Differential gene expression analysis

04-10-2022

Dr. Janou Roubroeks

Radboud University

What is differential gene expression analysis?

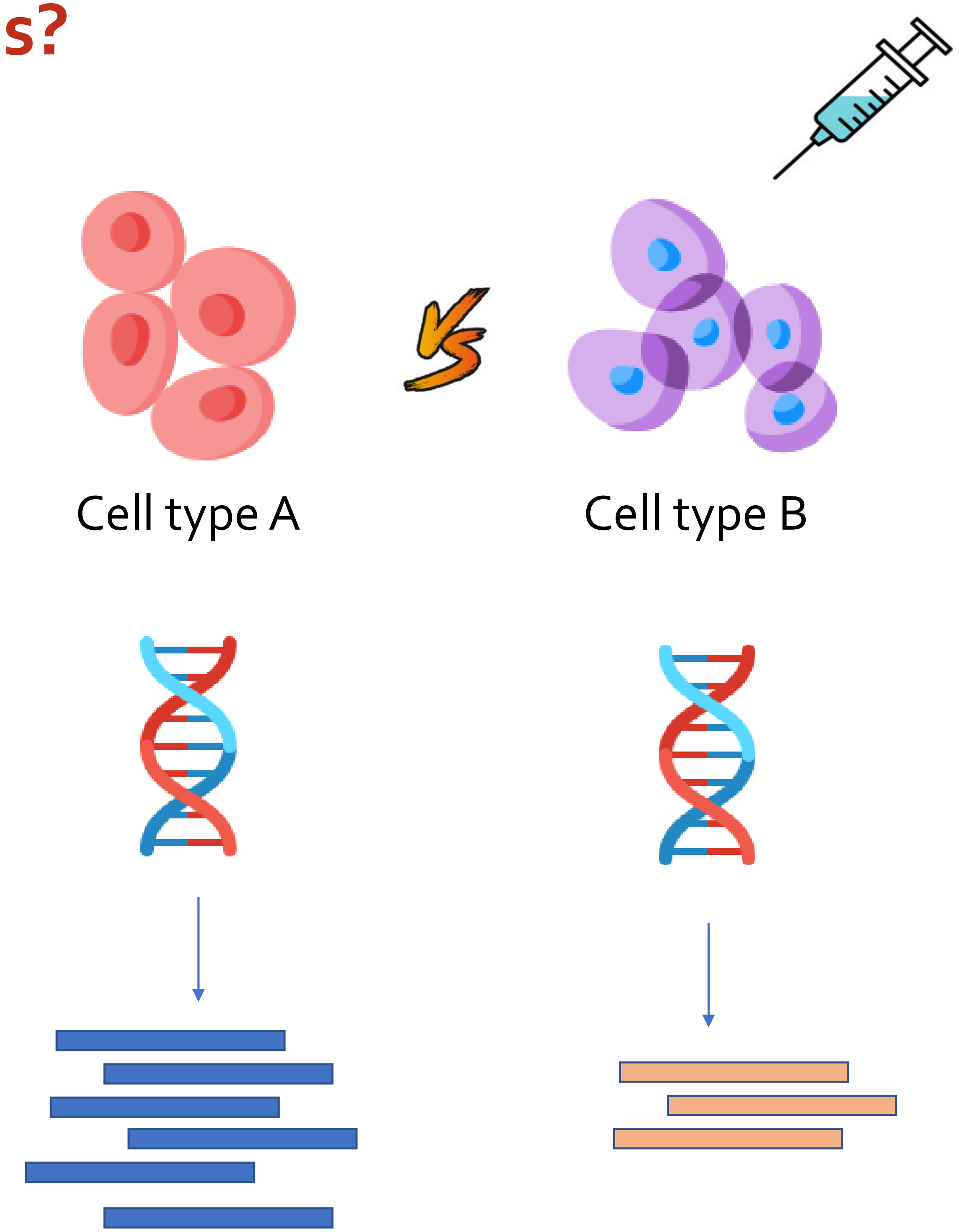
Goal:

Find genes that are *significantly* higher or lower epxressed between groups of samples

- Quantify the proportion of change
- Assign a p-value to each comparison

We use the raw count values as starting point:

	Gene 1	Gene 2
Sample A	4	4
Sample B	4	2



How do we quantify changes?

The difference in gene expression is usually expressed as a **fold change**

A fold change (FC) describes the factor of change between two quantities:

$$FC = \frac{A}{B}$$

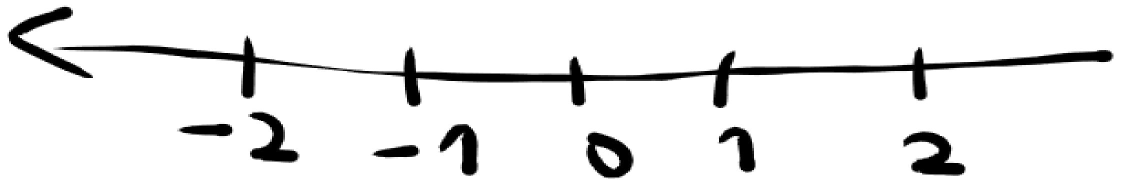
	Gene 1	Gene 2
Sample A	4	4
Sample B	4	2
FC (A/B)	1	2
Log_2 FC (A/B)	0	1

Gene 2 is **2-fold** upregulated in Sample A compared to sample B

Note: Watch out for FC **directionality**

FC (A/B) != FC (B/A)

FC is typically expressed on a Log2 scale



Symmetrical scale, centered at 0 (no change)

Challenge 1: Library size variation

Sequencing depth can vary between samples

- Sample A has double the read depth of Sample B
- The variation we see is technical, not biological

Raw counts:

	Gene 1	Gene 2	Total reads
Sample A	20	40	60
Sample B	10	20	30

Would it be fair to compare Sample A and B directly?

- No, we must adjust for different library sizes between samples first
- We need to **normalise** our data

Example: TPM (transcripts per million) normalisation

TPM (Transcripts per million)

I. RPK (reads per kilobase) -> Divide each gene by its size in kb

II. Scaling factor -> Sum up RPK per sample and divide by 10^6

III. TPM -> RPK / scaling factor (per sample)

TPM adjusts for gene length and library size.

TPM allows between-sample comparisons of proportional gene expression (total TPM counts are the same in each sample)

- Suitable for exploratory data analysis
- **Not** suited for DEG analysis

Raw counts:

	Gene 1	Gene 2	Total reads
Sample A	20	40	60
Sample B	10	20	30

TPM normalisation

	Gene 1 (10kb)	Gene 2 (20kb)	Total RPK (scaling factor)	TPM Gene 1	TPM Gene 2
Sample A	20/10 = 2	40/20 = 2	4 / 10	5	5
Sample B	10/10 = 1	20/20 = 1	2 / 10	5	5

Challenge 2: Library composition bias

The number of reads in a sequencing run is finite

Example:

Assume a gene is expressed in tissue **A** but not in tissue **B**?

- Sample A and Sample B have the same number of total reads
- Gene 3 is **not** transcribed in Sample B, but highly expressed in sample A
- The 60 **leftover** reads that would have been assigned to Gene 3 in Sample B are distributed to Gene 1 and Gene2

Real counts:

	Gene 1	Gene 2	Gene 3
Sample A	10	10	40
Sample B	10	10	0

Observed counts:

	Gene 1	Gene 2	Gene 3	Total
Sample A	10	10	40	60
Sample B	10 + 20	10 + 20	0	60

Gene 1 and Gene 2 appear overexpressed in Sample B

- This called a **composition bias**
- Library size normalisation is not enough
- We need to account for these genes during normalisation

→ TPM **does not** account for composition bias

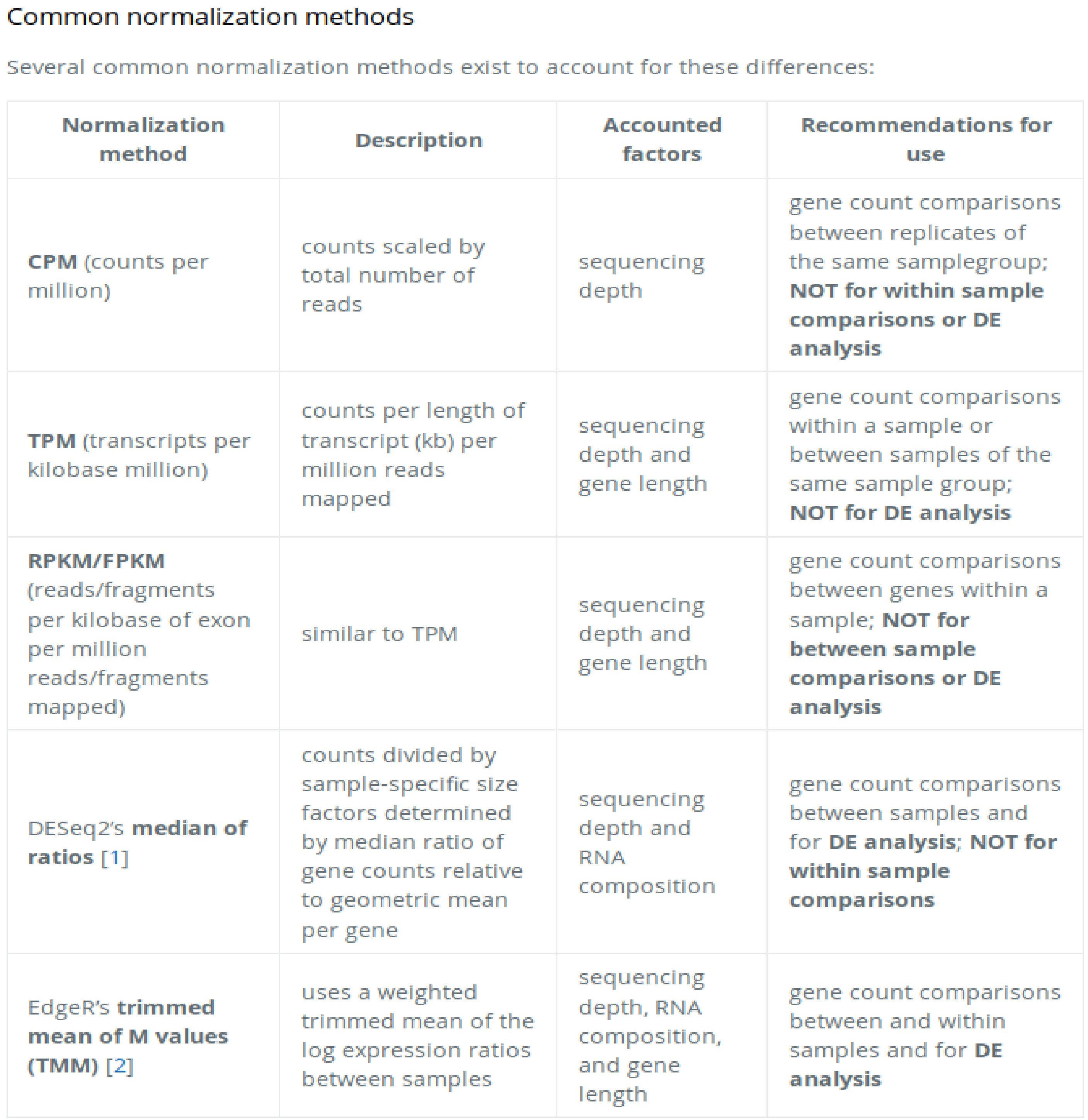
Normalisation for DE analysis

To test differential expression we use **median of ratios** or **TMM**

- Between sample normalisation
- Accounts for sequencing depth & library composition

DE analysis tools incorporate normalisation in their pipeline, like **DESeq2**

- Incorporates information from biological replicates to control variance
- The more replicates, the better!



Common normalization methods

Several common normalization methods exist to account for these differences:

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

To sum it up:

$$True\ gene\ expression \approx observed\ gene\ expression - technical\ noise$$

Correct preprocessing is needed to remove noise and enable fair comparisons!

- Between-sample normalisation
- Statistical modeling & hypothesis testing
- Multiple testing correction (e.g. Bonferroni, FDR)

Your data:

- 6 conditions, with each 3 biological replicates (18 samples)

Conditions:	DMSO	TA 100 nM	TA 1 µM	RU 1 µM	RU 1 µM + TA 100 nM	RU 1 µM + TA 1 µM
Your samples	3	3	3	3	3	3
Technical replicates:	3	3	3	3	3	3

Merged

- A Seq2science pipeline has been run that merges the technical replicates for each sample
    - You can find the multiQC report for this run here: https://mbdata.science.ru.nl/ghe_2022/day2/

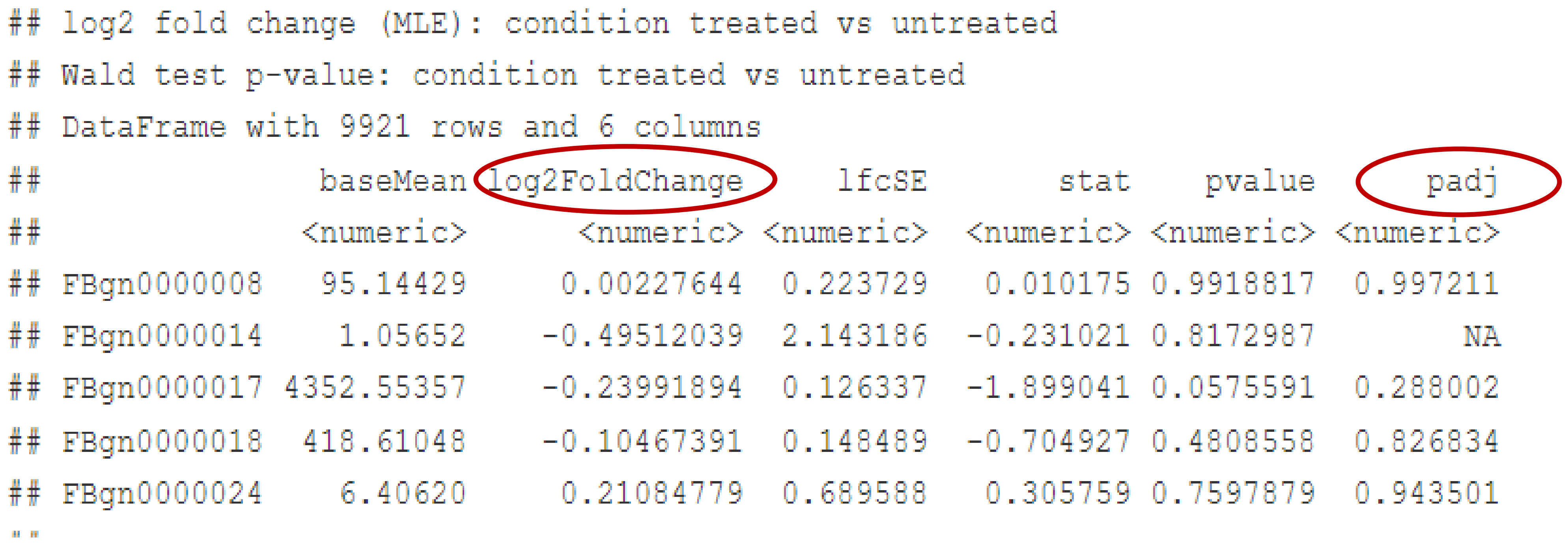
- We will run DESeq2 on this data to find genes that are differentially expressed between different conditions

How do we run DESeq2?

We make use of an R script (run on the mbscourse server) that takes as input:

- A contrast you want to run (e.g. DMSO vs. TA 100nM)
- A samples file that tells DESeq2 which samples belong to which groups (.tsv)
- Count table (always use raw counts)
- A path to a directory where the results can be stored

DESeq2 output example:



```
## log2 fold change (MLE): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 9921 rows and 6 columns
##               baseMean log2FoldChange     lfcSE      stat    pvalue      padj
##              <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## FBgn0000008   95.14429     0.00227644  0.223729  0.010175 0.9918817  0.997211
## FBgn0000014    1.05652    -0.49512039  2.143186 -0.231021 0.8172987        NA
## FBgn0000017 4352.55357    -0.23991894  0.126337 -1.899041 0.0575591  0.288002
## FBgn0000018  418.61048    -0.10467391  0.148489 -0.704927 0.4808558  0.826834
## FBgn0000024    6.40620     0.21084779  0.689588  0.305759 0.7597879  0.943501
##
```

Statistical testing assumes $Log2FoldChange = 0$ (No change in gene expression)

DE genes can be extracted by applying filtering conditions to columns

Use multiple-testing corrected *p*-value for more accuracy (recommended)

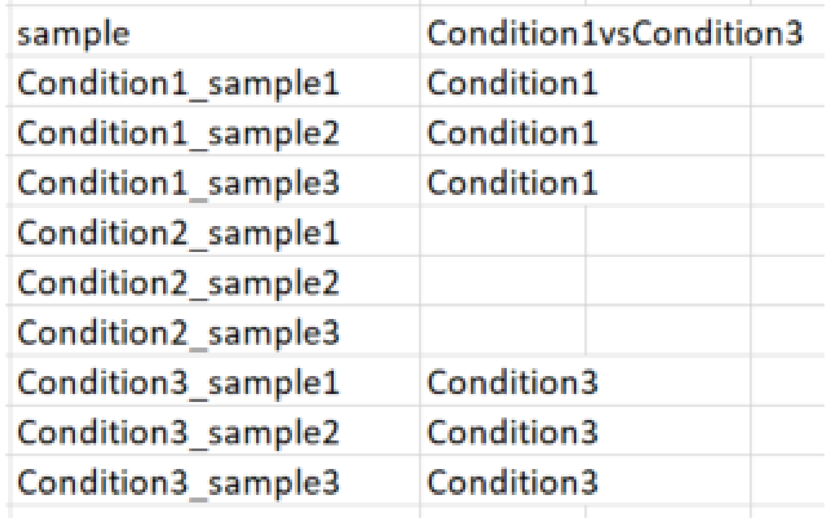
How do we run DESeq2?

1. Navigate to your group's folder on the server, and create a new folder for DESeq2

2. Copy the `samples_DESeq2.tsv` file in `/scratch/ghe_2022/analysis/all_samples/` to your own folder

3. Add a column with the contrast (comparison you want to run)

   - This tells DESeq2 which samples to use in your comparison



sample	Condition1vsCondition3
Condition1_sample1	Condition1
Condition1_sample2	Condition1
Condition1_sample3	Condition1
Condition2_sample1	
Condition2_sample2	
Condition2_sample3	
Condition3_sample1	Condition3
Condition3_sample2	Condition3
Condition3_sample3	Condition3

4. You can then run the DESeq2 script by running

```
/scratch/ghe_2022/scripts/deseq2.R --help
```

   - Using the --help flag will show you how to correctly pass the input files to your script!

   - The counts file you need is located in: `/scratch/ghe_2022/analysis/all_samples/results/counts/`

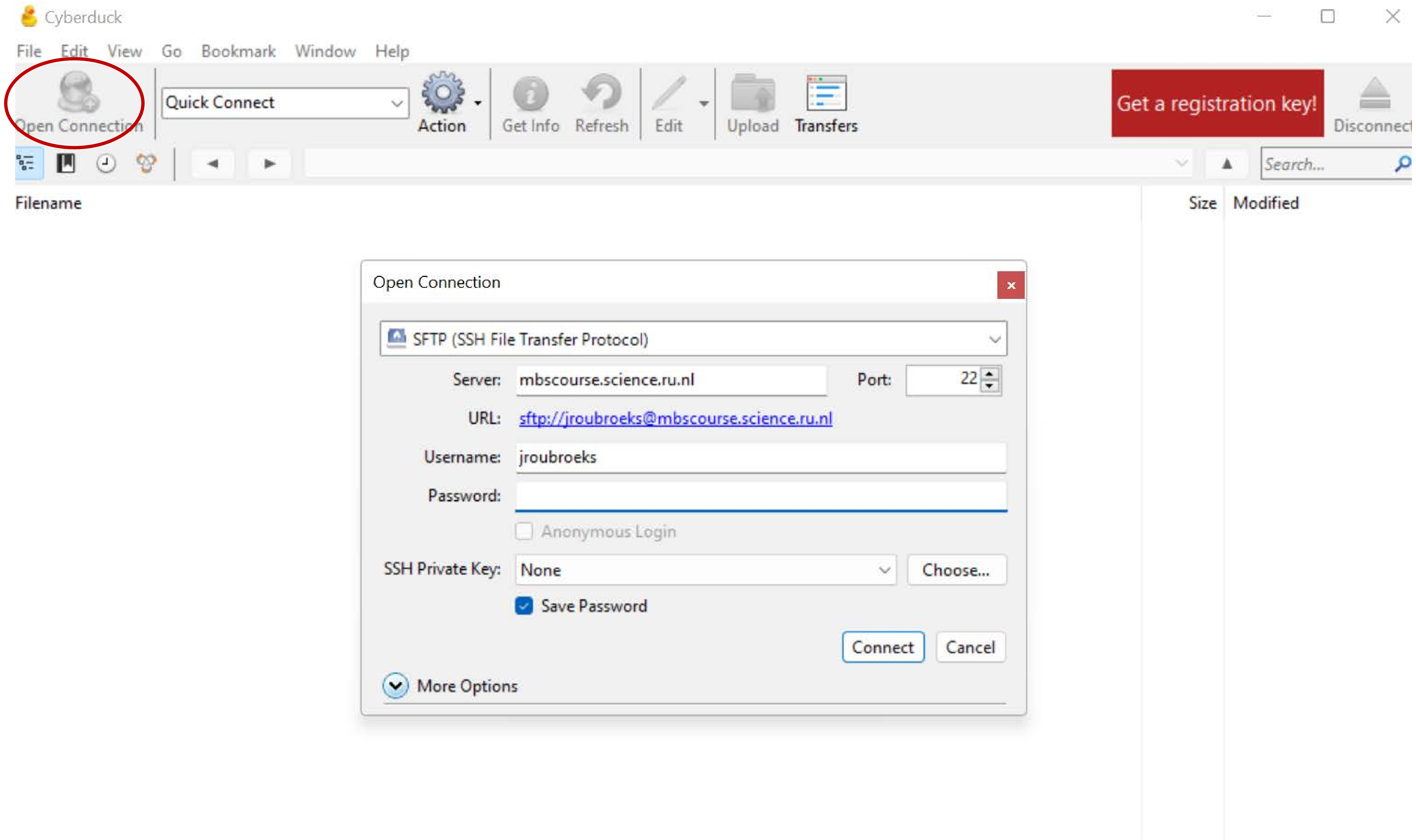
     (this folder also contains the TPM files you can use later)

   - Make sure to save your output to your own directory!

https://cyberduck.io/download/

https://mbdata.science.ru.nl/ghe_2022/day2/

Cyberduck setup:



Open Connection

SFTP (SSH File Transfer Protocol)

Server: mbscourse.science.ru.nl Port: 22

URL: sftp://jroubroeks@mbscourse.science.ru.nl

Username: jroubroeks

Password:

□ Anonymous Login

SSH Private Key: None

☑ Save Password