

Bioinformatics I

Differential gene expression with RNAseq

Contact: Lukas.Schrader@wwu.de

Contents

1 Biological background	1
Insects as partners in symbioses	1
2 Technical background	3
Quantifying gene expression with high-throughput sequencing	3
3 Practical Part	5
3.1 Open the count data in Excel	5
3.2 Get an idea of the data set	5
3.3 Transforming raw counts to comparable expression values	8
3.4 Data exploration	9
3.5 Statistical Testing	11
3.6 FDR correction	13
3.7 Differentially expressed genes	15
4 A brief introduction into R	15

List of Figures

1	Figure 1: <i>Westeberhardia cardiocondylae</i> , a vertically transferred endosymbiont of ants.	2
2	Figure 2: The principle of quantifying gene expression with RNAseq	3
3	Figure 3: Workflow of an RNAseq analysis	4
4	Figure 4: readCounts tab in the excel file <i>Westeberhardia_RNAseq.xlsx</i>	5
5	Figure 5: Total read counts per sample.	6
6	Figure 6: Number of expressed genes per Sample.	7
7	Figure 7: Total average number of reads of all expressed genes.	8
8	Figure 8: Total average number of reads of all expressed genes.	9
9	Figure 9: Overview of <i>Westeberhardia</i> expression and expression variation in queens and larvae.	12
10	Figure 10: MA plot comparing <i>Westeberhardia</i> gene expression in queens and larvae.	14
11	Figure 11: Updated MA-plot (FDR corrected)	15

1 Biological background

Over the course of 4 billion years, biological evolution has created millions of species on this planet. Although this is mostly due to differentiation of preexisting species, species differentiation was not the only process that produced today's biodiversity. The tendency of nature to combine two species also promotes biodiversity. Combinations include predator-prey relationships, parasitisms, and symbioses. Among these, symbioses have had major consequences for evolution. The magnitude of how symbioses have impacted life on earth is immediately evident in light of the origin of eukaryotic cells, which are chimeras of several prokaryotes evolving about 2 billion years ago.

Insects as partners in symbioses

As multicellular organisms, insects as a group seem to be most tolerant of foreign organisms and live together with many different microorganisms, both inside and outside their bodies, in a variety of ways. In this sense, insects provide great systems for studying the evolutionary significance of interspecific symbioses. Insects have one of the most successful lifestyles on earth. One important factor in their success is that they have adapted to a wide variety of diets. Such flexible feeding habits have evolved, at least in part, through symbiosis where bacterial endosymbionts are associated with the insect host's digestive tract.

Insects display the full scope of endosymbiosis. Some cases seem to be very new in an evolutionary sense, because the association between host insect and symbiont is still only loose and opportunistic, whereas others are so intimate that symbionts seem to be tightly integrated into the physiology of the host. In the latter cases, endosymbionts appear to be reduced to the level of a cell's organelle, and their significance for the host is no less than that of mitochondria for a eukaryotic cell.

One important point to keep in mind is that symbiosis is not necessarily a fair association in which both partners benefit equally. Mutualism is not always maintained on a 50-50 basis. One associate often takes more, sometimes much more, than the other. This is observed typically in the endosymbioses between insects and microorganisms. For example, aphids appear to profit disproportionately more from the symbiosis with *Buchnera* symbionts compared to the bacterium, yet the symbiosis has lasted for over 200 million years. On the other hand, bacterial symbionts of the *Wolbachia* genus take full advantage of insect hosts so as to propagate their progeny, whereas the hosts seem to only receive little reward (if any). Yet, many insects have hosted these selfish *Wolbachia* passengers for an evolutionarily significant length of time.

Many insect species harbor intracellular symbionts that live inside the cells of their host and are vertically transmitted (usually from mother to daughter) over generations. If all cases of endosymbiosis are included, the percentage of insect species that contain intracellular symbionts likely exceeds 70%. Intracellular symbiosis is the most intimate association between two different organisms, and it is generally reasoned that the association is maintained over generations because the host and symbiont equally benefit from the association. In reality, however, it does not seem to apply in intracellular symbiosis between insects and microorganisms.

Many intracellular symbioses of insects with microorganisms are characterized by a specialized internal organ – the bacteriome – in the insect that has differentiated for the purpose of harboring the endosymbionts. In such so-called bacteriome symbioses, the benefits to the host insects are often easy to decipher. By contrast, it is much less clear how the endosymbionts truly benefit from the association. These bacteriome symbionts seem to be domesticated by the host insects. In this respect, their relation to insects somewhat resembles that of livestock to us.

In this practical, we will focus on the bacteriome symbiosis of the ant *Cardiocondyla obscurior* and the endosymbiont *Westeberhardia cardiocondylae* (Fig. 1, see <https://doi.org/10.1038/ismej.2015.119>). In this

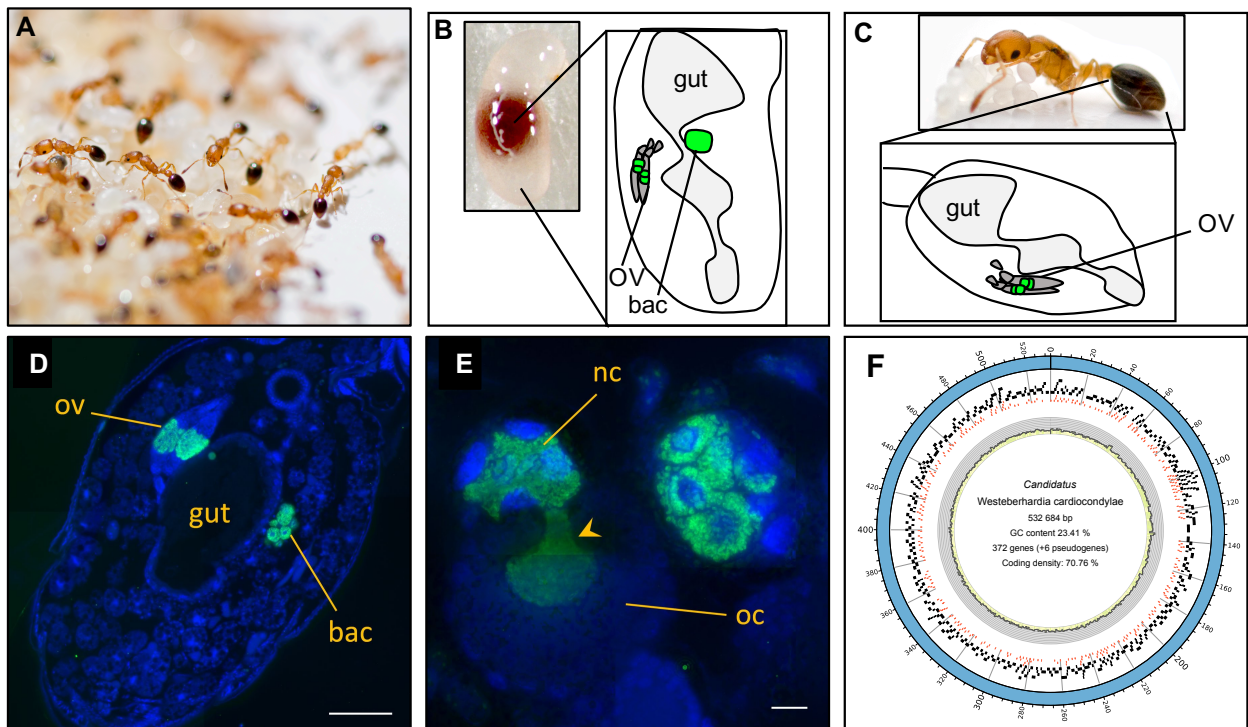


Figure 1: *Westeberhardia cardiocondylae*, a vertically transferred endosymbiont of ants.

(A) A laboratory colony of *C. obscurior* at the University of Münster. (B) Schematic of the localization of *Westeberhardia cardiocondylae* in the ant larvae. The endosymbionts are found in specialized organs in the insect abdomen, the so-called bacteriome (bac). This gut-associated organ is an adaptation to the mutualistic interaction between the ant and its symbiont. The bacteria are also found in the developing ovaries (OV) (C) Schematic of the localization of *Westeberhardia cardiocondylae* in the mature ant queen. The bacteriome has degenerated and the endosymbiont is only found in the ovaries (OV) (D) Histological detection of the endosymbiont in a queen pupa. OV = ovaries, bac = bacteriome. (E) Histological detection of the mother-offspring inheritance of the endosymbiont. Developing oocytes are infected with the bacteria during a process called nurse cell depletion, during which the endosymbiont is flushed into the oocyte. nc = nurse cell, OC = oocyte. (F) A summary of the simple genome of *Westeberhardia cardiocondylae*. The genomes of endosymbionts are usually much smaller than genomes of related, free-living organisms. This makes it feasible for us to analyze expression of an entire organism in this practical.

supposed mutualism, the bacteria are inherited from one generation to the next by passage inside the eggs - which is a typical mode of *vertical transfer* in bacteriome insect-bacteria symbioses. The contribution of *Westeberhardia cardiocondylae* to the mutualism is suspected to be 4-hydroxyphenylpyruvate, which is produced by the bacteria in the skikimate-pathway and then converted by the host to DOPA, an important ingredient for insect cuticles. Consequently, it has been hypothesized that *Westeberhardia cardiocondylae* is particularly important during the development of the ants, during which the cuticle is formed.

As part of this practical, you will analyze expression of all protein-coding genes of the endosymbiont *Westeberhardia cardiocondylae* and compare two different groups of samples (adult queens and developing larvae), to identify genes that are up- or downregulated between these groups. Your task will be to compare expression profiles of *Westeberhardia* in third instar larvae to expression profiles of *Westeberhardia* in adult

queens of *C. obscurior*. The working hypothesis here is that the the endosymbiont changes expression of several important genes in the transition from being in a developing larva to being in a mature queen. Any endosymbiont genes differentially expressed between queens and larvae of the ant are likely important for the mutualistic interaction between host and bacterium (e.g. genes of the shikimate pathway that help producing DOPA). You will run a (simplified) analysis to figure out which genes are differentially expressed using data produced by RNAseq (i.e. "RNA sequencing"), which is the method of choice for quantifying gene expression across all genes encoded in an organism's genome.

2 Technical background

Quantifying gene expression with high-throughput sequencing

RNAseq is a method to determine the identity and abundance of RNA sequences in biological samples. The molecular methods involve isolation of RNA from cell, tissue, or whole-animal samples, preparation of libraries that represent the entire suit of RNAs in the samples, chemical sequencing of the library, and subsequent bioinformatic data analysis where RNA sequences are mapped onto a reference genome. Finally expression levels for individual genes are inferred and compared (Fig. 2). The hallmark innovations of RNAseq are the possibility to sequence with high throughput, the sensitivity of identifying individual transcripts, and the ability to discover novel transcripts, gene models, and small non-coding RNA species. RNAseq is derived from high-throughput DNA sequencing technologies ("next generation sequencing"), utilizing a "sequencing-by-synthesis" approach in a massively parallel format, so that the number of sequencing reactions in a single run can be in millions. A typical sequencing run consists of 6000 M sequencing reactions, each yielding 100 nucleotides of sequence information. These 100-nucleotide-long sequences are called reads and each read represents a single RNA/DNA molecule that was present in the sample.

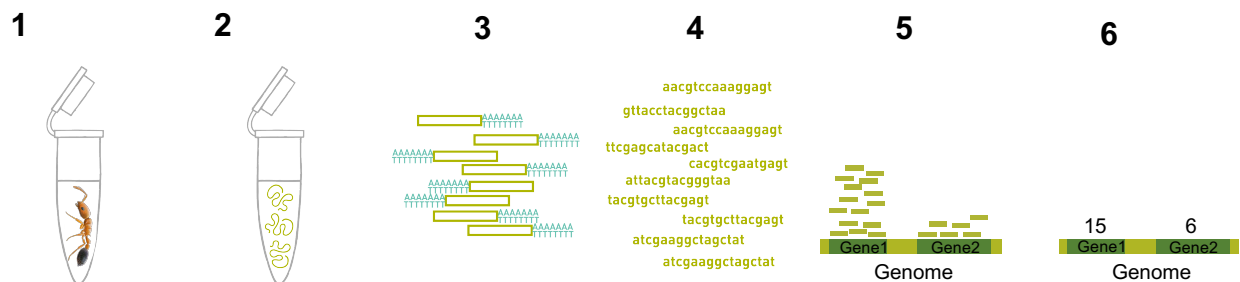


Figure 2: The principle of quantifying gene expression with RNAseq

The figure shows the basic steps in an RNAseq experiment. (1) Sample collection. (2) RNA extraction. (3) mRNA isolation. (4) high-throughput sequencing of mRNA reads. (5) Alignment of mRNA read sequences against the reference genome sequence. (6) Counting of aligned mRNA reads per gene.

In this practical, you will work with a real data set. You will start with a table that contains the counts for each gene and each sample. These counts are simply the number of reads that could be mapped to a specific gene in a given sample. So generally, the higher the number in the table, the higher the gene was expressed in the sampled individual. The dataset contains gene expression values for **seven individual larvae** and **seven individual queens**. The samples are named Larva 1 to Larva 7 and Queen 1 to Queen 7. Remember, that we only look at genes expressed by the endosymbiont.

The basic workflow of the bioinformatic analysis of an RNAseq experiment is summarized in Figure 3. We have performed the preprocessing (i.e. aligning of reads against the genome and counting the reads per gene) for you already, so that you will start with the counts per gene per sample.

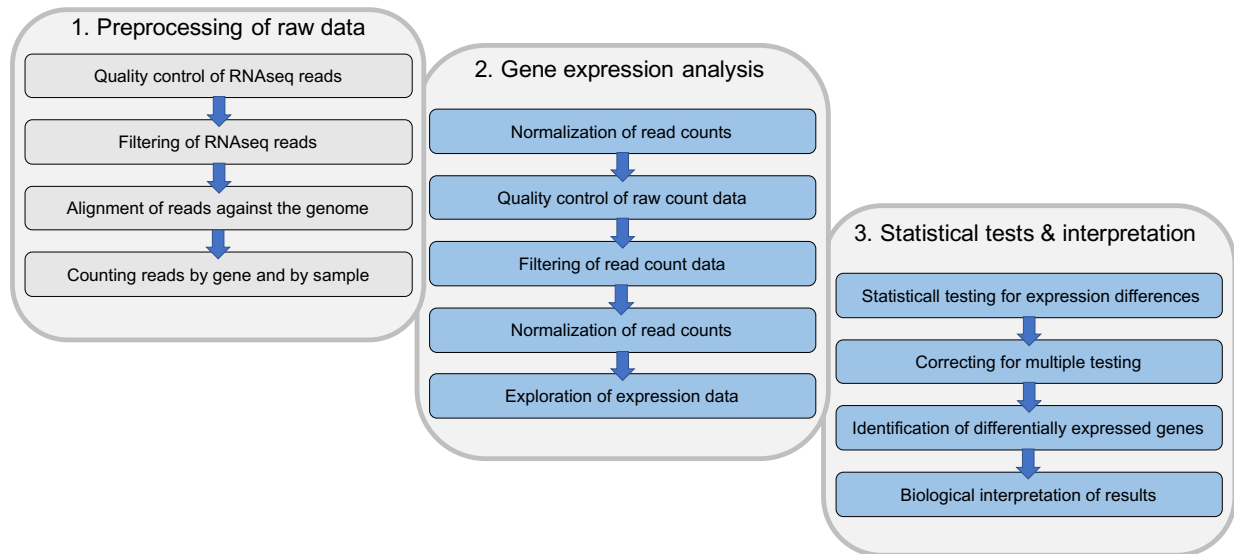


Figure 3: Workflow of an RNAseq analysis

We will provide you with a preprocessed dataset, ready for you to analyse. In this practical, you will conduct the *Gene expression analysis* (2.) and the *Statistical tests & interpretation* (3.) part of the analysis.

3 Practical Part

Please note that copy-pasting from this pdf into excel will not work well. Please type in the formulae in excel and then copy them from within the excel file. There is no need to manually enter a formula in each cell as there are plenty of handy ways to dynamically populate an excel sheet with the correct formulae. Feel free to ask or google for them.

3.1 Open the count data in Excel

Open the excel file `Westeberhardia_RNAseq.xlsx` and have a look at the first tab (`readCounts`) in the file. It should look something like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC						
1	gene	Queen1	Queen2	Queen3	Queen4	Queen5	Queen6	Queen7	Larva1	Larva2	Larva3	Larva4	Larva5	Larva6	Larva7	Feature	Start	Stop	Score	Strand	Frame	symbol	func												
2	WEOB_001	700	510	697	965	555	940	610	104	56	117	37	156	5	42	CDS	216	2117	.	+	.	mmnG	NAD-binding protein involved in the addition of a carboxymethylaminomethyl (cmnm) group at t												
3	WEOB_002	330	250	243	573	279	518	205	28	20	29	15	28	1	25	CDS	2427	3251	.	+	.	atpB	Key component of the proton channel; it plays a direct role in the translocation of protons across												
4	WEOB_003	50	33	40	90	33	67	20	3	1	6	0	1	0	6	CDS	3310	3549	.	+	.	atpE	Key component of the F(0) channel; it plays a direct role in translocation across the membrane. A												
5	WEOB_004	274	180	229	418	196	397	183	40	17	39	9	31	1	10	CDS	3731	4201	.	+	.	atpF	Component of the F(0) channel, it forms part of the peripheral stalk, linking F(1) to F(0)												
6	WEOB_005	270	186	313	452	247	334	223	33	32	30	19	20	1	24	CDS	4226	4759	.	+	.	atpH	This protein is part of the stalk that links CF(0) to CF(1). It either transmits conformational chang												
7	WEOB_006	1270	1008	1382	2012	1056	1528	1024	152	101	174	63	109	12	68	CDS	4798	6348	.	+	.	atpA	Produces ATP from ADP in the presence of a proton gradient across the membrane. The alpha ch												
8	WEOB_007	771	519	739	1304	588	912	614	56	57	69	28	60	3	50	CDS	6391	7257	.	+	.	atpG	Produces ATP from ADP in the presence of a proton gradient across the membrane. The gamma												
9	WEOB_008	1056	907	985	1495	928	1545	819	103	64	145	40	86	4	94	CDS	7296	8672	.	+	.	atpD	Produces ATP from ADP in the presence of a proton gradient across the membrane. The catalytic												
10	WEOB_009	24	25	30	55	27	64	24	2	1	1	1	5	0	1	CDS	8703	8957	.	+	.	atpC	Produces ATP from ADP in the presence of a proton gradient across the membrane												
11	WEOB_010	335	276	362	607	302	426	318	30	30	35	8	50	2	21	CDS	9801	#####	.	-	.	ptsN	Seems to have a role in regulating nitrogen assimilation												
12	WEOB_013	239	189	247	428	222	304	212	18	16	29	13	30	0	24	CDS	#####	#####	.	+	.	glX	Catalyzes the attachment of glutamate to tRNA(Glu) in a two-step reaction; glutamate is first ac												
13	WEOB_014	62	57	68	133	56	77	61	10	8	15	7	6	1	7	CDS	#####	#####	.	-	.	nupC	Transports nucleosides with a high affinity except guanosine and deoxyguanosine. Driven by a pr												
14	WEOB_015	109	93	81	172	80	120	75	4	6	8	5	13	2	3	CDS	#####	#####	.	+	.	mntH	Hi(+)-stimulated, highly selective, manganese uptake system												
15	WEOB_016	44	33	40	87	44	59	39	3	5	7	4	4	0	8	CDS	#####	#####	.	+	.	rppH	Accelerates the degradation of transcripts by removing pyrophosphate from the 5'-end of triphos												
16	WEOB_017	189	143	179	303	108	311	141	46	34	40	31	36	5	16	CDS	#####	#####	.	-	.	atp	Transfers the 5'-end diphosphate group to what will become the 3' terminal nucleotide of mRN												

Figure 4: readCounts tab in the excel file `Westeberhardia_RNAseq.xlsx`

The `readCounts` tab contains 373 rows – one row for each gene (plus a header in row 1). Columns `Queen1` to `Queen7` and `Larva1` to `Larva7` show the counts of reads per gene for each of the queens and larvae that we sampled. These are the counts that provide the raw measure of gene expression in RNAseq experiments. In addition, the tab contains the following columns with information for each gene:

- Feature:** What kind of genetic element is this? Here, we only look at coding sequence (CDS).
- Start:** Where in the genome does the gene start?
- Stop:** Where in the genome does the gene stop?
- Score:** How well is the gene annotated? (You can ignore this.)
- Strand:** Is the gene encoded on the + or - strand?
- Frame:** In what frame are the amino acids encoded? (You can ignore this.)
- symbol:** What is the short name of the gene?
- func:** What is the putative function of the gene?

Have a look at the first gene `WEOB_001` in the table. This NAD-binding protein has a much higher count in the `Queen` samples than in the `Larva` samples.

Questions

- Are read counts higher in queens or in larvae in gene `groL`?
- How many and which genes have a function associated with the term "shikimate"? (Use Excel's search function.)

3.2 Get an idea of the data set

Go to the second tab (`Summary`). Your task is now to calculate some quality statistics of the different samples. RNAseq experiments are complex experiments and there is always a chance that some of the samples have worse quality than others or that systematic errors disturb the analysis.

You will now use simple excel formulas to calculate three different statistics. First, you will calculate how many reads could be sequenced for each sample. For this, we can simply calculate the sum of of read counts across all genes, for each sample.

The excel formula for this follows the following logic:

```
=SUMME(ZELLE1:ZELLE100)
```

In excel, the colon (:) means that the entire range from ZELLE1 to ZELLE100 should be used to compute the formula.

Enter the following formula in cell **B2** to calculate the total number of reads generated for sample **Larva1**:

```
=SUMME(readCounts!B2:B373)
```

Let us quickly take apart the structure of the text in parentheses: `readCounts!B2:B373`.

`readCounts!` This tells excel to lookup values in the tab `readCounts`.
`B2:B373` This is the range that excel will use to calculate the sum. Here this is all the read count data for sample **Queen1**, which is found in column B in rows 2 to 373.

Once you entered the above formula in cell **B2**, you can copy it (`ctrl-c` / `ctrl-v`) to cells **C2** to **O2** to calculate total read counts for each sample.

If everything worked, the figure "Total read count per sample" should look like Figure 5:

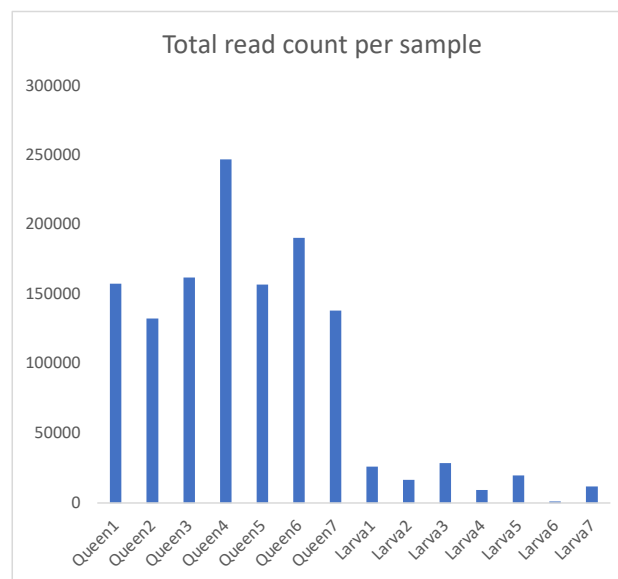


Figure 5: Total read counts per sample.

You can now easily see that we have much more data in the queens than in the larvae. Can you find a biological explanation for this pattern?

Using a similar approach, you should now calculate for each sample: (1.) the total number of expressed genes and (2.) the average number of reads per gene. The excel formulas are a bit more complicated, but otherwise the approach is the same.

Here's the formula you can use to calculate the number of expressed genes per sample.

```
=ZÄHLENWENN(BEREICH, SUCHKRITERIUM)
```

For Queen1, BEREICH would be again `readCounts!B2:B373` and SUCHKRITERIUM would be `">0"` (i.e. genes with more than zero reads.)

So

```
=ZÄHLENWENN(readCounts!B2:B373;">0")
```

Once you entered the formula for each sample, the plot "Number of expressed genes per sample" should look like Figure 6.

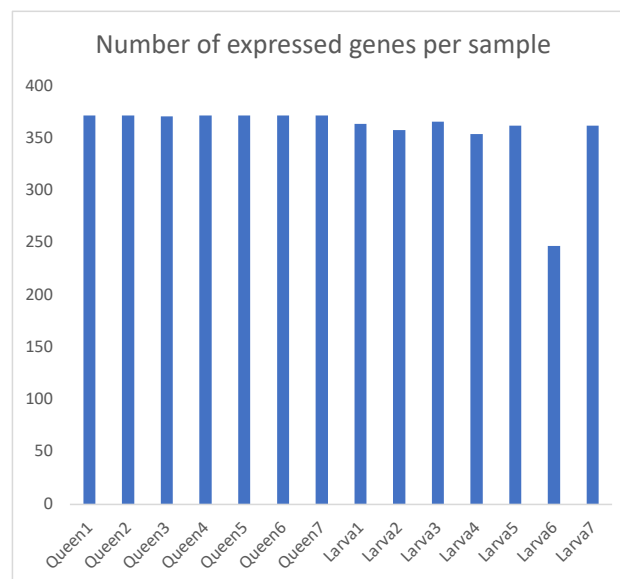


Figure 6: Number of expressed genes per Sample.

For calculating the average number of reads of all the expressed genes, we simply divide the total number of reads by the number of expressed genes. Thus, for `Queen1` the formula is

```
=B2/B3
```

Once you computed this statistic for each sample, the plot "Average reads per gene per sample" looks like Figure 7.

After analyzing these overall statistics for each sample, it becomes apparent that in sample `Larva6` the sequencing worked much worse than in the rest of the samples. In fact, it is so bad that it will likely mess up the gene expression analysis. Therefore, we should remove the sample from our dataset. That one or more samples simply do not work well is something that happens quite regularly in RNAseq experiments. So it is important to check for such poorly sequenced samples in the beginning of any analyses.

Questions

1. Which queen sample and which larva sample have the highest total read counts?
2. What could be the biological explanation for having so much more reads in adult queens compared to the larvae (Figure 1 might help finding an answer)?

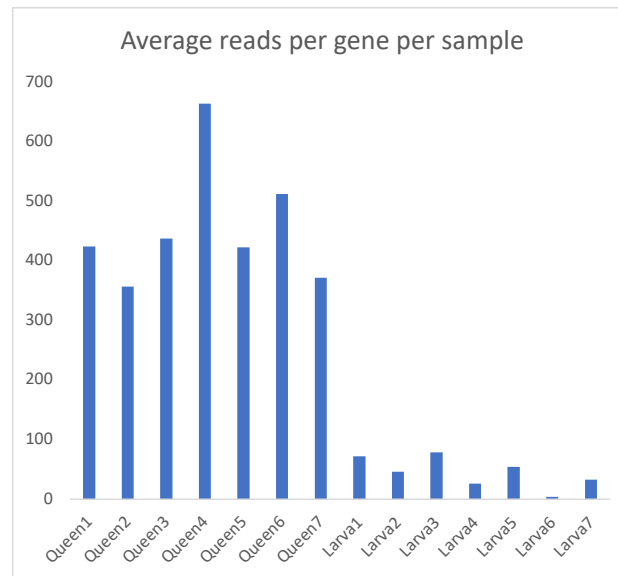


Figure 7: Total average number of reads of all expressed genes.

Tab `Filter Data` contains the same data as tab `readCounts`, but sample `Larva6` has been removed.

3.3 Transforming raw counts to comparable expression values

Now switch to the tab `FPKM`, where you will normalize the raw counts. This is important, because (as you have seen in the previous section) different samples can yield different amounts of RNAseq data. We need to account for these differences in order to compare gene expression levels across different samples. Similarly, we need to account for differences in gene length, because longer genes have a higher chance of being sequenced, simply because they produce longer mRNAs (see Figure 8).

Accounting for these two factors can be done with the so called "FPKM" normalization: "fragments per kilobase per million reads". The idea is simple: for each gene in each sample, we first divide the raw read count by the total number of reads per sample and then divide this by the length of the gene (in kilo bases). In excel, this FPKM normalization for gene `WEOB_001` in sample `Queen1` is calculated as follows:

First, calculate the length of each gene in column `P` `Gene length`.

```
=ABS(readCounts!Q2-readCounts!R2)
```

ABS in the above equation just means that the absolute value should be computed, so we don't get negative gene lengths.

Next, you can calculate the FPKM normalized expression with the following formula:

```
='Filter Data'!B2/($P2/1000)/(Summary!B$2/10^6)
```

Here's what this does: 1. Dividing the raw read counts (`'Filter Data'!B2`) by the gene length in kilobases (`($P2/1000)`) and divide the result by the total number of reads from that sample in millions

((Summary!B\$210^6)).

Note! Make sure you reference the right library size value for Larva7, as you removed one column from the tab `Filter Data!`

Figure 8 illustrates the concept of FPKM normalization.





Sample1		Sample2	
			
Gene1 15	Gene2 9	Gene1 15	Gene2 9
0.2 kbp	0.5 kbp	0.2 kbp	0.5 kbp
Total reads: 100 Million		Total reads: 30 Million	
Control for differences in total number of reads			
15/100	9/100	15/30	9/30
Control for differences in gene length			
15/(100*0.2)	9/(100*0.5)	15/(30*0.2)	9/(30*0.5)
Calculate			
0.75	0.18	2.5	1.5

Figure 8: Total average number of reads of all expressed genes.

Assume you want to compare expression of genes between two samples. Sample 1 has 100 Million sequenced RNA reads, Sample 2 has only 30 Million reads. Now to compare gene expression between these two samples, you need to account for these differences in sequencing success. This is done for each gene by dividing the raw counts per gene (e.g. 15 for Gene 1 in Sample 1) by the total number of all reads (in million, e.g. 100 for Sample 1). Next, we need to account for different gene lengths, because longer genes (e.g. gene 2) will yield more reads, simply because they are longer, but not because they are higher expressed. Correcting for this bias is done by dividing the expression values by gene length (in kilobases). In this example Gene 1 first seemed to have the same expression levels in Sample 1 and 2 (15 reads each). After correcting using FPKM we see that Gene 1 is much less expressed in Sample 1 (0.75) than in Sample 2 (2.5).

Questions

- Are FPKM values higher in queens or in larvae for gene *groL*?
(Extra: Which is the longest gene?)
(Extra: Which gene has the highest expression in Queen1?)

3.4 Data exploration

Expression averages

Now that you have calculated reliable measures of gene expression, we can start comparing the two groups of samples (larvae and queens). Open the tab `Data exploration` in the excel file.

We will begin with calculating the mean expression per gene for each group, using the MITTELWERT function

in Excel. The mean will be calculated first for all queen and then for all larva samples from the FPKM values you calculated before.

The structure of the MITTELWERT function is simply:

```
=MITTELWERT (BEREICH)
```

So, to calculate the mean expression of gene **WEOB_001** in all the queen samples, you simply enter the following in cell **B2**

```
=MITTELWERT (FPKM!B2:H2)
```

Use this to calculate the mean expression of all the genes in the queen in the column **Queen Mean**. Do the same for mean expression values in larvae in the column **Larva Mean**. After that, instead of calculating the mean, you should also calculate the median using the MEDIAN formula:

```
=MEDIAN (BEREICH)
```

Use this to calculate the median expression for all genes for both the queens and the larvae in the columns **Queen Median** and **Larva Median**.

Expression variation

It is important to also look for variation of expression in each group to understand how strongly expression levels vary within each group. For this, we use a metric called *Coefficient of Variation (CV)*, which is calculated as the standard deviation divided by the mean. This is done using the following expression in Excel:

```
=STABW (BEREICH) /MITTELWERT (BEREICH)
```

For gene **WEOB_001** the Coefficient of Variation is calculated in cell **D2** as

```
=STABWA (FPKM!B2:H2) /MITTELWERT (FPKM!B2:H2)
```

Calculate the Coefficient of Variation for all the genes and for both groups, queens and larvae in columns **Queen CV** and **Larva CV**.

If everything worked out, you should see four figures appear that look similar to those in Figure 9.

These figures summarize the expression data in dot plots with **each dot representing one gene**, and should give you a good idea of how different the expression profiles between queens and larvae actually are. In Figure 9A you see a comparison between mean expression values in both groups. The red dotted line gives the regression of this cloud of data points and the slope of this regression is almost exactly 1. So, on average genes are expressed at the same level in queens and larvae (all those dots falling on and around the red line). So, we can conclude that *Westerberhardia* in queens and in larvae do not differ fundamentally in their expression profiles.

Figure 9B shows how the Coefficients of Variation are correlated between queens and worker. Here, we see that the slope is not one, but that many genes fluctuate in their expression much stronger in larvae than in queens (i.e. the slope is larger than 1). For example, the gene with the highest CV in queens (WEOB_171) has a much higher CV in larvae ($CV_{Larva}=1.99$) than in queens ($CV_{Queen}=0.86$). A similar pattern can be observed for almost all genes: CV_{Queen} is smaller than CV_{Larva} . So, we can conclude that gene expression levels are much more stable in queens than in larvae. Figures 9C and D also confirm this pattern. While in queens (Fig. 9C) CV is constantly relatively low regardless of expression level, in queens we see that weakly expressed genes show extreme variation with CV values up to 2.0.

While it is possible that these effects are caused by an actual biological phenomenon (i.e. the bacteria regulate their expression much more accurately when they are in queens), a technical explanation is much more likely: Because the sequencing produced much less data for the larva samples, the effects of technical errors and random noise in our expression estimates are much stronger. This is also apparent more generally, as weakly expressed genes in both data sets queens and larvae tend to have higher CV values. Again, such effects are common in RNAseq experiments, which makes it so important to have a good overview of your data set at all stages of your analysis.

Questions

1. Why is it reasonable to expect a slope of 1 for the correlation of average expression between queen and larvae (Fig. 9A)?
2. What group of genes would you generally expect to fluctuate stronger in their expression? Genes involved in basic functions of the cell or genes involved in responding to the environment?

3.5 Statistical Testing

Open the tab **Statistical Testing**. Here, you will test whether there are genes that show **significant** expression differences between queens and larvae (at a significance level of $p < 0.05$). For this, we will here use a very simple t-test, as this is one of the few tests implemented in Excel.

Before we get to the statistical test, let's calculate overall averages of gene expression for visualization. First, calculate the overall (i.e. across all queens and larvae) expression mean (column **Overall Mean**), by calculating the mean of **Larva Mean** and **Queen Mean** for each gene. E.g. for calculating the overall mean for **WEOB_001** enter the following in cell **F2**:

```
=MITTELWERT (B2;D2)
```

Continue with calculating the overall means and medians for all genes.

Next, you will calculate for each gene the ratio of expression levels between queens and larvae. This measure (called **log fold change** or logFC) is a very well established measure in RNAseq. It is calculated for each gene by first dividing the expression level of one group by the expression level in another group and then calculating the \log_2 .

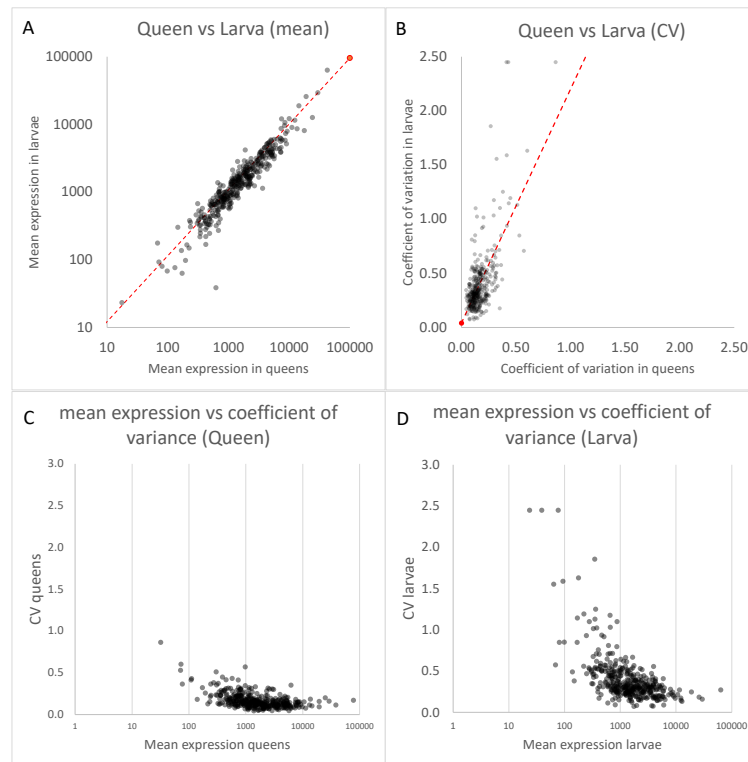


Figure 9: Overview of *Westeberhardia* expression and expression variation in queens and larvae.

(A) Correlation of expression levels for *Westeberhardia* genes in queens (y-axis) and larvae (x-axis). (B) Correlation of Coefficient of Variation for *Westeberhardia* genes in queens and larvae. (C-D) Plot showing the relationship between mean gene expression levels (x-axis) and Coefficient of Variation (y-axis) for larvae (C) and queens (D).

So:

$$\log FC = \log_2 \left(\frac{\text{mean expression group A}}{\text{mean expression group B}} \right) \quad (1)$$

This expression gives an easily understandable measure of expression differences. A gene that is expressed twice as strong in group A than in group B has a logFC of 1. A gene that is expressed half as strong in group A than in group B has a logFC of -1.

$$\log FC = \log_2 \left(\frac{1000}{500} \right) \Rightarrow \log FC = \log_2(2) \Rightarrow \log FC = 1 \quad (2)$$

$$\log FC = \log_2 \left(\frac{500}{1000} \right) \Rightarrow \log FC = \log_2(1/2) \Rightarrow \log FC = -1 \quad (3)$$

Start by calculating the fold change, i.e. the ratio of mean expression values in queens and larvae. E.g. for `WEOB_001` in cell `H2` enter

```
=B2/D2
```

From this, calculate the binary logarithm using the following formula. E.g. for `WEOB_001` in cell `I2` enter

```
=LOG(H2;2)
```

Use these formulae to calculate logFC values for each gene.

Next, we will test for statistically significant differences in *Westeberhardia* gene expression between queens and larvae using t-tests. Here, we will compare for each gene the FPKM expression values from all the larvae samples with the FPKM expression values for all the queen samples. T-tests can be run easily in Excel using the following basic formula:

```
=T.TEST('DATEN GRUPPE 1';'DATEN GRUPPE 2';SEITEN;TYP)
```

Here we will set `SEITEN=2` and `TYP=3`. Setting `SEITEN=2` tells Excel that we want to test whether the mean expression level in one group is significantly *different* (i.e. higher or lower) from the other and not only whether expression is significantly higher in one group. The `TYP=3` tells Excel that our samples are independent (as they come from different individuals) and that variance might not be the same between both groups.

Calculate the t-test statistics for `WEOB_001` in cell `J2` using

```
=T.TEST(FPKM!B2:H2;FPKM!I2:N2;2;3)
```

Calculate the t-test statistics for each gene in the column `p-value (T-test)`.

After successfully calculating these values, column `significance` will show an asteriks (*) in each row, where the t-test was significant with a p-value smaller than 0.05. In addition, the figure should look somewhat like Figure 10.

Questions

1. Are more genes overexpressed in queens or in larvae?
2. What is the highest and the lowest logFC value shown in the MA plot?

3.6 FDR correction

When everything worked out so far, you should have 112 genes that show a significant t-test result. However, this is misleading, because we conducted the same statistical test several hundred times (once for each gene), so it is very likely that just by chance some of these tests appear to be significant. After all, the 0.05 we used as a significance threshold already imply that there is a 5 % chance that the differences we observed are just a coincidence and not really a significant, biologically meaningful difference.

Say you have a set of hypotheses that you wish to test simultaneously. The first idea that might come to mind is to test each hypothesis separately, using 0.05 as the level of significance. At first glance, this doesn't seem like a bad idea. However, consider a case where you have 20 hypotheses to test, and a significance level of 0.05. What's the probability P of observing at least one significant result just due to chance?

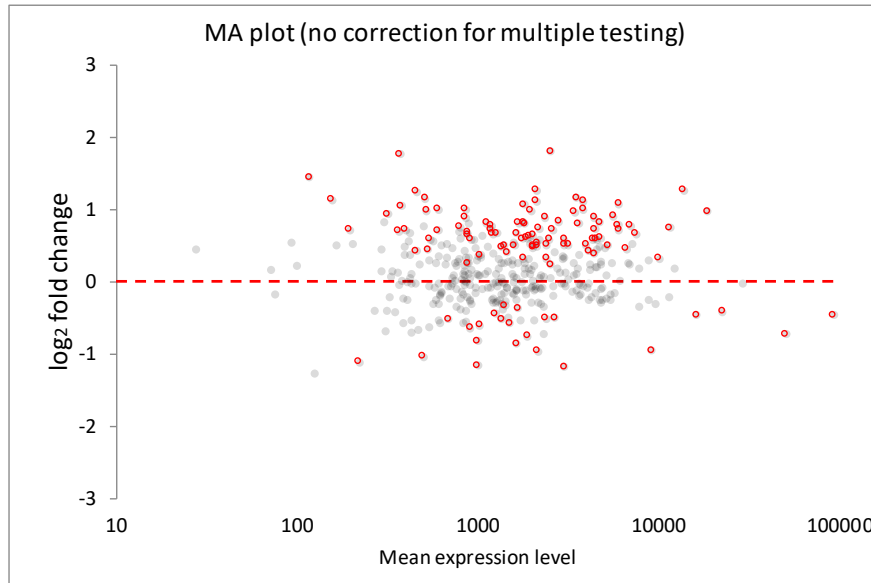


Figure 10: MA plot comparing *Westeberhardia* gene expression in queens and larvae.

This Figure shows an MA-plot, which is also a frequently used plot in RNAseq experiments. For each gene the overall average expression (x-axis) and the \log_2 fold change expression difference (as expression in larvae relative to expression in queens, y-axis). Red colored points show genes where the t-test was significant at $p < 0.05$. Shown here are only genes with a \log_2 FC between -3 and 3.

$$\begin{aligned}
 P(\text{at least one significant result}) &= 1 - P(\text{no significant results}) \\
 &= 1 - (1 - 0.05)^{20} \\
 &= 0.64
 \end{aligned}
 \tag{4}$$

So, with 20 tests being considered, we have a 64 % chance of observing at least one significant result, even if there are no actual differences between groups in the 20 tests. In RNAseq you usually test several hundreds to thousands of genes and the probability of getting a significant result simply due to chance keeps going up. Methods for dealing with multiple testing adjust for this in some way, so that the probability of observing at least one significant result due to chance remains below your desired, corrected significance level.

Here, we use an approach called "false discovery rate" correction ("FDR") and the underlying logic is that we adjust our test statistics so that we end up with 5 % of false-positive results. The math behind this is rather complex and we won't cover that in this practical. Have a look at the tab FDR correction, if you are curious as to how it's done.

The FDR-corrected test statistics are found in tab `Statistical Testing 2`. Open this tab and have a look at the updated MA-plot (Fig. 11). Less genes should be labelled as significant, because we now apply a more stringent cut-off where only p-values < 0.000401862 are considered significant. The value 0.000401862 was calculated in the `FDR correction` tab. Altogether, from the previous 112 genes only 14 remain significant after correcting for multiple testing.

Now that we have confidently identified all the significant differences, what can you say about the general pattern of differential expression? Are the significant genes overexpressed in queens or in larvae?



Figure 11: Updated MA-plot (FDR corrected)

3.7 Differentially expressed genes

In the last step, go through the remaining tabs `Significant genes`, `Upregulated in Queens`, and `Upregulated in Larvae`. All genes with significant expression differences (after correcting for multiple testing) are marked with three asterisks (***) . Can you identify any patterns in the differential expression? Are there specific functionally related genes that appear to be upregulated in queens? What about genes of the shikimate pathway?

Questions

1. How many genes are significantly overexpressed in larvae (before and after FDR correction)?
2. What are the putative functions of the genes overexpressed in queens? Can you come up with a hypothesis why these genes might be significantly overexpressed?
3. Is *groL* expression significantly different?
4. Which gene is the most significantly different? Is it the gene with the largest logFC change? If no, why not?
6. What percentage of genes are expressed significantly different between queens and larvae?

4 A brief introduction into R

If you still have time, go to <https://github.com/schraderL/bioinfo1> and follow the instructions there. It will give you a very superficial and rough introduction into the programming language R. This is just to give you an idea how large data sets are actually analyzed in a scientific setting.