Bachelor Thesis - LIACS

Identifying key genes in CRC patients to predict efficacy of CAPOX treatment

Koen van der Burg

Supervisor:
Katy Wolstencraft
Second supervisor:
Christina Plattner

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

20/07/2022

Abstract

Colorectal cancer (CRC) has been identified as one of the most common cancers in the world; in America more than 150,000 new cases are expected in 2022. Since prevention is difficult, patients often undergo harsher treatments such as CAPOX (capacitabine and oxaliplatin). The treatment works well, but 15% of the patients still relapse after 2 years. This study will therefore be focusing on identifying key genes in CRC patients to predict efficacy of the CAPOX treatment. Two RNA-sequencing data sets from patients diagnosed with CRC and treated with CAPOX have been analyzed. A R-pipeline has been constructed to perform a Differential Expression (DE) and Gene Ontology (GO) analysis. The pipeline identified 234 significant DEG's (232 down-regulated, 2 up-regulated) in the first data set, and 3 up-regulated significant DEG's in the second data set. A total of 15 genes were identified as having a potential relationship with cancer, where a singular gene (*DRAXIN*) was directly related to CRC. Two of the identified genes (*MYTL1L* and *EML6*) are suggested to be down-regulated due to the CAPOX treatment, inviting further research. The lower quality and quantity of samples in the first data set, and group imbalance in the second data set might have caused biased results. Improvement of the pipeline is crucial for finding more definitive results.

Contents

1 Introduction

RNA-sequencing (RNA-seq), a technique for accurately measuring gene expression levels, has replaced the previously used micro-arrays to examine cell tissue expression with high accuracy, providing new insights. This new technique makes it possible to examine cell tissue expression with high accuracy, providing new insights. Identifying which genes are expressed enables the identification of genes may cause, or at least involved are in, the growth of cancer cells.

Colorectal cancer (CRC) is one of the most common malignant diseases in the world, and its incidence increased with age. According to estimates by the American Cancer Society, the number of new colorectal cancer cases in America for 2022 is 151.030 (106.180 colon cancers, 44.850 rectal cancers) [3]. Most CRC cases were related to old age and lifestyle factors, with only a fraction of cases caused by underlying genetic disorders [10]. Although multiple attempts have been made to grasp the underlying genetic mechanism for initiation and progression of CRC, the prevention of early on-set CRC is still not in sight [24]. Since prevention is not always a possibility, different treatments such as CAPOX have been used to treat stage II and III patients [11]. These later staged patients are harder to treat since the cancer has invaded the healthy tissue.

This paper focuses on the gene expression of cells in CRC tissue, from patients that were diagnosed with CRC and subjected to the CAPOX treatment. The name CAPOX is derived from the combination of drugs provided during treatment;Capecitabine and Oxaliplatin. After three years of follow up, patients reported on the success of the treatment; they had either relapsed or stayed cancer-free. The patients data, including a tissue sample and treatment outcome were collected and used for further research. Two different data sets are used, the first data set contains 85 samples, and the second data set 62.

The RNA-seq data can be used as input for a differential expression (DE) analysis. The DE analysis ultimately outputs differential expressed genes (DEG's), which can be identified as either up or down-regulated when compared to normal gene expression levels. The DEG's give a rough idea on the biological pathways affected in the cancerous tissue. Some of the more explored canonical pathways include; cell cycle, Hippo, Myc, Notch, Nrf2, PI-3-, Kinase/Akt, RTK-RAS, TGFb signaling, p53 and b-catenin/Wnt [21]. The biological purpose of the DEG's can be elucidated by executing a gene ontology (GO) analysis, consisting of gene enrichment and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway exploration.

The results from executing the DE and GO analysis should be reproducible when repeated at a later point in time. Both analyses require many small decisions, each affecting the final results. The results of the experiment cannot be reproduced if the same detailed structure is not applied. To streamline the processing of the data set and thus improve reproducibility, a pipeline has been created using Rstudio [7]. This pipeline takes in a variety of data from patients for pre-processing. During this phase the lower quality samples were identified and removed. The remaining samples were subjected to DE and GO analysis.

Besides reproducibility, it would be interesting is to identify a transcriptional profile that may predict the efficacy of the CAPOX treatment beforehand. There are many side effects when undergoing the CAPOX treatment which can become increasingly distressing when combined with other treatments such as radiotherapy [1]. If a predictive transcriptional profile can be identified, it may allow the selection of patients with the highest chance to respond to CAPOX. This would decrease the number of patients that relapse after going through intensive therapy. Therefore, the aim of this study is to construct a three-phase R-pipeline to identify genes possible of determining the efficacy of the CAPOX treatment on stage III CRC-patients beforehand.

To identify the genes of interest and construct a biological overview from the acquired data set multiple steps were performed. First a RNA-seq analysis is performed, the resulting data is then is put through pre-processing; in order to remove samples with low gene counts or low tumor cell content. After the pre-processing the differential expression analysis is ran using the DESeq2 package [18].This package estimates variance-mean dependence in count data and can be used to test for differential expression based on a model using the negative binomial distribution. From these estimates a selection of differentiated genes is generated. Using the derived DEG's a gene ontology analysis is done to biologically interpret the results. The GO analysis is performed utilizing the ClusterProfiler package [33]. This package is designed to facilitate semi-automated gene enrichment and KEGG analysis, working together with the DESeq2 package.

This paper will first provide background information on important topics in the context of already published research knowledge. Then the methods; pre-processing, DE, GO and the different phases of the pipeline will be fully explained in order to get a better grasp on the design of the pipeline. Subsequently, the results from executing the pipeline will be shown. These results are from two different data sets each going through, a slightly personalized, pipeline. The pipeline had to be slightly adapted, correcting for difference in input data. Finally, the acquired results are discussed at the end of the paper.

1.1 Thesis overview

Chapter 2, background information;
Chapter 3, methods description;
Chapter 4, results from executing the pipeline;
Chapter 5, conclusion;
Chapter 6, discussion and future outlook;
Chapter 7, description of definitions;

2 Background information

CAPOX is now a standard treatment for CRC patients, and has been proven to be highly effective for two years after administration [16]. Although the study lasts for 2 years, a year shorter than the data set currently used, it has shown the promising results of the CAPOX treatment.

In a recent study from, 2019 designed to identify potential key protein interaction networks and genes in early-onset CRC, 12 patients with CRC were included as well as 10 healthy control tissues [35]. A total of 131 DEG's were identified (108 up- and 23 down-regulated). These DEG's were subjected to a gene ontology functional enrichment analysis and KEGG pathways analysis. This functional enrichment analysis showed the classes of the genes or proteins that were over-represented and belong to a group with similar biological function. A KEGG pathway analysis is applied to determine the pathways the DEG's in the over represented group has effect on. They identified several genes which were suggested to be strongly implicated in CRC (*ACTA2*, *ACTG2*, *MYH11*, *CALD1*, *MYL9*, *TPM2* and *LMOD1*). These genes were involved in muscle contraction, vascular smooth muscle contraction and cGMP-PKG signalling pathway.

Another RNA-seq study on patients whom received oxaliplatin-based chemotherapy, called FOLFOX ,revealed several other important biomarkers. The analysis suggested that the expression of 58 genes correlated, negatively or positively, to oxaliplatin response. These genes were found to be mainly enriched in Wnt/ β-catenin signaling and EMT pathways [17]. It is known that Wnt/ β-catenin activation and malignant transformation of inflammatory bowel disease are the two major causes of colorectal cancer [13]. Moreover, the induced EMT pathway has an important role in chemo resistance of different types of cancer cells. The study suggested that high gene expression of the *FZD5* and *HNF1A* (TCF1) genes may be a predictor of the activation of upstream β-catenin. Furthermore, it suggested that a high expression of the NOTCH1 gene causes increased activation of the EMT pathway.

3 Methods description

The construction of the pipeline starts with the collection of the data. From each patient a singular tissue sample is taken during surgery on the CRC tissue. Each sample is analyzed and the sequence is saved as a FASTQ file [31]. The FASTQ files are then used for the nf-core pipeline [5]. Specifically, the nf-core/RNA-seq pipeline [6] is used to calculate the transcription counts of each expressed gene in the tissue for all of the patients. This step is not included in the R pipeline, but rather outsourced to the services of the nf-core pipeline. In the nf-core pipeline the FASTQ files are to be processed in multiple stages; pre-processing, genome alignment & quantification, pseudo alignment & quantification, post-processing and Multi-QC. The pipeline outputs a report where transcription counts of all the expressed genes can be found. The Multi-QC tool uses the Salmon method to quantify the expression of transcripts using the RNA-seq data. The nf-core pipeline provides a preliminary report with the Multi-QC tool [4], which supplies the user with a check to see if the RNA-seq went well and if further analysis is possible.

The transcription counts of 60,000 genes for all of the patients is needed for the DE analysis. The annotation table, containing the names of the samples corresponding to the transcription files, is also required. The annotation table matches the response to the treatment amongst various other details of each of the patients. Some of these details include information such as the minimum and maximum tumor cell content in the sample, or the gender of each patient. The response is needed to create biological replicates. Since each of the patients only has one sample taken and this sample has been tested once no technical replicates can be used. These replicates are necessary to achieve statistical power [2]. Since RNA-seq has been more reliable than the micro-assay, technical replicates are less of an issue. Biological replicates are needed for statistical power, as only one sample was taken, natural biological variation could not be assessed. Therefore 2 different response types (relapse, no relapse) have been used to provide this variation. A minimum of 10 patients in each group provides enough biological replicates for the analysis. With the correct input files for the pipeline ready, the data is processed in three different steps: quality control, differential analysis and gene ontology analysis.

3.1 Quality Control

The first phase is the quality control check. This first stage is implemented to remove samples that do not meet the set standards, and thus remove noise from the results. The set standards are based on a balance of quality and quantity. Samples in the first data set were removed when transcription counts were lower than 200,000 and the tumor cell content had not reached a minimum of 50%. The final filtering step is the selection of columns which hold important information. Quality control can be left out, however, is recommended to ensure better results.

To get a better understanding of the data we are working with three different plots are made; total count bar plot, response bar plot and a Principal Component Analysis (PCA) plot. The bar plots the samples that passed the standards from high to low, based on the total gene transcription counts in said sample. The response plot shows the different responses (relapse or no relapse) for each of the samples. The samples are ordered from the highest minimal tumor cell content to the lowest. Finally the PCA plot is generated [26]. Where different principal components (PC) can be plotted against each other. For example, the different responses or genders to discover whether certain components influence the clustering of data points.

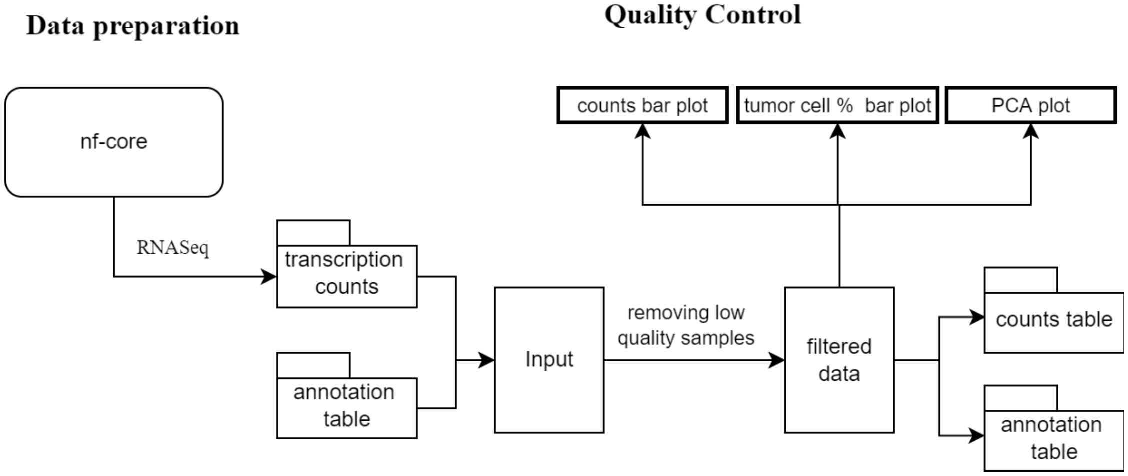

Figure 1: Workflow Quality Control

3.2 Differential Expression Analysis

Once finished with the quality control the second phase begins: differential expression analysis. The function of this part in the pipeline is to create a DEseq object for different plotting (PCA, Heatmap, Dispersion, Volcano & MA). If possible, analyze if there are significant DEG's present. DESeq2 is seen as one of the most reliable packages for RNA-seq data processing [22], making it one of the most popular packages.

The DE analysis begins with the DEseq object for two reasons; this object is used for visualization but also the processing of data. The input for the DESeq statistical model is non-normalized counts, because only the count values will allow the assessment of the measurement precision correctly. The DEseq2 model handles the normalization of the values internally, therefore, scaling the library beforehand will disturb the assessing capabilities of the model. The normalization is needed to cover deviances from sequencing depth and RNA composition. DESeq2 applies the median of ratios method to cover this. On the user-end, one line is needed to activate multiple steps on the back-end [27]. To create a DESeqDataSet (dds) object a design needs to be supplied. The design of the experiment will express the variables that are used for the modelling. Considering the objective is to find a difference between the two response groups (relapse, no relapse), the response variable design is chosen. By pre-filtering samples that do not meet the requirements memory usage by the dds object is lowered. Lower memory usage translates directly into increased speed and performance, for calculations and plotting.

With the dds object PCA plotting is also performed to observe possible changes. A heat-map is provided for the exploration of the count matrix. Here the samples are divided into two response groups. It will show whether there is a similar expression between the samples in the same group. Similar samples will have the same shade of coloring.

After examining the normalized data an adjusted p-value filter is added as a multiple testing correction is required. Each DEG is treated as a separate experiment, correcting for occurrences of false positives. The DESeq2 object provides the adjusted p-values by using the Wald test [8]. These attained values are corrected for multiple testing using the Benjamini and Hochberg method, to lower the false discovery rate (FDR) [14]. Then, a package called IHW is used to implement the method of Independent Hypothesis Weighting. This provides more power detecting genes, whilst controlling the FDR [15]. After the filter, a log fold shrinkage is applied. The shrunken log fold changes are used for ranking and visualization, without needing the arbitrary filters for the low count genes. A dispersion plot as well as MA plots are made to visualize the ranking and the differences between measurements taken in two samples, respectively.

Separating the significantly DEG's from the normally expressed ones is based on a pre-determined cut-off. This cut-off is based on the adjusted p-value and the log2FoldChange of the gene. Thus any gene where the adjusted p-value is under the set $\alpha = (0.05)$ and has a higher or lower log2FoldChange of 1 or -1 is either a up or down regulated DEG respectively. The selection is made and saved to the designated folder whilst running the pipeline. To visualize the DEG's a volcano plot with the same cut-off is made.

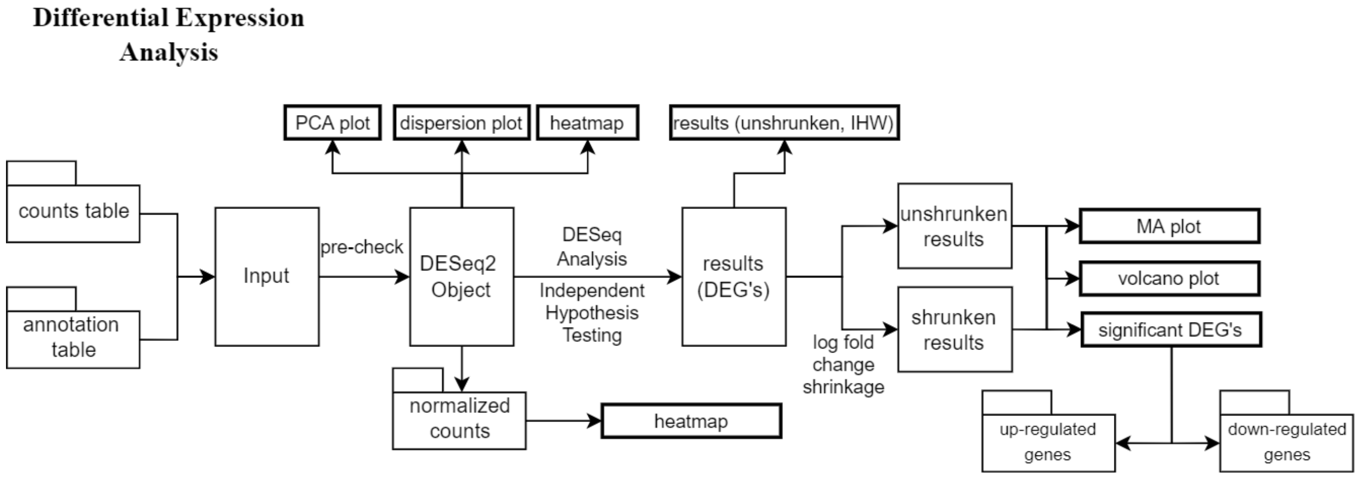

Figure 2: Workflow Differential Expression Analysis

3.3 Gene Ontology

The DE analysis delivers the DEG's needed for the last phase of the pipeline; the Gene Ontology analysis. The GO analysis, is used to derive groups with similar functionality from the DE analysis results. These DEG's tell us nothing just by name and individual function. Therefore, two different parts of the GO analysis are used for interpreting gene expression data; gene set enrichment and KEGG pathway analysis. Gene set enrichment derives its power from focusing on gene sets. In particular, groups of genes that share a common biological function, chromosomal location, or regulation. The enrichment score is the weighted Kolmogorov-Smirnov statistic, comparing the ranks of genes in the data set with the uniform distribution. These gene sets have been grouped *a priori*, based on their treatment response. KEGG pathway analysis uses a collection of manually hand drawn pathway maps which represent our knowledge of the molecular interaction, reaction and relation networks on multiple processes and systems. Some of these include; metabolism, genetic/environmental information processing, cellular processes, organismal systems, human diseases and drug development.

For both of the different analysis approaches different plots are made to visualize the results from the DEG's; dot plot, cnet plot, ridge plot and an emap plot. The dot plot shows the activated and suppressed pathways. It does this for the top ranking pathways, based on the adjusted p-value. Whilst the dot plot is used to display the most significant or selected enriched terms, we still want to know which genes are involved. The cnet plot is used to depict the linkages of genes and biological concepts known as a network, capable of considering the potentially biological complexities to which a gene may belong. In addition, an enrichment plot is created utilizing the emap plot module. A ridge plot is also provided to visualize the changes in distribution over time and space. Plotting is done by using the gseGO (gene enrichment) and gseKEGG (KEGG pathway) objects. Input preparation is needed to line up the encoded genes names into ENSEMBL gene names. Further transformation of ENSEMBL type to ENTEZID type is needed for KEGG analysis.

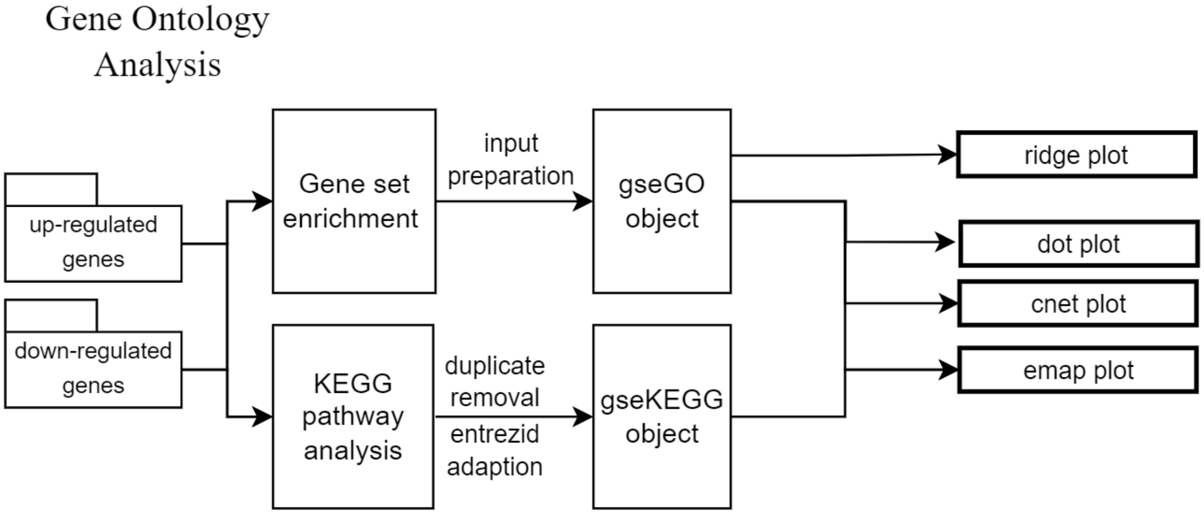

Figure 3: Workflow Gene Ontology Analysis

4 Results from executing the pipeline

4.1 Quality control

The first phase of the pipeline is to check the usability of the data set. This only applies to the first data set, the quality of the second data set was sufficient. To have statistical power enough samples (¿=10) need to be provided for each group (relapse, no-relapse). The total transcription counts bar plot (Figure 4) is a visualization of the samples and their corresponding transcription counts over all of the expressed genes. From a total of 85 samples 43 were left.

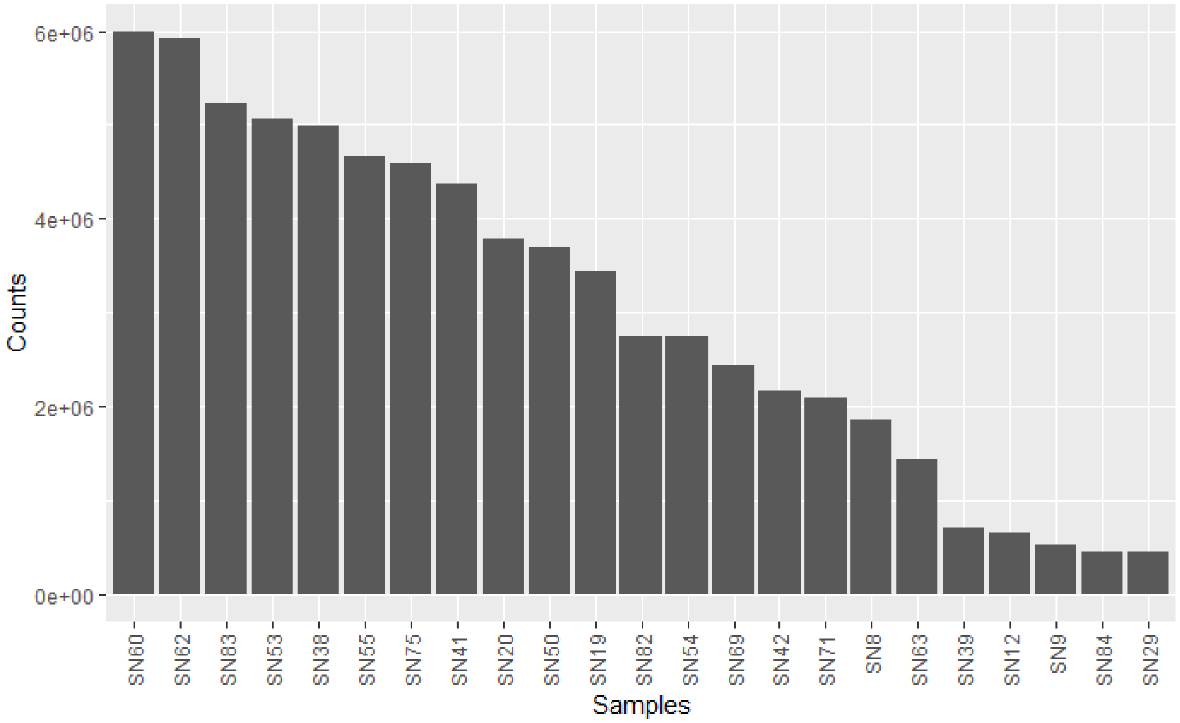

Figure 4: Total gene transcription count per sample, ranked highest to lowest.

Since 20 of the samples had low tumor cell content, they had to be removed. As seen in the bar plot (Figure 5), the total number of samples left are from 15 patients whom relapsed and 8 patients whom did not.

The PCA plot (Figure 6) supplies an early indication of division of the groups based on their response to the treatment. Here PC1 is plotted against PC2, where grouping of the patients with no relapse is detected, surrounded by a formation of relapse samples. The less explaining principal components (PC3 - PC8) showed no different results.

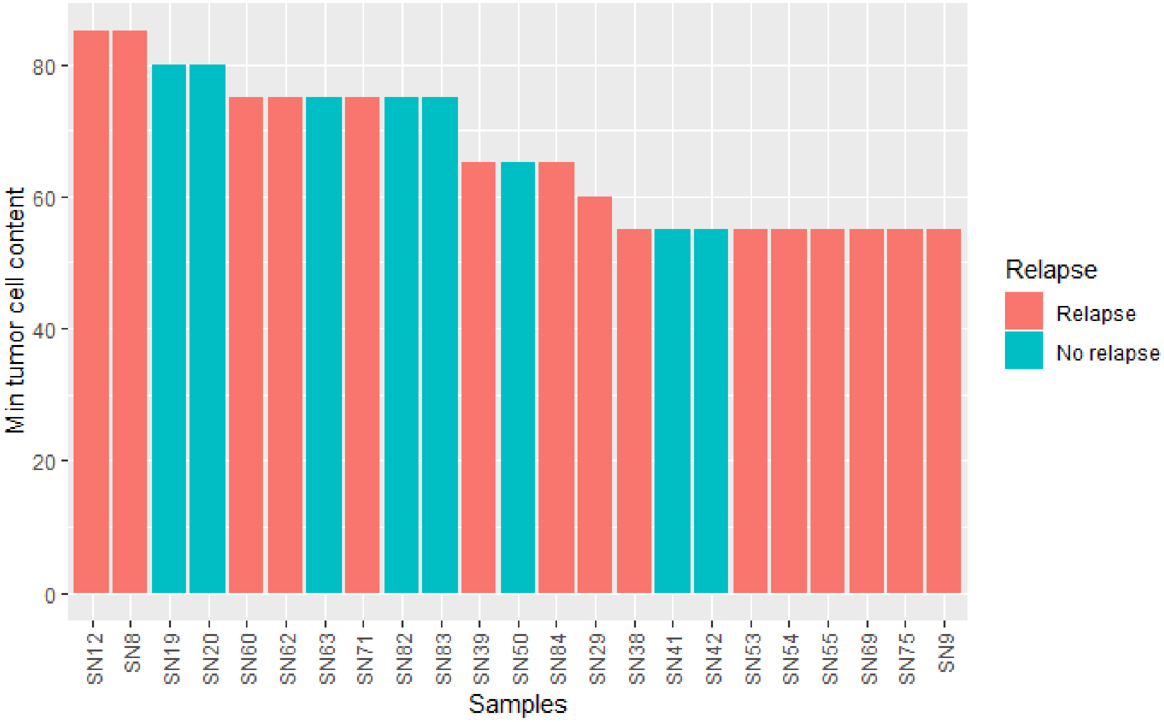

Figure 5: Samples are ranked on their minimal tumor cell content, colored by their response to the treatment. A minimum of 50% is required, and a maximum of 85% is reached.

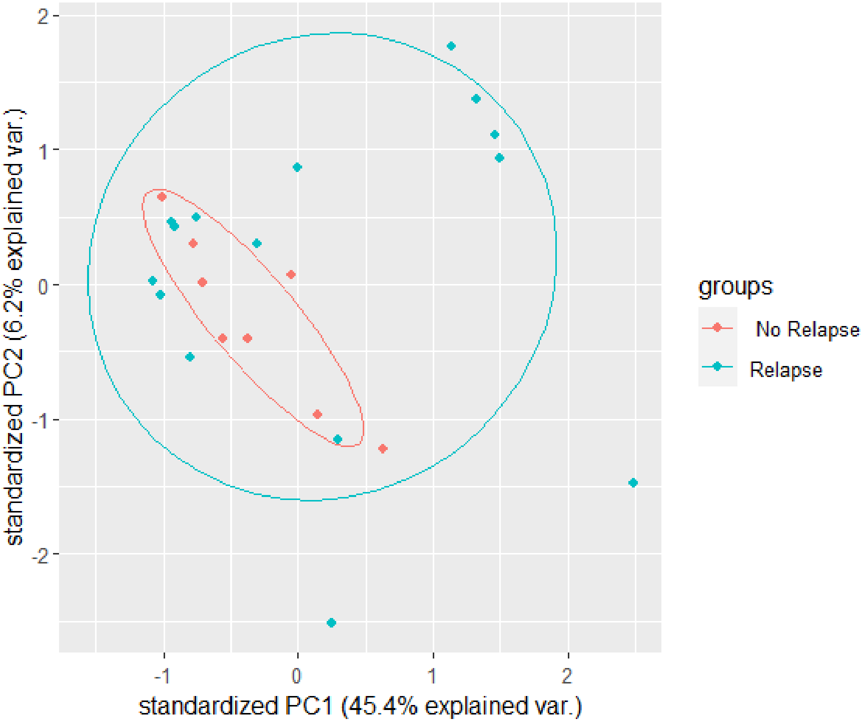

Figure 6: PCA plot displaying PC1 (45%) against PC2 (6.2%), based on patient response to treatment.

4.2 Differential analysis

The second phase is designed to retrieve the differentially expressed genes. The DESeq 2 object, mentioned in methods (3.2), is used to initially visualize 3 different plots; PCA, heatmap, and a dispersion plot.

PCA plotting has been done in the Quality Control phase, but since the DESeq2 object performs internal normalization and processing it is interesting to see if this has changed anything. Instead of separating the groups based only on their response (Figure 7), other variables such as gender (Figure 8) or the minimal tumor quality (Figure 9) of the sample, were also tested.

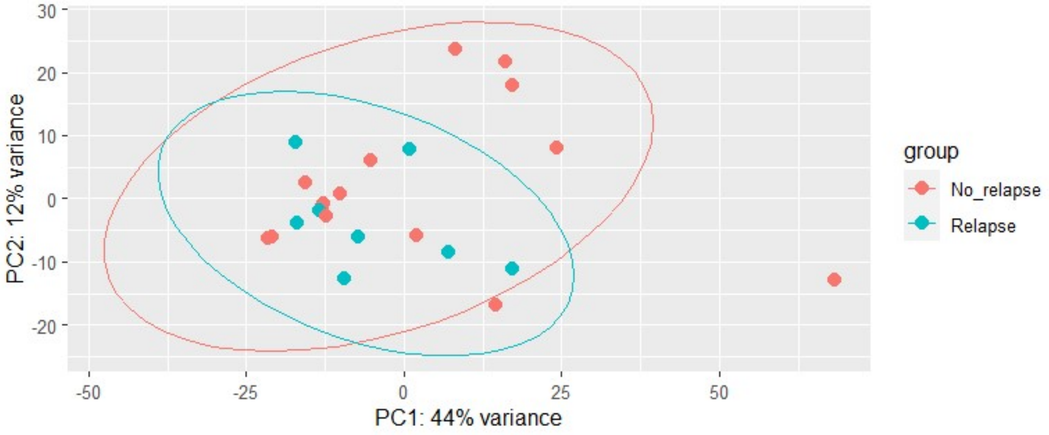

Figure 7: PCA - Response to treatment (relapse - no relapse)

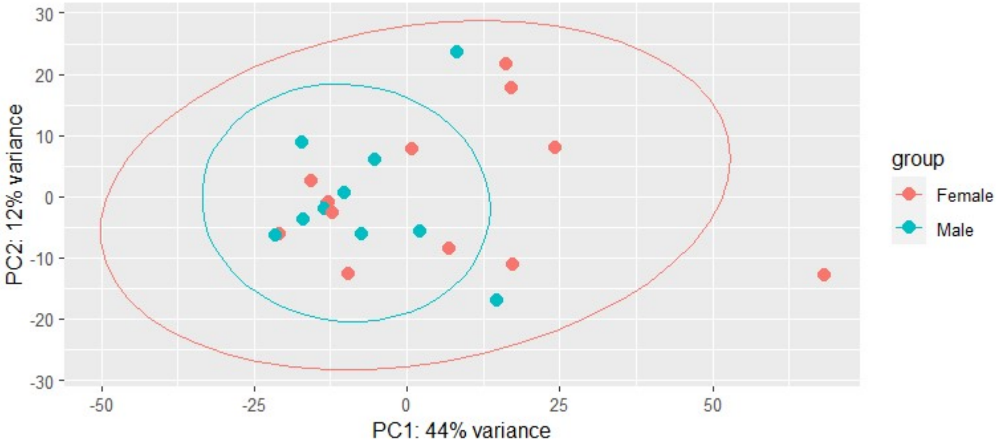

Figure 8: PCA - Gender (male - female)

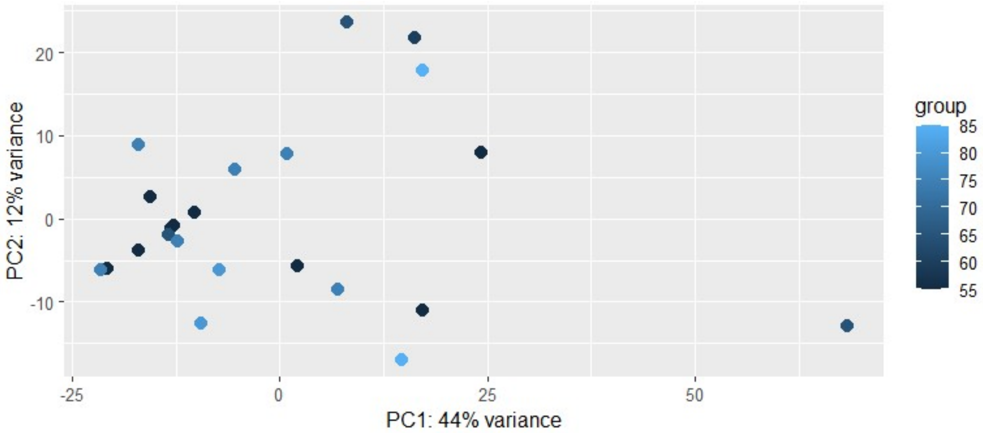

Figure 9: PCA - Minimum tumor cell content (%)

The heatmap (Figure 10) utilizes normalized gene expression counts but transforms these using variance stabilizing transformation (VST) to visualize the data. The r-log matrix is extracted from the transformed data to calculate the pairwise correlation between samples based on their response. A red color indicates a high correlation and blue a low correlation. The response of each sample is also denoted to show potential grouping. Some similar expression is seen in samples residing in the lower right corner, but no distinction between the two groups can be created, some relapse samples are included in the larger no relapse group.

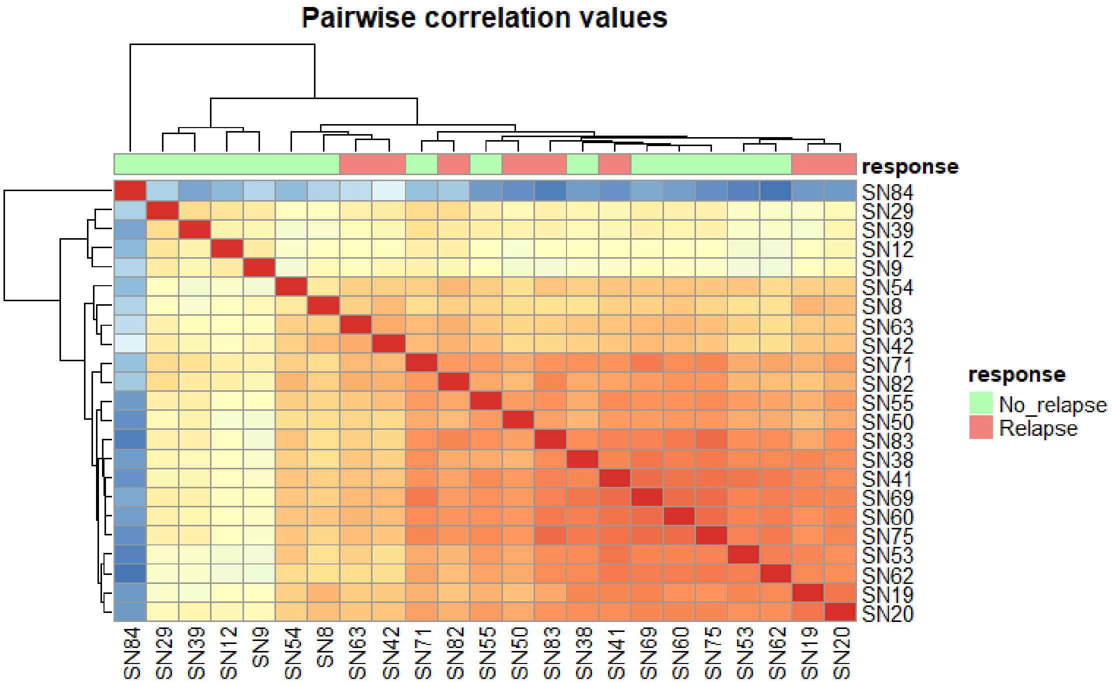

Figure 10: Heatmap based on pairwise correlation between samples.

To see the variance in gene expression a dispersion plot is created. The DESeq2 dispersion estimates are inversely related to the mean and directly related to variance. Based on these calculations the dispersion will be higher for smaller mean counts and lower for higher mean counts. If the mean count is identical, the dispersion will only differ based on their variance. Thus, the dispersion estimates reflect the variance in gene expression for a given mean value. The dispersion plot corresponding to the first data is shown below (Figure 11). Here a general scatter of genes around the curve (fitted line) is observed. A decrease in dispersion is seen with increasing expression levels, indicating that this particular data set neither contains contamination nor includes outlier samples anymore.

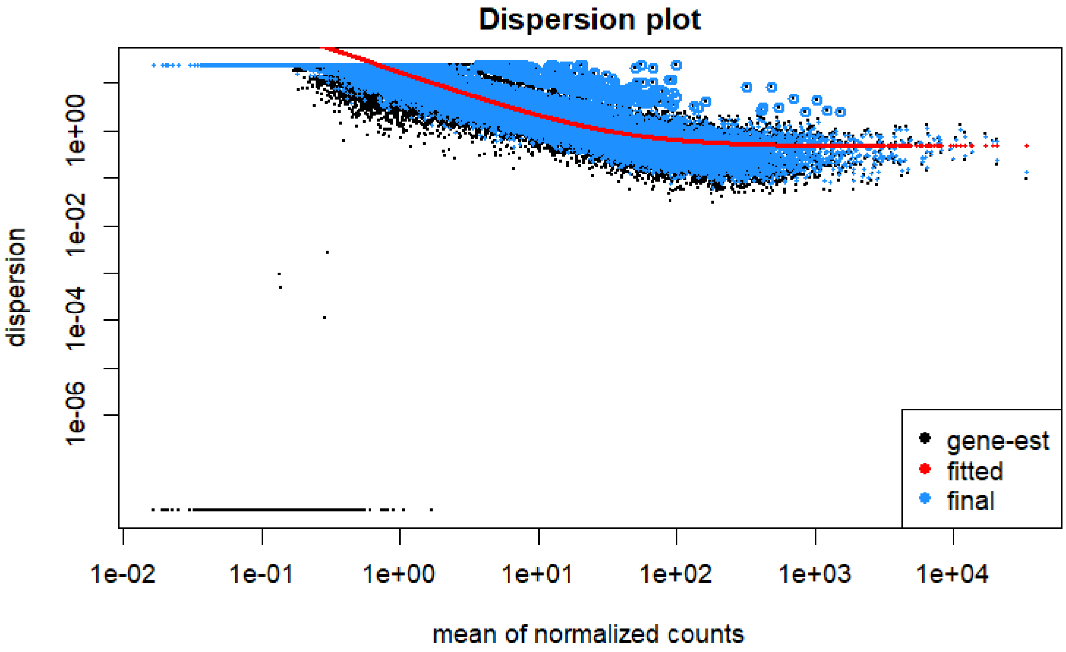

Figure 11: Dispersion plot

After extracting the results using IHW a MA plot is used. The plot shows the log fold-change against the mean expression based on treatment response. Each data point represents the gene and gives an indication of the up and down regulated genes. A comparison is made between the shrunken (13) and unshrunken (12) results, where the shrunken results had log fold change shrinkage applied. The shrinkage is applied to better select the differentially expressed genes. Large fold changes with high statistical power are not shrunk, whereas imprecise fold changes are shrunk. The unshrunken MA plot shows a lot of genes having a log fold change above and below 2 or -2 respectively, indicating that these are all significant DEG's. After applying shrinkage the genes are better fitted, creating a better distinction between non-significant and significant DEG's.

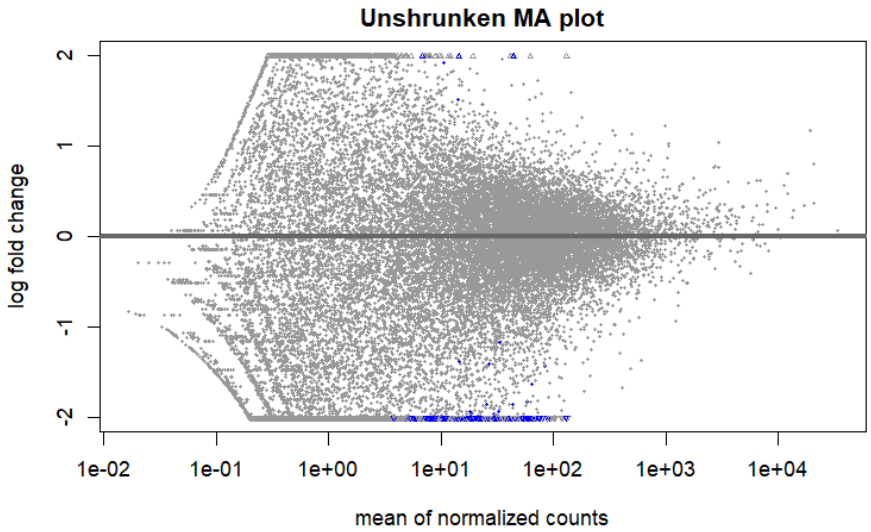

Figure 12: MA plot - no logfoldchange applied

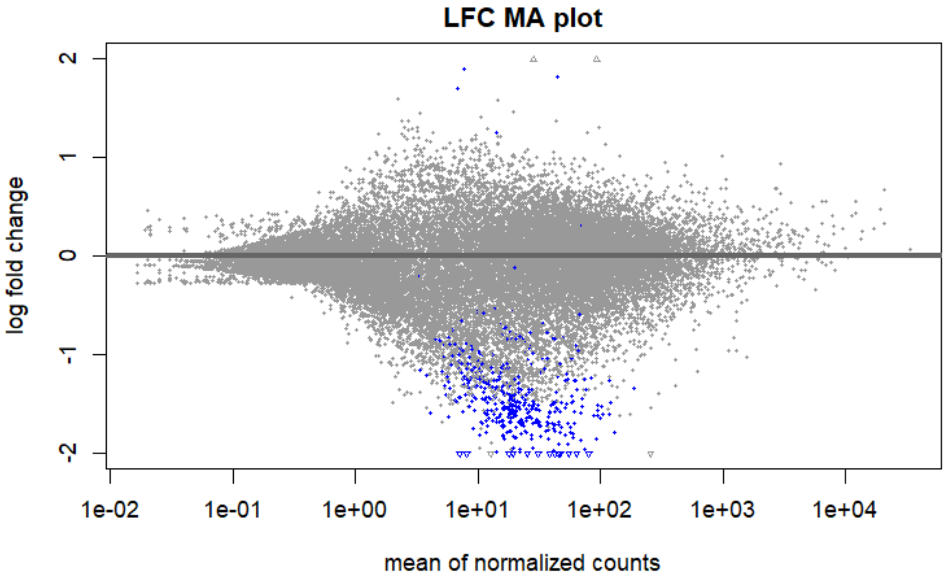

Figure 13: MA plot - shrunk by logfoldchange

The MA and dispersion plots have demonstrated that the data can be used for further analysis. Thus, significant DEG's can now be extracted. Using the pre-determined cut-offs ($\alpha = 0.05$, log2foldchange $= > 1$ or < -1) a selection has been made. This selection is better visualized in the Volcano plot (14). The cut-offs are illustrated by the dotted lines, the y-axis represents the p-value and the x-axis the log2foldchange for each gene. In total there are 28,973 genes plotted, where the significant DEG's are labelled red. Genes labelled green made the log2foldchange cut-off, but not the p-value. The genes labelled grey made neither cut-off. Thus, genes appearing in the upper left section represent the down-regulates genes, whereas genes in the upper right section represent the up-regulated genes.

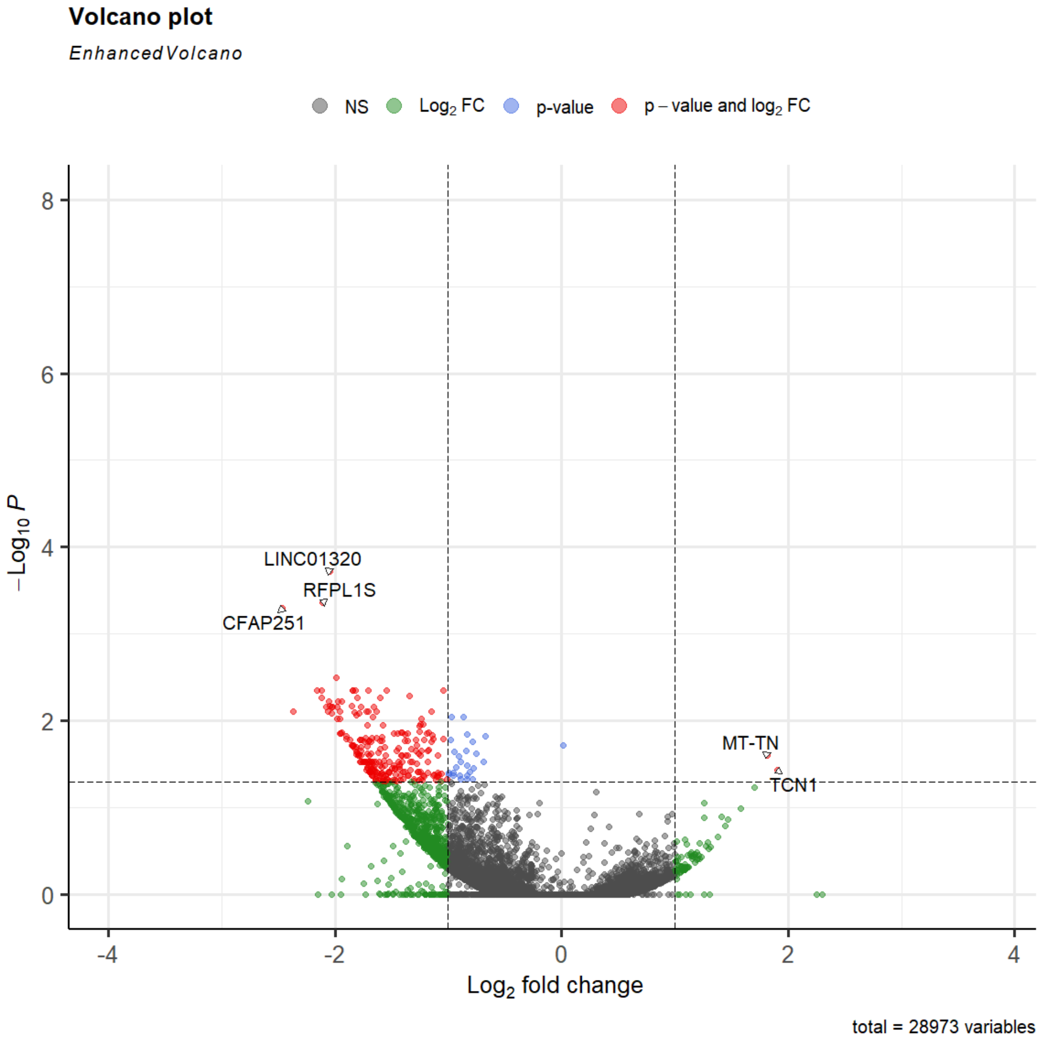

Figure 14: Volcano plot showing down and up-regulated genes in red, which can be seen in the upper left and right corner respectively. Genes in green or grey are not significant enough.

After extracting the significant DEG's, a heatmap (Figure 15) can be constructed using the normalized counts of their expressions. The expression is shown as a Z-score value, where a red cell represents an above average expression across all of the samples, and a blue cell an under average expression of genes. The top annotation shows the response of each patient. A very similar

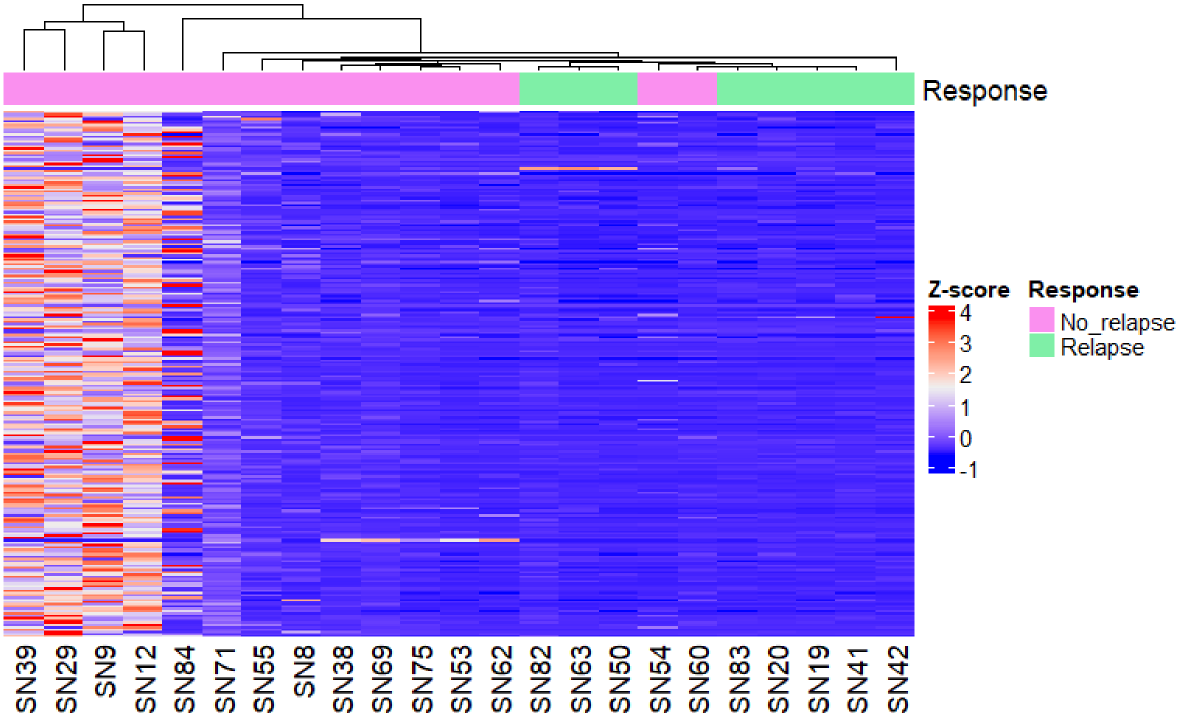

Figure 15: Heatmap - Pairwise correlation values

expression can be seen in the relapse group, however, since a part of the no-relapse group has comparable expression no distinction can be made.

4.3 Gene ontology analysis

4.3.1 Gene Set Enrichment

As mentioned before, gene set enrichment is one of the possible ways to analyze the significant DEG's. They gather their power from focusing on groups of genes that share a common biological function, location or regulation. A dot plot (Figure 16) is applied to gather the processes that are either activated or suppressed the most. Where the size of each dot represents the amount of genes involved and the color the adjusted p-value.

The ridge plot (17) visualizes the same processes as the dot plot, and focuses on the enrichment distribution specifically. The height of each wave represent the number of genes involved with that process at a certain logfoldchange.

The dot plot and ridge plot have presented some of the more important processes involved. To establish which genes and processes interact with each other an emap plot (Figure 18) and a cnet plot (Figure 19) are created. The emap plot uses the most involved processes and shows the relation between each of them. Where size and color represent the number of genes involved and adjusted p-value respectively. From these important processes the most represented are chosen, and their genes are made visible. Where the size of the processes represents the amount of connected genes, and each individual gene is colored by their logfoldchange.

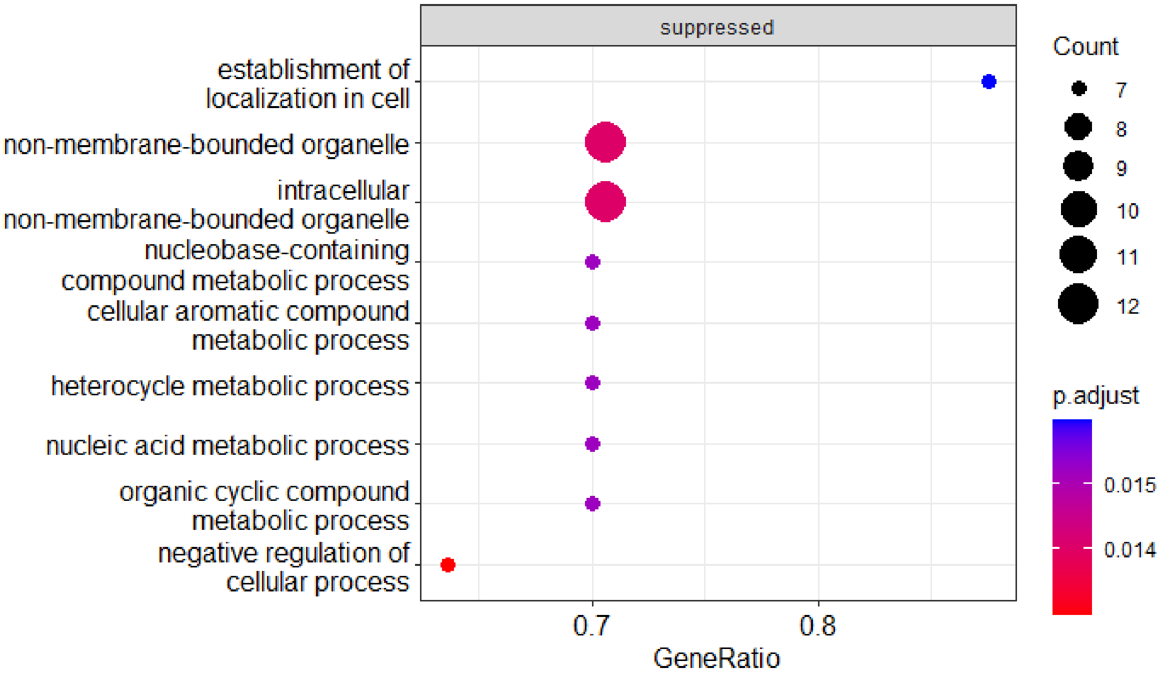

Figure 16: Dot plot, using the gseGO object. Showing the top 8 affected processes.

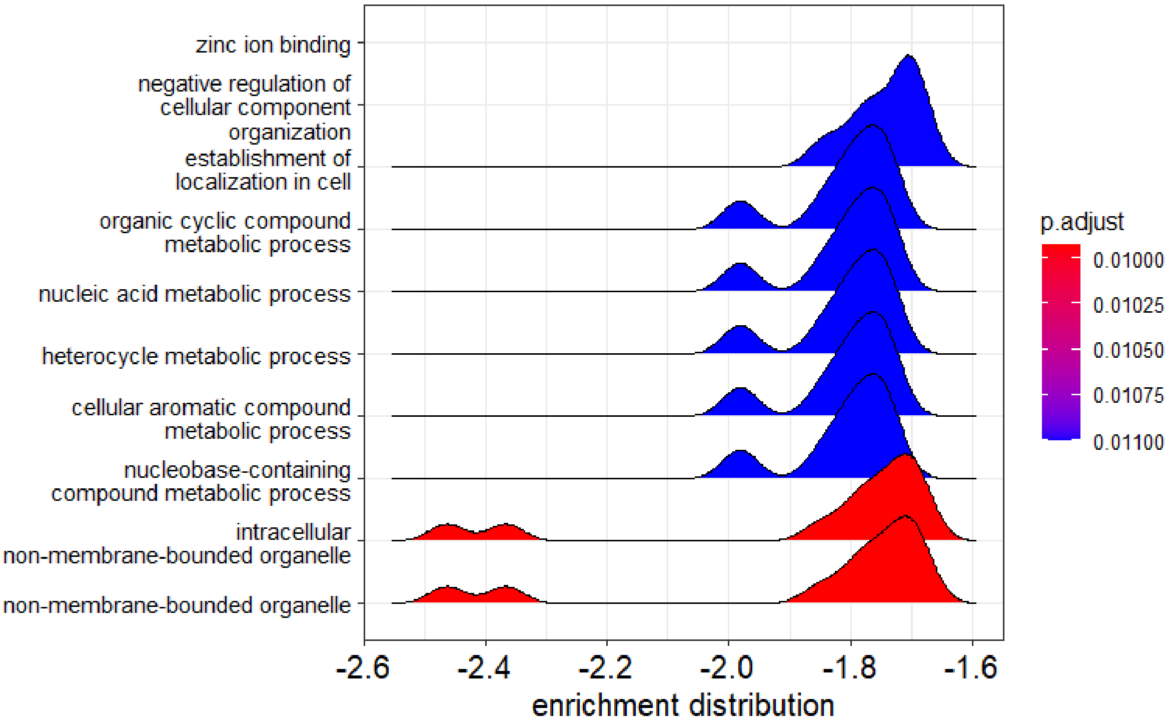

Figure 17: Ridge plot, using the gseGO object. Showing the top 10 affected processes.

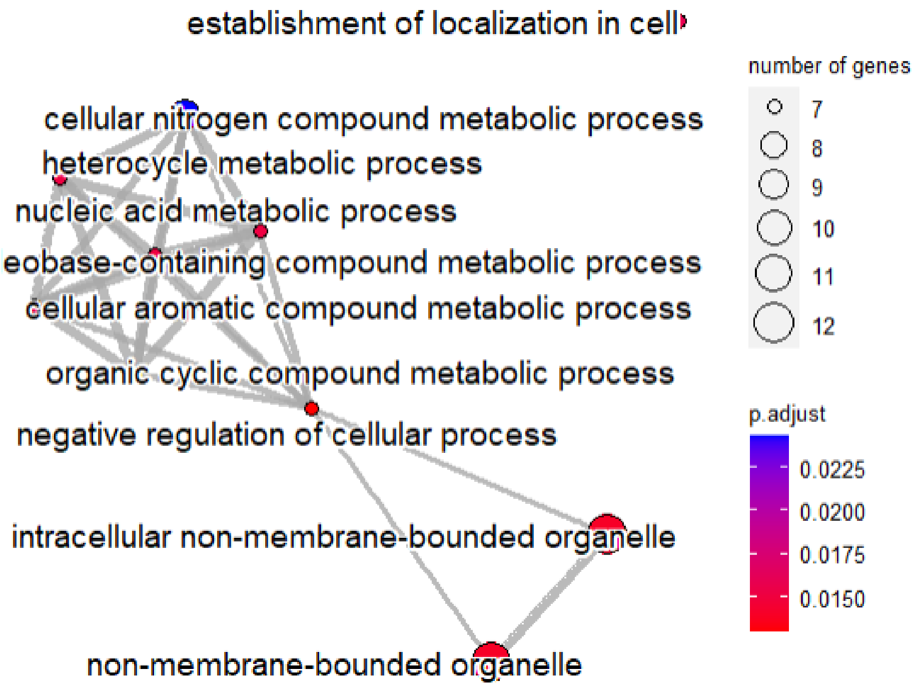

Figure 18: Emap plot, using the gseGO object.

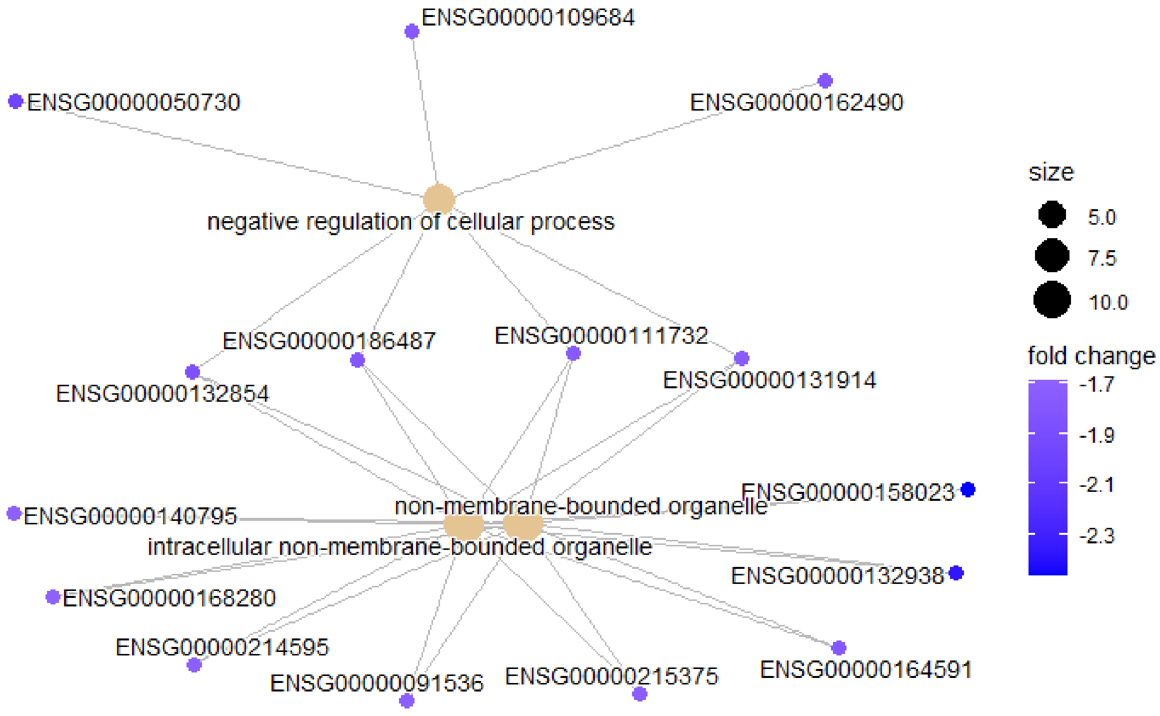

Figure 19: Cnet plot, using the gseGO object.

From the GO enrichment analysis, a few interconnected genes have been identified. These genes were related to two processes: the negative regulation of cellular processes (NRCP) and processes in the (intracellular) non-membrane-bounded organelle (NMBO). In the table below (Table 1) the identified genes are displayed together with their function and connected process.

Gene	**Function**	**Process**
TNIP3	Supression in NF-kB signaling and sustained NF-kB activity.	NRCP
CLNK	Positive regulation of immunoreceptor signaling.	NRCP
DRAXIN	Negative regulation of canonical Wnt signaling pathway and neuron apoptotic process.	NRCP
KANK4	Regulation actin polymerization + cell motility (PI3k / Akt pathway).	NRCP + NMBO
MYT1L	Repress expression YAP1 (proliferation / glioblastoma growth).	NRCP + NMBO
AICDA	Chronic inflammation and skin cancer.	NRCP + NMBO
LIN28A	Deviated regulation of this gene is reportedly involved in cancer development.	NRCP + NMBO
MYLK3	Phosphorylation cardiac myosin heavy chains. Associated with improved overall survival in patients with low residual disease.	NMBO
KIF5C	Transport of cargo within central nervous system.	NMBO
EML6	Part of EML family, control of oocyte meiotic division.	NMBO
MYO15A	Codes for unconventional myosin. Associated with hearing impairment.	NMBO
MYL5	Codes for myosin light chains, overexpression promotes metastasis in cervical cancer models.	NMBO
MYOZ3	Important role in modulation of calcineurin signaling.	NMBO
MTUS2	Breast cancer tumor suppressor gene.	NMBO
CFAP251	Formation protein-protein complexes. Cause morphological abnormalities of the flagella.	NMBO

Table 1: Down-regulated genes identified by the cnet plot during Gene Enrichment Analysis

4.3.2 KEGG Pathway

KEGG pathway analysis represent our knowledge of the molecular interaction, reaction and relation networks on multiple processes and systems. Analyzing the data via KEGG is relatively similar to the Gene Enrichment Analysis (GEA), as demonstrated in the GO phase of the pipeline. A dot plot (Figure 20) showing activated and suppressed pathways. An emap plot (Figure 21) and a cnet plot (Figure 22) are displayed to show the connected pathways.

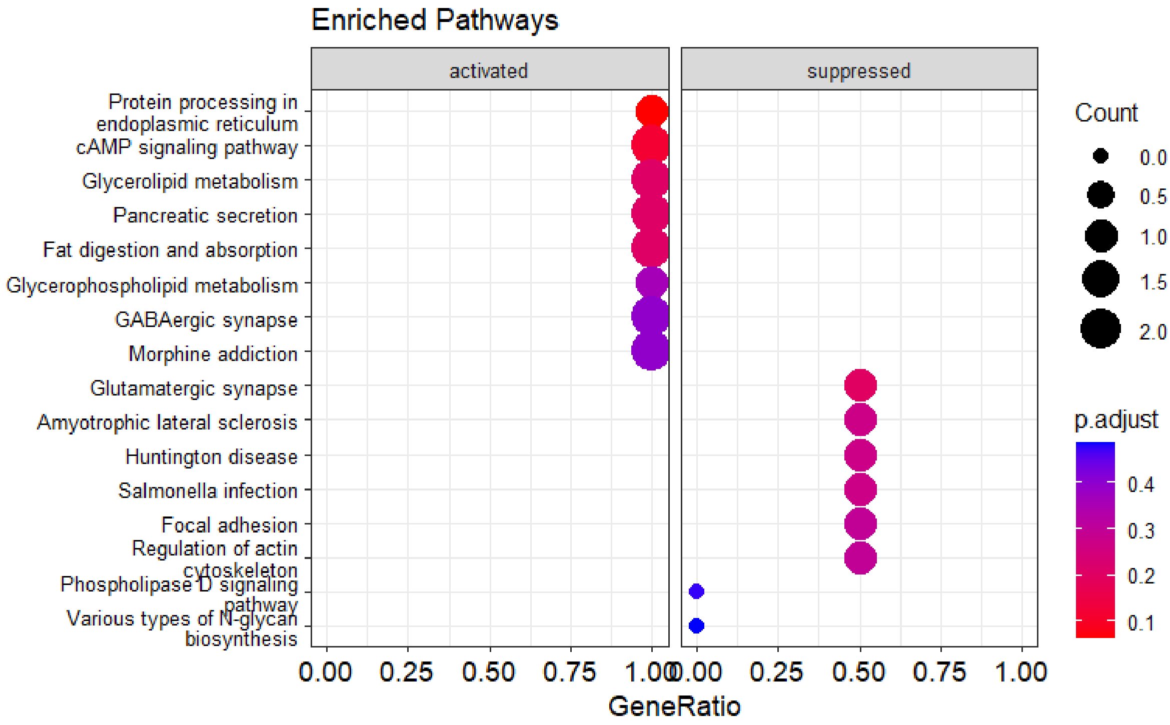

Figure 20: Dot plot, using the gseKEGG object.

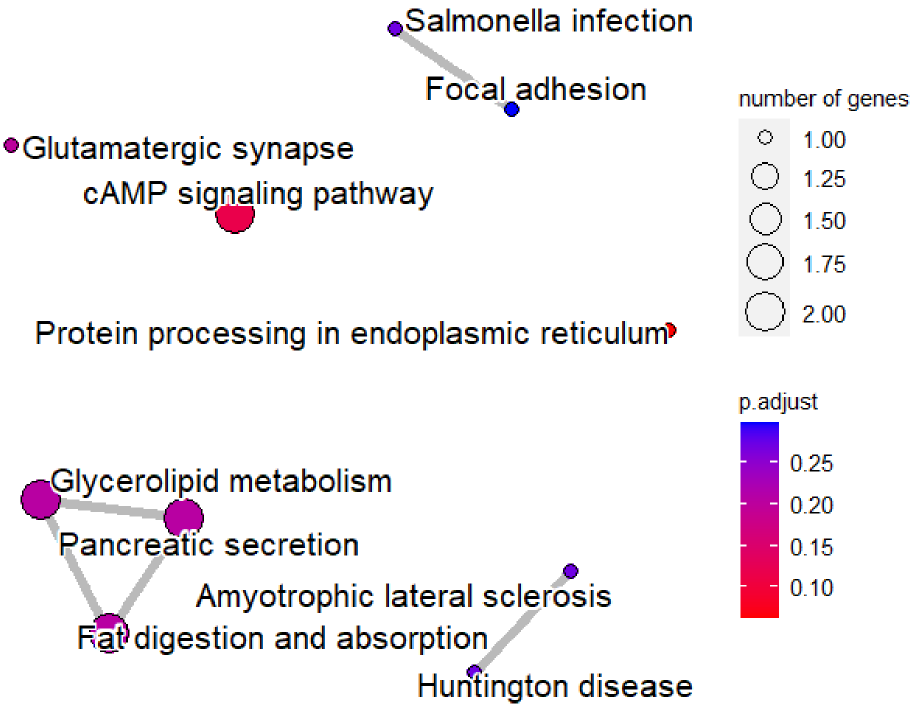

Figure 21: Emap plot, using the gseKEGG object.

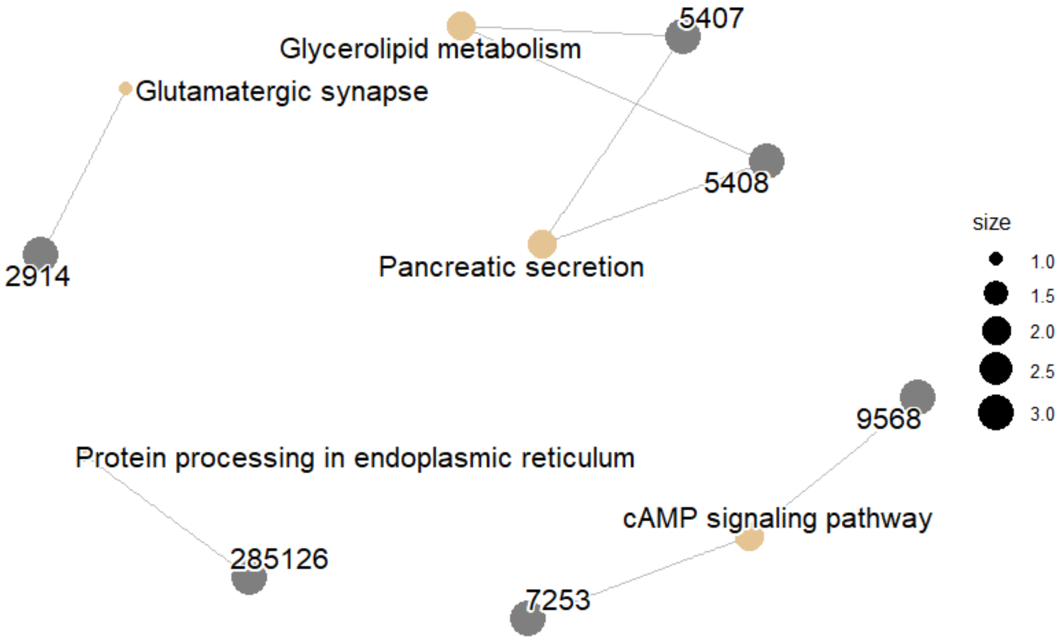

Figure 22: Cnet plot, using the gseKEGG object.

4.4 Analysis of second data set

The results presented originate from the first data set. The second data set, which had no sample loss during the quality control, identified 3 DEG's; *CCDC88B*, *RNU2-63P*, *HMGB1P21*. The leading gene codes for a protein hook, linking organelles to micro-tubules. The latter two are pseudogenes, which are non-functional segments of DNA. No pathways could be identified because only 1 functional DEG has been identified, which is not related to a cancerous activity or pathway.

4.5 Pipeline performance

The final pipeline consists of approximately 500 lines in Rstudio, producing reliable and reproducible results when performed on multiple occasions. The speed depends mostly on the size of the data set and hardware. Performing a single run on the first data set (23 samples, 62,000 involved genes) has a maximum duration of 5 minutes, making it a quick tool to use. The second data set, containing almost tree times more samples than data set 1, has a duration increase of 15%. There were less DEG's to process, thus the GO analysis was performed quicker. The pipeline provides multiple checkpoints for the user to detect whether there are mistakes in the data causing it to be noisy. It can be used interchangeably with other data sets, with small changes needed in the early processing (data preparation, quality control). These changes are mostly due to the difference in data input, where one data set might require a quality control check, and others having different informative annotations (gender, age, pre/post-treatment). Further customization for plotting is possible where needed, having the fundamental processes already put in place (DESeq2, gseGO, gseKEGG).

5 Conclusion

The pipeline was provided with two different data sets. Data set 1 contained a total of 85 samples (27 relapse, 58 no-relapse) and data set 2 a total of 62 samples (10 relapse, 52 no-relapse). Since a part of the samples in the first data set were of a low quality, the data was subjected to pre-processing, leaving 23 samples (8 relapse, 15 no-relapse). After running the pipeline a total of 234 significant DEG's (232 down-regulated, 2 up-regulated) were identified in the first data set. The data set 2 provided a total of 3 up-regulated DEG's. After the inspection of their potential relationship with cancer, no DEG's remained. Therefore, only results from data set 1 have been displayed and discussed.

The GO analysis showed a high suppression of pathways concerning the (intracellular) non-membrane-bounded organelle and the establishment of localization in the cell. Some metabolic processes (nucleobase-contraining compound, cellular aromatic compound, heterocycle, nucleic acid, organic cyclic) were also suppressed, but the number of genes involved was lower. Furthermore, pathways concerning protein processing, cAMP signalling, glycerolipid metabolism and glycerophospholipid metabolism were activated. These pathways are crucial for intracellular signalling processes and protein regulation. Pathways concerning muscle control (amyotrophic lateral sclerosis, actin regulation), the nervous system (glutamatergic synapse) and the phospholipase D signaling pathway (intracellular protein trafficking, cell proliferation, cell survival) were all suppressed. Related work indicated that certain genes involved with muscle contraction are implicated in CRC, but these genes differ from the genes found in this data set.

Genes found in the gene set enrichment (Table 1) were not related directly to the genes reported by others in CRC ([35], [13]). Related work demonstrates that genes involved with the Wnt signalling activation can be a major cause of CRC. *DRAXIN*, a gene found in the analysis, is involved in the negative regulation of Wnt signaling [9]. Genes such as *LIN28A* [30], *TNIP3* [23], *KANK4* [20], *CLNK* [29], *MYT1L* [19], *MYL5* [34] and *MTUS2* [12] are, to some extent, involved with other cancers or pathways involved with these types of cancers, but do not come up in literature as CRC involved genes.

Patients underwent a CAPOX treatment, thus it is of importance to inspect the possible resulting disruptions. Capecitabine and Oxaliplatin, the main working ingredients of this treatment, both affect the tumor cells processes. In particular, it disrupts cells repairing their DNA and interferes with the development of DNA in the cell, stopping cell division [17] [28]. *MYT1L* [19] and *EML6* [32]are involved with proliferation and oocyte meotic division, respectively. Both of these genes are down-regulated, suggesting that these processes are suppressed.

6 Discussion & Future outlook

Although the constructed pipeline works well, improvements can be made. To start, being able to increase the size of the data sets without running out of memory. The current set-up uses a significant amount of memory, which is not a problem when running a single data set of 60 samples. But when raising this to 100 samples the pipeline requires too much memory from Rstudio and the hardware, causing it to terminate the pipeline. Increasing the available memory for Rstudio is a great theoretical solution, but this is not practical since this needs to be done individually, making it less user-friendly. The better approach is to increase the pre-processing to a degree where samples that do not provide enough information can be excluded. Genes where there is none or almost no expression could also be excluded to reduce memory needs.

The second improvement is to further streamline the process. Some calculations are done twice and steps could be left out if shown not to be of importance earlier on. Further improved streamlining is to regulate the importing and exporting of files more structured. There is no easy toggle or code chunk that can be run to save or import files. This is equally true for the input and output pathways, which are not strictly defined in the beginning, decreasing user-friendliness.

The last improvement is to expand the functionality of the pipeline and explore other options. There are many packages for R to view and process the data, each with their own advantages and disadvantages. Increasing the number of plots generated in the pipeline can provide the user with more information, making use of the pipeline more generally. An option is to explore immunotherapy pathways. A package that has been developed, called RImmPort, enables ready-for-analysis immunology research data. This package is designed for immunological data, thus a similar package could be designed for this analysis.

As seen in the results, some DEG's have been identified to be associated with cancer, and fewer directly with CRC. The same can be stated for the identified pathways, where even fewer examples could be retrieved. The division of samples could be part of the reason why many non-cancer related genes have been found. DESeq2 uses between sample normalization to identify significantly expressed genes, and prefers to have at least 10 biological replicates per group. The first data set was split into a group of 15 and 8, slightly unbalanced, but possibly enough to still have some noise in the data. Because the relapse group contained less than 10 samples, it might have prevented us from seeing which other genes are actually expressed differently. The second data set had an bigger difference (10 vs 52), which might be the cause for finding almost no DEG's. A solution could be to run a 10 v 10 on the second data set, keeping the groups balanced.

Not only was the imbalance a problem, but the quality of the first data set was also sub-optimal. A minimum of 50% of the contents were actually tumor cells in some of the samples. This leaves room for error, but increasing the quality would decrease the quantity too much. Having a setup where both data sets could be run together would produce enough samples to get results. The related work identified genes 2 years after the treatment, whereas the data used in this study identified them after 3 years. This difference could cause the dissimilarity in results.

It was not possible to identify key genes in CRC patients that would help understand when the CAPOX treatment would be effective. The pipeline being able to process the data and produce informative visuals was given the highest priority, therefor further research was beyond the scope of this paper. Genes and important pathways have been identified that are related to other published work, inviting further research. The next step is to trace back the genes found, to the samples that highly expressed them differently. If a difference in expression is seen between the two groups, some indication of a transcriptional profile can be constructed. Since these genes do not explain much or relate to other work, priority should be given to upgrading the existing pipeline to handle bigger data sets.

7 Description of definitions

CAPOX - CAPOX (which also goes under the name XELOX) is used to treat bowel cancer and thus CRC. The name originates from the combination of therapy drugs which is given. The CAPOX treatment is made up of capecitabine (CAP) and oxaliplatin (POX).
Technical replicates - Technical replicates are repeated measurements of the same sample that represent independent measures of the random noise associated with protocols or equipment.
Biological replicate - Samples that have been obtained from biologically separate samples. This can mean different individual organisms (e.g., tissue samples from different mice), different samplings of the same tumour. i.e., if there are triplicate non-relapse samples, a biological replicate would be testing samples 1, 2 and 3 of the non-relapse group
IHW testing - Hypothesis weighting improves the power of large-scale multiple testing. We describe independent hypothesis weighting (IHW), a method that assigns weights using covariates independent of the P-values under the null hypothesis but informative of each test's power or prior probability of the null hypothesis. IHW increases power while controlling the false discovery rate and is a practical approach to discovering associations in genomics, high-throughput biology and other large data sets.
RNA-sequencing - RNA-sequencing is a relatively new method to detect and quantify transcriptome-wide gene expression. Where micro-arrays were first used, this new replacement proved to be easier and cheaper to implement. Furthermore, the quality off the new sequencing method was significantly better. [25]

References

[1] Cancer information and support - xelox (or capox). https://www.macmillan.org.uk/cancer-information-and-support/treatments-and-drugs/xelox-or-capox.

[2] Difference between technical and biological replicates. https://altogen.com/difference-technical-biological-replicates/.

[3] Key statistic for colorectal cancer. https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html.

[4] Multiqc tool. https://multiqc.info.

[5] Nf-core. https://nf-co.re.

[6] nf-core rnaseq. https://nf-co.re/rnaseq.

[7] R studio workbench. https://www.rstudio.com.

[8] Wald test. https://en.wikipedia.org/wiki/Wald_test.

[9] Giasuddin Ahmed. Draxin inhibits axonal outgrowth through the netrin receptor dcc. https://www.jneurosci.org/content/31/39/14018.short.

[10] Yunhan Ma Bin Zhao, Zulqarmain Baloch. Identification of potential key genes and pathways in early-onset colorectal cancer through bioinformatics analysis. https://www.cancer.org/cancer/colon-rectal-cancer/about/key-statistics.html.

[11] Amy Kaji Carrie Luu, Jonathan Velasquez. Effectiveness of capecitabine plus oxaliplatin for advanced colon cancer: A public hospital experience. https://ascopubs.org/doi/10.1200/jco.2013.31.4_suppl.570.

[12] Michael R. Green Diane Fru, Victoria Ruhl. Abstract b29: Characterizing candidate breast cancer tumor suppressors: Mtus2 and lhx8. https://aacrjournals.org/cancerpreventionresearch/article/5/11_Supplement/B29/29968/Abstract-B29-Characterizing-candidate-breast.

[13] Yongchen Guo. Identification of key candidate genes and pathways in colorectal cancer by integrated bioinformatical analysis. https://www.mdpi.com/1422-0067/18/4/722/htm, March 2017.

[14] Winston Haynes. Benjamini-hochberg method. https://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_1215.

[15] Nikolaos Ignatiadis. Dge count normalization. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4930141/#:~:text=Independent%20hypothesis%20weighting%20(IHW)%20is,of%20p-values%20and%20covariates.

[16] Ashok V Kalidindi. Efficacy and safety of capecitabine and oxaliplatin (capox) treatment in colorectal cancer: An observational study from a tertiary cancer center in south india. https://pubmed.ncbi.nlm.nih.gov/33402602/, January 2022.

[17] Xin-Xiang Li. Rna-seq identifies determinants of oxaliplatin sensitivity in colorectal cancer cell lines. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4128987/, June 2014.

[18] Michael Love. Package: 'deseq2'. https://bioconductor.org/packages/devel/bioc/manuals/DESeq2/man/DESeq2.pdf, June 2022.

[19] Tiffany A. Melhuish. Myt1 and myt1l transcription factors limit proliferation in gbm cells by repressing yap1 expression. https://www.sciencedirect.com/science/article/abs/pii/S1874939918302566.

[20] N.Kakinuma. Kank proteins: structure, functions and diseases. https://link.springer.com/article/10.1007/s00018-009-0038-y.

[21] Franscisco Sanchez-Vega. Oncogenic signaling pathways in the cancer genome atlas. https://www.cell.com/cell/pdf/S0092-8674(18)30359-3.pdf.

[22] Nichoal J. Shurch. Optimization of an rna-seq differential gene expression analysis depending on biological replicate number and library size. https://www.frontiersin.org/articles/10.3389/fpls.2018.00108/full.

[23] Libing Song. mir-486 sustains nf-b activity by disrupting multiple nf-b-negative feedback loops. https://www.nature.com/articles/cr2012174.

[24] Amdt Stahler. Amphiregulin expression is a predictive biomarker for egfr inhibition in metastatic colorectal cancer: Combined analysis of three randomized trials. https://aacrjournals.org/clincancerres/article/26/24/6559/83059/Amphiregulin-Expression-Is-a-Predictive-Biomarker.

[25] Rory Stark. Rna sequencing: the teenage years. https://www.nature.com/articles/s41576-019-0150-2.

[26] BioTuring Team. How to read pca biplots and screen plots. https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/#:~:text=A.

[27] HBC training. Dge count normalization. https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html.

[28] Cancer Research UK. Capecitabine (xeloda). https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/cancer-drugs/drugs/capecitabine#:~:text=skin%20and%20nail%20problems%20such,and%20whites%20of%20the%20eyes.

[29] Oliver Utting. Immune functions in mice lacking clnk, an slp-76-related adaptor expressed in a subset of immune cells. https://journals.asm.org/doi/full/10.1128/MCB.24.13.6067-6075.2004.

[30] Tianzhen Wang. Aberrant regulation of the lin28a/lin28b and let-7 loop in human malignant tumors and its effects on the hallmarks of cancer. https://link.springer.com/article/10.1186/s12943-015-0402-5.

[31] Wikipedia. Fastq format. https://en.wikipedia.org/wiki/FASTQ_format.

[32] Hong Yin. Participation of eml6 in the regulation of oocyte meiotic progression in mice. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7007726/.

[33] Han Y He Q Yu G, Wang L. clusterprofiler. https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html, 2012.

[34] Lan Zhang. The bidirectional regulation between myl5 and hif-1 promotes cervical carcinoma metastasis. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667347/.

[35] Bin Zhao. Identification of potential key genes and pathways in early-onset colorectal cancer through bioinformatics analysis. https://journals.sagepub.com/doi/full/10.1177/1073274819831260, February 2019.