Analysis of RNAseq data

Rachel Steward
Postdoc researcher, Runemark Lab
Lund University

Česky Krumlov 2024
Things I do for fun (other than work, because I love work)

Current fave: Imperial Radch Trilogy, Ann Leckie Audiobook! Narrated by Adjoa Andoh
Gene expression

The selective activity of certain genes is a highly regulated process

Gene expression is a characteristic of space (e.g., cell type, tissue, etc.) and time (e.g., developmental stage, time after event)
Gene expression

The selective activity of certain genes is a highly regulated process

Gene expression is a characteristic of space (e.g., cell type, tissue, etc.) and time (e.g., developmental stage, time after event)
What are some questions we can answer with bulk RNAseq data?

- How many genes are being expressed?
- Which genes are uniquely expressed?
- Are patterns of gene expression different among samples?
- Are patterns of expression different among genes?
- Does gene expression differ between groups or in response to a certain variable?
- What are the functional roles of groups of differently expressed genes?
Lab activities

Part 1: Exploring patterns in RNAseq data
Part 2: Differential gene expression analysis
Part 3: Functional enrichment of gene sets

Structure:
- Short background
- Open work time
- Review
Gene expression analysis

- Biological samples/Library preparation
- Sequence reads
- Read quality check
- Trimming (adaptors and low quality bases)
- Mapping to genome or transcriptome
- Count reads associated with features

De novo transcriptome assembly and quality assessment

Compare expression patterns
Identify differentially expressed features
Quality control

Reads: To trim or not to trim?
- genome annotation, variant calling, transcriptome assembly: Trim!
- Anything else, maybe trim lightly?
  - adapters + low quality score (Q10-15)

Reference genome considerations:
- What maps where:
  - Recent duplications?
  - Highly repetitive content?
  - Missing content?

Annotation considerations:
- What features have been annotated?
- Was RNAseq data used in the annotation?
  - What RNA? Life stage? Sex?
RNA sequence alignment to a reference

What are some challenges when aligning RNA-seq reads to the reference genome?
Splice-aware sequence alignment

Transcript

Trimmed short reads

Alignment to reference genome

Genome
Counting reads as a measure of expression

Two common counting tools are `featureCounts` and `htseq`.

Total read count associated with a gene *(meta-feature)* == the sum of reads associated with each of the exons *(feature)* that are a part of that gene.

```
genosics@ip-172-31-11-182:~/workshop_materials/differential_expression.refs]$ head Pca_annotation.gtf
LG1  AUGUSTUS transcript 22193 24413 . - . transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1  AUGUSTUS exon 22193 22320 . - . transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1  AUGUSTUS exon 23838 24048 . - . transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1  AUGUSTUS exon 24390 24413 . - . transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1  AUGUSTUS CDS 22193 22320 . - . 2 transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1  AUGUSTUS CDS 23838 24048 . - . 0 transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1  AUGUSTUS CDS 24390 24413 . - . 0 transcript_id "Polcal_g1.t1"; gene_id "Polcal_g1";
LG1  AUGUSTUS transcript 79912 80136 . - . transcript_id "Polcal_g2.t1"; gene_id "Polcal_g2";
LG1  AUGUSTUS exon 79912 80136 . - . transcript_id "Polcal_g2.t1"; gene_id "Polcal_g2";
LG1  AUGUSTUS CDS 79912 80136 . - . 0 transcript_id "Polcal_g2.t1"; gene_id "Polcal_g2";
```
What should count??

HTSeq manual
Read count matrix

Output of counting = A **count matrix**, with features as rows and samples as columns

Each column is a sample

<table>
<thead>
<tr>
<th>GENE ID</th>
<th>KD.2</th>
<th>KD.3</th>
<th>OE.1</th>
<th>OE.2</th>
<th>OE.3</th>
<th>IR.1</th>
<th>IR.2</th>
<th>IR.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2-SBSRNA4</td>
<td>57</td>
<td>41</td>
<td>64</td>
<td>55</td>
<td>38</td>
<td>45</td>
<td>31</td>
<td>39</td>
</tr>
<tr>
<td>A1BG</td>
<td>71</td>
<td>40</td>
<td>100</td>
<td>81</td>
<td>41</td>
<td>77</td>
<td>58</td>
<td>40</td>
</tr>
<tr>
<td>A1BG-AS1</td>
<td>256</td>
<td>177</td>
<td>220</td>
<td>189</td>
<td>107</td>
<td>213</td>
<td>172</td>
<td>126</td>
</tr>
<tr>
<td>A1CF</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A2LD1</td>
<td>146</td>
<td>81</td>
<td>138</td>
<td>125</td>
<td>52</td>
<td>91</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>A2M</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>A2ML1</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A2MP1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>A4GALT</td>
<td>55</td>
<td>37</td>
<td>107</td>
<td>118</td>
<td>65</td>
<td>49</td>
<td>52</td>
<td>37</td>
</tr>
<tr>
<td>A4GNT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AA06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAA1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAAS</td>
<td>2288</td>
<td>1363</td>
<td>1753</td>
<td>1727</td>
<td>835</td>
<td>1672</td>
<td>1389</td>
<td>1121</td>
</tr>
<tr>
<td>AACS</td>
<td>1586</td>
<td>923</td>
<td>951</td>
<td>967</td>
<td>484</td>
<td>938</td>
<td>771</td>
<td>635</td>
</tr>
<tr>
<td>AACS1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>AADAC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AADAACL2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AADAACL3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AADAACL4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AADAT</td>
<td>856</td>
<td>539</td>
<td>593</td>
<td>576</td>
<td>359</td>
<td>557</td>
<td>521</td>
<td>416</td>
</tr>
<tr>
<td>AAGAB</td>
<td>4648</td>
<td>2550</td>
<td>2648</td>
<td>2356</td>
<td>1481</td>
<td>3265</td>
<td>2790</td>
<td>2118</td>
</tr>
<tr>
<td>AAK1</td>
<td>2310</td>
<td>1384</td>
<td>1859</td>
<td>1602</td>
<td>980</td>
<td>1675</td>
<td>1614</td>
<td>1108</td>
</tr>
<tr>
<td>AAMP</td>
<td>5198</td>
<td>3061</td>
<td>3179</td>
<td>3137</td>
<td>1721</td>
<td>4061</td>
<td>3304</td>
<td>2623</td>
</tr>
<tr>
<td>AANAT</td>
<td>7</td>
<td>7</td>
<td>12</td>
<td>12</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>AAR5</td>
<td>5570</td>
<td>3323</td>
<td>4782</td>
<td>4580</td>
<td>2473</td>
<td>3953</td>
<td>3339</td>
<td>2666</td>
</tr>
<tr>
<td>AARS</td>
<td>4451</td>
<td>2337</td>
<td>2321</td>
<td>2131</td>
<td>1340</td>
<td>2492</td>
<td>2024</td>
<td>1553</td>
</tr>
</tbody>
</table>
Some problems with raw counts...

Some samples consistently have more reads, some have fewer: **systematic biases**

Each column is a sample

<table>
<thead>
<tr>
<th>GENE ID</th>
<th>KD.2</th>
<th>D.3</th>
<th>OE.1</th>
<th>OE.2</th>
<th>OE.3</th>
<th>R.1</th>
<th>IR.2</th>
<th>IR.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2-SBSRNA</td>
<td>57</td>
<td>41</td>
<td>64</td>
<td>55</td>
<td>38</td>
<td>45</td>
<td>31</td>
<td>39</td>
</tr>
<tr>
<td>A1BG</td>
<td>71</td>
<td>40</td>
<td>100</td>
<td>8</td>
<td>41</td>
<td>77</td>
<td>58</td>
<td>40</td>
</tr>
<tr>
<td>A1BG+AS1</td>
<td>256</td>
<td>177</td>
<td>220</td>
<td>187</td>
<td>107</td>
<td>213</td>
<td>172</td>
<td>126</td>
</tr>
<tr>
<td>A1CF</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A2LD1</td>
<td>145</td>
<td>81</td>
<td>138</td>
<td>129</td>
<td>52</td>
<td>91</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>A2M</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>A2ML1</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A2MP1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A4GALT</td>
<td>55</td>
<td>37</td>
<td>107</td>
<td>11</td>
<td>62</td>
<td>49</td>
<td>52</td>
<td>37</td>
</tr>
<tr>
<td>A4GNT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AA06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAA1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAAS</td>
<td>2288</td>
<td>1363</td>
<td>1753</td>
<td>172</td>
<td>835</td>
<td>1672</td>
<td>1389</td>
<td>1121</td>
</tr>
<tr>
<td>AAC5</td>
<td>1585</td>
<td>923</td>
<td>951</td>
<td>96</td>
<td>484</td>
<td>938</td>
<td>771</td>
<td>635</td>
</tr>
<tr>
<td>AACP1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>AADAC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AADACL2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AADACL3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AADACL4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AADAT</td>
<td>855</td>
<td>539</td>
<td>593</td>
<td>57</td>
<td>359</td>
<td>557</td>
<td>521</td>
<td>416</td>
</tr>
<tr>
<td>AAGAB</td>
<td>4648</td>
<td>2550</td>
<td>2648</td>
<td>235</td>
<td>1481</td>
<td>3255</td>
<td>2790</td>
<td>2118</td>
</tr>
<tr>
<td>AAK1</td>
<td>2310</td>
<td>1384</td>
<td>1859</td>
<td>160</td>
<td>980</td>
<td>1675</td>
<td>1614</td>
<td>1108</td>
</tr>
<tr>
<td>AAMP</td>
<td>5198</td>
<td>3011</td>
<td>3179</td>
<td>313</td>
<td>1721</td>
<td>4061</td>
<td>3304</td>
<td>2623</td>
</tr>
<tr>
<td>AANAT</td>
<td>7</td>
<td>7</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>AARS</td>
<td>5570</td>
<td>3323</td>
<td>4782</td>
<td>458</td>
<td>2473</td>
<td>3953</td>
<td>3339</td>
<td>2666</td>
</tr>
<tr>
<td>ARH5</td>
<td>6561</td>
<td>3237</td>
<td>3281</td>
<td>320</td>
<td>1410</td>
<td>2402</td>
<td>2034</td>
<td>1652</td>
</tr>
</tbody>
</table>
Solution: normalization

- **Normalization is NOT** fitting a normal distribution or transforming data.
- **Normalization aims to** identify and account for the nature and magnitude of **systematic biases**

The main factors often considered during normalization:
- **Sequencing depth (aka library size)**
- **RNA composition**
- **Gene length (some methods)**
Normalization

Sequencing depth

Sample A Reads

Sample B Reads
Normalization

Sequencing depth

Sample A Reads

Sample B Reads
Normalization

RNA composition

• A few highly differentially expressed genes

• Can skew some normalization methods
### Median of ratios (MRN) normalization

- Used by DESeq2 (DGE analysis tool we will use today)
- Generates a **scaling factor** for each sample to account for variation in library size

<table>
<thead>
<tr>
<th>Gene</th>
<th>sampleA</th>
<th>sampleB</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF2A</td>
<td>1489</td>
<td>906</td>
</tr>
<tr>
<td>ABCD1</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Raw counts**

<table>
<thead>
<tr>
<th>Gene</th>
<th>sampleA</th>
<th>sampleB</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF2A</td>
<td>1489/1.3 = 1145.39</td>
<td>906/0.77 = 1176.62</td>
</tr>
<tr>
<td>ABCD1</td>
<td>22/1.3 = 16.92</td>
<td>13/0.77 = 16.88</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Normalized counts**

Normalized counts are not whole numbers!
Exploring patterns in RNAseq data

Clustering of samples
• Dimension reduction analysis (e.g., PCA, PLS, MDS)
• Clustering (e.g., hierarchical clustering, k-means clustering)

Clustering of features
• Same as above, just focusing on features
• Weighted co-expression analysis (WGCNA, correlation among features)
Properties of RNA-seq count data

Most genes are expressed at low levels (small counts), lower limit of 0

No upper limit of expression, and some genes are highly expressed leading to a long right tail
Data transformations for clustering and visualization

• Pseudo-log: $y = \log_2(n + n_0)$
  • $n_0$ is a constant, like 1
  • Variance not stable at low values (does not scale with expression)

Instead, we want to transform the data to remove the trend (variances roughly similar across mean values)

• Variance stabilizing transformation \textit{DESeq2::vst()} 
• Regularized log transformation \textit{DESeq2::rlog()}

Anders & Huber 2010 \textit{Nature}, Love et al. 2023
“Analyzing RNA-seq data with DESeq2”
Effect of transformations on variance

Pseudo-log

VST

rlog

Love et al. 2023 “Analyzing RNA-seq data with DESeq2”
Today's lab: *Polygonia c-album*
2 Our questions

1. Do patterns of gene expression differ between larvae reared on different host plants?
2. Which genes are differently expressed between larvae reared on different host plants?
3. What are the functions of differentially expressed gene sets?

3 Background

Today’s tutorial walks through a reference-based differential gene expression (DGE) analysis. This means our reads have been aligned to an existing reference genome for P. c-album, rather than a de novo transcriptome generated from the RNA-seq data. The three main steps of reference-based DGE analysis are 1) alignment, 2) quantification and 3) analysis (Fig. 2). In this tutorial, we will focus on step 3) analysis.

This tutorial has three units:

- Exploring patterns of gene expression among samples
- Identifying differentially expressed genes
- Evaluating functional enrichment of DE gene sets

Each unit has core exercises you should try to finish during the lab. If you finish the core exercises, there are additional challenge exercises at the end of each unit.

Occasional blue boxes give background on the analyses. Feel free to gloss over these – you can come back to them later if you are curious or want to learn more.

4 Unit 1: Exploring patterns of gene expression among samples

Everything in this tutorial will be done in RStudio.

4.1 Set the working directory

Open RStudio and start by checking (getwd()) and setting (setwd()) your working directory. The activity is designed to be run in the “RNAseq_analysis” directory.

Alternatively, you can set the working directory using the RStudio interface. Click on the Files tab. Navigate by clicking on the directories you want to enter (workshop_materials, then RNAseq_analysis). Once inside the working directory, use the More drop-down menu (next to the little blue gear) and select Set As Working Directory.

Take a look at the contents of the directory and subdirectory. You can do this using the list.files() command with the recursive = T option, or by selecting Go To Working Directory from the More drop-down menu on the Files tab.
Exploring patterns in RNAseq data

Part 1

Core tasks:

• Load raw count matrix
• Transform for visualization
• PCA of samples
• Hierarchical clustering of samples

Challenge exercises
Open work time (25 min)

Five more minutes!

5 minutes
Review
Lab activities

Part 1: Exploring patterns in RNAseq data

Part 2: Differential gene expression analysis

Part 3: Functional enrichment of gene sets
Differential expression analysis

Normalized expression level

Analysis

Significant difference between group means

Global mean
Condition A sample
Condition A group mean
Condition B sample
Condition B group mean

No significant difference between group means

features (e.g. genes)
samples

Input

<table>
<thead>
<tr>
<th>Gene_id</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polcal_g1</td>
<td>17</td>
<td>10</td>
<td>5</td>
<td>23</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Polcal_g2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Polcal_g3</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Polcal_g4</td>
<td>17</td>
<td>11</td>
<td>5</td>
<td>21</td>
<td>10</td>
<td>12</td>
</tr>
</tbody>
</table>
Differential expression analysis

1. Raw read counts
2. Normalization
3. Modeling counts for each gene
4. Testing for differential expression
DESeq2 package

METHOD | Open Access | Published: 05 December 2014

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love, Wolfgang Huber & Simon Anders

*Genome Biology* 15, Article number: 550 (2014) | Cite this article

450k Accesses | 34853 Citations | 131 Altmetric | Metrics
Modeling raw counts for each gene

Step 1. Normalization (aka estimation of size factors) → done!

Step 2. Estimate gene-wise dispersion

• To accurately model sequencing counts, we need to generate accurate estimates of within-group variation for each gene (aka dispersion)
  • need to choose the right distribution
Statistical modeling of count data

Which probability distributions are suitable for modeling count data?

Poisson distribution?

A property of Poisson distribution is that the mean = variance.
Statistical modeling of count data

Which probability distributions are suitable for modeling count data?

Poisson distribution?

A property of Poission distribution is that the mean = variance.

Poisson distribution is not suitable to model count data across the biological samples.
Statistical modeling of count data

The distribution that fits best is the **Negative Binomial (NB)** distribution.

- Two parameters, one for the mean and one for the variance
- Flexibility to estimate the amount of **dispersion** for each gene across samples.
How does the dispersion relate to our model?

Variation is an important part of model fitting and hypothesis testing.

Estimates of variation for each gene are often unreliable.

DESeq2 shares information across genes to generate more accurate estimates of variation:

Fitted dispersion curve = expected dispersion for genes of a given level of expression (e.g., mean normalized count)
Model fitting and hypothesis testing

\[ y = \beta_0 + x_1 \beta_1 \]

- \( y \) = normalized expression level
- \( \beta_0 \) = intercept (the estimated expression for the base level, condition A (red))
- \( x_1 \) = a binary indicator variable for (0 if part of the red group, 1 if part of the blue group)
- \( \beta_1 \) = coefficient for condition B (blue)

\[ y = \beta_0 + 0 \ast \beta_1 \]
\[ y = \beta_0 \]
\[ y = \beta_0 + 1 \ast \beta_1 \]
\[ y = \beta_0 + \beta_1 \]

Step 4. Generalized Linear Model fit for each gene
Model fitting and hypothesis testing

Step 4. Generalized Linear Model fit for each gene

\[ y = \beta_0 + \beta_1 \]
\[ y - \beta_0 = \beta_1 \]
\[ \log_2 (expression_{blue}) - \log_2 (expression_{red}) = \beta_1 \]
\[ \log_2 \left( \frac{expression_{blue}}{expression_{red}} \right) = \beta_1 \text{ "log}_2\text{ Fold Change"} \]
Specifying contrasts

\[
\log_2 \left( \frac{\text{expression}_{\text{Ribes}}}{\text{expression}_{\text{Urtica}}} \right) = \beta_1 \quad \text{"log}_2\text{ Fold Change"}
\]

Pca_dd <- DESeqDataSetFromMatrix(countData = Pca_counts,
colData = Pca_metadata,
design = ~ condition)

contrast_U_R <- c("condition", "Urtica", "Ribes")

# extract the results for your specified contrast
Pca_res_table_U_R <- results(Pca_dd_filt, contrast=contrast_U_R)
### Output of DESeq2

**log2 fold change (MLE):** condition Urtica vs Ribes

**Wald test p-value:** condition Urtica vs Ribes

Dataframe with 10253 rows and 6 columns:

<table>
<thead>
<tr>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>lfcSE</th>
<th>stat</th>
<th>pvalue</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;numeric&gt;</td>
<td>&lt;numeric&gt;</td>
<td>&lt;numeric&gt;</td>
<td>&lt;numeric&gt;</td>
<td>&lt;numeric&gt;</td>
<td>&lt;numeric&gt;</td>
</tr>
<tr>
<td>Polcal_g10</td>
<td>89.7562</td>
<td>0.2644909</td>
<td>0.164662</td>
<td>1.606262</td>
<td>0.108216</td>
</tr>
<tr>
<td>Polcal_g100</td>
<td>128.7307</td>
<td>0.0751998</td>
<td>0.120094</td>
<td>0.626174</td>
<td>0.531201</td>
</tr>
<tr>
<td>Polcal_g1000</td>
<td>80.8697</td>
<td>-0.0682283</td>
<td>0.117253</td>
<td>-0.581890</td>
<td>0.560641</td>
</tr>
<tr>
<td>Polcal_g10000</td>
<td>18.4347</td>
<td>0.0794954</td>
<td>0.237090</td>
<td>0.335296</td>
<td>0.737402</td>
</tr>
<tr>
<td>Polcal_g10006</td>
<td>19.1902</td>
<td>0.4310584</td>
<td>0.295618</td>
<td>1.458158</td>
<td>0.144797</td>
</tr>
<tr>
<td>Polcal_g9993</td>
<td>15.1301</td>
<td>-0.181906</td>
<td>0.356393</td>
<td>-0.51041</td>
<td>0.6097642</td>
</tr>
<tr>
<td>Polcal_g9994</td>
<td>16.6881</td>
<td>0.402894</td>
<td>0.294354</td>
<td>1.36874</td>
<td>0.1710811</td>
</tr>
<tr>
<td>Polcal_g9996</td>
<td>84.0056</td>
<td>0.140555</td>
<td>1.025049</td>
<td>0.13712</td>
<td>0.8909358</td>
</tr>
<tr>
<td>Polcal_g9998</td>
<td>2.9282</td>
<td>-1.638792</td>
<td>0.745256</td>
<td>-2.19897</td>
<td>0.0278803</td>
</tr>
<tr>
<td>Polcal_g9999</td>
<td>4.0105</td>
<td>-1.006017</td>
<td>0.598296</td>
<td>-1.68147</td>
<td>0.0926717</td>
</tr>
</tbody>
</table>

1. **baseMean**: mean of normalized counts for all samples
2. **log2FoldChange**: log2 fold change
3. **lfcSE**: standard error
4. **stat**: Wald statistic
5. **pvalue**: Wald test p-value
6. **padj**: BH adjusted p-values – use a pre-defined cutoff for significance
When can we detect differential expression?

Liu et al. 2014. Bioinformatics
What do we do with DE genes?

• Visualize expression levels, log fold changes, and significance
• Identify up- and down-regulated genes
• Compare sets of DE genes
• Test for functional enrichment of DE gene sets
Differential gene expression

Part 2 Core tasks:
• Run a pairwise contrast
• Visualize differential expression with a volcano plot
• Extract the list of DE genes
• Visualize DE genes in a heatmap

Challenge exercises
Challenge questions
Open work time

Five more minutes!

5 minutes
Part 3: functional annotation

Differential expression or clustering analysis can produce large gene sets.

How can we figure out the functional consequences of these differences?

Gene set enrichment analysis:

Do functional terms occur in the target gene set more than expected by chance?

  - GO terms
  - KEGG pathways
  - Reactome pathways
Additional slides
## Links to other DE/DS tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Use</th>
<th>Link to best resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGCNA (R package)</td>
<td>Weighted gene coexpression analysis groups genes into modules/clusters by expression patterns across samples</td>
<td>Horvath lab website: <a href="https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/">https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/</a></td>
</tr>
<tr>
<td>DEXSeq (R package)</td>
<td>Differential exon expression within the DESeq2 framework from exon count data</td>
<td>Vignette: <a href="https://bioconductor.org/packages/release/bioc/vignettes/DEXSeq/inst/doc/DEXSeq.html">https://bioconductor.org/packages/release/bioc/vignettes/DEXSeq/inst/doc/DEXSeq.html</a></td>
</tr>
<tr>
<td>LeafCutter (python &amp; R scripts)</td>
<td>Differential splicing analysis specifically focused on differential intron retention from junction count data</td>
<td>Github page: <a href="https://davidaknowles.github.io/leafcutter/">https://davidaknowles.github.io/leafcutter/</a></td>
</tr>
</tbody>
</table>
Median of ratios (MRN) normalization

- Used by DESeq2 (DGE analysis tool we will use today)

Let’s see how the normalization works...
Step 1. Create a pseudo-reference sample for each gene (row-wise geometric mean)

<table>
<thead>
<tr>
<th>Gene</th>
<th>sampleA</th>
<th>sampleB</th>
<th>Pseudo-reference sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF2A</td>
<td>1489</td>
<td>906</td>
<td>sqrt(1489*906) = 1161.5</td>
</tr>
<tr>
<td>ABCD1</td>
<td>22</td>
<td>13</td>
<td>sqrt(22*13) = 16.9</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Step 2. Calculates ratio of each sample to the reference

<table>
<thead>
<tr>
<th>Gene</th>
<th>sampleA</th>
<th>sampleB</th>
<th>Pseudo-reference sample</th>
<th>Ratio of sampleA/ref</th>
<th>Ratio of sampleB/ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF2A</td>
<td>1489</td>
<td>906</td>
<td>1161.5</td>
<td>1489/1161.5 = 1.28</td>
<td>906/1161.5 = 0.78</td>
</tr>
<tr>
<td>ABCD1</td>
<td>22</td>
<td>13</td>
<td>16.9</td>
<td>22/16.9 = 1.30</td>
<td>13/16.9 = 0.77</td>
</tr>
<tr>
<td>MEFV</td>
<td>793</td>
<td>410</td>
<td>570.2</td>
<td>793/570.2 = 1.39</td>
<td>410/570.2 = 0.72</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
The figure below illustrates the median value for the distribution of all gene ratios for a single sample (frequency is on the y-axis).

The median of ratio methods makes the assumption that not ALL genes are differentially expressed; therefore, the normalization factors should account for sequencing depth and RNA composition of the sample (large outlier genes will not represent the median ratio values).
Step 3. Calculate the normalization factor for each sample (size factor)

<table>
<thead>
<tr>
<th>Gene</th>
<th>sampleA</th>
<th>sampleB</th>
<th>Pseudo-reference sample</th>
<th>Ratio of sampleA/ref</th>
<th>Ratio of sampleB/ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF2A</td>
<td>1489</td>
<td>906</td>
<td>1161.5</td>
<td>1489/1161.5 = 1.28</td>
<td>906/1161.5 = 0.78</td>
</tr>
<tr>
<td>ABCD1</td>
<td>22</td>
<td>13</td>
<td>16.9</td>
<td>22/16.9 = 1.30</td>
<td>13/16.9 = 0.77</td>
</tr>
<tr>
<td>MEFV</td>
<td>793</td>
<td>410</td>
<td>570.2</td>
<td>793/570.2 = 1.39</td>
<td>410/570.2 = 0.72</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\[
\text{median}(c(1.28, 1.3, 1.39, 1.35, 0.59, \ldots)) = 1.3
\]

\[
\text{median}(c(0.78, 0.77, 0.72, 0.8, 0.73, \ldots)) = 0.77
\]
Step 4: calculate the normalized count values using the normalization factor

Raw counts:

<table>
<thead>
<tr>
<th>Gene</th>
<th>sampleA</th>
<th>sampleB</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF2A</td>
<td>1489</td>
<td>906</td>
</tr>
<tr>
<td>ABCD1</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Normalized counts

<table>
<thead>
<tr>
<th>Gene</th>
<th>sampleA</th>
<th>sampleB</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF2A</td>
<td>1489/1.3 = 1145.39</td>
<td>906/0.77 = 1176.62</td>
</tr>
<tr>
<td>ABCD1</td>
<td>22/1.3 = 16.92</td>
<td>13/0.77 = 16.88</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Normalized counts are not whole numbers!