RNA-Seq and DE analysis

Bioinformatics Core Facility, Centre for Molecular Medicine, CEITEC Masaryk University

[1]CEITEC-Central European Institute of Technology, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic

[2]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic

Václav Hejret[1,2] vaclav.hejret@ceitec.muni.cz

Acknowledgement

- We request co-authorship (and will assist with manuscript preparation) if we have developed novel tools, algorithms, or pipelines, participated in experiment design planning, or have contributed to a biological question addressed in a manuscript. We ask that, at a minimum, you acknowledge us in a publication to which we have contributed routine analysis without further support and discussions, data management or conversion, or data submission services.
- Core Facility Bioinformatics of CEITEC Masaryk University is gratefully acknowledged for the obtaining of the scientific data presented in this paper.

Description

- The main goal of the experiment is to perform Differential Expression analysis.
- In general, we have 10 human samples. We have 2 conditions (WT, Mut), those conditions are sequenced in five replicates, which is sufficient enough to obtain statistical significance.
- Conditions to compare are:

WT_vs_Mut

- Samples were sequenced using Lexogen Quantseq FWD kit, therefore a lot of reads in UTR regions are expected as well as huge 3' sequencing bias. These data can only be used for differential expression analysis.
- Samples contained 6bp long UMIs used for detection of PCR duplicates in library

Agreed task(s)

1. General samples QC.
2. RNA-Seq analysis.
3. DE analysis.

Samples

- Brief sample description and number of raw and preprocessed reads are summarized in Table 1.

Table 1: Sample description - condition assignment and number of raw and preprocessed reads.

Original name	Sample	Condition	raw reads	preprocessed reads
Mut-1	Mut_rep1	Mut	10,729,594	10,721,500
Mut-2	Mut_rep2	Mut	10,190,321	10,182,322
Mut-3	Mut_rep3	Mut	11,309,496	11,299,605
Mut-4	Mut_rep4	Mut	10,304,741	10,296,211
Mut-5	Mut_rep5	Mut	11,986,309	11,969,557
WT-1	WT_rep1	WT	10,731,040	10,722,069
WT-2	WT_rep2	WT	10,088,717	10,079,613
WT-3	WT_rep3	WT	10,511,508	10,504,764
WT-4	WT_rep4	WT	11,060,603	11,049,278
WT-5	WT_rep5	WT	12,926,867	12,912,520

Analysis

- Raw reads were quality checked (**FastQC, MultiQC, minion, swan**), preprocessed (**Trimmomatic, FastQC, MultiQC**) and mapped (**STAR, Samtools, MultiQC**) to the reference genome with gene annotation (**genome version: Ensembl GRCh38, gene annotation: Ensembl v94**)
- Mapped reads were counted and summarized to genes (**featureCounts**)
  - Only uniquely mapped and uniquely assigned reads were counted
  - “Forward” counts were considered because of the both library prep. kit strand-specificity
- Following checks were performed to estimate the overall sample quality
  - rRNA content estimate (**fastq_screen**)
  - Read duplication rate (**dupRadar, Picard tools**)
  - Sequenced (targeted) regions (**RSeQC, Picard tools**)
  - 5’/3’ coverage bias (**Picard tools**)
  - Expressed gene biotypes (**featureCounts**)
  - Library strandness (**RSeQC**)
  - Other quality checks are available upon request
- Complete settings, used commands, tool versions and methodology part for a publication can be provided upon request

Results

General samples QC

- The sequencing depth varies from ~10M of reads for sample WT_rep2 to ~13M of reads for sample WT_rep5.
- Initial quality check for analysis did not show presence of adapters (<0%), no adapter trimming was performed
- Other initial quality checks look OK

Alignment and splices

- The single end reads were aligned to human reference genome with the help of the gene annotation (GRCh38; Ensembl 94) using STAR aligner.
- The number and percentage of uniquely- and multi-mapped reads are [redacted] samples ~79% of uniquely mapped and we are getting approximately the same values across all samples, same applies for number of mapped reads, which varies between ~8M and ~10M (Table 2, Figure 1).

Table 2: Number and percentage of uniquely- and multi-mapped reads

Original name	Sample	uniquely mapped	uniquely mapped (%)	multimapped	multimapped (%)
Mut-1	Mut_rep1	8,564,030	79.88	1,468,222	13.69
Mut-2	Mut_rep2	8,128,507	79.83	1,361,834	13.37
Mut-3	Mut_rep3	8,933,108	79.06	1,622,784	14.36
Mut-4	Mut_rep4	8,186,419	79.51	1,462,967	14.21
Mut-5	Mut_rep5	9,313,518	77.81	1,661,406	13.88
WT-1	WT_rep1	8,412,480	78.46	1,643,464	15.33
WT-2	WT_rep2	7,924,970	78.62	1,485,485	14.74
WT-3	WT_rep3	8,443,782	80.38	1,503,497	14.31
WT-4	WT_rep4	8,619,720	78.01	1,590,757	14.40
WT-5	WT_rep5	10,156,024	78.65	1,868,552	14.47

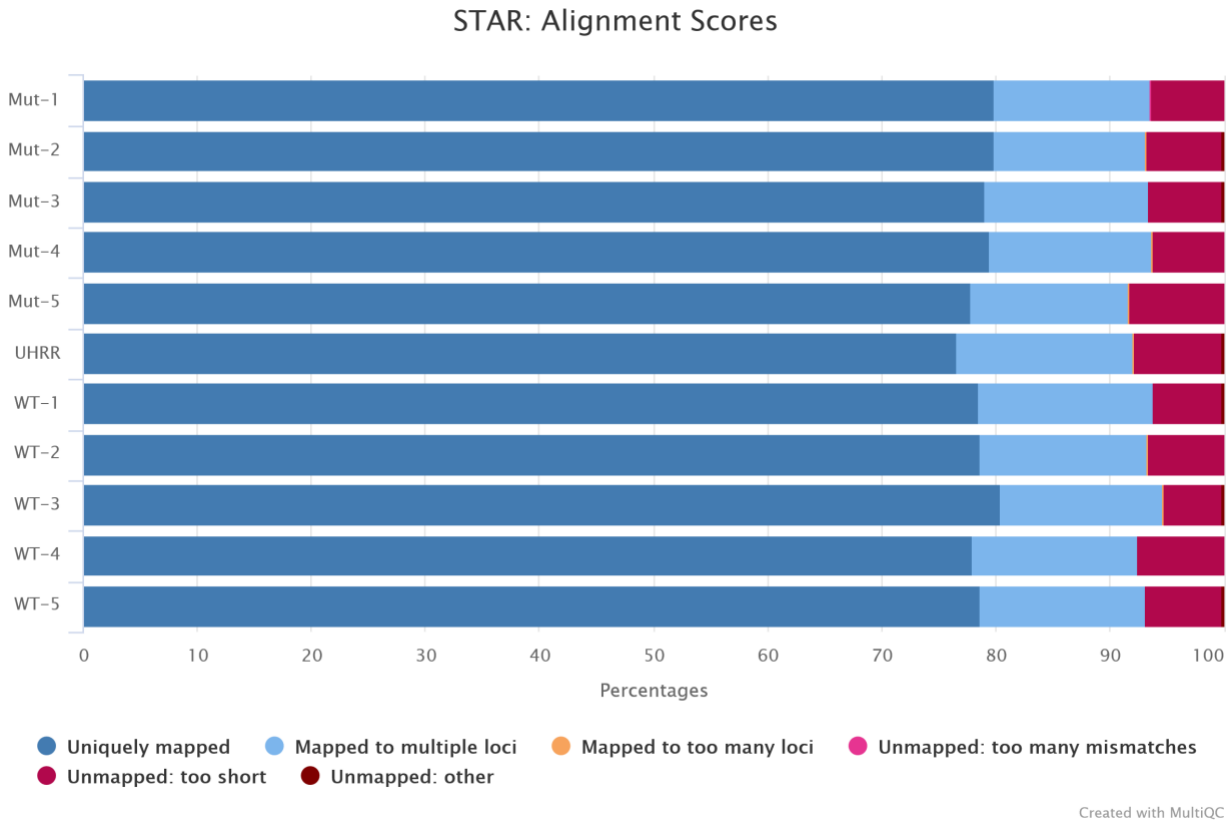

Figure 1: Percentages and distribution of different mapping types. For more details please see the STAR aligner manual.

rRNA contamination

- Samples had a little bit higher percentage of rRNA contamination up to 6%
  - For polyA selection, we would expect ~ 1%, for rRNA depletion could be up to 10%
  - This can be due to less effective polyA selection procedure, but overall we still have a lot of usable reads and information to perform Differential expression analysis
  - The percentages are summarized in Figure 2

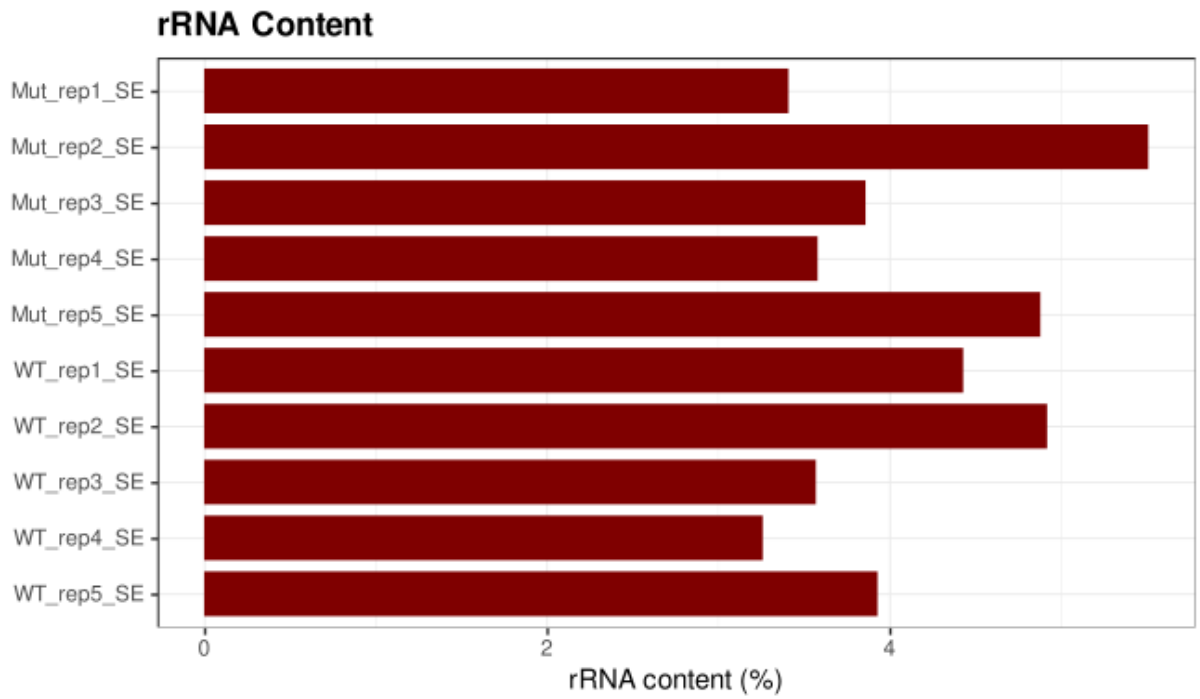

Figure 2: rRNA estimate.

Mapped regions

- Evaluation of the mapped genomic regions is very comparable
- All samples had approximately almost no content of intronic reads ~12% counted over bases as well as ~12% counted over reads which is expected for polyA selection library protocols
- All samples had majority of reads [redacted] counted over bases as well as [redacted] counted over reads mapped within coding regions (Figure 3 and 4).
- All samples are good quality and can be safely used for Differential expression analysis.

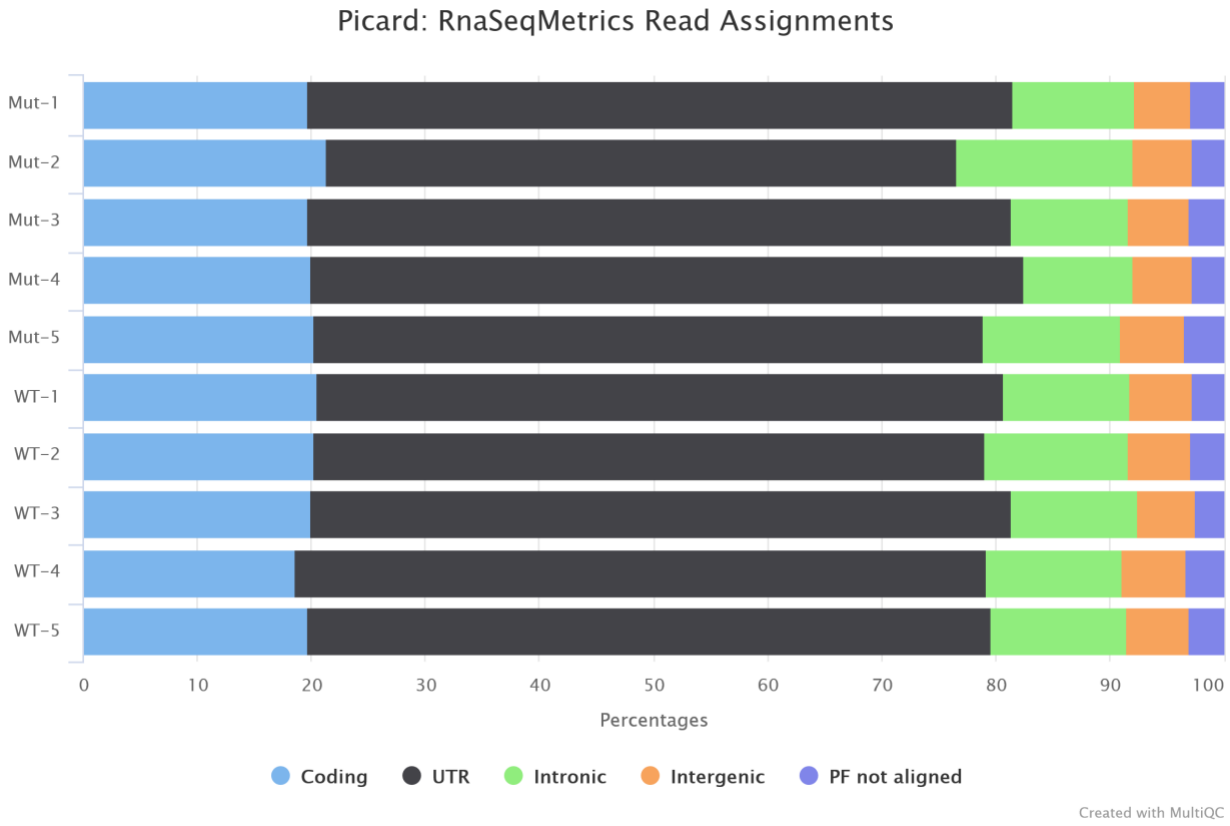

Figure 3: Mapped locations (Picard; in bases).

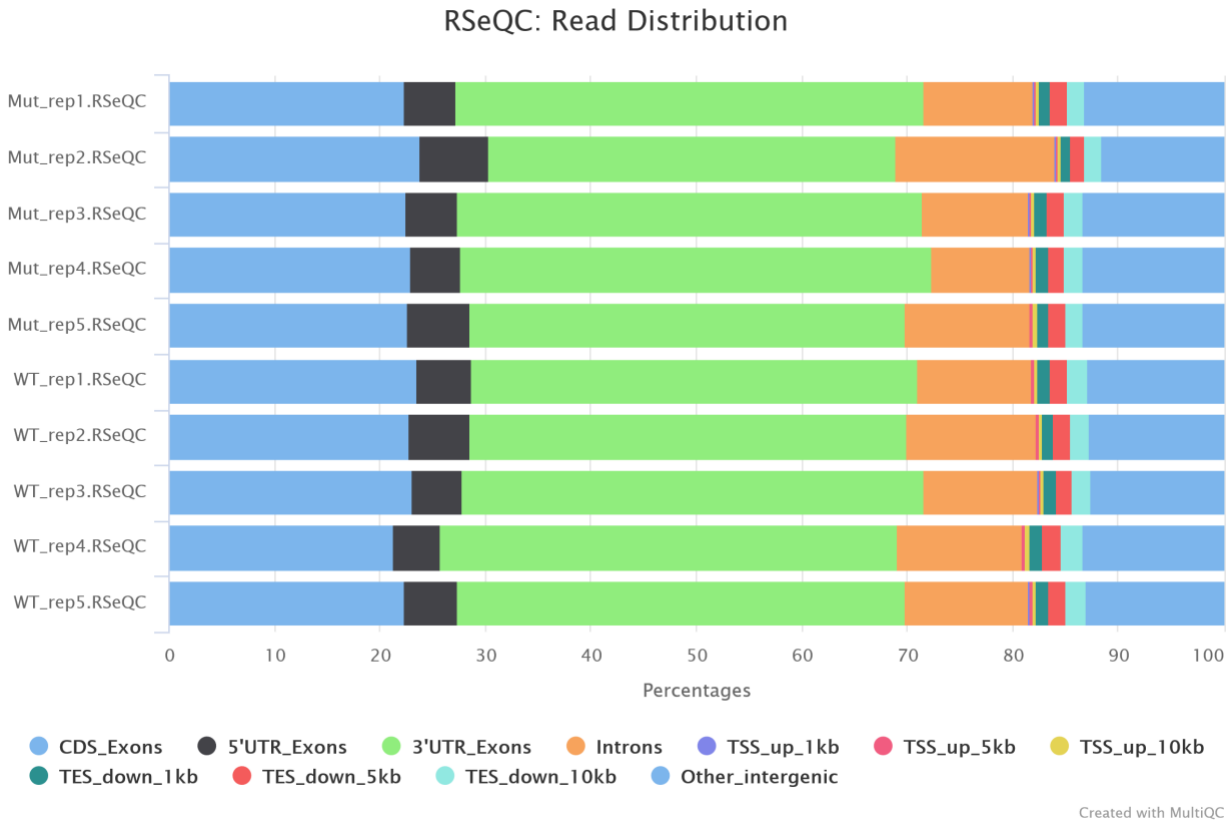

Figure 4: Mapped locations (RSeQC; in reads).

Read coverage distribution

- The gene body coverage is very similar in all the samples.

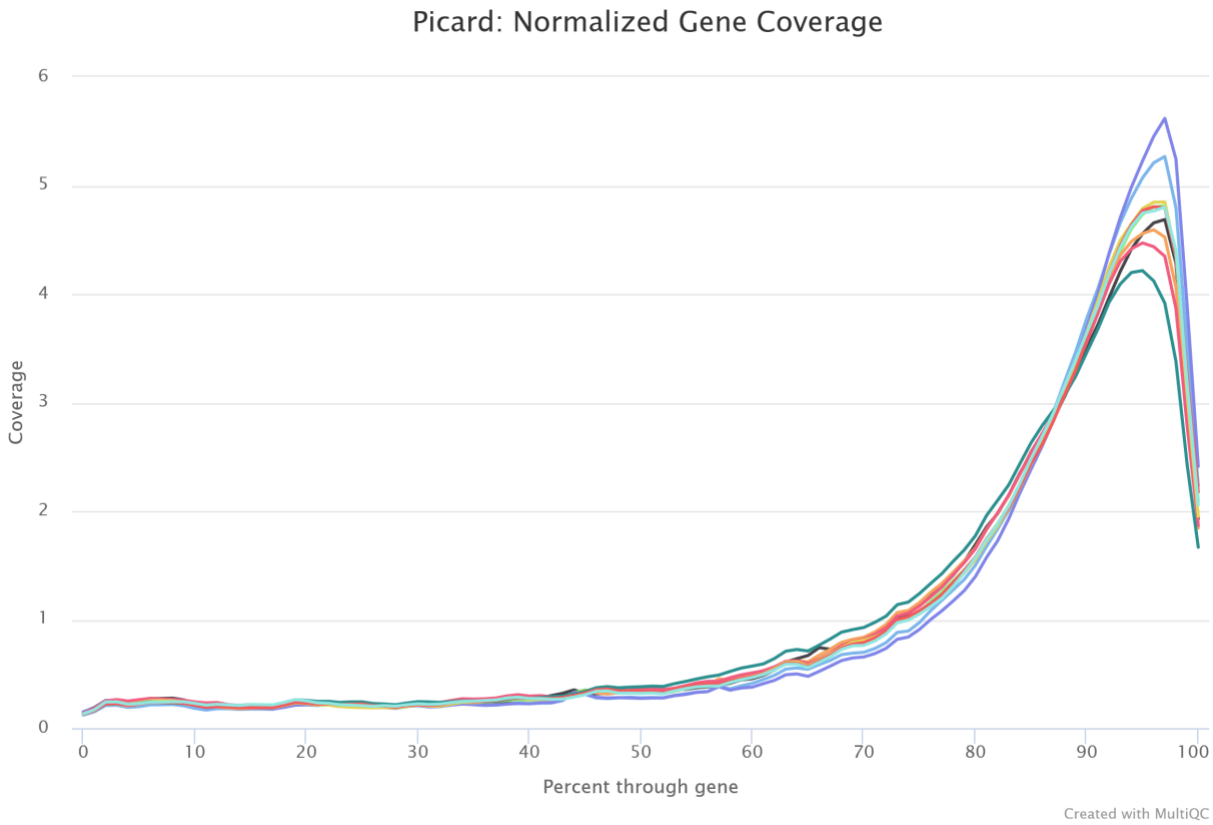

Figure 5: Gene body mapping distribution.

Strand specificity

- The analysis showed the strand-specificity of the library cannot be determined with a full confidence.
  - All the samples have ~86% of the forward-strand specificity, ~1% of the reverse-strand specificity but ~13% mappings cannot be determined.
    - These are usually mappings to intergenic regions, unannotated regions or regions where two genes overlap at the same strand (Figure 6).
  - We consider mapping as stranded (forward), as we got the highest number of suitable reads for DE analysis.

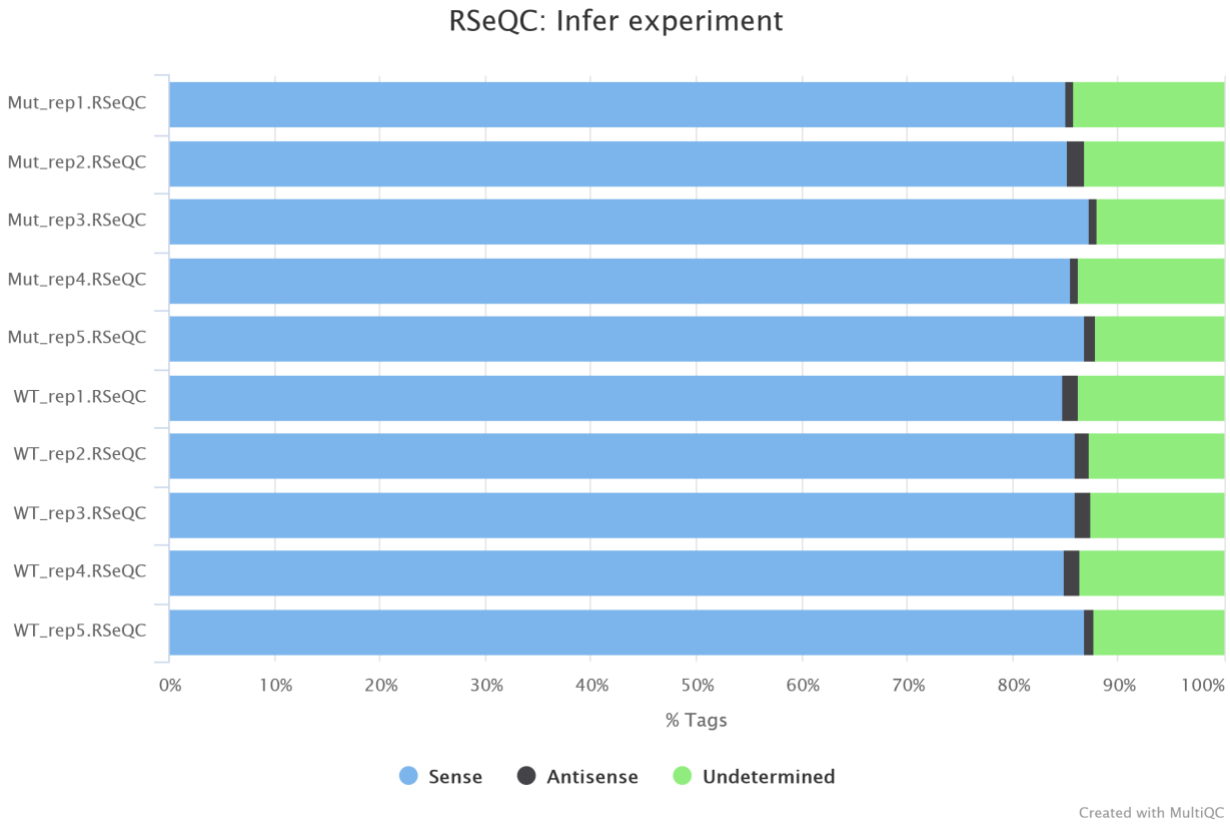

Figure 6: Library strand specificity.

Read count assignment to genes

- Read count assignment from **featureCounts** for RNA-Seq is quite high (between 52% and 57%, calculated [redacted])
  - Number of assigned reads is above limits of the recommended number of reads (~5-6M).
  - Read assignment rates are summarized in Table 3 and Figure 7.

Table 3: Summary of read-to-gene assignment. Percentages counted only from featureCounts.

Sample	Assigned	Assigned (%)
Mut-1	5,380,841	57.33
Mut-2	4,586,049	52.99
Mut-3	5,624,381	55.09
Mut-4	5,260,029	55.99
Mut-5	5,516,216	54.35
WT-1	5,251,339	52.69

Sample	Assigned	Assigned (%)
WT-2	4,798,645	52.25
WT-3	5,048,669	54.61
WT-4	5,354,256	53.14
WT-5	6,250,925	53.61

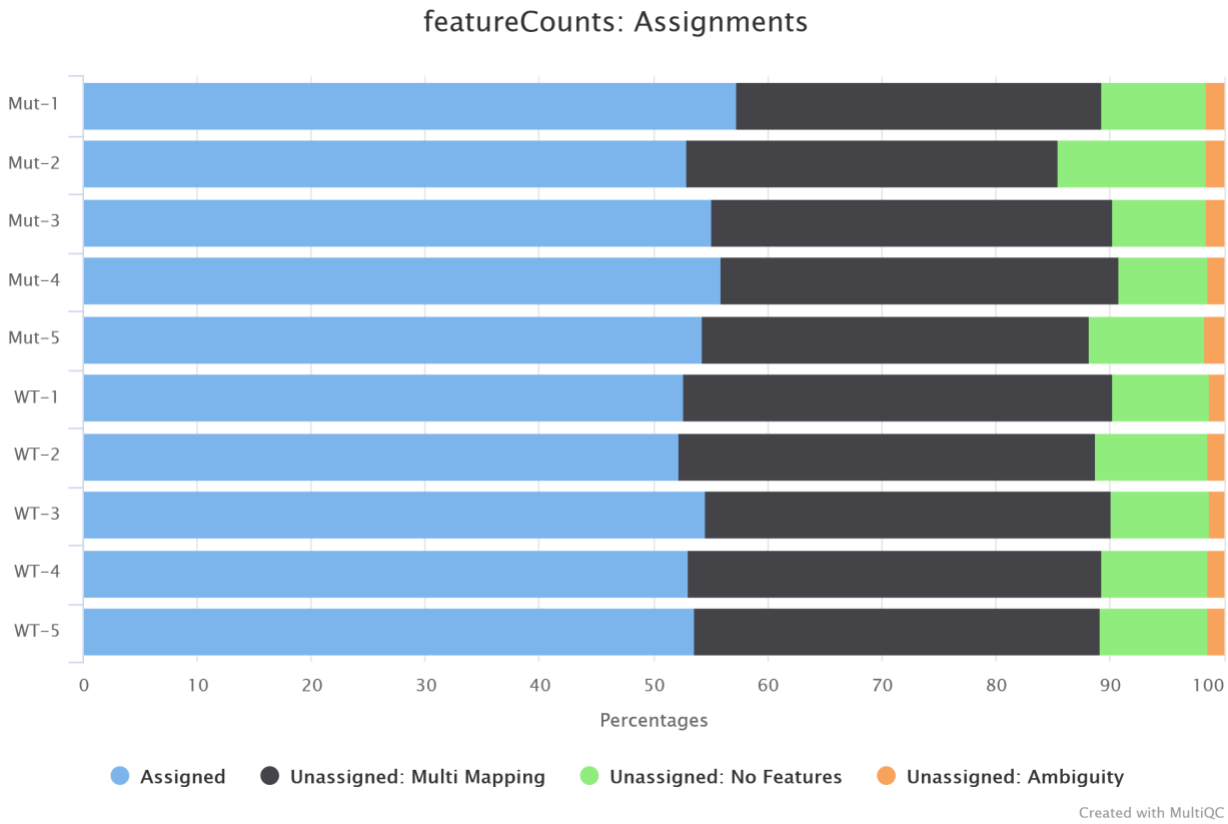

Figure 7: Read count assignment to genes (featureCounts).

Biotypes

- All the samples captured mainly protein-coding genes (Figure 8).
  - This is expected from polyA selection based libraries.

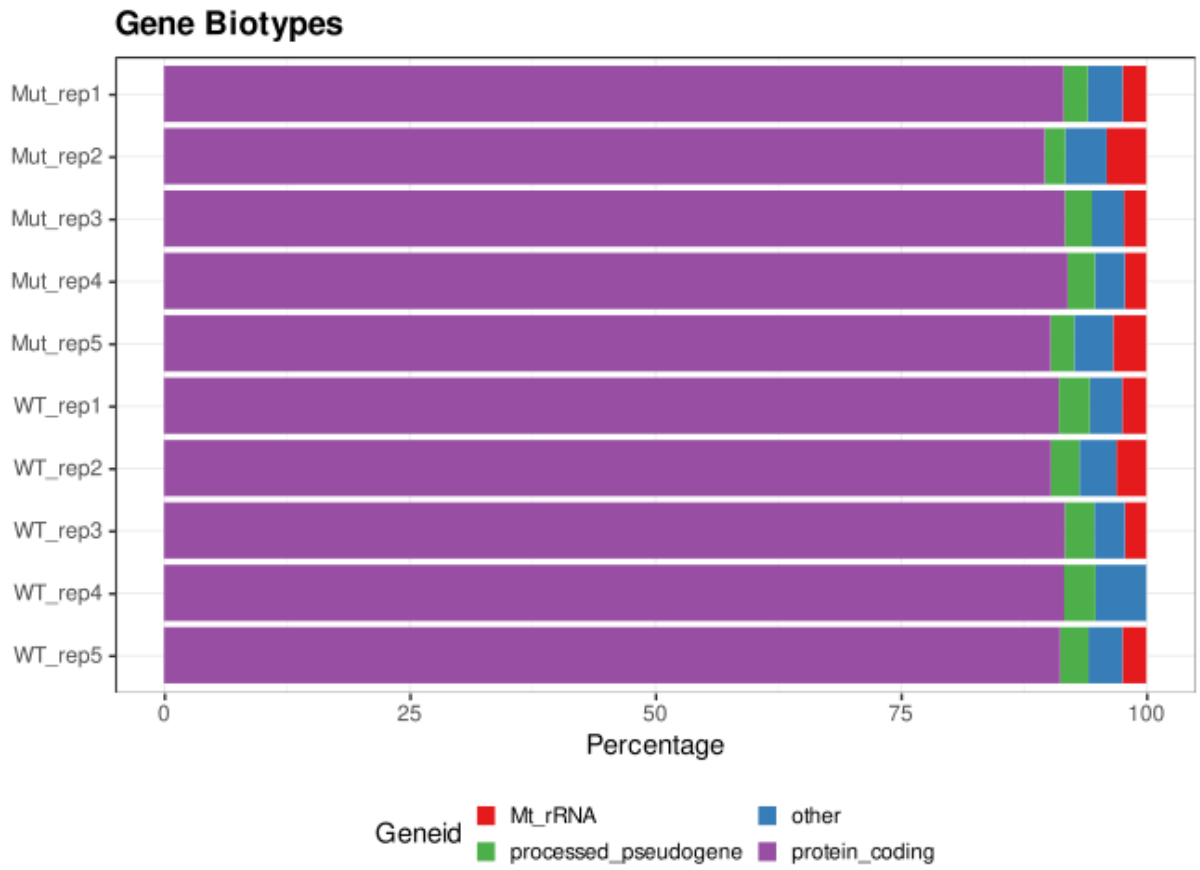

Figure 8: Captured gene biotypes.

Contamination check - Biobloom tools

- To check contamination we used basic model organisms eg. Mouse, Rat, Yeast, Dog, *D. Melanogaster*, *A. thaliana*, *B. napus* and *C. elegans*
- All samples have about 78% of reads mapped only to human, which is quite good and expected number
- The percentages are summarized in Figure 9.

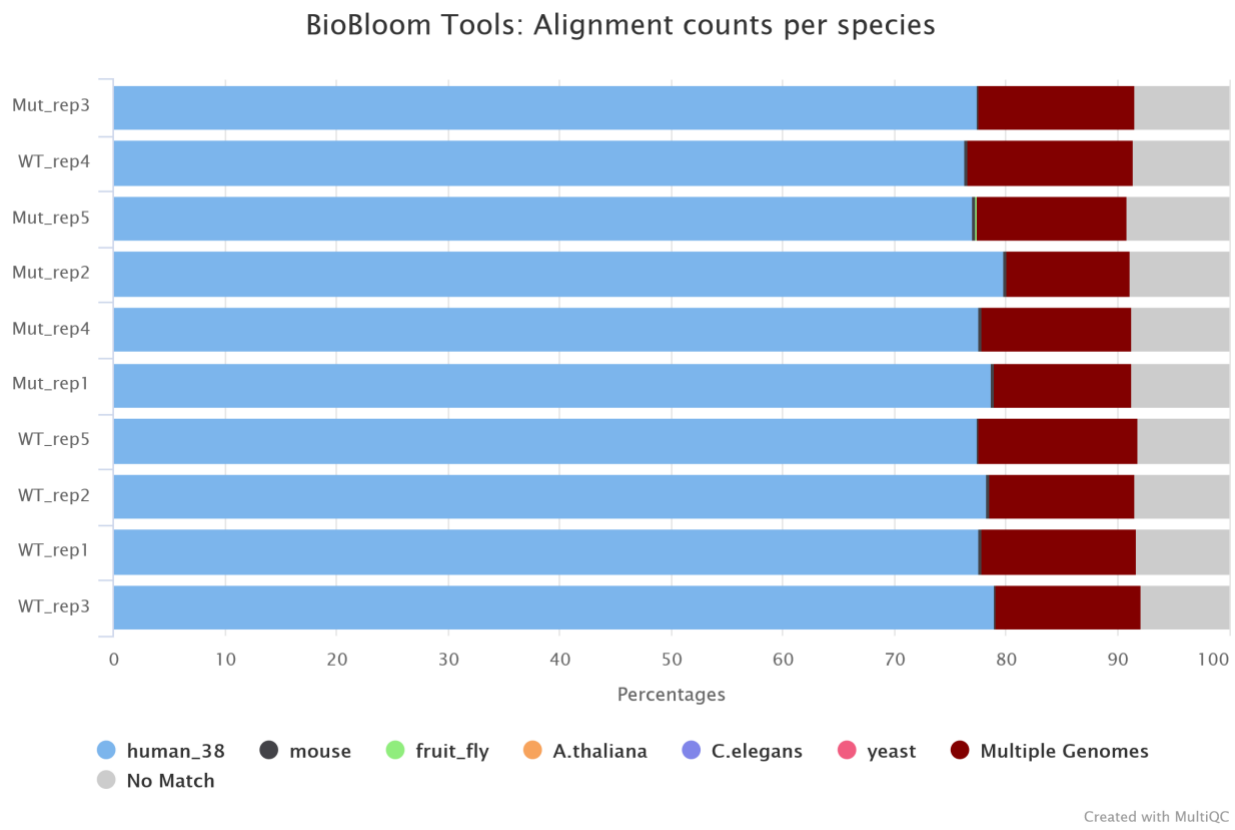

Figure 9: Captured read contamination.

Notes and comments

- We strongly recommend confirming all the events manually in genome/alignment browsers such as IGV, Tablet or Savant.
  - All the alignments and references are provided.

Visualization

IGV

- You can upload the provided aligned bam files, the provided genome annotation and the gene annotation to the IGV and browse the alignments.
- The IGV tutorial is available here
- References are available to download (for example for IGV) here.
- Used references
  - Homo_sapiens.GRCh38.dna.primary_assembly.fa
  - Homo_sapiens.GRCh38.dna.primary_assembly.fa.fai
  - Homo_sapiens.GRCh38.94.sorted.gtf
  - Homo_sapiens.GRCh38.94.sorted.gtf.idx
- All the index files (.bai, .fai, .idx) must be in the same folder as their “parent” files otherwise IGV wouldn’t open them.

DE analysis

- For the DE analysis we used gene counts from featureCounts tool.
- Read counts for coding genes (mostly protein-coding genes) distribution is depicted in Figure 10.
- Post-normalization counts (~normalized) seem to be OK (Figure 10).

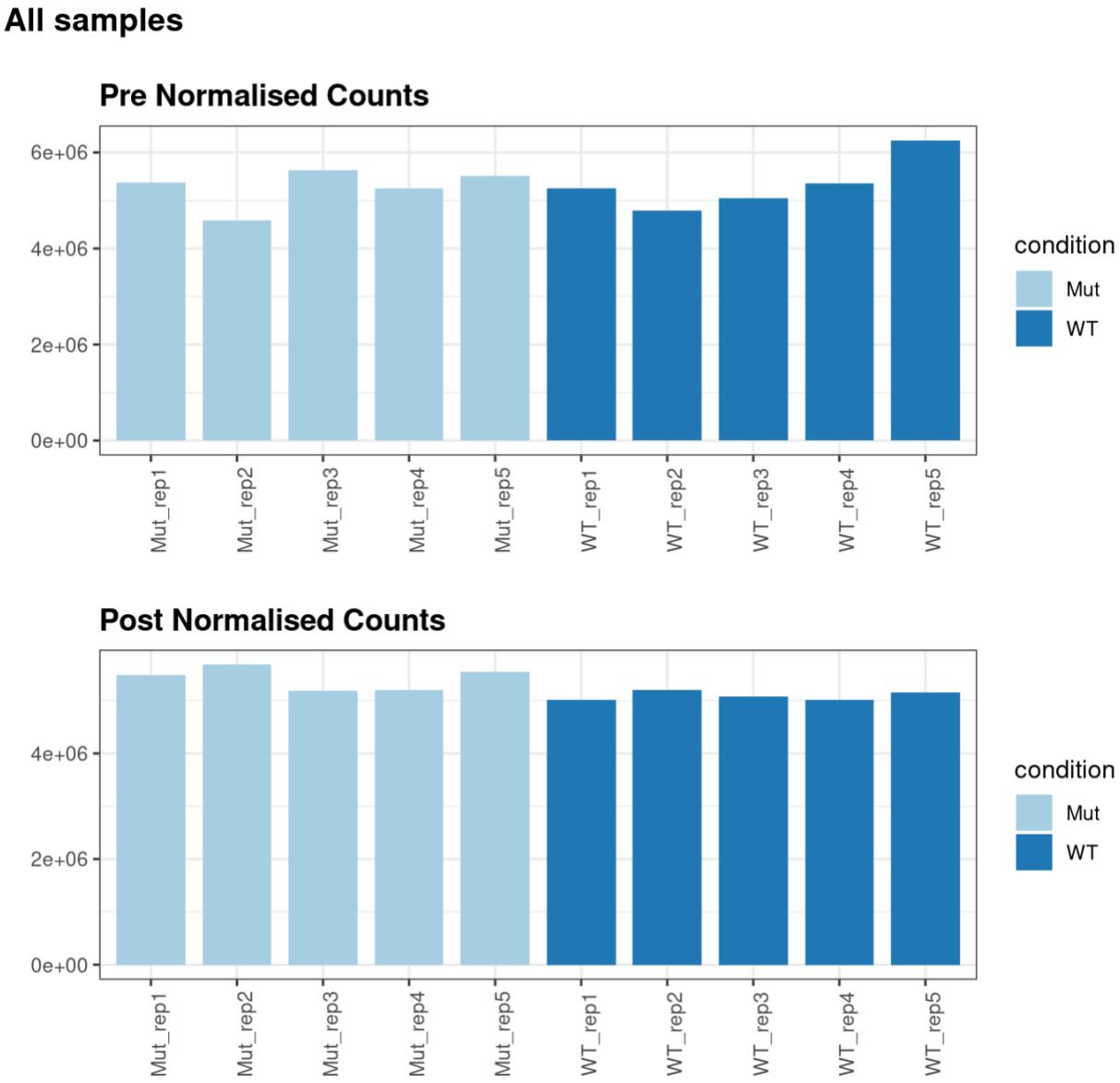

Figure 10: Number of reads assigned to coding genes for all the samples before normalization (raw counts). Coding genes depth for all the samples after normalization (normalized counts).

- Clustering of the samples (based on [redacted]) shows that all the samples cluster together by their condition.
  - This is summarized in heatmap in Figure 11.

- PCA shows a similar effect as the previous clustering.
  - PCA is depicted in Figure 12.

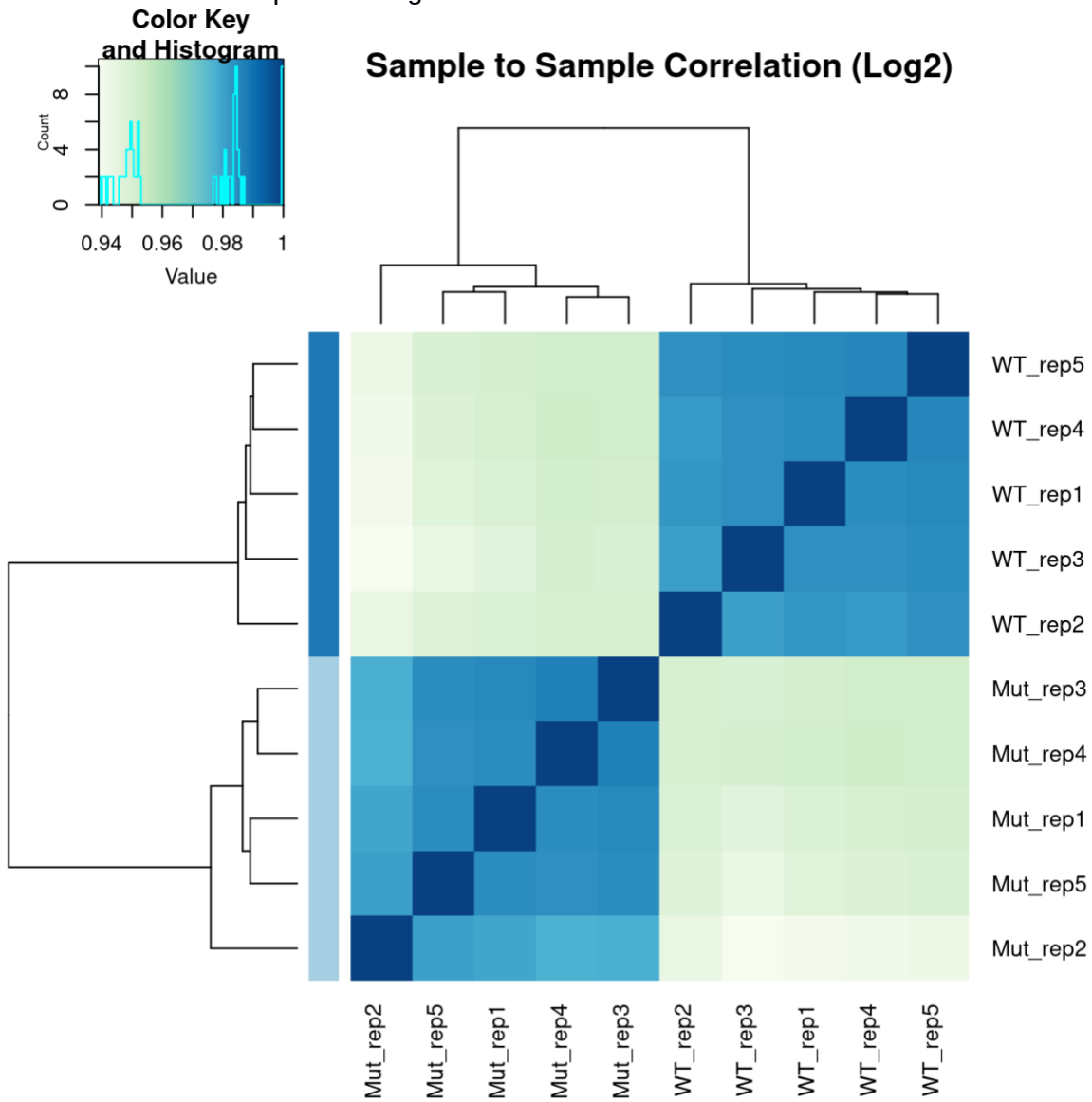

Figure 11: Clustering of samples (DESeq2 log2 normalization).

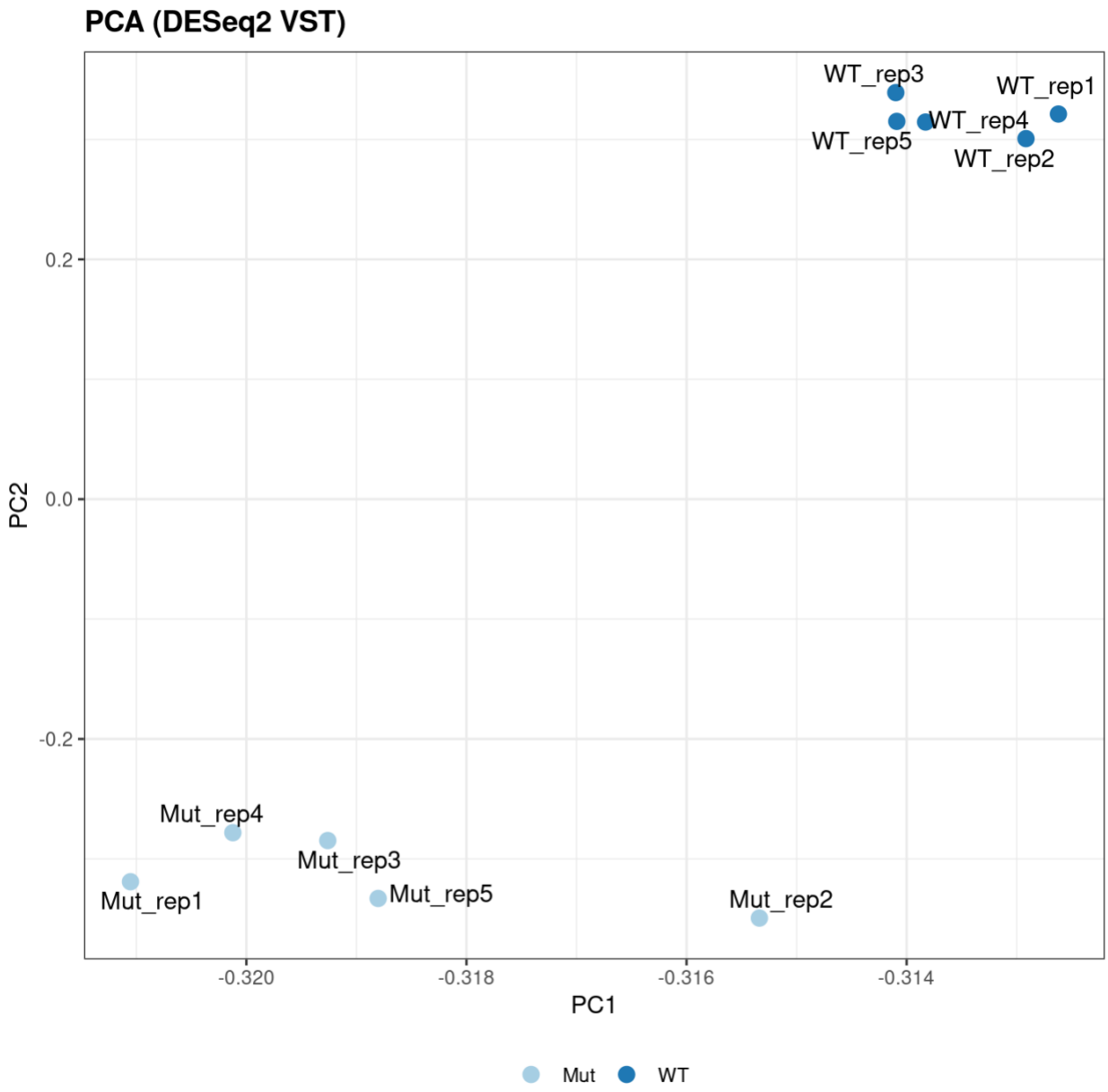

Figure 12: PCA (first two components) visualization of all samples (DESeq2 VST normalization).

Differential expression results

- Please note that for visualization purposes we had to set a predefined cut-off value for some of the parameters (adj.p-value and logFC)
  - **Please read the disclaimer at the beginning of the report for more information about selected cut-off values!**
- In this section, we might provide several results as the number of compared conditions is variable between experiments
- Differential gene expression was calculated by two separate tools - edgeR and DESeq2

  - Both of the tools have different performances: edgeR is usually more sensitive and less specific, DESeq2 is usually more specific and less sensitive; edgeR is more suitable for fewer replicates (<12) as it is not as conservative as DESeq2
  - Generally, DESeq2 is more conservative than edgeR.
  - If your goal is an exploratory analysis with a low number of replicates and with low expressed genes, edgeR might be your choice.
  - If you need a selection of genes which are strongly differentially expressed and you want to remove as many false-positive results as possible for a price of some false-negative results DESeq2 might be your choice.
  - If you want to perform some sort of prioritization, you might select an overlap between the two tools.
  - **Note for DESeq2**: By default, DESeq2 applies independent filtering which aims at removing genes which are potential outliers or show “strange” behavior (~large variance). Filtered genes are then marked by NA in either p-value or adj.p-value column. This filtering might be in some cases too strict and might cause a loss of interesting results. For this reason, we also make DESeq2 results without the independent filtering (DESeq2_noIndFilt). If you choose DESeq2 results, I recommend to start with filtered results but look at the end of the table at the filtered genes (genes with NA in p-value/adj.-value are at the bottom of the table). In case you see some interesting results there you might consider switching to the unfiltered results. In edgeR analysis, only genes with very low expression are excluded (1 read-per-million reads in at least 3 samples).
- The full description of a DE results is given after in the Output files section
- Number of DE genes with default cut-off values is summarized in Table 4
  - Please note this is only a tentative view on approximate differences between the conditions and not final results.
- Volcano plot visualization visualizes log2FC (x-axis) and adj.p-value (y-axis).
  - Colored dots highlight genes above the defined adj.p-value cut-off value, blue lines show defined log2FC cut-off value.
  - Volcano plot from DESeq2 results can be found in Figure 13.
- MA plot visualizes mean expression (x-axis) and log2FC (y-axis).
  - Colored dots highlight genes above the defined cut-off values (both adj.p-value and log2FC).
  - MA plot from DESeq2 results can be found in Figure 14.
- An example of maximum top 20 most DE genes (by adj.p-value) can be found in Figure 15.
  - The selection of a maximum 20 genes is always based on compared conditions but visualizes expression values from all other conditions if applicable.

Table 4: Summary or number of DE genes based on default cut-off values from all three DE calculations. The first two columns are counts with both adj.p-value and logFC cut-off, the other two columns are only with adj.p-value cut-off. DESeq2 results, DESeq2 results without independent filtering (noIndFilt) and edgeR results are shown.

Comparison	Analysis	Up (adj.pval & logfc)	Down (adj.pval & logfc)	Up (adj.pval)	Down (adj.pval)
WT_vs_Mut	DESeq2	1,715	1,436	3,600	3,462

Comparison	Analysis	Up (adj.pval & logfc)	Down (adj.pval & logfc)	Up (adj.pval)	Down (adj.pval)
WT_vs_Mut	DESeq2 noIndFilt	1,669	1,404	3,439	3,291
WT_vs_Mut	edgeR	1,537	1,355	3,610	3,613

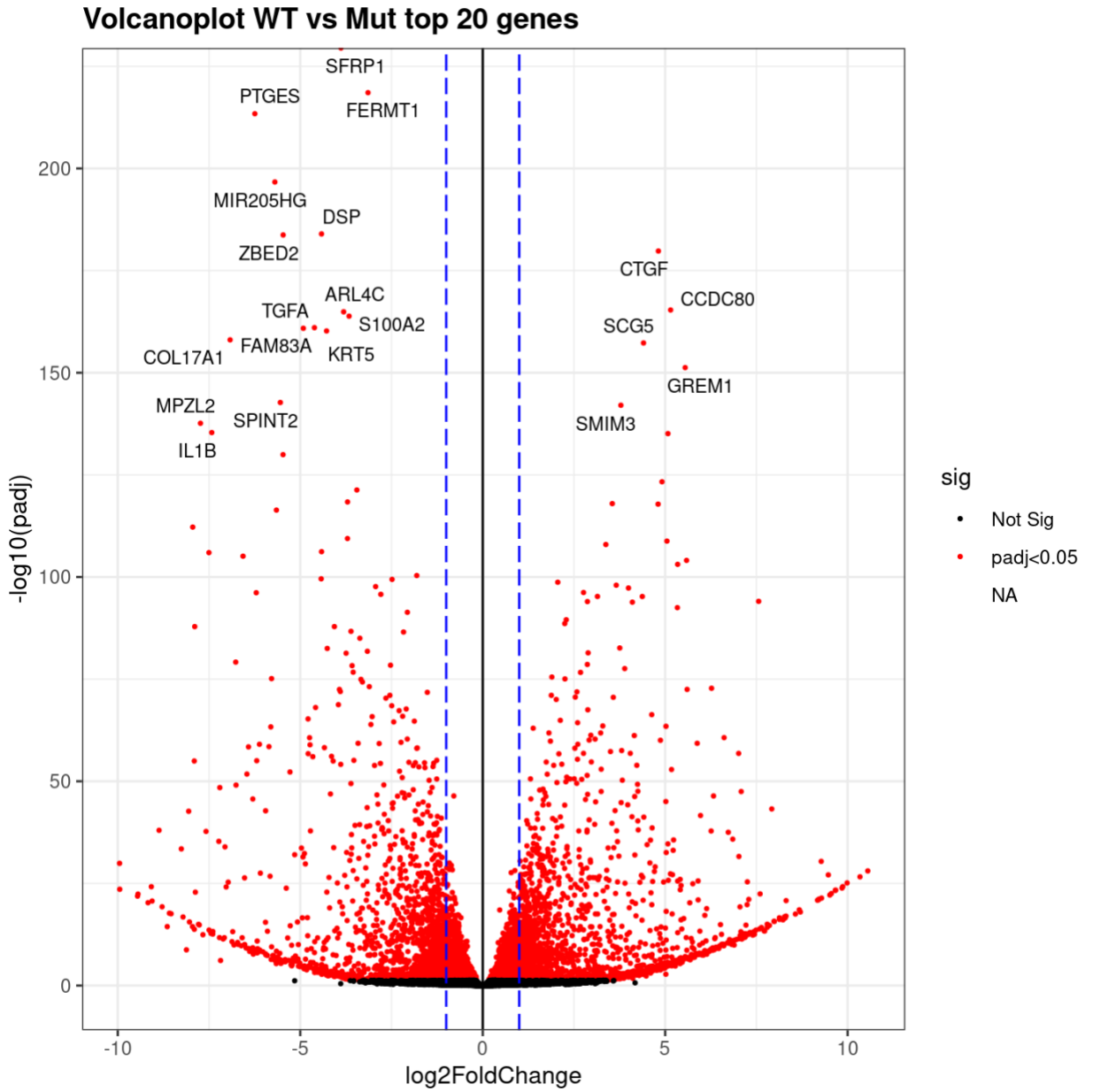

Figure 13: Volcano plot (DESeq2 results with independent filtering on). Maximum of 20 most DE genes (by adj.p-value) are named.

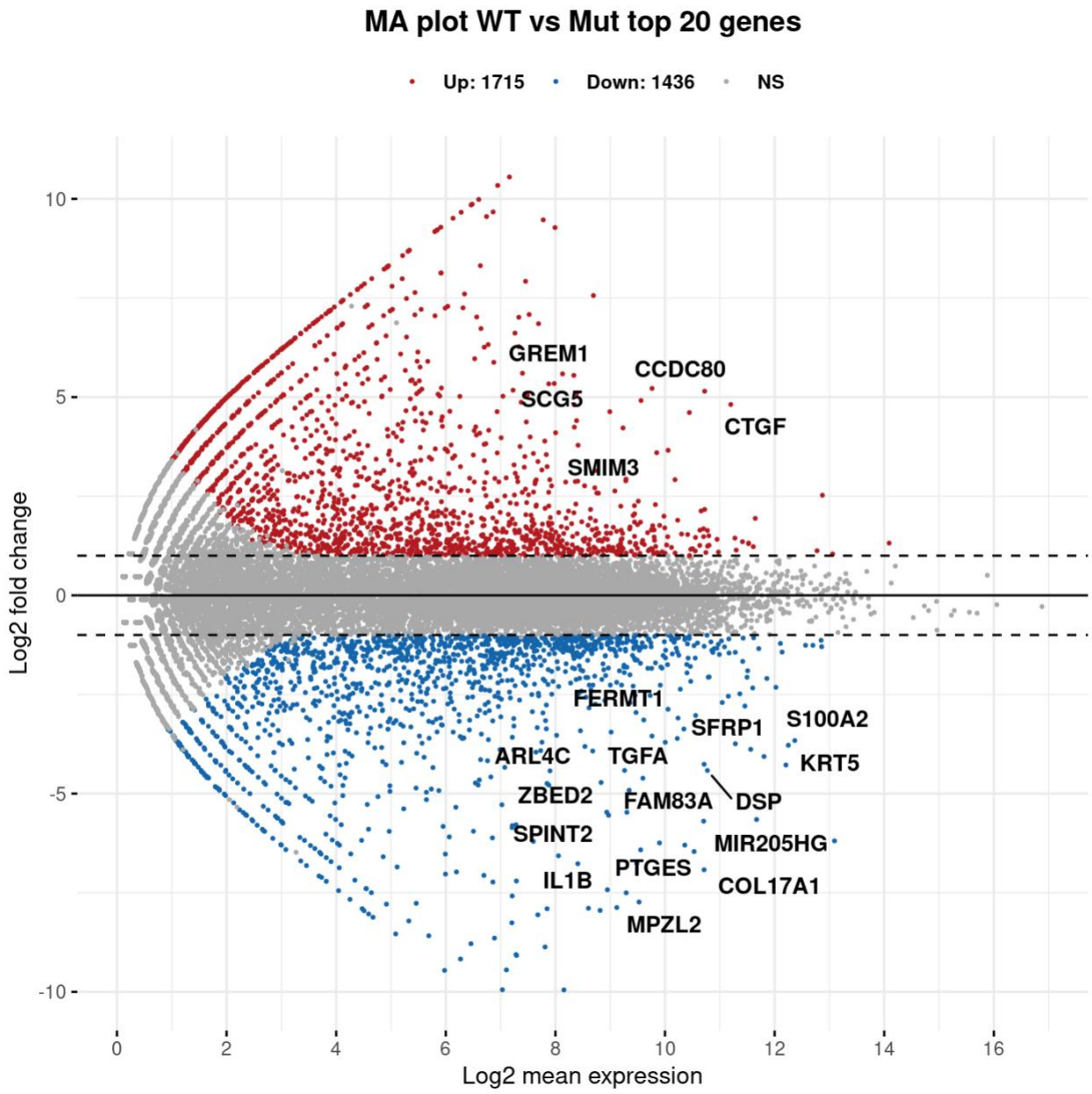

Figure 14: MA plot (DESeq2 results with independent filtering on). Maximum of 20 most DE genes (by adj.p-value) are named.

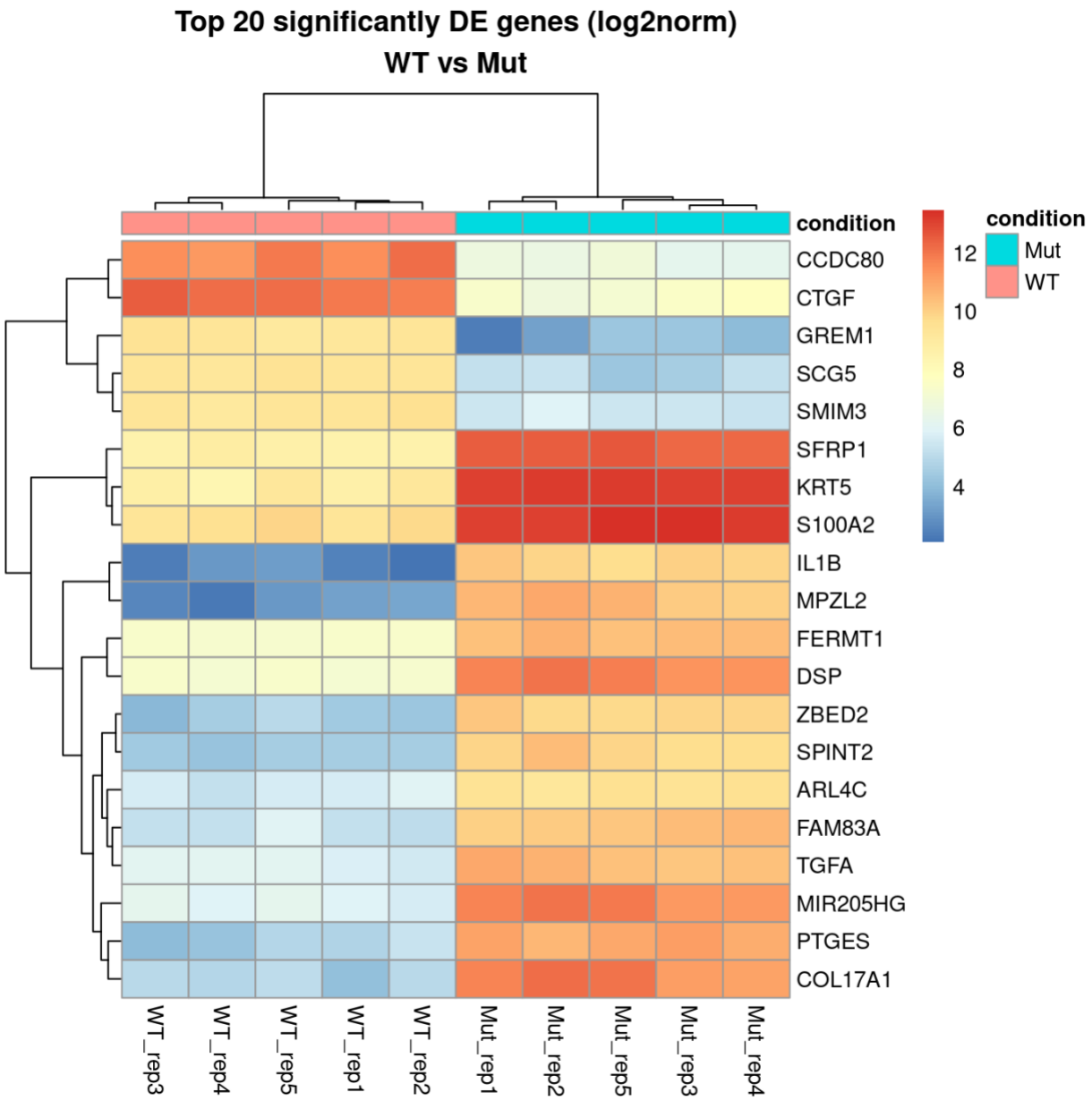

Figure 15: Heatmap of maximum top 20 differentially expressed genes (DESeq2 results with independent filtering, adj.p-value < 0.05, logFC >= +/-1).

Used statistics and selection of differentially expressed genes

Log2FC and adjusted p-value

- Differences between the conditions are expressed in **log2FC** (log2 of fold-change)
  - This is calculated by comparing average expression one condition vs the other and then the log2 transformation
  - This makes the results symmetric around 0 and it's easier to understand
  - For example: -1 is a 2-fold decrease of the expression, +1 is a 2-fold increase of the expression.

  - Please note it always depends on the “direction” of the comparison - if you are not sure just compare normalized counts of one gene to see from which “direction” the log2FC was calculated.
- Calculation of DE genes consists of a very high number of comparisons (thousands of genes = thousands of comparisons)
  - For this reason, we need to correct raw p-values for false discovery rate which emerges from the number of comparisons
  - This results in **adjusted p-values** (adj.pval, adj. p-values, adj. pval)
    - In case you want to use some filtering or selection based on the statistical evidence you should ALWAYS USE adjusted p-values and never raw p-values
    - Raw p-values can be used only in very special cases and always require a discussion with bioinformatician or statistician

Note on selecting of differentially expressed genes (DEG)

- Selection of differentially expressed genes always has to be primarily based on the biological background of the experiment and the hypothesis
- A lot of people select DEG based on a combination of cut-off values
- This might be tricky as no predefined cut-off values exist and it depends on the experiment itself, on the amount of false-positive results you are willing to accept and on the minimum level of changes if any
- Commonly used statistical values, such as logFC and adj.p-value, should help you to confirm the hypothesis, results or the observations and not lead the discovery itself
- If you decide to go for some cut-off values, do not blindly select only genes above the thresholds!
  - For example, the frequently used cut-off value of adj.p-value 0.05 at which you will reject the hypothesis the gene is not expressed the same: one gene will have adj.p-value 0.0499 which would pass this cut-off whereas the second gene will have adj.p-value of 0.0501 and would not pass the cut-off. The difference between the genes is negligible and still, the first gene would and the other gene would not be accepted if a strict cut-off value would be set.
  - Another “issue” with adj.p-value is when the groups we are comparing and heterogeneous (for example clinical samples). In this case, adj.p-value are generally much higher. In this situation, we should strongly focus on biological effect and hypothesis and use statistics to either confirm the observation or to function for prioritization of the results
  - The similar thing applies to logFC
  - In the case of logFC, we also have to consider the effect of the expression of individual genes. For example, the frequently used cut-off value of 1 (2x fold-change): A gene with expression 10 in one group and 5 in other group results in logFC 1 and would pass the cut-off. The overall expression of the gene is negligible and most likely the change doesn’t have any biological effect. Another gene has an expression of 10000 in one group and 5001 in another group. This gene would not pass the cut-off value but most likely the change will have some kind of biological effect.
  - Sometimes small logFC might be more important than high logFC and the same applies to adj.p-value.

- We should always choose the genes based on the overall properties of the gene and the comparison between conditions

Output files

- Results are organized by compared conditions
  - The compared conditions are always stated in the name of the folder
  - All the results in the corresponding folder are based on the currently compared conditions but in some of the cases other samples and conditions as added to the visualization/summaries as well
  - For example: **cond1_vs_cond2** compares cond1 to cond2. In the DE results then - positive log2FC signalizes more expression in cond1 (or less in cond2) and negative log2FC signalizes less expression in cond2 (or more in cond1).
- Main output files are **DESeq2.tsv**, **DESeq2_noIndFilt.tsv**, and **edgeR.tsv**
- All three main results contain several columns:
  - **first column (no name)** - Ensembl gene id
  - **baseMean** - an average expression of genes from samples from compared conditions
  - **log2FoldChange** - log2FC of the gene expression difference. The difference is calculated by comparing average expression of samples of one condition with the other samples from the other condition
  - **stat and lfcSE** (in DESeq2); LR and tgw.Disp (in edgeR) - results of the statistical test performed by the tool and variation/dispersion of the gene
  - **pvalue** - a raw p-value of the comparison
  - **padj** - multiple testing correction (Benjamini-Hochberg correction) p-value of the comparison. In case you are comparing and analyzing RNA-Seq use this as the evaluation of the statistical significance of the comparison
  - **gene_name** - common gene name
  - **gene_biotype** - biotype of the gene
  - **_normCounts** - normalized gene expression for each sample (not only the samples in the current comparison)
  - **_rawCounts** - raw read counts for each sample (not only the samples in the current comparison)
- The full description is given below
- Additional visualizations and/or a different combination of visualization is available upon request after discussion

General/common outputs

- **xxx** in the name of some pdf files in the description below represents compared conditions
- **background.txt** – list of genes that were expressed in your samples (=gene was expressed in at least one of the samples). The first column is Ensembl gene ID, second is common gene name.

- **norm_counts.xlsx** - DESeq2 normalized counts (~expression) for all genes and all samples. Can be used for visualization of the data.
  - **norm_counts.tsv** - the same information as in norm_counts.xlsx but in a text form.
- **all_sig_genes_normCounts.pdf** – significantly DEG (here, "default cut-off" adj.p-value 0.05 and logFC 1 is considered) expression visualization. Selection of DE genes is based on the particular comparison but all conditions/samples are plotted. This is based on DESeq2 results.
- **counts_barplot.pdf** - visualization of used (raw) read counts per sample
- **DESeq2.tsv** – main results from differential gene expression by DESeq2. Description of the columns is given above. tsv is a tab-separated source file for the xlsx.
- **DESeq2_noIndFilt.tsv** – results from differential gene expression without independent gene filtering (see Independent filtering and Cooks cut-off in DESeq2 manual). tsv is a tab-separated source file for the xlsx.
- **DESeq2_de_genes_check.txt** and **DESeq2_de_genes_check_noIndFilt.txt** – basic summary of differential expression. "Default" adj.p-value of 0.05 and logFC of 1 are used. The upper part of the table considers only adj.p-value cut-off value, the lower part considers adj.p-value and logFC. This servers only for a demonstration of the possible effect, not as a final result
- **DESeq2_design_control.txt** - a control file for used sample design
- **edgeR.tsv** – main results from differential gene expression by edgeR. tsv is a tab-separated source file for the xlsx.
- **edgeR_de_genes_check.txt** - same as in case of DESeq2 - "quick and dirty" check of the effect to get a general idea about the differences.
- **heatmap_selected_orderBaseMean.pdf** – heatmap of maximum top 20 most differentially expressed genes (based on adj. p-value) ordered by an average expression. Blue means low expression, red means high expression. The maximum top 20 genes are selected based on the compared groups but all samples are visualized. This is based on DESeq2 results.
- **heatmap_selected_orderBaseMeanCluster.pdf** – heatmap of maximum top 20 most differentially expressed genes (based on adj. P-value, same as above) clustered by rows and columns. Blue means low expression, red means high expression. This is based on DESeq2 results.
- **heatmaps_samples.pdf** - heatmap visualization of sample clustering based on gene expression. Several normalizations are applied but the first one usually looks the best.
- **MAplot_xxx_vs_xxx_ggpubr.pdf** - MA plots for the gene expression. Maximum top 20 most DEG (by adj.p-value) are named. This is based on DESeq2 results.
- **MAplot_xxx_vs_xxx_noIndFilt_ggpubr.pdf** - the same as above but without independent filtering.

- **MDS_plot.pdf** - MDS-based visualization of sample clustering by expression
- **overlap_DESeq2_edgeR_de.xlsx** and overlap_DESeq2_edgeR_venn.pdf – overlap between DE genes between DESeq2 and edgeR calculation. In default, genes with adj. p-value <0.05 and logFC +/- 1 are included. Columns DESeq2 and edgeR contain values TRUE or FALSE and this represents whether the resulting gene was included in the results by an individual tool.
- **pre_post_norm_counts.pdf** - visualization of used raw (top figure) and normalized (bottom) read counts per sample
- **sample_to_sample_PCA.pdf** - PCA-based visualization of sample clustering by expression. The first figure shows first two PCA components, the second shows all combinations of first three components (~3D look)
- **volcanoplot_xxx_vs_xxx_ggplot2.pdf** – volcano plot of the results from DESeq2. Maximum top 20 most DEG (based on adj.p-value) are depicted. This is based on DESeq2.
- **volcanoplot_xxx_vs_xxx_noIndFilt_ggplot2.pdf** - same as above but without independent filtering.
- Please note the differences between MA and Volcano plots - different X and Y axes giving you a different view on the results.
- Alignments can be visualized in genome browsers such as IGV, Tablet or Savant
  - For more details please read the manual or contact your bioinformatician
- If the mapping files were shared through the online storage please download them as soon as possible and let us know so we can delete them

Output files (alphabetical order)

Folders

Raw_fastq

- Provided raw sequenced fastq files ### Alignment
- Provided alignments in bam format ### QC_general
- html file summarizing most of the QC done ### FeatureCounts_expression
- Gene expression raw counts for each sample ### RSEM_expression
- Gene and transcript expression for each sample
- The normalized expressions are summarized in TPM column #### Files
- .genes.results.tsv - gene expression estimates
- .isoforms.results.tsv - transcript expression estimates ### UCSC_signals
- Alignment signal for UCSC browser
- Two signal “types” per sample are provided:
  - Unique - signal only from uniquely mapped reads

  - UniqueMultiple - signal from both uniquely and multi-mapped reads
- Each set is then split into two:
  - str1 - signal coming from the alignment to the plus strand of DNA
  - str2 - signal coming from the alignment to the minus strand of DNA
- If you have sense/forward specific library, the signal strand corresponds to the strand of the annotated gene (mapping —-> corresponds to the + strand annotation)
- If you have antisense/reverse specific library, the signal strand corresponds to the opposite strand gene annotation (mapping <—- corresponds to the + strand annotation)

DE_analysis

- All files related to Differentially expressed genes analysis

Gene ontology, GSEA and pathway analysis

- This section summarizes possible follow-up analyses/interpretations of the differential expression results
- In the following analyses the input can sometimes be specified as Ensembl ID (usually provided in the results unless noted otherwise), some tools required common gene names (usually provided in the results as well) or Entrez ID (NCBI/RefSeq)
  - This is always specified in the manual of the tool
- GO and/or Pathway analysis can be performed for up-/down-regulated genes as well as for the whole list of de-regulated genes
  - Both of the analyses give a different view of the results
    - Analysis of separate up-/down-regulated genes gives you a direct answer on the change, in particular, GO or Pathway and the “direction” of the regulation
    - Analysis of all de-regulated genes (up-/down-regulated together) gives you a more broad view of the total change
  - Both of the results are helpful but you have to realize what kind of information you can get and what kind of answer they provide

Gene Ontology

- Input for gene ontology (GO) testing is usually a list of differentially expressed genes
  - You can either use filtering by some cut-off value(s) or select top X genes sorted by adj.p-value/logFC/…
- You can either analyze all DEG or separately down-/up-regulated
  - Please see the section above for more details
- A second input is gene background/universe
  - This is a list of all genes expressed in the comparison
  - This list is used as a “reference” for the statistical evaluation of significantly changed GO groups
- For “common” organisms like human or mouse, a very nice tool is GOrilla with very nice figures and direct export to Excel
  - It also contains links to REViGO which can help you to summarize the GO groups into more general categories
- To simply browse GO categories you can use QuickGO or Ontology Lookup Service

- Another option to filter down GO categories is AmiGO2.

- If, for whatever reason, you don't have gene ontology annotations but just sequences of your genes you can still get GO analysis. Blast2GO is a nice service combining BLAST similarity search with GO annotation to produce GO analysis. There is a free and paid version where the free version gives you basic annotations and have some pretty visualizations. A tutorial is for example here https://www.youtube.com/watch?v=GqSqS_izIYg&t=335s.
- Other tools include g:Profiler and many more

Pathway analysis

- Input data are very similar to GO (mentioned above)
- PANTHER and Reactome provide quite nice analysis for the Pathway analysis and exploration.
  - PANTHER recently started to include GO testing as well so it became a multi-purpose tool
- KEGG is another commonly used tools for pathway analysis but it is not that easy to interpret the data unless you know which pathways should be involved

Gene Set Enrichment

- Input data to Gene Set Enrichment are a bit different than in previous cases
- You do NOT subsample your results but you take a whole list of gene expression results and sort it according to specified criteria (logFC, adj.p-value, …)
  - In the case of logFC, you have the most positively changed genes on the top and the most negatively changed genes at the bottom and non-changed genes are in the middle but they are still in the SAME list
  - The enrichment then compares different parts of the list and looks for patterns or similarities
- One of the most used is probably GSEA
  - It's a standalone tool which takes some time to learn but provides reasonable results
- GOrilla also offers gene enrichment analysis
- Another option is to use DAVID which is also a multipurpose tool but I do not recommend it too much since the interpretation of the results might be difficult and very subjective. But it can still be used if there is no other option or you find it easy to work with.

Protein-protein interactions

- HIPPIE provides reliable and meaningful human protein-protein interaction networks. It is suitable more for gene-by-gene exploration.
- STRING is a database of known and predicted protein-protein interactions supported by SIB, CPR - NNF and EMBL. This includes physical, functional as well as predicted interactions.
- BioGRID is an interaction repository with data compiled through comprehensive curation efforts. Contains interactions extracted from publications for major model organisms

Multipurpose tools and other

- There are other tools/approaches to help with the secondary analysis

- Enrichr is nice and simple to run the tool to identify enriched pathways, different gene summaries, and many others. It provides many various plots and tables
- GeneMANIA visualizes genes and their interaction, co-expression as well as Gene Ontology, etc. It is suitable more for gene-by-gene exploration.
- ConsensusPathDB-human integrates several different approaches including pathway analysis
- BioCyc Database Collection is a collection of 9387 Pathway/Genome Databases (PGDBs), plus software tools for understanding their data.
- KEGG database is a database resource that integrates genomic, chemical and systemic functional information. In particular, gene catalogs from completely sequenced genomes are linked to higher-level systemic functions of the cell, the organism, and the ecosystem. KEGG accepts common gene names and NCBI gene/transcript ID and UniProt ID. For the analysis, you can use http://www.genome.jp/kegg/tool/map_pathway2.html or http://www.genome.jp/kegg/tool/map_pathway1.html. KEGG looks for genes in your list and tries to highlight the pathway where selected genes have some kind of association.
- Venny is a nice tool for a simple visualization of overlaps between up to three sets of data.

R/Bioconductor

- If you know how to use **R** you can check **clusterProfiler** which provides many different analyses types and has very nice manual and tutorials
- **topGO** is another R package which offers both gene ontology and gene set enrichment analysis
- **ReactomePA** allows for analysis of pathways with a slightly different source than KEGG
- To add GO categories (and other interesting information) to your list of genes, you can try **biomaRt** which can directly load the Ensembl database