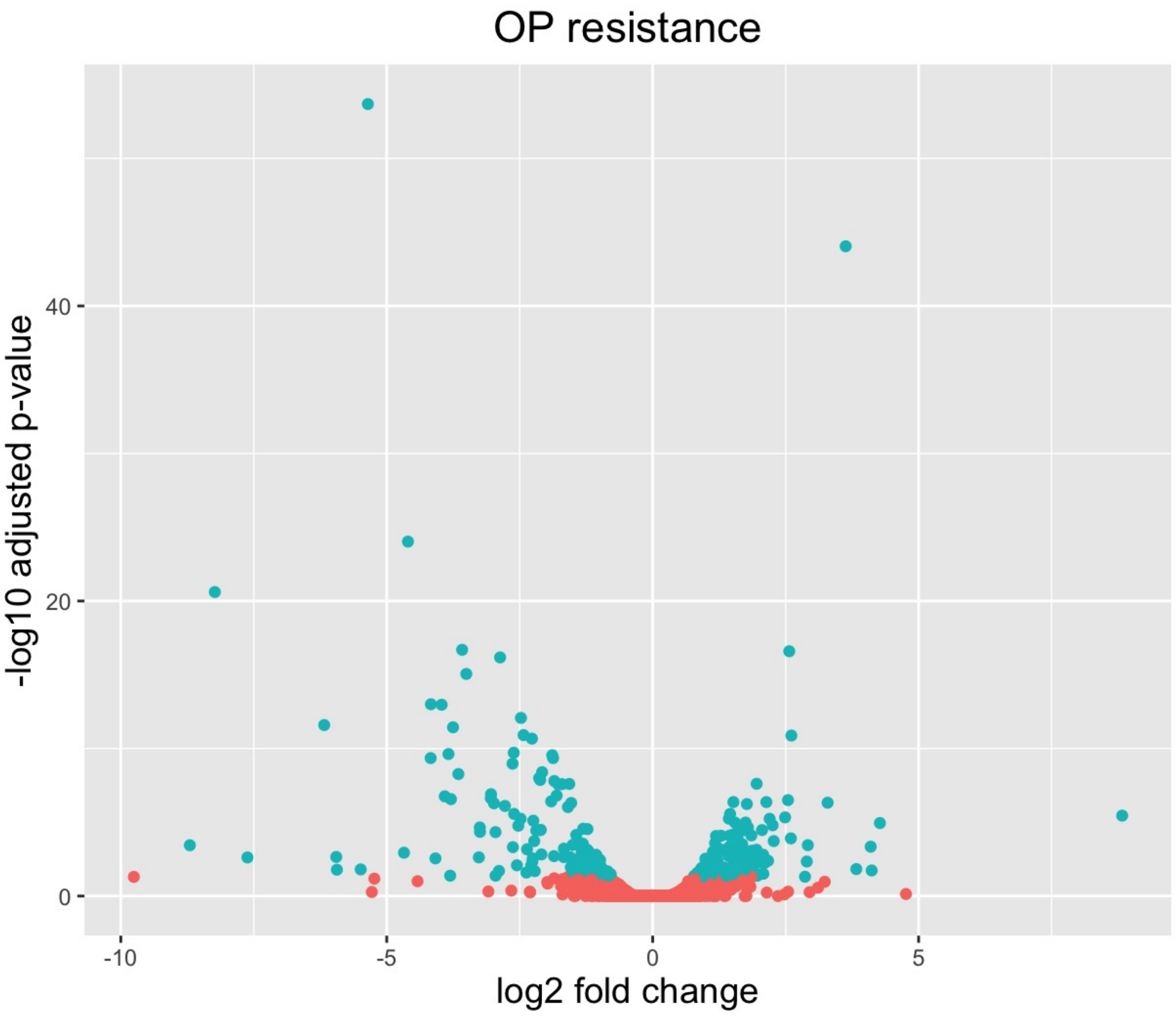
RNA-Seq: Differential Gene Expression

Exploring Results



RNA-Seq: Differential Gene Expression

Exploring Results



Next steps

Functional Analysis with GO

- **Gene Ontology (GO) Analysis**
    - Categorizes genes based on biological process, molecular function, and cellular component.
    - Helps interpret gene expression changes in a biological context.
- **Pathway Enrichment Analysis**
    - Identifies overrepresented pathways (e.g., KEGG, Reactome).
    - Provides insights into affected biological mechanisms.

Next steps

Functional Validation

- **Experimental Validation of Candidate Genes**
  - **RNAi (RNA interference)**: Knockdown of gene expression to assess phenotypic effects.
  - **CRISPR/Cas9**: Gene knockout or targeted mutagenesis to confirm gene function.
  - **Overexpression Studies**: Testing functional effects by increasing gene expression.

Next steps

Integrating Multi-Omics Data

- **Combining Transcriptomics with Other Data**
    - Genomics: Identifying regulatory variants affecting gene expression.
    - Proteomics: Correlating mRNA levels with protein abundance.
    - Metabolomics: Linking gene expression to metabolic changes.

Next steps

Conclusion

- RNA-Seq provides powerful insights into gene expression..
- Integration with multi-omics enhances interpretation.
- Future directions: single-cell RNA-Seq, spatial transcriptomics, and regulatory network analysis.

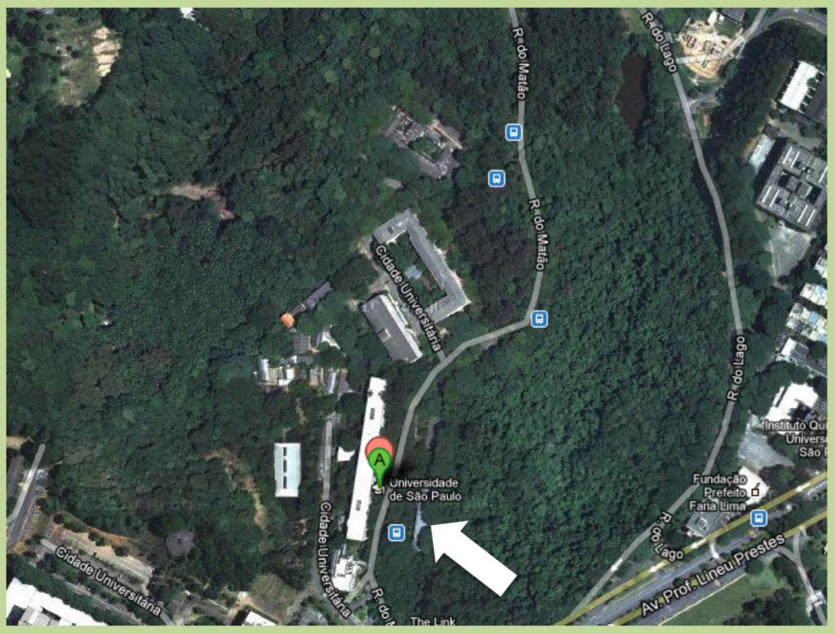
Questions?

¡Gracias por su atención!

- Ha sido un placer compartir este curso con ustedes.
- ¡Espero que sigan explorando el fascinante mundo de la transcriptómica!
- ¡Mucho éxito en sus investigaciones y proyectos futuros!
- ¡Vengan a visitarme a São Paulo!

¡Gracias por su atención!

- Ha sido un placer compartir este curso con ustedes.
- ¡Espero que sigan explorando el fascinante mundo de la transcriptómica!
- ¡Mucho éxito en sus investigaciones y proyectos futuros!
- ¡Vengan a visitarme a São Paulo!



http://torres.ib.usp.br/