Most of the slides are taken from tutorial videos by Chipster available at https://www.youtube.com/playlist?list=PLjiXAZO27elBj3KYi7AScgOxlNkNOxPc and from a book P.N. Robinson, R.M. Piro, M. Jäger: Computational Exome and Genome Analysis, CRC Press, 2019.

RNA-seq data analysis workflow

I. Quality control (QC) of raw reads
II. Preprocessing if needed
III. Alignment (= mapping) to reference genome
IV. Alignment level QC
V. Quantitation
VI. Describing the experiment with phenodata
VII. Experiment level QC
VIII. Differential expression analysis
IX. Visualization of reads and results in genomic context

Exercise 2. RNA-seq hands-on tutorial using Chipster: Drosophila dataset

In this tutorial you start with a ready-made read count table, and perform experiment level quality control. You then detect differentially expressed genes using DESeq2 and edgeR, and learn how to take confounding factors into account in differential expression analysis. Finally, you filter data based on a given column and play with different visualizations. For example, you learn how to compare gene lists using the interactive Venn diagram.

We use Drosophila data from an RNAi knock-down experiment of the pasilla splicing factor gene. The experiment is a two-group comparison with 4 untreated samples and 3 RNAi-treated samples. Unfortunately, some samples were sequenced single end and some paired end, and it is your job to correct for this in the differential expression analysis!

VII. Experiment level QC

1. **Launch Chipster**. Open new session. Select **Open example session** and **course_RNAseq_drosophila**. Inspect the session description and check the phenodata file. In particular, pay attention to the group, description and readtype columns of the phenodata.

2. Check the experiment level quality with DESeq2 Select the file **pasilla_counts.tsv** and tool **Quality control / PCA and heatmap of samples with DESeq2.** Choose **Phenodata column for coloring samples in PCA plot = treatment_description Phenodata column for the shape of samples in PCA plot = readtype_description**

-Do the groups separate along the first principal component (PC1)? How much variance does this principal component explain? How much variance is explained by PC2? Do the single end and paired end samples separate along PC2?

-According to the heatmap, do there seem to be subgroups within the treated and untreated samples which are more similar to each other?

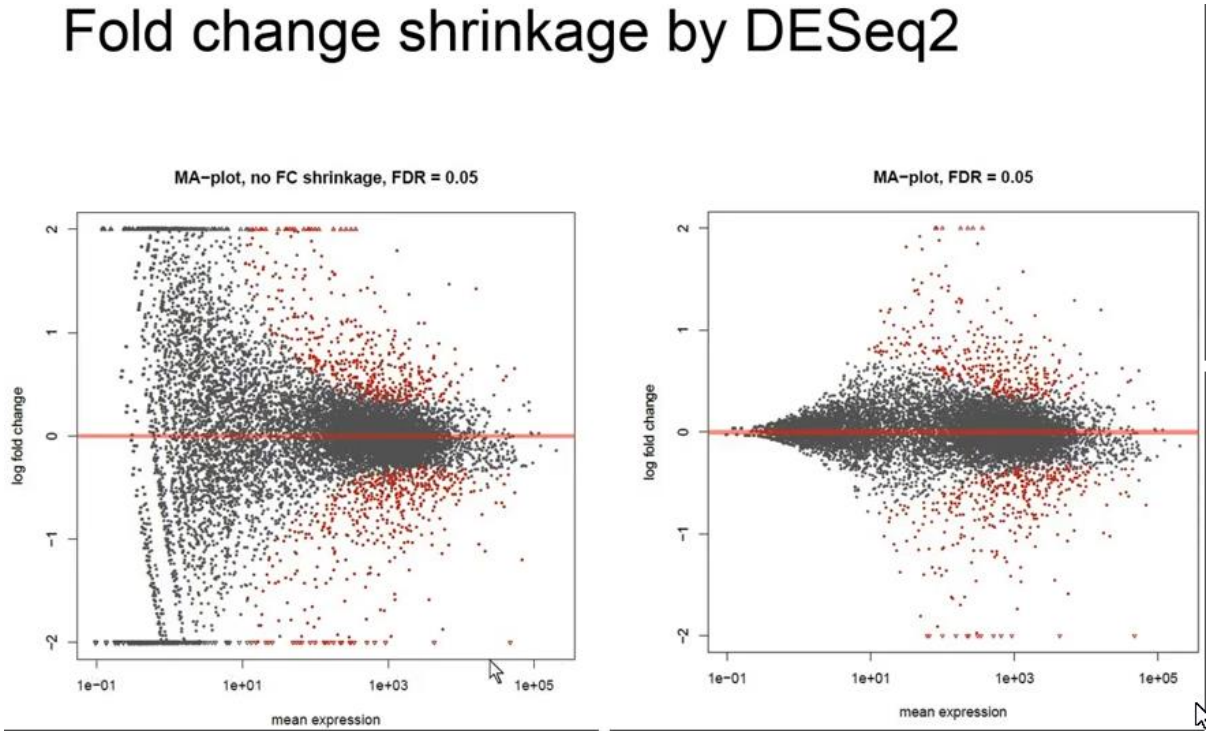
VIII. Differential expression analysis

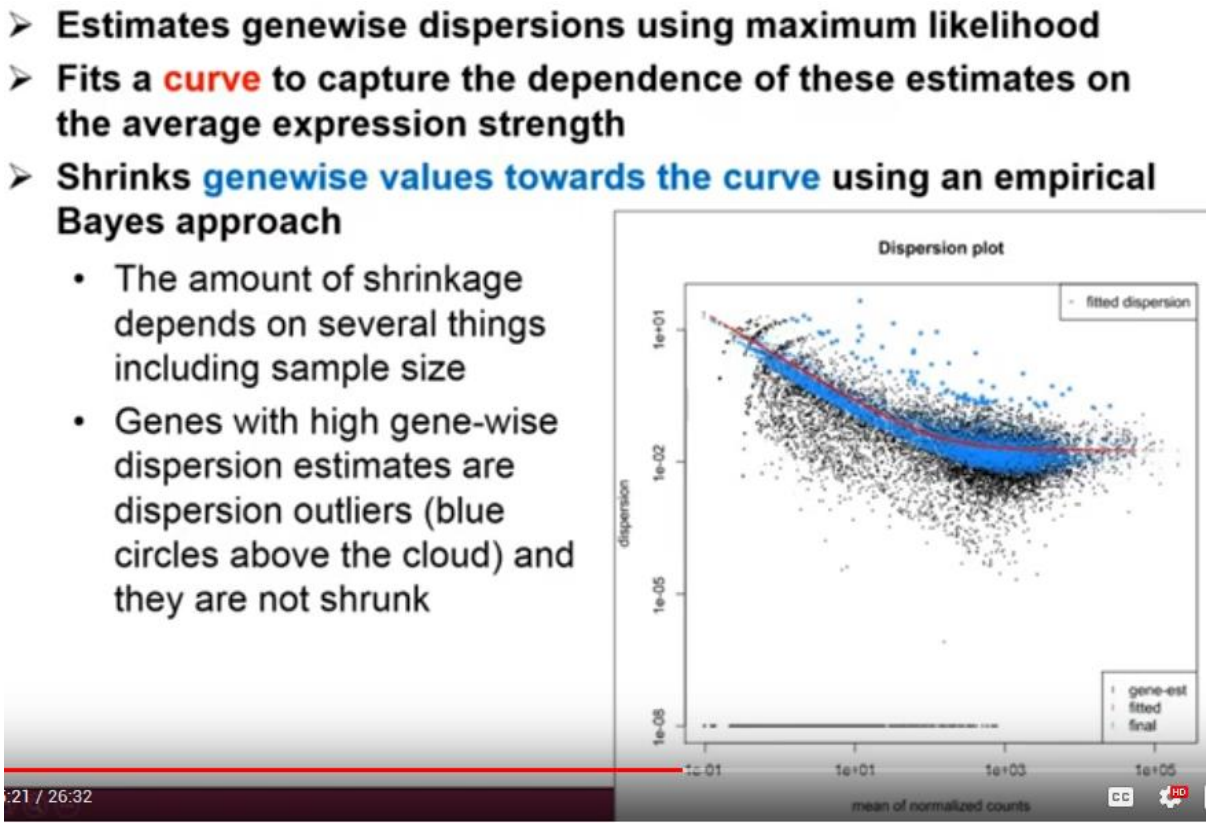
3. Analyze differential expression with edgeR. Select the file **pasilla_counts.tsv** and run the tool **RNA-seq / Differential expression using edgeR** so that you set **Filter out genes which don't have counts in at least this many samples = 3**.
-Why do we use the criteria of 3 samples in filtering? Help: "Analyze only genes which have at least 5 counts in at least this many samples. You should set this to the number of samples in your smallest experimental group."
-How many differentially expressed genes do you get?
-Look at the **MA plot**. How big fold change is required for a gene to be considered statistically significantly differentially expressed?

- **When comparing gene's expression levels between groups, it is important to know also its within-group variability**
- **Dispersion = (BCV)^2**
  - BCV = gene's biological coefficient of variation
  - E.g. if gene's expression typically differs from replicate to replicate by 20% (so BCV = 0.2), then this gene's dispersion is $0.2^2 = 0.04$
- **Note that the variability seen in counts is a sum of 2 things:**
  - Sample-to-sample variation (dispersion)
  - Uncertainty in measuring expression by counting reads

Statistical testing

- **edgeR**
  - Two group comparisons
    - Exact test for negative binomial distribution.
  - Multifactor experiments
    - Generalized linear model, likelyhood ratio test.
- **DESeq2**
  - Shrinks log fold change estimates toward zero using an empirical Bayes method
    - Shrinkage is stronger when counts are low, dispersion is high, or there are only a few samples
  - Generalized linear model, Wald test for significance
    - Shrunken estimate of log fold change is divided by its standard error and the resulting z statistic is compared to a standard normal distribution





"This dispersion plot is typical, with the final estimates shrunk from the gene-wise estimates towards the fitted estimates. Some gene-wise estimates are flagged as outliers and not shrunk towards the fitted value."

Multidimensional scaling (**MDS**) is a means of visualizing the level of similarity of individual cases of a dataset. MDS is used to translate "information about the pairwise 'distances' among a set of n objects or individuals" into a configuration of n points mapped into an abstract Cartesian space.

4. Analyze differential expression with DESeq2. Select the file **pasilla_counts.tsv** and run the tool **RNA-seq / Differential expression using DESeq2**
-How many differentially expressed genes do you get?
-Inspect **summary.txt**. How many genes had some reads mapping to them? How many of those genes had too low read counts and were hence left out of the analysis? What was the low count threshold that DESeq2 decided?
-Does the MA plot look different from the one made by edgeR? Why?
-Are the final dispersion values (blue) always higher than the original ones (black)?

5. Analyze differential expression with DESeq2 so that you take read type into account. Select the file **pasilla_counts.tsv** and the tool **RNA-seq / Differential expression using DESeq2**, and set the parameter **Column describing additional experimental factor = readtype.** Rename the resulting DE list to **de-list-deseq2-rt.tsv.**
-How many differentially expressed genes do you get now? Was it a good idea to include the readtype?
-Did DESeq2 decide to use the same low count threshold as before?

6. Compare the gene lists from exercises 3, 4 and 5 using a Venn diagram. Select the DE lists from exercises 3, 4 and 5 by keeping the ctrl/cmd key down. In the visualization panel select the method **Venn diagram**.
-How many genes do the lists have in common? In order to practice visual selections, select the genes found only by edgeR. Go to the **Selected** tab and click **Create dataset from selected**.

7. Check how many genes have changed their expression more than 4-fold up and visualize their profiles. Select the file **de-list-deseq2-rt.tsv** and run the tool **Utilities / Filter table by column value** by setting the parameters as follows:
-**Column to filter by** = **log2FoldChange**
-**Does the first column have a title** = **no**
-Cutoff = 2 (remember that 2 in log2 scale means 4 in linear scale)
-Filtering criteria = larger-than
-How many genes have a fold change higher than 4? Visualize them as an interactive **expression profile**.

//----------
//----------
Let us explain expressions used in the headers of columns of tsv-files:
- **logFC**: For quantities A and B, the **fold change** (**FC**) of B with respect to A is B/A.
//----------
- **logCPM**: **Counts per million** (**CPM**) mapped reads are counts scaled by the number of fragments you sequenced (N) times one million:

$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

i.e., the accounted factor is a sequencing depth.
//----------

//----------

- **p-value**:

Hráč provedl 300 hodů hrací kostkou, aby vyzkoušel, zda všechna čísla padají se stejnou pravděpodobností. Chce testovat na hladině významnosti α = 0,05. Četnosti jednotlivých výsledků jsou 58, 46, 39, 61, 35 a 61. Test dobré shody vykonejme pomocí softwaru R zadáním příkazů:

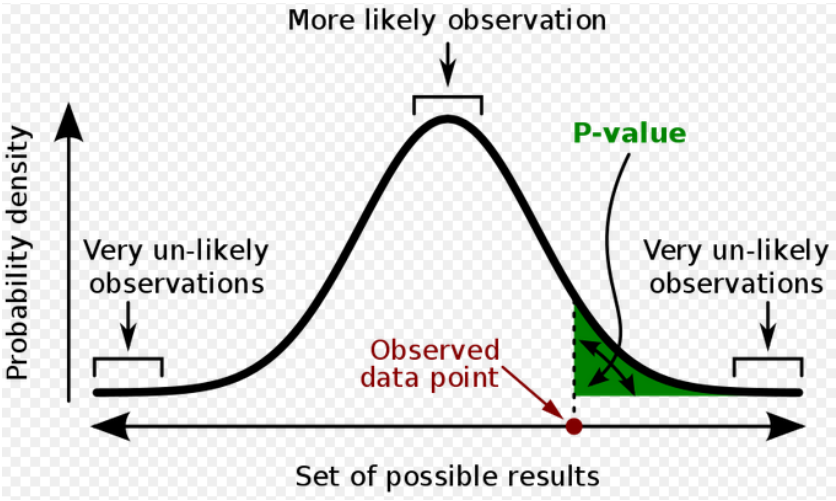
```
kostka <- c(58, 46, 39, 61, 35, 61)
chisq.test(kostka)
```

Výstup je potom:

```
        Chi-squared test for given probabilities
data:  kostka
X-squared = 13.36, df = 5, p-value = 0.02023
```

Poněvadž vypočítaná *p*-hodnota 0,02023 je ***menší*** než zvolená hodnota = 0,05, na hladině 0,05 *zamítáme* nulovou hypotézu stejné pravděpodobnosti všech výsledků a na základě naměřených dat máme za to, že hrací kostka je „cinknutá“.

p-hodnota (**p-value**) je číselná hodnota používaná při statistickém testování hypotéz. Testujeme-li na daném statistickém souboru nulovou hypotézu **H_0** na hladině významnnosti **α** pomocí testové statistiky **T**, p-hodnota je nejmenší hladina významnosti, při které ještě zamítneme **H_0**. V praxi se **p-hodnota** používá tak, že si předem stanovíme hladinu významnnosti **α**, poté spočítáme pomocí statistického programu **p-hodnotu** a porovnáme ji s **α**. Vyjde-li **p-hodnota** menší než **α**, nulovou hypotézu **H_0** zamítneme, zatímco v opačném případě prohlásíme, že na základě zkoumaných dat ji s použitím daného testu zamítnout nelze. Čím menší tedy je **p-hodnota**, tím se nulová hypotéza jeví za jinak stejných podmínek nevěrohodnější. By convention, is commonly set to 0.05, 0.01, 0.005, or 0.001.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

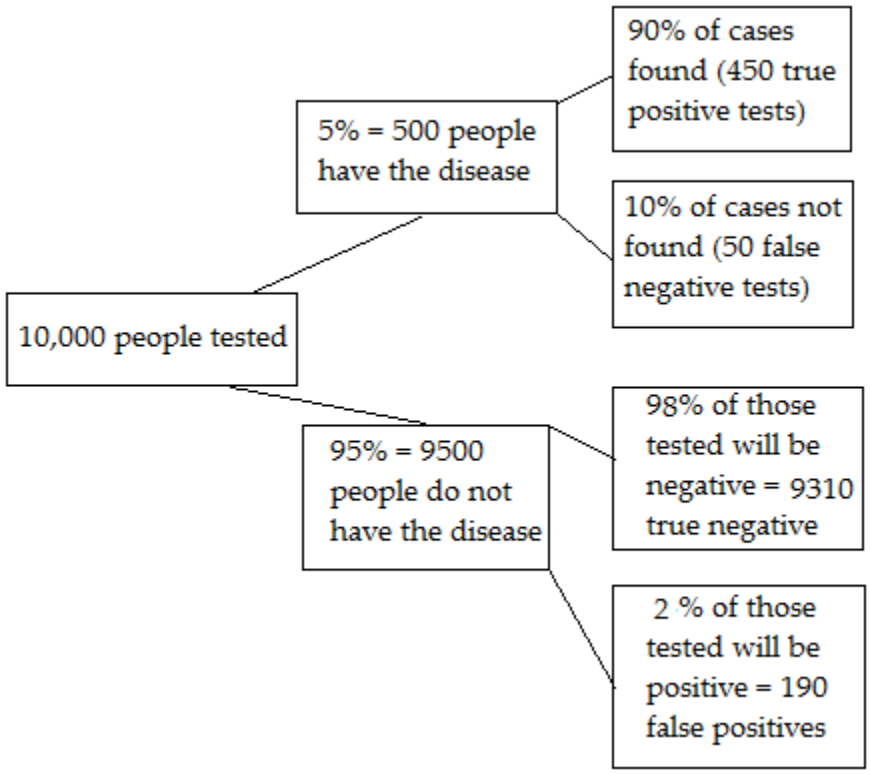
In inferential statistics, the **null hypothesis** is a general statement or default position that there is nothing new happening, like there is no association among groups, or no relationship between two measured phenomena.

//----------

A small p-value means there is a small chance that the observed results occurred by chance. FDR (False Discovery Rate) is a modification of the p-value.

//----------

- **FDR** (**padj**): The image below shows a medical test that accurately identifies 90% of real diseases/cases. The **false discovery rate** (FDR) is the ratio of the number of false positive results to the number of total positive test results. Out of 10,000 people given the test, there are 450 true positive results (box at top right) and 190 false positive results (box at bottom right) for a total of 640 positive results. Of these results, 190/(450+190) = 190/640 = 0.296 are false positives so the false discovery rate is 30%.



If you repeat a test enough times, you will *always* get a number of false positives. One of the goals of multiple testing is to control the FDR: the proportion of these erroneous results. For example, you might decide that an FDR rate of more than 5% is unacceptable. Note though, that although 5% *sounds* reasonable, if you're doing a lot of tests (especially common in medical research), you'll also get a large number of false positives; for 1000 tests, you could expect to get 50 false positives by chance alone. This is called the multiple testing problem, and **the FDR approach is one way to control for the number of false positives**.

The FDR approach adjusts the p-value for a series of tests. A p-value gives you the probability of a false positive on a single test; If you're running a large number of tests from small samples (which are common in fields like genomics and protoemics), you should use q-values instead.

- **A p-value of 5% means that 5% of all tests will result in false positives.**
- **A q-value of 5% means that 5% of *significant* results will be false positives.**

The procedure to control the FDR, using q-values, is called the Benjamini-Hochberg procedure, named after Benjamini and Hochberg (1995), who first described it.

Multiple testing correction

- **We tests thousands of genes, so it is possible that some genes get good p-values just by chance**
- **To control this problem of false positives, p-values need to be corrected for multiple testing**
- **Several methods are available, the most popular one is the Benjamini-Hochberg correction (BH)**
  - largest p-value is not corrected
  - second largest p = (p *n)/ (n-1)
  - third largest p = (p * n)/(n-2)
  - …
  - smallest p = (p * n)/(n- n+1) = p * n
- **The adjusted p-value is FDR (false discovery rate)**

DESeq2 result table

- baseMean = mean of counts (divided by size factors) taken over all samples
- **log2FoldChange = log2 of the ratio meanB/meanA**
- lfcSE = standard error of log2 fold change
- stat = Wald statistic
- pvalue = raw p-value
- **padj = Benjamini-Hochberg adjusted p-value**

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
FBgn0026562	47282.42	-2.4	0.08	-30.26	4.159e-201	3.309e-197
FBgn0039155	924.27	-4.46	0.16	-27.04	4.476e-161	1.781e-157
FBgn0029167	4287.44	-2.21	0.08	-26.75	1.107e-157	2.937e-154
FBgn0035085	654.94	-2.5	0.11	-22.08	5.278e-108	1.050e-104
FBgn0034736	231.7	-3.29	0.18	-18.28	1.261e-74	2.006e-71
FBgn0000071	359.53	2.6	0.14	17.98	2.741e-72	3.635e-69
FBgn0034434	153.84	-3.69	0.21	-17.26	9.008e-67	1.024e-63
FBgn0039827	342.77	-3.83	0.23	-16.54	1.742e-61	1.733e-58
FBgn0029896	513.08	-2.34	0.14	-16.29	1.168e-59	1.033e-56
FBgn0052407	220.26	-2.2	0.15	-14.99	8.597e-51	6.841e-48
FBgn0037754	299.03	-2.23	0.15	-14.94	1.916e-50	1.386e-47

So, for example:

scenario one - you wish to pull out significantly differentially expressed genes for some sort of ontology or enrichment analysis. Your primary goal is to identify biological processes, pathways or other ontology categories. So you may be fairly relaxed with your choice of cutoff in order to be sure to have sufficient genes to get a reasonably robust enrichment result. So, you may pick an FDR of < 0.05, or even 0.1 if you need to pad out your gene lists.

scenario two - you are trying to pick out genes as candidates for bio-assay development, so you'd like to find the least number necessary to characterize your system, and you need to be stringent about your risk of false positives (wasted money down the road if those fail to validate for your assay). So you now pick a more stringent FDR, maybe even going to < 0.01 if that gives you enough to continue with. Perhaps you simultaneously throw in a fold change cutoff as well, so only take genes with both an FDR < 0.05 and a log2 FC > 2 (picking only highly significant high expressors).

So, as with any choice of statistical criteria, you pick a cutoff that makes sense in light of your questions(s) and your system.
//----------
//----------

VIII. Visualization of reads and results in genomic context

8. Visualize the read counts of the gene which has the smallest padj-value. Select the file **de-list-deseq2-rt.tsv** and the tool **Utilities / Plot normalized counts for a gene**. In the parameters, indicate the **gene name** (FBgn0039155) and set **Show names in plot = yes.**
-Do the groups differ clearly in the read counts?