VILNIUS UNIVERSITY

Faculty of Medicine

Scientific report

RNR sequencing data analysis

Živilė Grublytė

Vilnius, 2020

Introduction

This report describes how transcriptome data analysis was performed on given data and gives brief interpretation of the results, explaining why particular steps of analysis were needed. To accomplish these tasks snakemake workflow was created which can be reached by this link: https://github.com/ZivileG/Transcriptomics_exercise .

Checking raw reads quality with FastQC and MultiQC

FastQC is a tool to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which can be used to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis. FastQC can import data from BAM, SAM or FastQ files (any variant) and give summary graphs and tables to quickly assess the data. MultiQC is a tool to create a single report with interactive plots for multiple bioinformatics analyses across many samples. Raw reads were checked using FastQC and report generated using MultiQC to check the quality of the data.

MultiQC report shows adapter content per base position in reads (Figure 1)

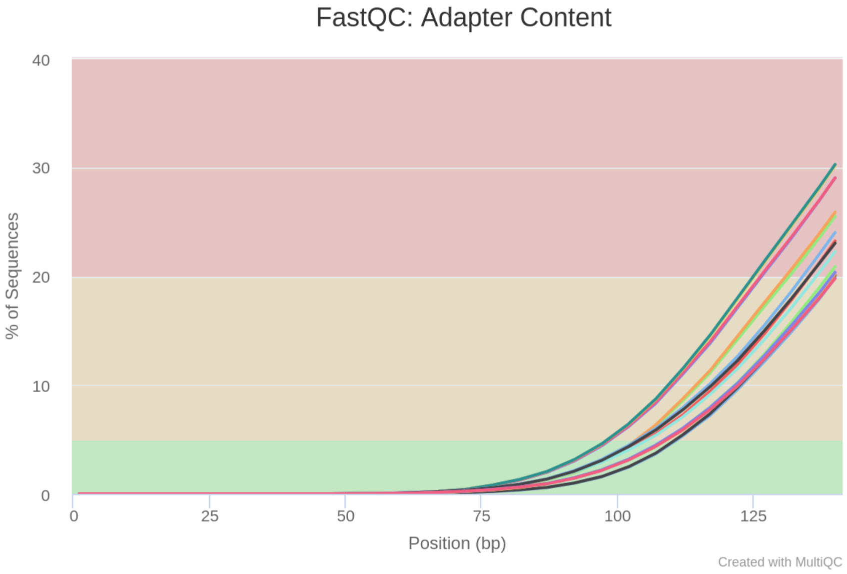

Figure 1. The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position.

Adapters have to be ligated to every single DNA molecule during library preparation. Adapter contamination will lead to NGS alignment errors and an increased number of unaligned reads, since the adapter sequences are synthetic and do not occur in the genomic sequence. To avoid this sequences were trimmed using bbduk tool.

The mean quality value across each base position (Figure 2) was above 30, data corresponds to the fact that normal random library typically have a roughly normal distribution of GC (Figure 3), so the conclusion was made, that data is quite good quality and analysis can be continued.

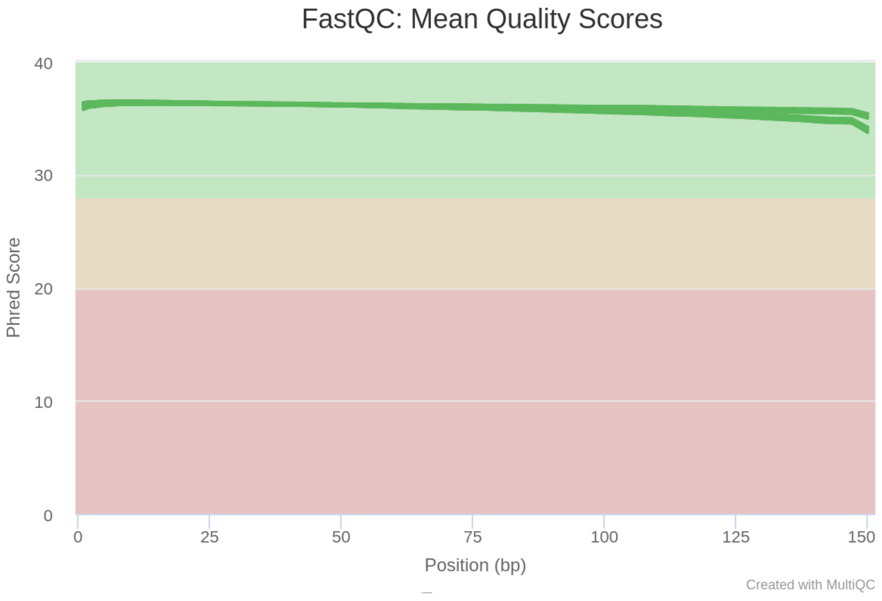

Figure 2 The mean quality value across each base position in the read.

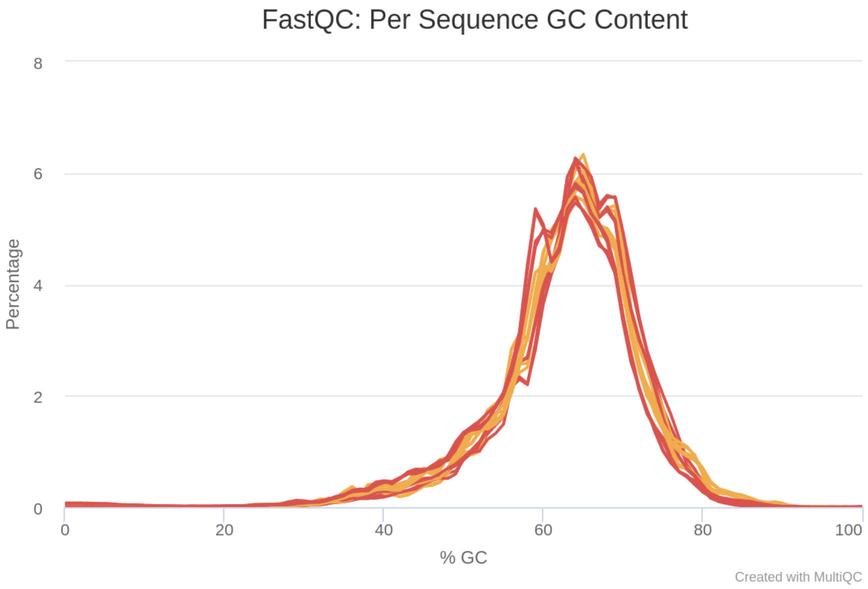

Figure 3 The average GC content of reads

Aligning reads to the reference genome

As the quality of reads was already explored read alignment could be done. Read alignment war performed to determine where in the genome the reads originated from. For this task read alignment tool STAR (Spliced Transcripts Alignment to a Reference) was used. STAR is an aligner designed to specifically address many of the challenges of RNA-seq data mapping using a strategy to account for spliced alignments.

Basic STAR workflow consists of 2 steps: Generating genome indexes files and Mapping reads to the genome. Generating STAR indexes you can choose `–sjdbOverhang` value which can be specified as ReadLength-1. I have chosen 100 which is recommended as a generally good value in the STAR documentation. This may have not been very wise, because from Figure 1 we can see, that most reads after trimming will be shorter than 100 bp. Maybe because of incorrect value chosen, from STAR Alignments Scores we can see, that most reads are unmapped because they are too short (Figure 4., Figure 5). Also it can be seen that we have much bigger number of reads for tissue type UHRR than HBR. But the percentages of mapped reads are quite similar between samples of different preparation methods and tissue types.

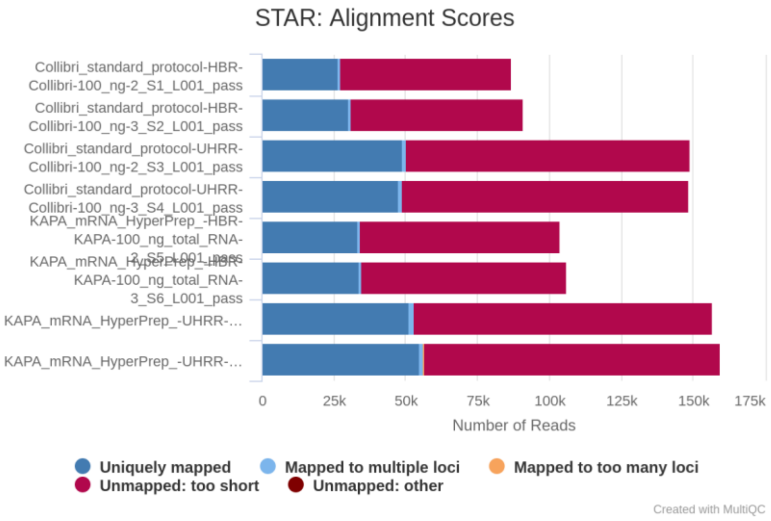

Figure 4 STAR Alignment Scores, number of reads

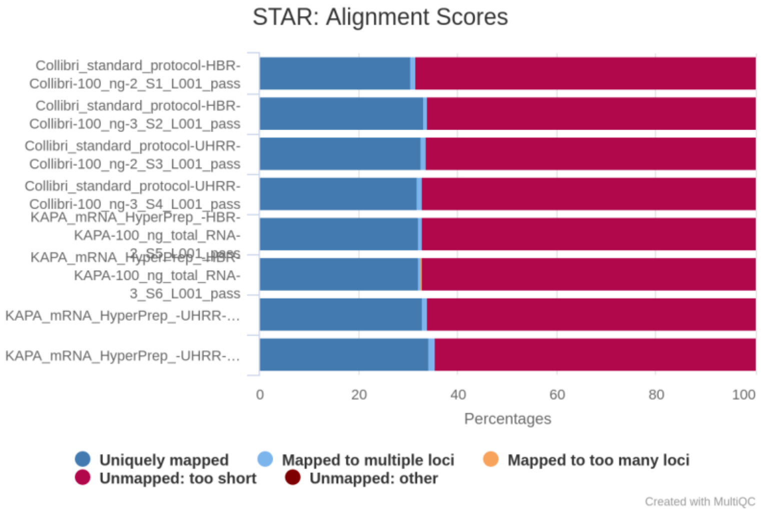

Figure 5Alignment scores, percentages

Counting step with featureCounts

FeatureCounts is a highly efficient general-purpose read summarization program that counts mapped reads for genomic features such as genes, exons, promoter, gene bodies, genomic bins and chromosomal locations. *featureCounts* takes as input SAM/BAM files and an annotation file including chromosomal coordinates of features. It outputs numbers of reads assigned to features (or meta-features). It also outputs stat info for the overall summrization results, including number of successfully assigned reads and number of reads that failed to be assigned due to various reasons. From Figure 6 you can see the percentage of assigned genes when the strand specificity with most feature counts was chosen (s=1 for Collibri preparation method data and s=2 for KAPA preparation method data). When incorrect settings are chosen we have much less assigned reads (see Figure 7)

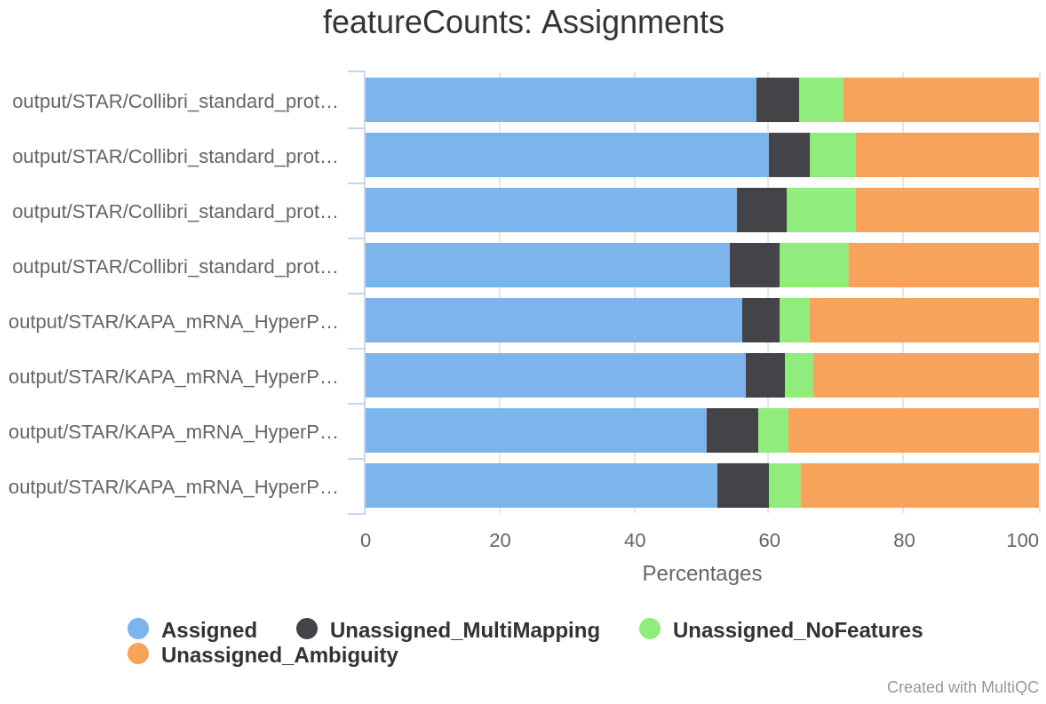

Figure 6 The percentage of assigned genes when s=1 for Collibri preparation method data and s=1 for KAPA preparation method data

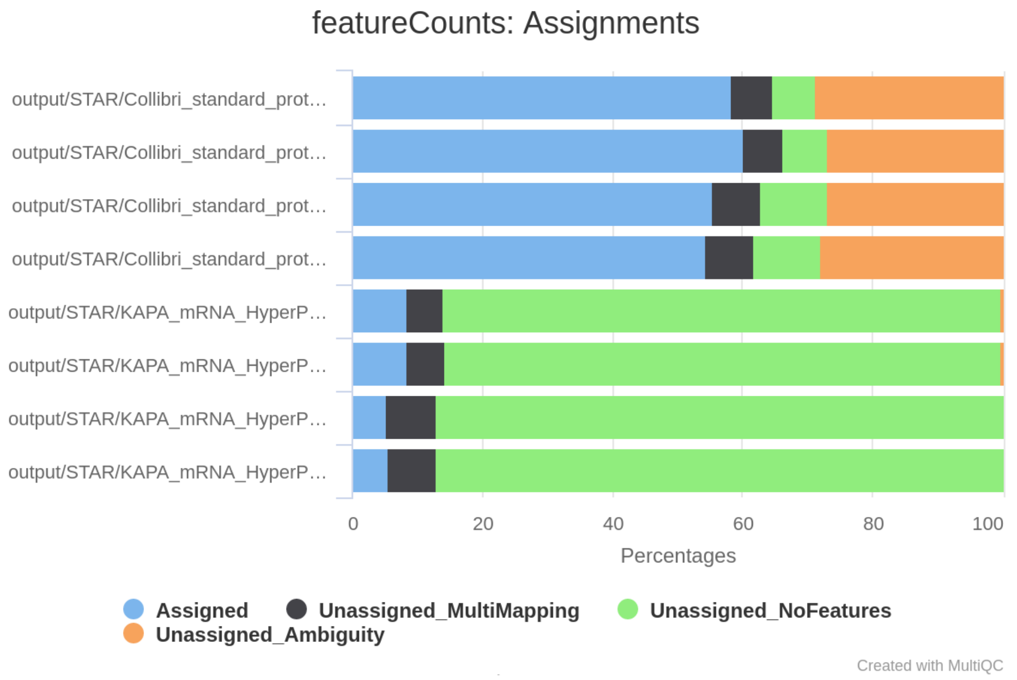

Figure 7 The percentage of asgined genes when s=1 for both Collibri and KAPA preparation method data

Differential gene expression analysis with DESeq2

Using tool DESeq2 we can find differentially expressed genes. This part wasn't included into snakemake workflow. Using snakemake feature counts were merged based on the sample preparation method and the following analysis were performed using R package DESeq2. This package provides methods to test for differential expression by use of negative binomial generalized linear models. Firstly, DESeq object was created. PCA plot was used on DESeq2 object. From it it can be seen that PC1 it self separates data into two groups based on tissue type (Figure 8)

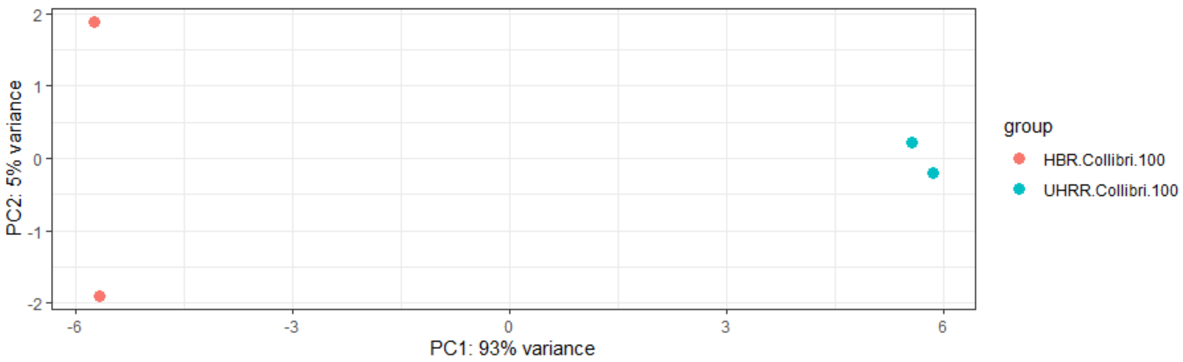

Figure 8 DESeq2 object for Collibri data

We can also plot counts of gene of lowest adjusted p-value (i.e gene with highest statistical significance between the two groups) (Figure 9). We can see that there is significant difference between samples from healthy people (HBR) and unhealthy (UHBR). This makes sense, because genes in unhealthy people samples may be mutated and therefore not mapped.

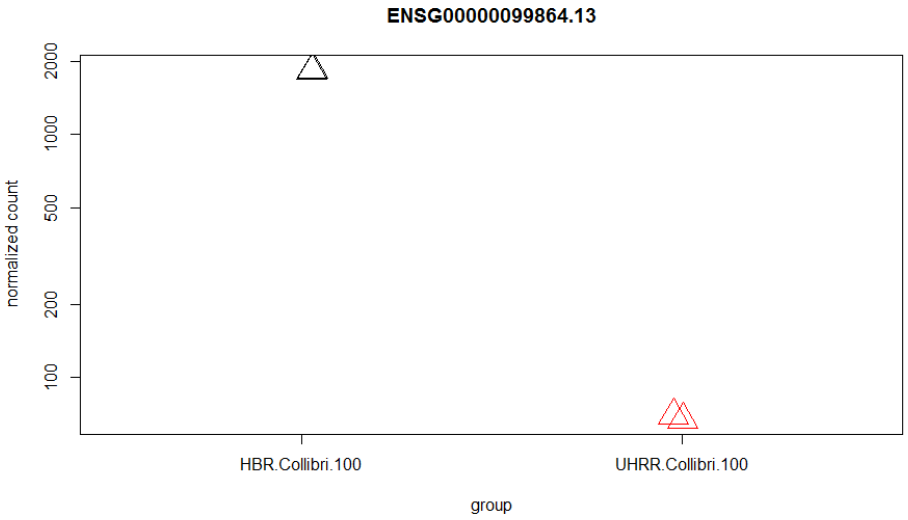

Figure 9 Counts of gene of lowest adjusted p-value. Collibri data

Finally we can plot Volcano plot Figure 10. Volcano plot is a type of scatter-plot that is used to quickly identify changes in large data sets composed of replicate data. It plots significance versus fold-change on the y and x axes, respectively. A volcano plot is constructed by plotting the negative log of the p value on the y axis (usually base 10). This results in data points with low p values (highly significant) appearing toward the top of the plot. So from this plot we can see, that some genes are differentially expressed in our groups taken to the consideration (HBR and UHRR).

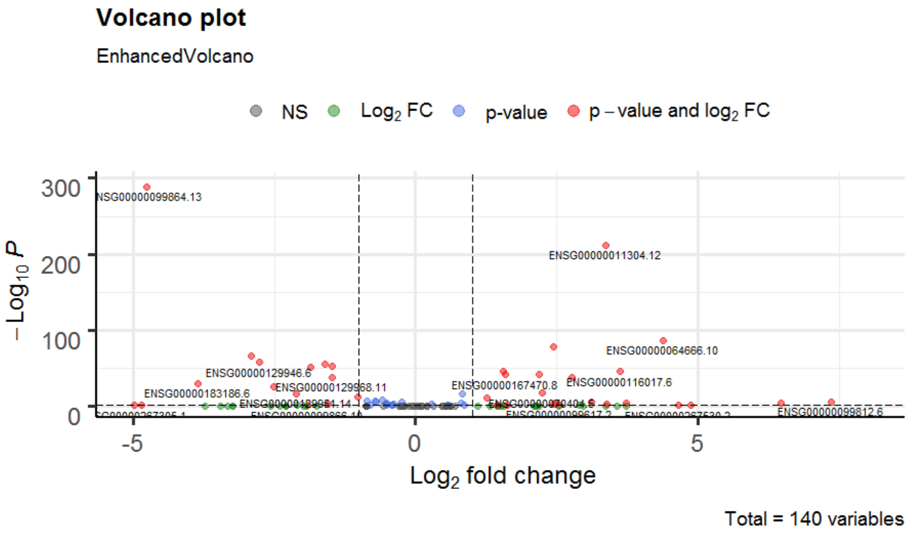

Figure 10 Volcano plot for Collibri data

The same steps as described before were also performed on KAPA preparation method data. You van see the Volcano plot in Figure 11. The genes with lowest adjusted p values are quite similar.

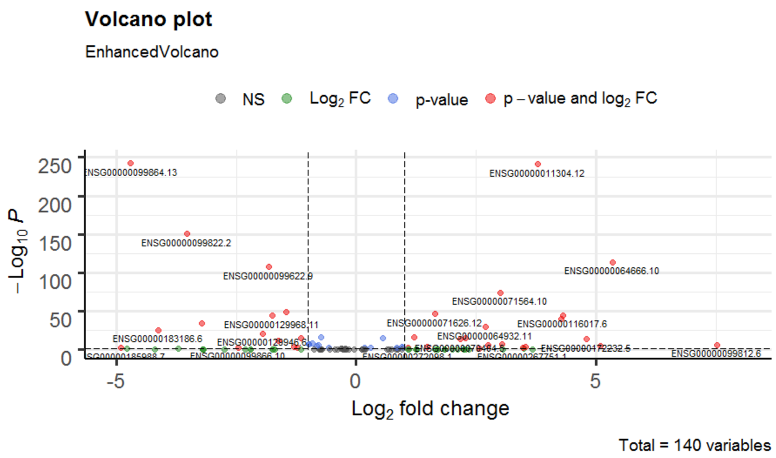

Figure 11 Volcano plot for KAPA data

That results are really similar can be proven by drawing Venn diagram. Here genes which has p value < 0.05 (are differently expressed in healthy and unhealthy tissues) were chosen to be compared. The diagram shows, that most of the differentially expressed genes are the same for both sample preparation methods.

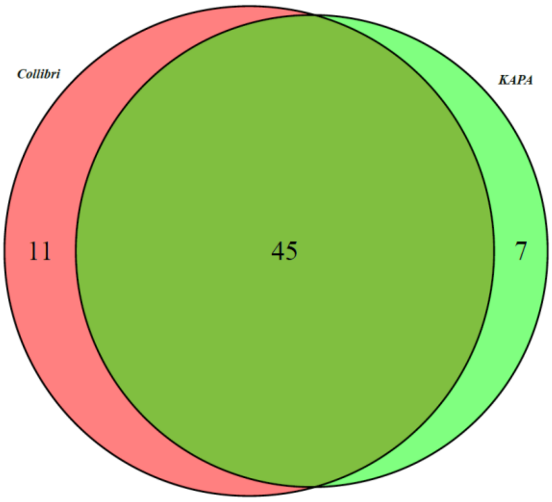

Figure 12Venn diagram illustrating differences for 2 sample preparation methods – KAPA and Collibri. Differentially expressed genes compared, p value <0.05

Gene Ontology (GO) term enrichment is a technique for interpreting sets of genes making use of the Gene Ontology system of classification, in which genes are assigned to a set of predefined bins depending on their functional characteristics. GOrilla is a tool for identifying and visualizing enriched GO terms in ranked lists of genes.

Conclusions

From RNA sequencing data we can find differentially expressed genes using tools described in this report (STAR, DESeq2).

References

https://digibio.blogspot.com/2017/11/rna-seq-analysis-hisat2-featurecounts.html

https://biobeat.wordpress.com/2013/05/27/how-to-draw-venn-diagrams-from-differential-gene-expression-data/

https://evodify.com/rna-seq-star-snakemake/

https://snakemake.readthedocs.io/en/stable/tutorial/tutorial.html