Predicting clinical outcomes of LIHC patients based on transcriptomic and epigenetic data

Group 4

Abhinav Mishra, Kristin Köhler, Sanket Gosavi, Utkarsha Kandale

Data Science in the Life Sciences, Summer Semester 2022

Course Code: 19405612

Contents

1 Abstract

As we know that molecular networks in a cancer cell are interwoven and complicated. Several cancer histology studies revealed that clinical features often fail to provide the bigger picture of liver and intrahepatic bile ducts cancer, since it is too narrow. Here, we tried to predict the prognosis and outcomes of the same, to integrate multiple biomarker types, and look at how they interact and affect each other. Categorically, our work focuses on predicting the vital status, and tissue type using random forest classifier using differentially expressed genes, and differential methylated regions, with accuracy 38%, and 96% on the test set of samples, and alive/dead patients. Most of the identified GO-annotated features were responsible for downregulation of immunogenic pathways. Moreover, the significant correlation of genes with RNA metabolism, and regulation processes, especially the ones involving miRNAs, may help in discovering potential targets for therapeutic interventions.

2 Introduction

2.1 Source

TCGA-LIHC project from the TCGA database: dbGaP accession *phs000178.v11.p8* found here: https://portal.gdc.cancer.gov/projects/TCGA-LIHC [1]. The following data types were used:

1. **Transcriptome profiling** - Gene Expression Quantification using RNA-Seq in *tsv* format (371 cases)
2. **DNA methylation** - Methylation beta values using methylation array in *txt* format (377 cases)
3. **Copy number variants** - GISTIC_Peaks values obtained using the curatedTCGA-Data package (373 cases)

2.2 Disease

Liver Hepatocellular Carcinoma (Adenomas and Adenocarcinomas) in *Homo Sapiens* [2].

2.3 Scientific Question

Are there potential epigenomic and transcriptomic bio-markers associated with the overall survival rate of patients with hepatocellular carcinomas?

After exploratory data analysis with some cleanup and normalization, to identify potential bio-markers, we extracted differentially expressed genes (DEGs) and differentially methylated regions (DMRs) with *R* packages, *DESeq2* for DEGs and *DMRcate* with *limma* for DMRs, respectively. The annotation of DEGs and DMRs for enriched pathways was done with *clusterProfiler*, and later, same for the machines learning features from *machine learning* model in *python* [3].

Additionally, we implemented a *random forest* model to investigate the association between the survival time of patients and the potential bio-markers, with regards to sample tissue type as a parallel to survival analysis.

The time period in which the whole project was carried out was six weeks.

3 Methods

3.1 Differential Gene Expression Analysis using *DESeq2*

DESeq2 integrates methodological advances with several novel features to facilitate a more quantitative analysis of comparative RNA-seq data using shrinkage estimators for dispersion and fold change [4]. It effectively controlled type-I errors, maintaining a median false positive rate just below the chosen critical value in a mock comparison of groups of samples randomly chosen from a larger pool. The steps involved in the script are as following:

1. **Querying**: Gene expression quantification values from transcriptome profiling for TCGA-LIHC project based on RNA-Seq were queried in GDC data portal using *GDCquery* function in *TCGAbiolinks* package [5].
2. **Downloading**: Using *GDCdownload* function in *TCGAbiolinks* package, the expression data downloaded in the directory with *api* method. We only used *unstranded* in *SummaryExperiment:assays()* for the expression matrix after using *GDCprepare()* on the downloaded data.

3. **Preprocessing (quality control, transformation, and normalization)**:

   - Quality assess and clean raw sequencing data
   - Align reads to a reference
   - Count the number of reads assigned to each contig/gene
   - Extract counts and store in a matrix
   - Create column metadata table

   Removing the irrelevant definition type *recurrence*, calculating NA values, checking the order of the column data, and count matrix, passing the grouping variable *definiton* as factor after relevelling. Variance stabilization transformation and generating top 5/10 PCA plots (pairplot) using *PCAtools* after removing 10% variance [6].

4. **Hierarchical Clustering**: Calculation of *dissimilarity matrix* between the samples, and *heatmap* visualization.

5. ***DESeq* object**: Differential expression analysis based on the Negative Binomial (commonly known as, Gamma-Poisson) distribution:

   (a) Choose the *design formula*( ) to run on raw counts (manual, rest is automated).

   (b) *Estimate size factors*: if you use *estimateSizeFactors()*, then it uses median of ratios method.

   (c) *Estimate gene-wise dispersions*: The dispersion estimates are inversely related to the mean and directly related to variance. Based on this relationship, the dispersion is higher for small mean counts and lower for large mean counts. The dispersion estimates for genes with the same mean will differ only based on their variance. Therefore, the dispersion estimates reflect the variance in gene expression for a given mean value. However, for low mean counts, the variance estimates have a much larger spread; therefore, the dispersion estimates will differ much more between genes with small means.

   (d) Fitting Curve to gene-wise dispersion estimates.

   (e) *Mean disperson relationship*: Shrink gene-wise dispersion estimates toward the values predicted by the curve.

   (f) Final GLM fit for each gene.

6. ***Gene Count***: Plot of normalized counts for a single gene on log scale: either with or without 'technical replicates', i.e. multiple sequencing runs of the same library, grouped by *sample_id*'s.

7. ***ENSEMBL* to *HGNC* ID conversion**: It was done using *biomaRt* package in *R* [7]. The *ENSEMBL* gene identifiers were converted to canonical forms from isoforms to get a match.

8. **Up/Down Regulated Genes**: Sub-setting the results from step 5, and sorting it by adjusted $p - value < 0.1$, and then ordering it by log fold change values to get the significant genes.

9. **Diagnostic plots**: *Dispersion estimate* plots the per-gene dispersion estimates together with the fitted mean-dispersion relationship. After *quantile binning* of mean counts, the plot of ratio of $p-values$ to check the distribution for the results in *DESeq* object. A group comparison was done using *MA plot* for all *LFC shrinkage* estimators, including hypothesis testings:

   (a) A two tailed test with $|\beta| > logFC$

   (b) A two tailed test with $|\beta| < logFC$

   (c) $\beta > logFC$

   (d) $\beta < logFC$

10. **Shrinkage, and outlier detection with MA plots**: *apeglm* provides Bayesian shrinkage estimators for effect sizes using approximation of the posterior for individual coefficients while removing the noise and preserving large differences. *ashr* implements an empirical Bayes approach for large-scale hypothesis testing and false discovery rate (FDR) estimates. *ashr* performs the best in controlling false positives, compared to *normal* and *apelgm*, but *apelgm* is the smoothest in noise reduction [8]. The count outlier detection was done using *cook's distance* [9]. *MA plots*, a plot of log-intensity ratios (M-values) versus log-intensity averages (A-values) was based on $logFC\varepsilon[1.5, -1.5]$ & $p - value < 0.05$

11. **Volcano plots**: A scatterplot that shows statistical significance ($p - value$) versus magnitude of fold change ($logFC$), showing the changes in large dataset containing replicate data (differential expression). The genes selected for the plots were based on these four parametric criterion:

   (a) $p - value < 0.01$ & $logFC > 2$

   (b) $p - value < 0.01$ & $logFC > 3$

   (c) $p - value < 0.05$ & $logFC > 2$

   (d) $p - value < 0.05$ & $logFC > 3$

12. **Accessing top *x* up/down genes**: Writing a function to fetch the results of step 8 by specifying the wanted number of genes, and calling it in a *for* loop, which saves the results in *.csv* format for top 100, 200, and 300 for both upregulated and downregulated genes.

3.2 Differential Methylated Loci and Regions using *limma* and *DMRcate*

To investigate the molecular changes in liver cancer on an epigenomic level, we further analysed and integrated DNA methylation data provided on TCGA. We compared the methylation data of 50 patients with tumor tissue samples and matched normal tissues samples to find differential methylated CpG loci and regions.
We started by downloading txt.-files containing beta-values from TCGA using TCGAbiolinks [5]. The data was generated from Illumina HumanMethylation450 arrays and preprocessed with the R-tool *SeSAMe* [10] that calculates and normalizes the beta-values while also implementing methods for quality control and correction of common experimental errors. After filtering for patients containing both tumor and normal tissue samples, we removed critical CpG's including loci with missing values, loci mapping to the X or Y chromosome to remove some sex bias and loci that overlap with common SNP's (MAF>0.01) which resulted in 288556 remaining CpG loci. We visualized the the beta-value distribution (see Figure 7a) to check the data quality. Moreover, we applied MDS on the samples and plotted the first two components (see Figure 7b).
To extract differential methylated loci (DML) between the two conditions (normal/tumor tissue), we used the R-package *limma* that firstly fits a linear model and then applies empirical Bayes techniques to rank CpG loci in order of evidence for differential methylation. We visualized the results in a Volcano plot (see Figure 13) and filtered the loci regarding adjusted p-value (<0.005) and mean beta value difference between groups (>0.2) as described in the paper [11] resulting in 24457 DML's.
Next, we extracted differential methylated regions (DMR's) with the DMRcate [12] package that combines DML's closely located on the genome into regions. Since DMRcate annotates the loci to hg19, we changed the coordinates manually to match hg38. The extracted regions were filtered by FDR (<0.01) and mean beta difference (>0.2) resulting in 4003 DMR's. The most significant DMR's were visualized using Gviz [13] and overlapping genes were further investigated. Hyper- (n=297) and hypomethylated (n=3706) regions were extracted and visualized separately. We analysed locations and sizes of DMR's regarding chromosomes (see Figure 10) and annotated them to functional regions including gene promoters, exons,

introns, CpG islands and more (see Figure 11) to compare hyper- and hypomethylation. Lastly, we analysed the correlation of gene expression and methylation. Therefore, hyper- and hypomethlated DMR's overlapping with promoter regions of genes were extracted and intersected with DEG's extracted in our previous analyses. We found 70 genes overlapping with hypomethylated and 16 genes overlapping with hypermethylated promoters. which were compared regarding Log_2FC of the DEGs (see Figure 17).

3.3 Copy Number Variation

The copy number variation data was imported using the curatedTCGAData [14] package from BioConductor. The data was converted into a GRanges object to link the loci to chromosomes using the GenomicRanges [15] package. Since the data required no additional processing it was used as it is for further analysis.

3.4 Random Forest Classifier using *scikit-learn*

We trained several machine learning models to predict different outcomes like the tissue type of the sample (normal/tumor), tumor stage (I/II/III) and vital status (alive/dead) of the patients. We implemented random forest classifiers using scikit-learn's [16] RandomForestClassifier() that predicts the outcome by averaging over the results of multiple decision trees. The classifier predicting the tissue type was built on the features extracted from the differential expression and methylation analyses. Cases providing both methylation and expression data were intersected and split into trainings (2/3: n=276) and test set (1/3: n=136). The top 200 up- and downregulated DEG's and top 500 DML's were used to train the random forest model. A 10-fold cross validation using a random grid search was applied to optimize the model regarding number of trees, the maximal number of features and samples provided to each decision tree and the class weight method used to correct for imbalanced class outcomes. The optimal hyperparameters n_estimators=250, max_samples=0.4, max_features=0.1, class_weight='balanced_subsample' were used to build the classifier. Importantly, the 'balanced_subsample' method assigns weights to the classes and therefore, corrects for class imbalances within the decision tree's subsamples. Feature importances based on the Gini-criterion were extracted and used for functional analyses. Moreover, the model was retrained on a reduced number of features (top 100) to compare the performances and predictive power of less features.
Another random forest model classifier was built to predict the cancer stage of the tumor

samples (stage I: 196, stage II 86, stage III: 80, stage IV: 5). Since samples of stage IV were immensely underrepresented, they were removed and the remaining samples were split into train (2/3) and test set (1/3). Additional to the DEG and DML features, all information on copy number variants (CNVs) was included. Again, the model was optimized using a 10-fold cross validation.
The last model was build to predict the vital status *dead* (n=127) or *alive* (n=236) from DEG's, DML's and CNV's. Again, the samples were split into train (2/3) and test (1/3) set and the model's hyperparameters were optimized using a random grid search with a 10-fold cross validation.

3.5 DEGs and Features Annotation using *clusterProfiler*

The data obtained from deferentially expressed genes (DEGs) and the features obtained from machine learning algorithms were used to perform gene annotation. In this analysis, the genes are first classified based on their molecular functions, cellular location, etc. using an annotation database (for example, GO or KEGG). For the purpose of this study only the GO annotations were used. The algorithm works by classifying the genes with their respective GO terms. This data is then used to group the genes based on their classification. These groups are then simply ranked by using the ratio of number of genes from that group present in your sample to the total number of genes in that group. The top 'n' number of groups can then be visualised and then this information can be used to get a rough overview of what is happening in your sample (which metabolic, immune, proliferation, etc. pathways are enriched/repressed). This work was carried out using the clusterProfiler [17] package from Bioconductor. Apart from this, several supporting packages like tidyverse [18], AnnotationDbi [19], biomaRt [7], enrichplot [20], etc. to tidy the data, plot the data, etc. The DEGs were filtered to include genes expressed/repressed of magnitude higher than log to the base 2 change of 0.5. Since the features obtained from the machine learning algorithms were a mix of CpG methylation loci and genes they were filtered to remove any CpG methylation loci. This data was then subjected to annotation.

4 Results

4.1 Gene Expression

Analyses were conducted using the R Statistical language (version 4.2.1; R Core Team, 2022) on macOS Monterey 12.5

After the calculation for estimate dispersion trend and applying a variance stabilizing transformation to remove 10% variance in the biplot for principal component analysis in Figure 1a, only a minority cluster of normal tissues explains almost 13% of the variation in our dataset, and most of the tumour sample account for almost 9% in the Figure 1b.

The multiple biplots (or pairplots) for the top five principal components for rapidly skimming over the data revealed some overlaps for few points that resulted in large differences in the samples after the analysis (Figure 1b).

Figure 2 plots the normalized counts (on a $-log(p-value)$ scale) for a single gene on log scale, by removing or collapsing *technical replicates* i.e. multiple sequencing runs of the same library, grouped by *sample_id*'s. We can see clearly the overexpression in those genes.

The gene dispersion estimated is fitted well dominantly for higher mean counts considering the amount of significant genes for the *logFC* threshold and *p-value* that was defined. We performed quantile binning of mean counts, and observed that the plot of ratio of $p-values < 0.01$ holds the skewed normal distribution as we see in Figure 3a & 3b.

Setting $\alpha = 0.05$ (type I error rate) and threshold value of $logFC = 1.5$, alternative hypothesis testing showed the most parsimonious multi-modal results in Figure 4a while the self-evident Figure 4b was a result of scaled standard error from the closest boundary. In Figure 4c & 4d, the yielded outlier boundary for data points matched only at the edge of upregulation, and downregulation, supposedly.

Out of 49287 nonzero total read count with adjusted $p-value < 0.05$, the fitted negative binomial model of *DESeq2* had 3% upregulation ($logFC > 1.5$), and 0.35% downregulation ($logFC < -1.5$), i.e. 1490 upregulated genes, and 174 downregulated genes with 41% low counts ($meancount < 1$)(Figure 6a).

As a comparison for *logFC* shrinkage methods in Figure 5, *ashr* performs the best in controlling false positive rate, compared to *normal* and *apelgm*, but *apelgm* achieved smoothest noise reduction. As *apeglm* uses an adaptive student's t-prior distribution method, provides

bayesian shrinkage estimators for effect sizes in a linear model, in succession of approximation of the posterior for individual coefficients. In simpler terms, removing noise, and preserving large differences while smoothing the curve.

After re-balancing the y-axis, by defining global maxima and minima, *MA-plot* in Figure 6a visualized differentially expressed genes for controlled $p - value < 0.05$ (*blue dots*) in which up-regulated genes for $logFC > 1.5$ and down-regulated for $logFC < -1.5$. The solid lines represents the boundaries for *logFC*.

The most significant genes, both, in Figure 6a & 6b were based on the approximated range [-5,12] for *logFC* in the results of *DESeq2* analysis. While controlling type I error rate (p-value), *logFC* threshold acts as a type II error rate. For the likelihood concerning standard errors, as one gets lower, the other one gets higher.

4.2 Differential Methylated Loci and Regions

Figure 7a shows the beta-value distribution of the 50 tumor and 50 normal tissue samples. The minor peaks between 0 and 1 indicate tissue heterogeneity (differences in methylation) within some tumor samples which is not optimal but not critical for our analyses and thus, has not been corrected. Figure 7b shows the first two components after MDS of all 100 samples. While normal tissue samples show a strong similiarity regarding methylation, the tumor tissue samples differ largely and indicate between sample heterogeneity.

In Figure 13, the results of the differential methylation analyses using limma are shown in a Volcano plot. Since the default thresholds for logFC and adjusted p-value lead to a large number of DML's, we additionally filtered the loci by mean beta-value difference between normal and tumor tissue which led to 24457 DML's. The two most differential methylated CpG's which are hypomethylated are shown in Figure 8. We did not find any literature associating those two CpG's with liver cancer so far.

However, one of the most differential methylated regions (see Figure 9) is a known prognostic biomarker for hepatocellular carcinomas [21]. Hypermethylation within the overlapping TBX15 gene that is originally involved in mesodermal differentiation causes a decrease of its expression. Surprisingly, we did not see a significant difference in expression of TBX15 in our DESeq2 results.

The annotated DMR's are shown in Figure 10 and 11. The majority of DMR's is hypomethylated (n=3706), a much smaller amount is hypermethylated (n=297). Many differential

methylated regions are located on chromosome 1,7 and 10 (Figure 10). However, we did not look closer into reasons for that. Additionally, the majority of hypomethylated DMR's is located within introns, followed by exons. A minor part is located within promoter regions. Interestingly, the proportion of hypermethylated DMR's within promoter regions is much higher in comparison to exon and intron hypermethylated DMR's.

Finally, we compared differential expressed genes to differential methylated promoter regions. Figure 12 shows the Log_2FC of significantly (p.adj<0.1) DEG's that are associated with differential hypo- and hypermethylated regions. Not as expected, the boxplots show a similar distribution and do not reflect the suppression of expression by hypermethylation within promter regions. A negative Log_2FC indicates downregulation of genes which can by correlated with hypermethylation but can not be observed in our analysis.

4.3 Random Forest Classifier

The random forest classifier predicting the tissue type from the 400 DEGs and 500 CpGs shows a great performance on train and test set. After optimizing its hyperparameters, it achieves an accuracy of 0.99 and AUC of 1 on the test set (see Figure 14b). All tumor tissue samples are recognized, only one normal tissue sample is misclassified as a tumor sample (see Figure 14a). Even after reducing the features to the 100 most important ones, the performance of the classifier regarding accuracy, AUC and confidence did not change indicating that less features contain enough information to distinguish between both tissue types.

The three most important features are shown in Table 1. Those three genes are enriched in liver tissue, specifically in hepatic stellate cells. CLEC4M is involved in the progression of numerous cancer types and has been shown to be a biomarker in HCC [22]. CRHBP is also a known biomarker whose underexpression has been observed in hepatocellular carcinoma tissue [23]. FCN2 is known to play an important role in metastasis and epithelial-mesenchymal transitions of HCC [24]. The top 100 features also included the methylation status of 29 CpG loci. We did not use those for further functional enrichment analysis which might be a future task.

The classifier predicting the cancer stage of tumor tissue samples did not perform as well as the previous classifier. Although performing a hyperparameter optimization, the classifier seems to strongly overfit to the train data. On the train set it achieves an accuracy of 0.89, wherase the accuracy on the test set increases dramatically to 0.48. That trend is also

Feature	Importance	Function
ENSG00000104938.18 (CLEC4M)	0.063	involved in peripheral immune surveillance, enriched in liver
ENSG00000145708.11 (CRHBP)	0.050	binds CRF and inactivates it, enriched in liver
ENSG00000160339.16 (FCN2)	0.047	may function in innate immunity, enriched in liver

Table 1: Three most important features and their function

reflected by the confidence matrices (see Figure 15). The majority of samples in the test set is predicted as stage I cancer. Biologically interpreted, the features extracted by differential expression and methylation analyses only contain information to distinguish between normal and tumor tissue and do not facilitate the prediction of cancer stages. To improve the performance of the classifier, the design formulas to extract DEG's and DML's can be changed to compare the cancer stages and extract more informative features.

Our initial idea was to predict the vital status of patients by integrating DEGs, DMRs and CNVs. Therefore, we trained a classifier to predict the outcomes *alive* and *dead*. Unfortenately, the classifier was not able to learn any signal from the input data. Even on the train set, it classifies most samples as *alive*. On the test set only 2 out of 42 samples taken from donors that died are classified correctly indicating that the extracted features are not predictive for this analysis (see Figure 16a). The model classified the samples with an AUC of 0.57 (see Figure 16b) which is almost as bad as a random choice. To change that, the design formula of the differential expression and methylation analyses was changed to find DEGs and DMLs in comparison between alive and dead patients. However, no significant changes in expression and methylation have been found. Thus, we were not able to adjust our feature space regarding genetic and epigenetic characteristics. Important to mention is that all samples were taken from patients when they were alive which might be the reason for that. To improve the classifier's performance, one might include additional clinical variables like cancer stage at the time of diagnosis, treatment or age.

4.4 Annotation

The GO enrichment plots obtained from the DEGs and the machine learning features are slightly different. Most of the suppressed genes (see Figure 17) from the DEGs data are directly involved in the immune defenses. Similarly, the activated pathways are mostly

related to cell cycle and reproduction such as chromosome segregation, nuclear division, organelle fission, etc. However, some deviations from these observations are seen as well. For example, genes related to metabolism (like oxidoreductase activity, amino-acid betaine catabolism, caffeine oxidation) are repressed whereas some other genes related to gated channel activity are activated.

The GO enrichment plot (see Figure 18) of features data on the other hand is also similar in terms of the suppressed genes. Most of the the genes which are suppressed belong to the immune system such as antigen binding, complement activation, opsonisation, etc. However, the activated pathways are mostly related to ncRNA and sncRNA production and processing. Most of these molecules are involved in regulation of cellular functions such as transcription and translation.

Diving a bit deeper into the GO enrichment analysis we get to see a slightly different picture about the cellular processes. The downregulated genes (see Figure 19) from the DEGs data shows that majority of the affected cellular processes are metabolism related. Many of these genes are involved in important catabolic functions such as carboxylic acid, organic acid, small molecule, cellular amino acid, etc catabolic processes. Whereas most of the remaining affected cellular processes are a part of the immune system. The upregulated genes (see Figure 20) from the DEGs data show that majority of upregulated genes are involved in cell cycle/proliferation processes and in organ mophogenesis, appendage, embryonic skeletal system, and limb development processes. Similarly, the features data also shows that most of the cellular pathways are also related to cell cycle/proliferation, limb and appendage moprhogenesis and development, and the immune system.

5 Discussion

We were able to extract known and new biological markers associated with LIHC from differential expression and methylation analyses. Further analysis regarding the correlation of expression and methylation would be an interesting future task. Moreover, investigating the correlation of gene expression and CNV is an essential step to understand the causes of up- or downregulation of particular genes.
To improve the performance of the random forest classifier predicting the survival status, integrating more features like clinical variables might be a promising option allowing to perform survival analyses and put more focus onto important medical and biological characteristics that influence overall survival.

Most of the affected cellular processes obtained from the GO enrichment analysis are clearly known to play a role in cancer survival and proliferation. Genes related to cell cycle or proliferation and immune system suppression are easy to explain since cancer cells are known to proliferate continuously while evading the immune system. Interestingly genes related to metabolism (especially catabolism) and limb, appendage, organ, etc morphogenesis and development are also observed. One can expect a cancer cell to modify metabolism to keep up with the increased energy and necessary building block (amino acids, nucleotides, etc.) requirements due to its unregulated proliferation. However, we also saw an upregulation in a lot of genes related to catabolic processes. These genes can supplement the cell with extra energy but at the same time it might also reduce its stockpile of available building blocks which might affect its proliferation abilities. It seems as though the cancer cells are relying on the rest of the body to provide them with the required building blocks. Furthermore, we also see hints of de-differentiation occurring within the cancer cells such that they are regressing from a specialised liver cancer cell to a cell having more general embryonic/stem cell-like properties. This loss of identity from a liver cell to some random cell-type such as embryonic limb/appendage cell favours the cancer cells since it allows them to proliferate easily.

Biological systems, especially the ones as complicated as human cells are unrivaled in the precise and delicate balance that they maintain within themselves. Any deviation from this balance will most likely lead to lethal outcomes, or in some rare cases carcinogenicity. Looking from afar it seems as though cancer cells are inherently chaotic. The term 'chaotic' here is used to indicate that fact that cancer cells are random in their cellular processes such as gene expression, chromosome duplication/deletion, methylation, etc. We know that this is true since even within cancer tissues subclonal cancer populations arise which may be quite different from each other and the original cancer cells in terms of the above mentioned processes and more [25]. However, even cancer cells must pass the test of natural selection. Cancer cells must evolve over time to evade the body's defences or therapeutic interventions to survive and proliferate. Simply speaking, cancer cells are subjected to natural selection which are governed by external factors (selective pressures). Some of these selective pressures like therapeutic interventions are completely within our control whereas others like body's defences can at least be altered to varying extent. Furthermore, we also know that different distantly related organisms can evolve to gain similar or analogous characteristics due to similar environmental pressures. This is known as convergent evolution and it is quite common in nature. For example, carcinisation (a phenomenon where unrelated crustaceans evolve into crab-like organisms), evolution of some marsupials and placentals with similar characteristics, evolution of wings in bats and birds, etc [26]. This begs the question whether one can predict how cancer cells evolve based on the knowledge about its environment.

Furthermore, since these external pressures can either be fully controlled or at least partly regulated, can we control the evolution of cancer as well? Studying cancer evolution from this perspective might allow us to identify selective pressures which play an important role in cancer evolution and help develop better therapeutic interventions to fight cancer.

6 Code and Data Availability

All code used for our analyses is available at https://github.com/koehlek99/LIHC-project. The data used in this project is available at https://portal.gdc.cancer.gov/projects/TCGA-LIHC.

7 References

References

[1] D. A. Wheeler, L. R. Roberts, C. G. A. R. Network, et al. "Comprehensive and integrative genomic characterization of hepatocellular carcinoma". *Cell* 169.7 (2017), p. 1327.

[2] B. Erickson, S. Kirk, Y. Lee, O. Bathe, M. Kearns, C. Gerdes, K. Rieger-Christ, and J. Lemmerman. "Radiology data from the cancer genome atlas liver hepatocellular carcinoma [TCGA-LIHC] collection". *Cancer Imaging Arch* 10 (2016), K9.

[3] K. Köhler, A. Mishra, and S. Gosavi. *LIHC-project*. 2022. URL: https://github.com/koehlek99/LIHC-project.

[4] M. I. Love, W. Huber, and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". *Genome biology* 15.12 (2014), pp. 1–21.

[5] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, et al. "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data". *Nucleic acids research* 44.8 (2016), e71–e71.

[6] K. Blighe, M. Lewis, A. Lun, and M. K. Blighe. *Package 'PCAtools.'* 2019.

[7] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis". *Bioinformatics* 21.16 (2005), pp. 3439–3440.

[8] M. Stephens. "False discovery rates: a new deal". *Biostatistics* 18.2 (2017), pp. 275–294.

[9] A. Zhu, J. G. Ibrahim, and M. I. Love. "Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences". *Bioinformatics* 35.12 (2019), pp. 2084–2092.

[10] W. Zhou, T. J. Triche, P. W. Laird, and H. Shen. "SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions". *Nucleic Acids Research* (2018). DOI: 10.1093/nar/gky691. URL: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky691/5061974.

[11] N. K. Mishra, S. Southekal, and C. Guda. "Survival Analysis of Multi-Omics Data Identifies Potential Prognostic Markers of Pancreatic Ductal Adenocarcinoma". *Frontiers in Genetics* 10 (2019), p. 624. DOI: 10.3389/fgene.2019.00624. URL: https://www.frontiersin.org/article/10.3389/fgene.2019.00624/full.

[12] T. J. Peters, M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, R. V Lord, S. J. Clark, and P. L. Molloy. "De novo identification of differentially methylated regions in the human genome". *Epigenetics & Chromatin* 8.1 (2015), p. 6. DOI: 10.1186/1756-8935-8-6. URL: https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/1756-8935-8-6.

[13] F. Hahne and R. Ivanek. "Visualizing Genomic Data Using Gviz and Bioconductor". *Statistical Genomics*. E. Mathé and S. Davis (Eds.). Vol. 1418. Series Title: Methods in Molecular Biology. New York, NY: Springer New York, 2016, pp. 335–351. DOI: 10.1007/978-1-4939-3578-9_16. URL: http://link.springer.com/10.1007/978-1-4939-3578-9_16.

[14] M. Ramos, L. Geistlinger, S. Oh, L. Schiffer, R. Azhar, H. Kodali, I. de Bruijn, J. Gao, V. J. Carey, M. Morgan, et al. "Multiomic integration of public oncology databases in bioconductor". *JCO Clinical Cancer Informatics* 1 (2020), pp. 958–971.

[15] M. Lawrence, W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey. "Software for computing and annotating genomic ranges". *PLoS computational biology* 9.8 (2013), e1003118.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[17] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. "clusterProfiler: an R package for comparing biological themes among gene clusters". *Omics: a journal of integrative biology* 16.5 (2012), pp. 284–287.

[18] H. Wickham. "The tidyverse". *R package ver* 1.1 (2017), p. 1.

[19] P. H. "AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor". *R package version 1.58.0* (2022).

[20] G. Yu. "Enrichplot: visualization of functional enrichment result". *R package version* 1.2 (2019).

[21] Y. Morine, T. Utsunomiya, Y. Saito, S. Yamada, S. Imura, T. Ikemoto, A. Kitagawa, Y. Kobayashi, S. Takao, K. Kosai, K. Mimori, Y. Tanaka, and M. Shimada. "Reduction of T-Box 15 gene expression in tumor tissue is a prognostic biomarker for patients with hepatocellular carcinoma". *Oncotarget* 11.52 (2020), pp. 4803–4812. DOI: 10.18632/oncotarget.27852. URL: https://www.oncotarget.com/lookup/doi/10.18632/oncotarget.27852.

[22] L. Luo, L. Chen, K. Ke, B. Zhao, L. Wang, C. Zhang, F. Wang, N. Liao, X. Zheng, X. Liu, Y. Wang, and J. Liu. "High expression levels of CLEC4M indicate poor prognosis in patients with hepatocellular carcinoma". *Oncology Letters* (2020). DOI: 10.3892/ol.2020.11294. URL: http://www.spandidos-publications.com/10.3892/ol.2020.11294.

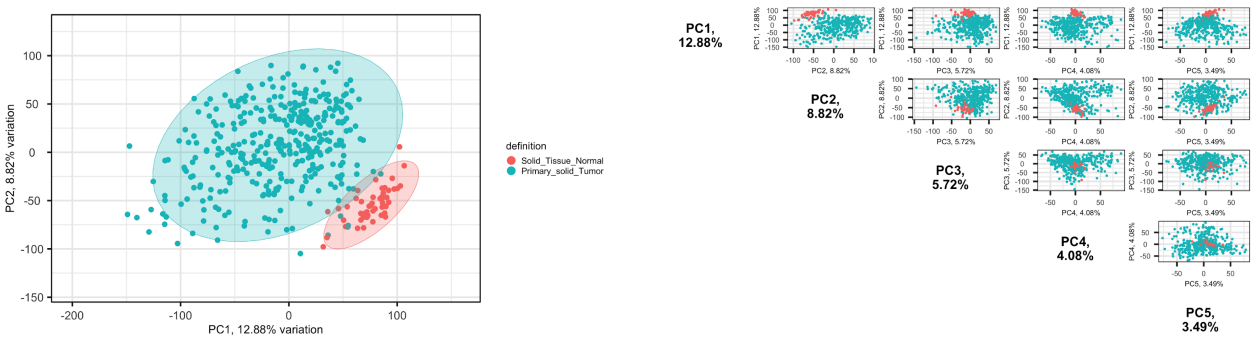
[23] H.-B. Xia, H.-J. Wang, L.-Q. Fu, S.-B. Wang, L. Li, G.-Q. Ru, X.-L. He, X.-M. Tong, X.-Z. Mou, and D.-S. Huang. "Decreased CRHBP expression is predictive of poor prognosis in patients with hepatocellular carcinoma". *Oncology Letters* (2018). DOI: 10.3892/ol.2018.9073. URL: http://www.spandidos-publications.com/10.3892/ol.2018.9073.

[24] G. Yang, Y. Liang, T. Zheng, R. Song, J. Wang, H. Shi, B. Sun, C. Xie, Y. Li, J. Han, S. Pan, Y. Lan, X. Liu, M. Zhu, Y. Wang, and L. Liu. "FCN2 inhibits epithelial–mesenchymal transition-induced metastasis of hepatocellular carcinoma via TGF-/Smad signaling". *Cancer Letters* 378.2 (2016), pp. 80–86. DOI: 10.1016/j.canlet.2016.05.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304383516303020.

[25] J. Liu, H. Dang, and X. W. Wang. "The significance of intertumor and intratumor heterogeneity in liver cancer". *Experimental & molecular medicine* 50.1 (2018), e416–e416.

[26] L. Gabora. "Convergent Evolution". *Brenner's Encyclopedia of Genetics (Second Edition)*. S. Maloy and K. Hughes (Eds.). Second Edition. San Diego: Academic Press, 2013, pp. 178–180. DOI: https://doi.org/10.1016/B978-0-12-374984-0.00336-3. URL: https://www.sciencedirect.com/science/article/pii/B9780123749840003363.

8 Figures



(a) Principal component (highest variance) *(b) Pairplot of top 5 PCs*

Figure 1: PCA plot

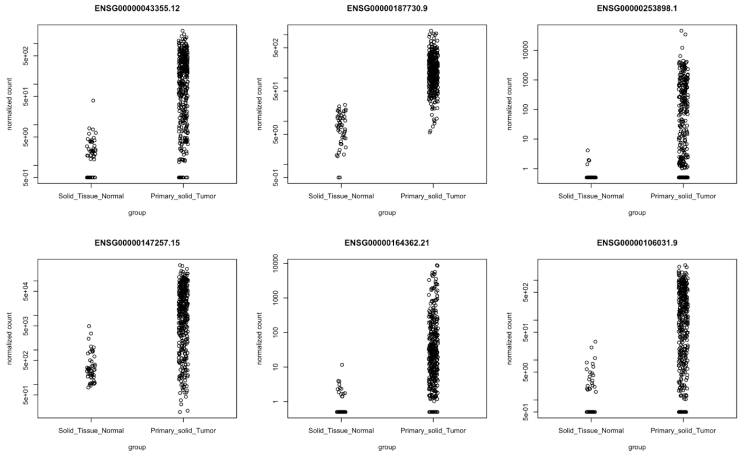
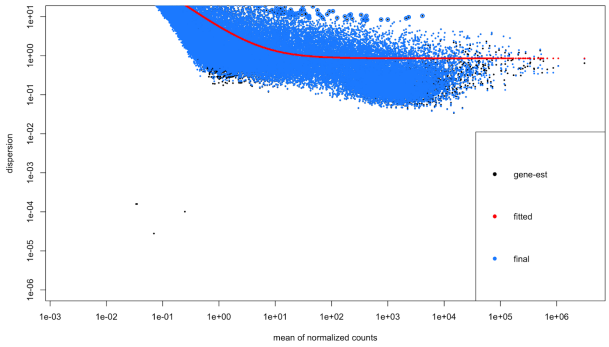
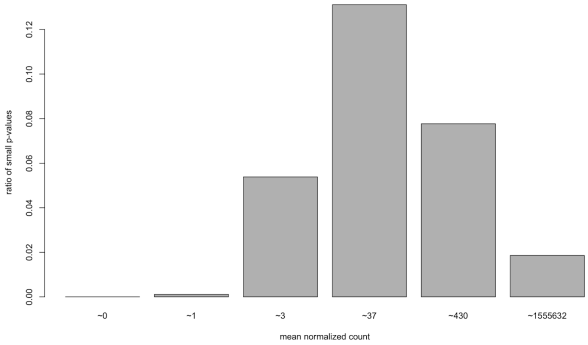


Figure 2: Gene Count plots: top 6 upregulated genes

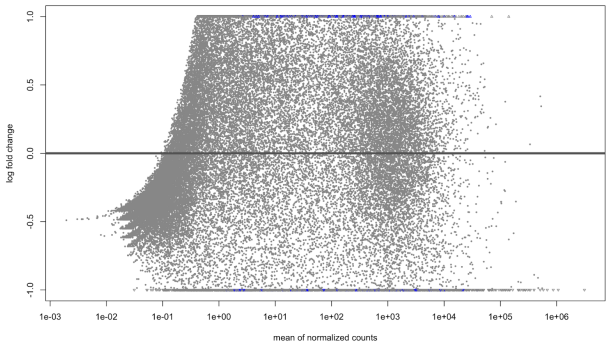


(a) Goodness of fit: Cook's distance

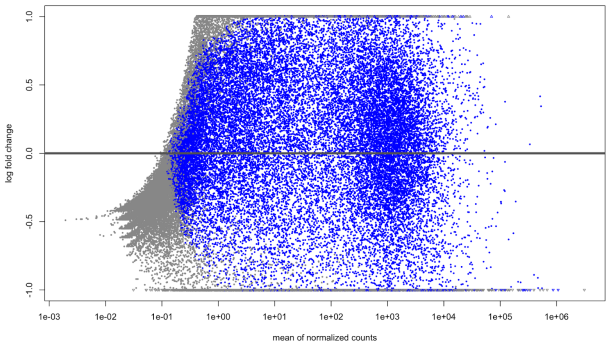


(b) Normal distribution: quantile binned ratios of p-values for each gene

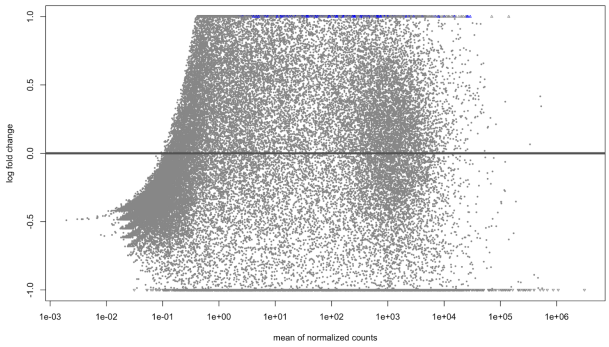
Figure 3: Gene dispersion estimates



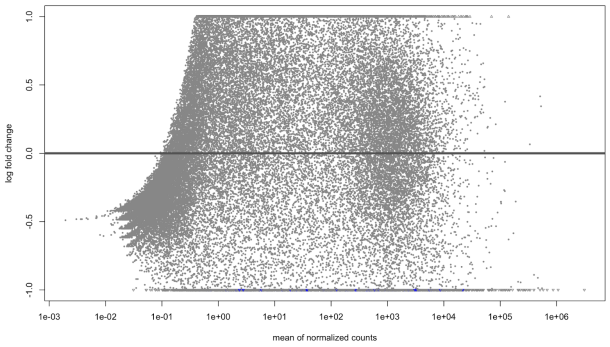
(a) $|\beta| > logFC$



(b) $|\beta| < logFC$



(c) $\beta > logFC$



(d) $\beta < logFC$

Figure 4: Alternative hypothesis: two-tailed test

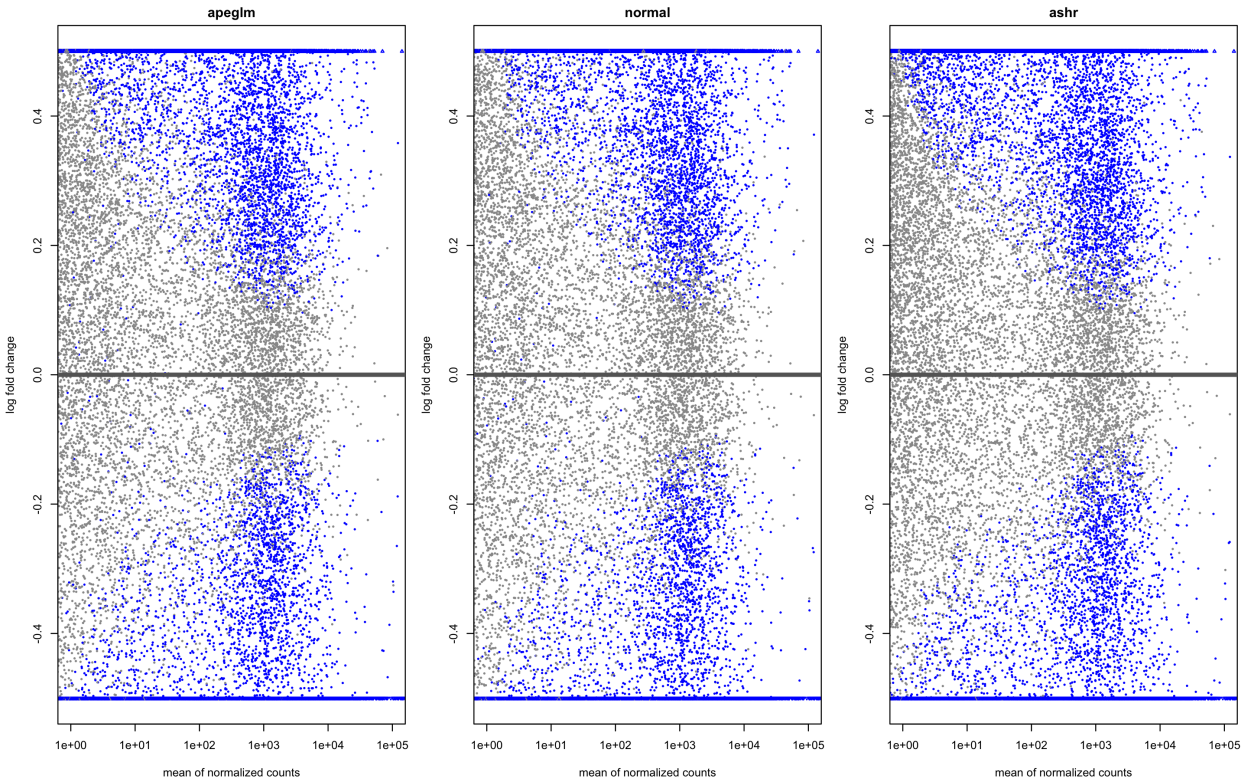
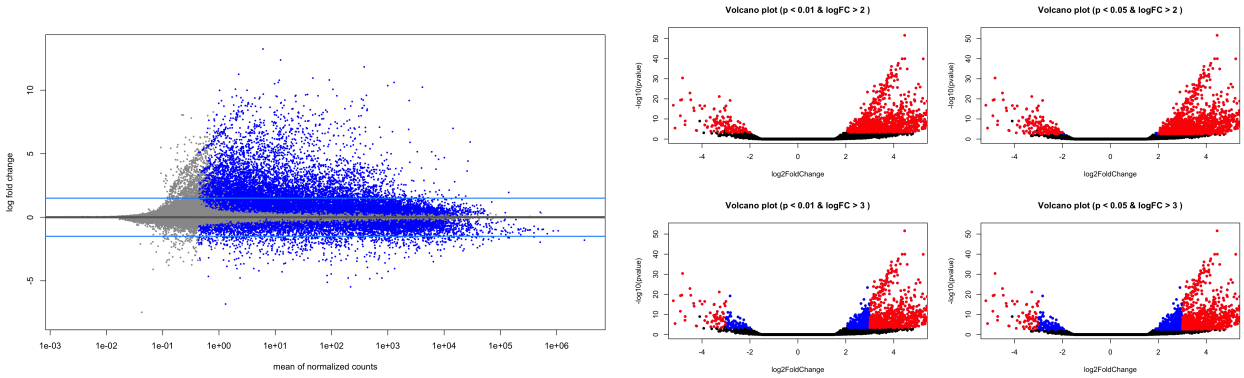
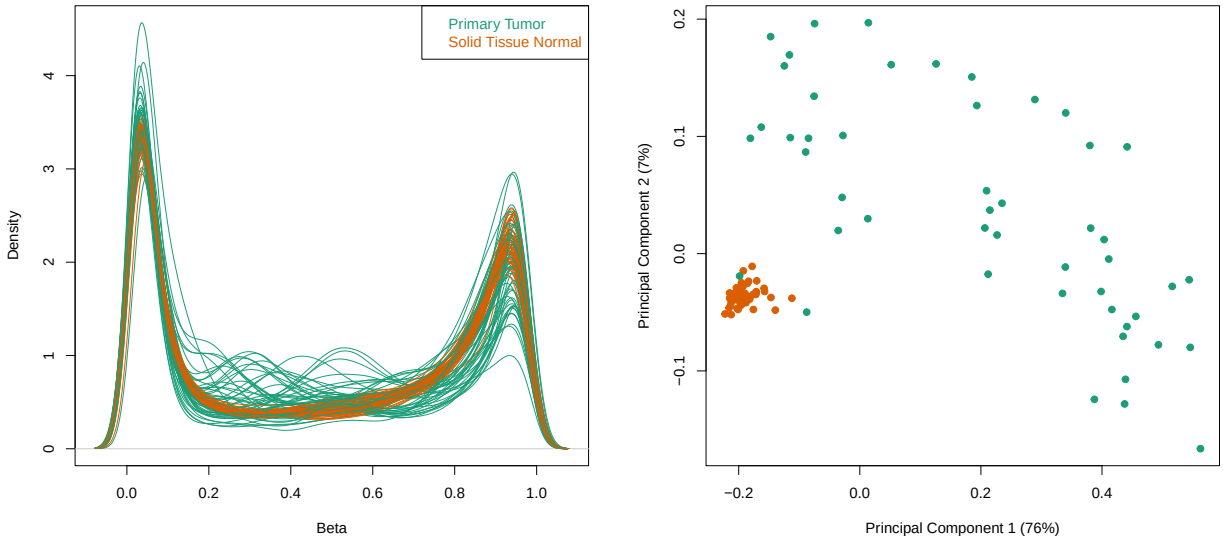


Figure 5: logFC shrinkage estimators (left to right: apeglm, normal, and ashr)



(a) MA plot: logFC $\varepsilon[1.5, -1.5]$ *(b) Volcano plot: controlled logFC & p-values*

Figure 6: Differential gene expression



(a) Beta density distribution *(b) PCA plot*

Figure 7: Quality control plots

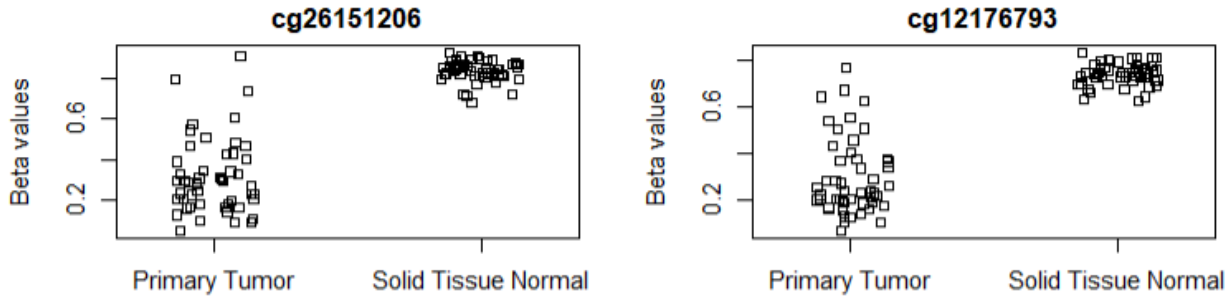


Figure 8: The two most differential methylated CpG's

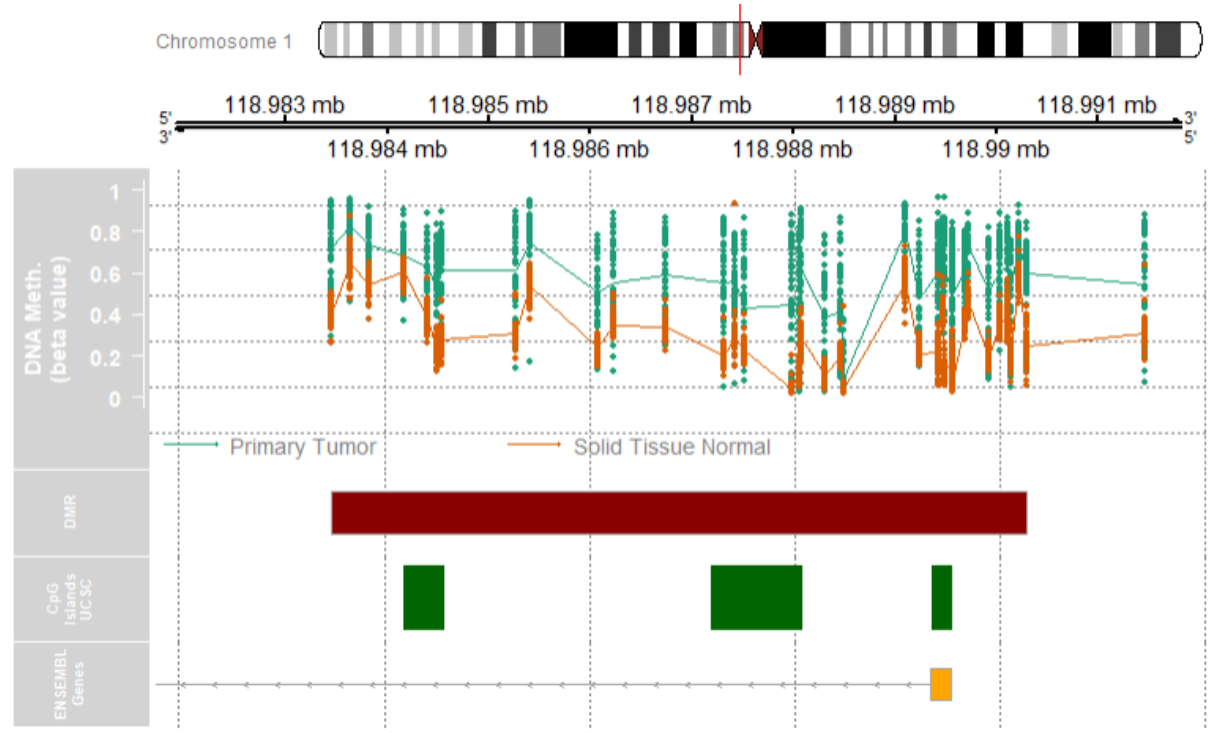


Figure 9: DMR overlapping with TBX15 gene

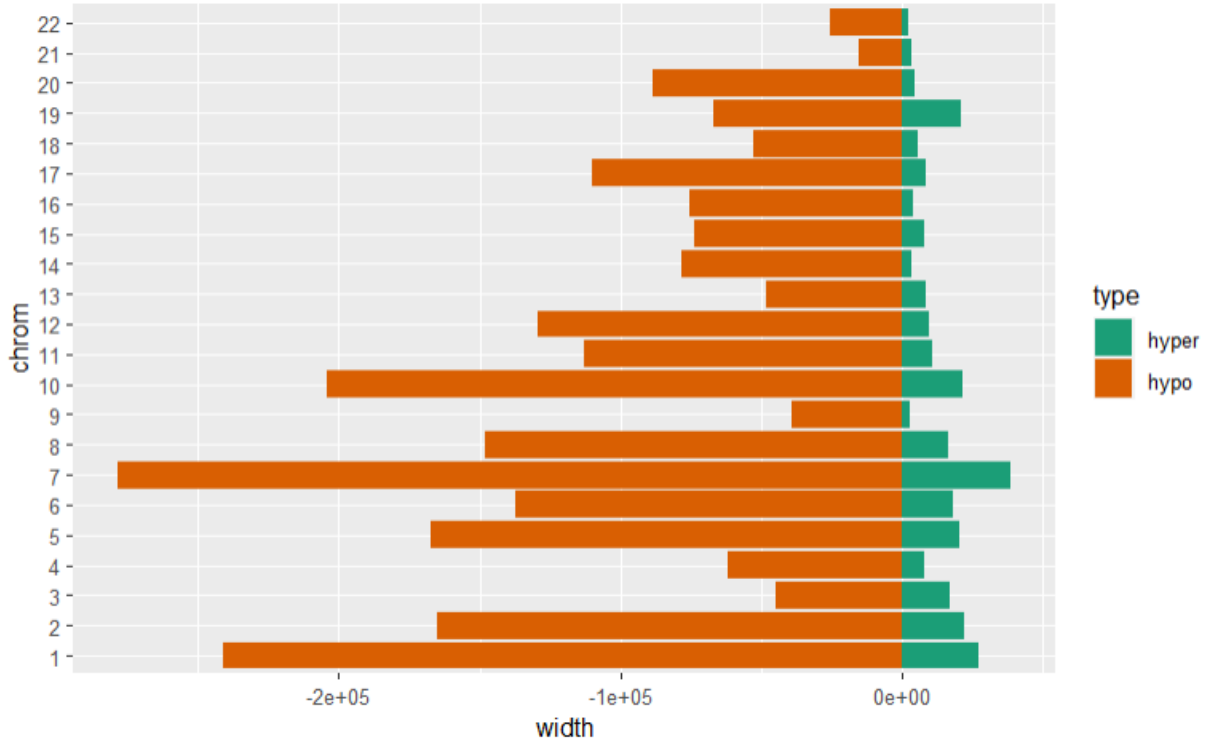


Figure 10: DMR lengths per chromosome

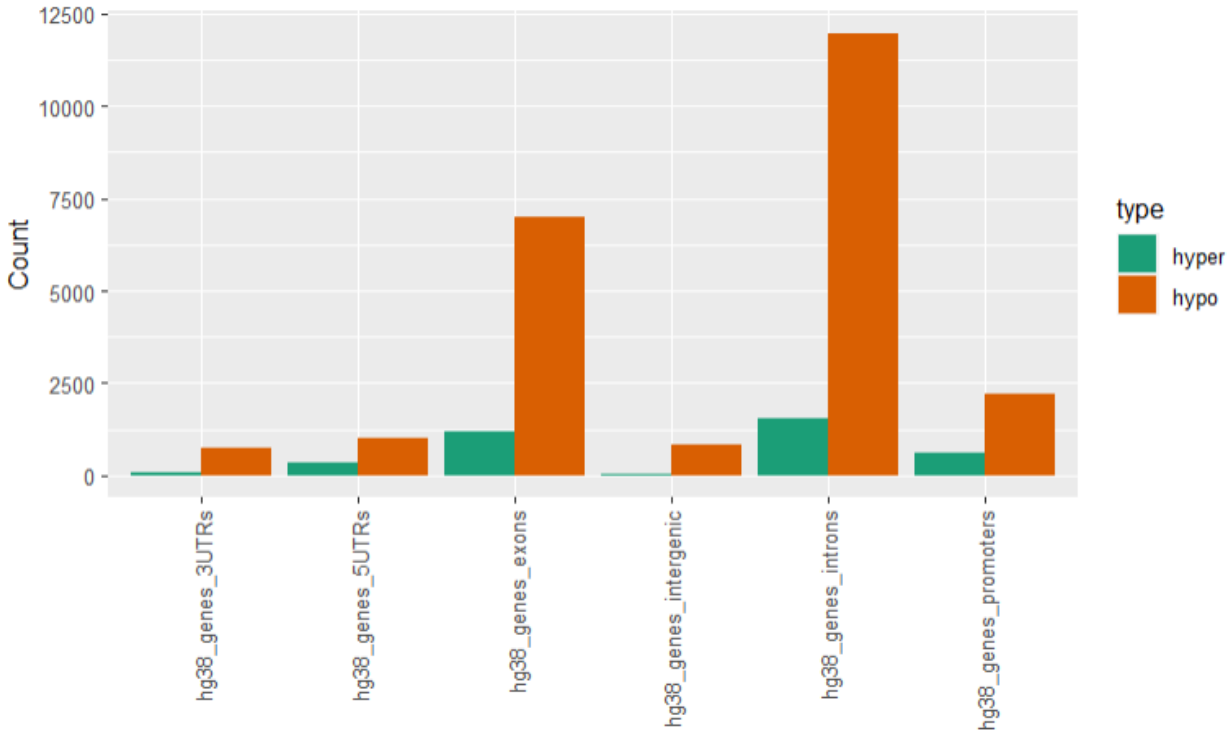


Figure 11: Functional annotation of DMRs

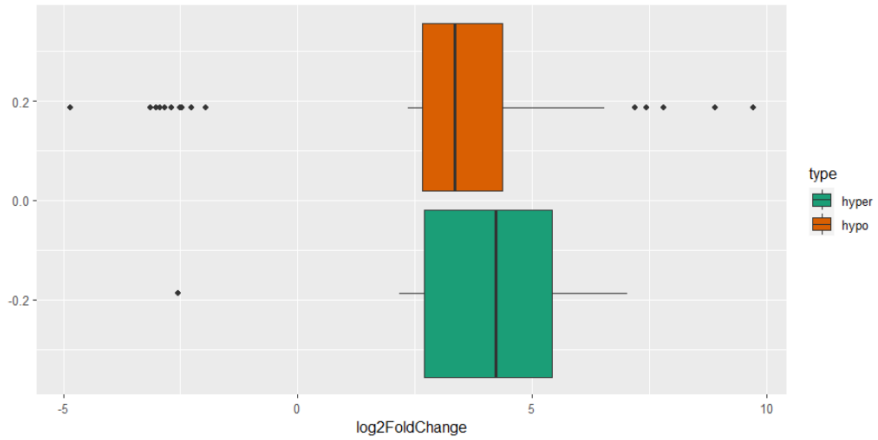


Figure 12: Log_2FC distribution of DEGs containing hyper-/hypomethylated promoters

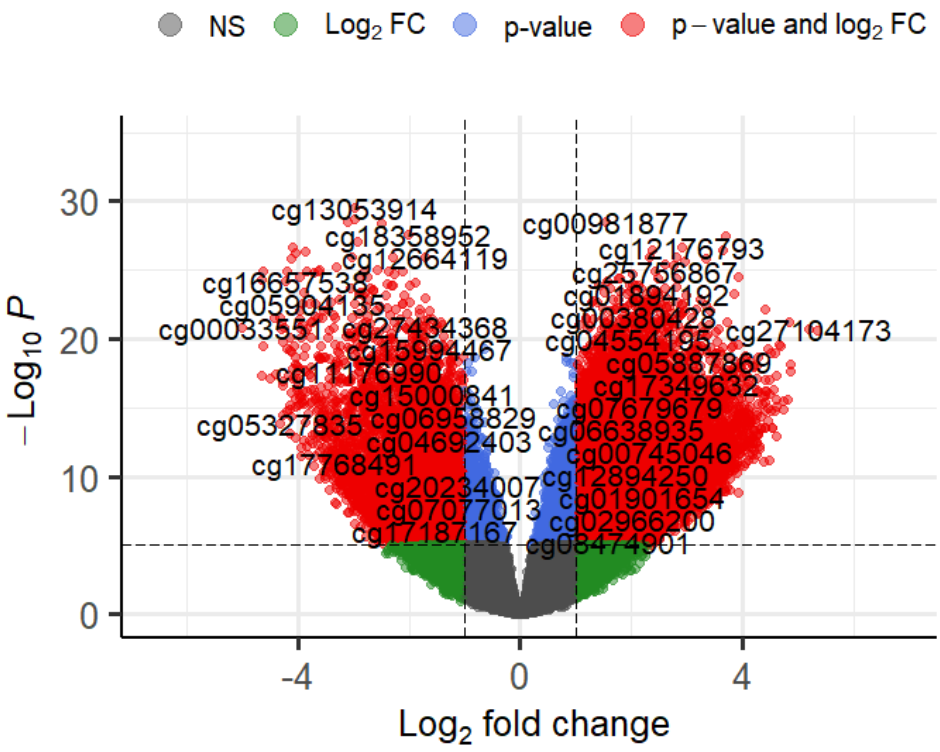
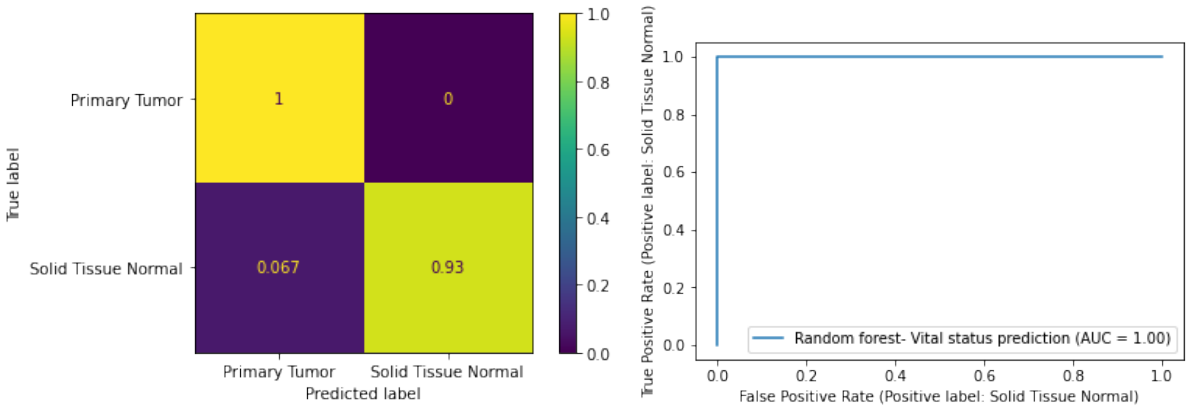
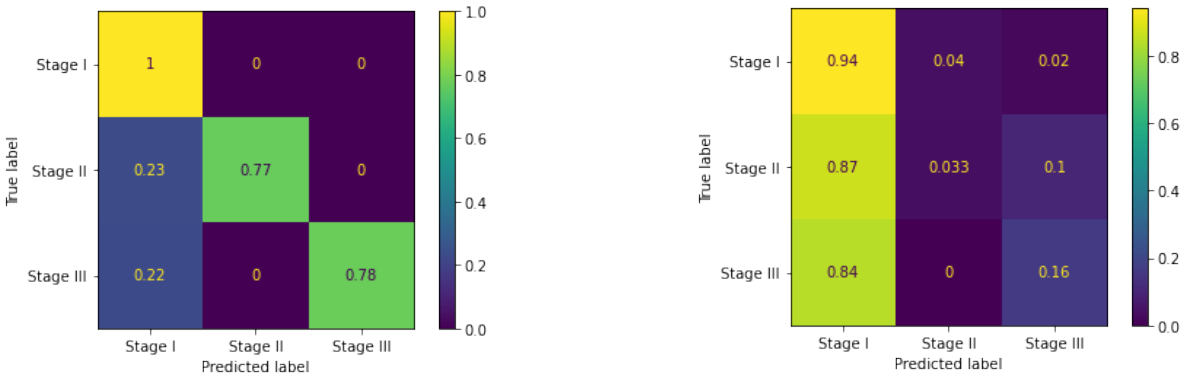


Figure 13: Volcano plot of limma results



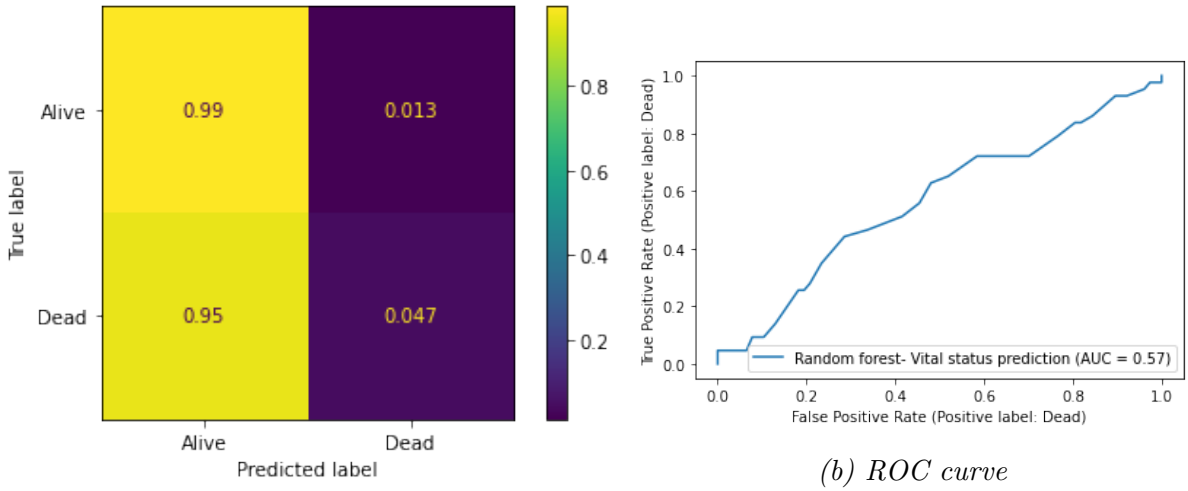
(a) Confusion matrix *(b) ROC curve*

Figure 14: Performance of the tissue classifier on the test set



(a) Confusion matrix on the train set *(b) Confusion matrix on the test set*

Figure 15: Performance of the cancer stage classifier



(a) Confusion matrix *(b) ROC curve*

Figure 16: Performance of the vital status classifier on the test set

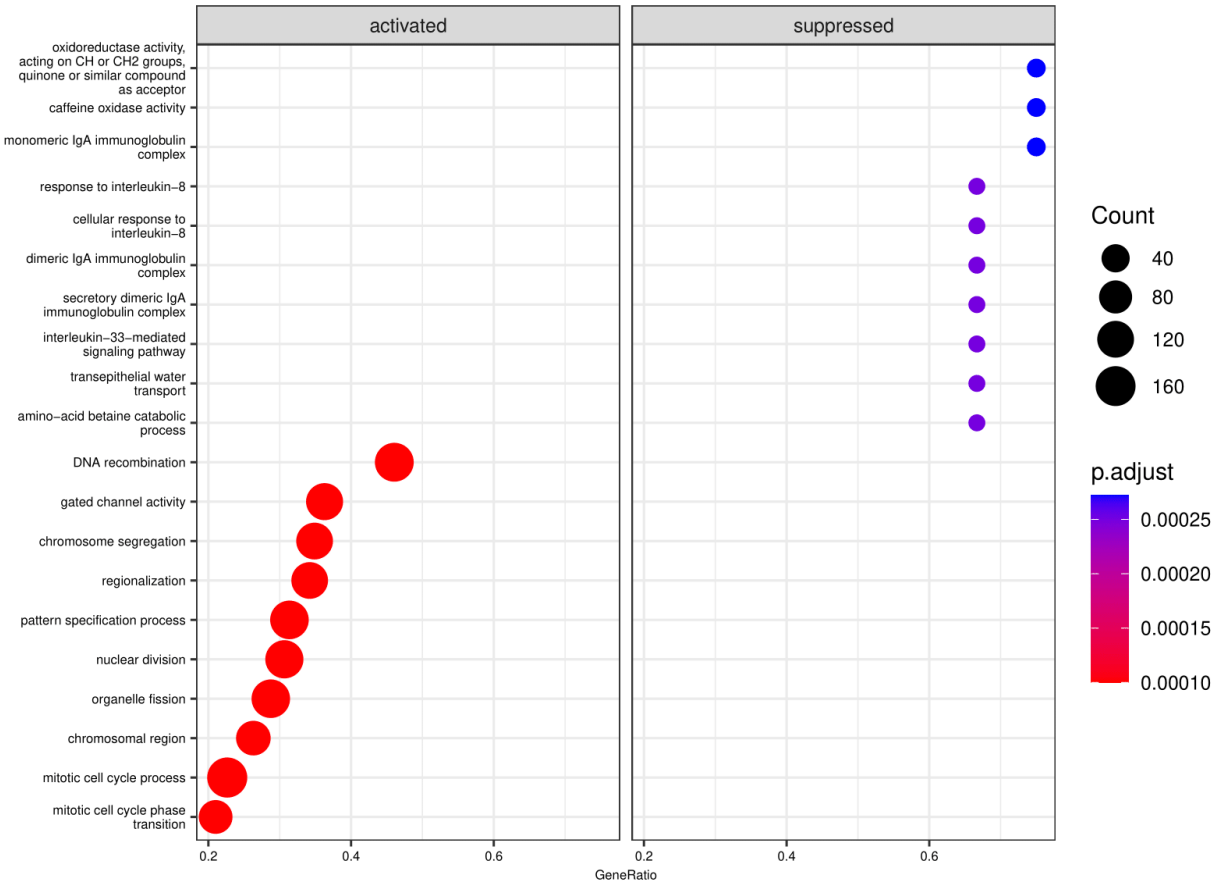


Figure 17: DEGs dotplot

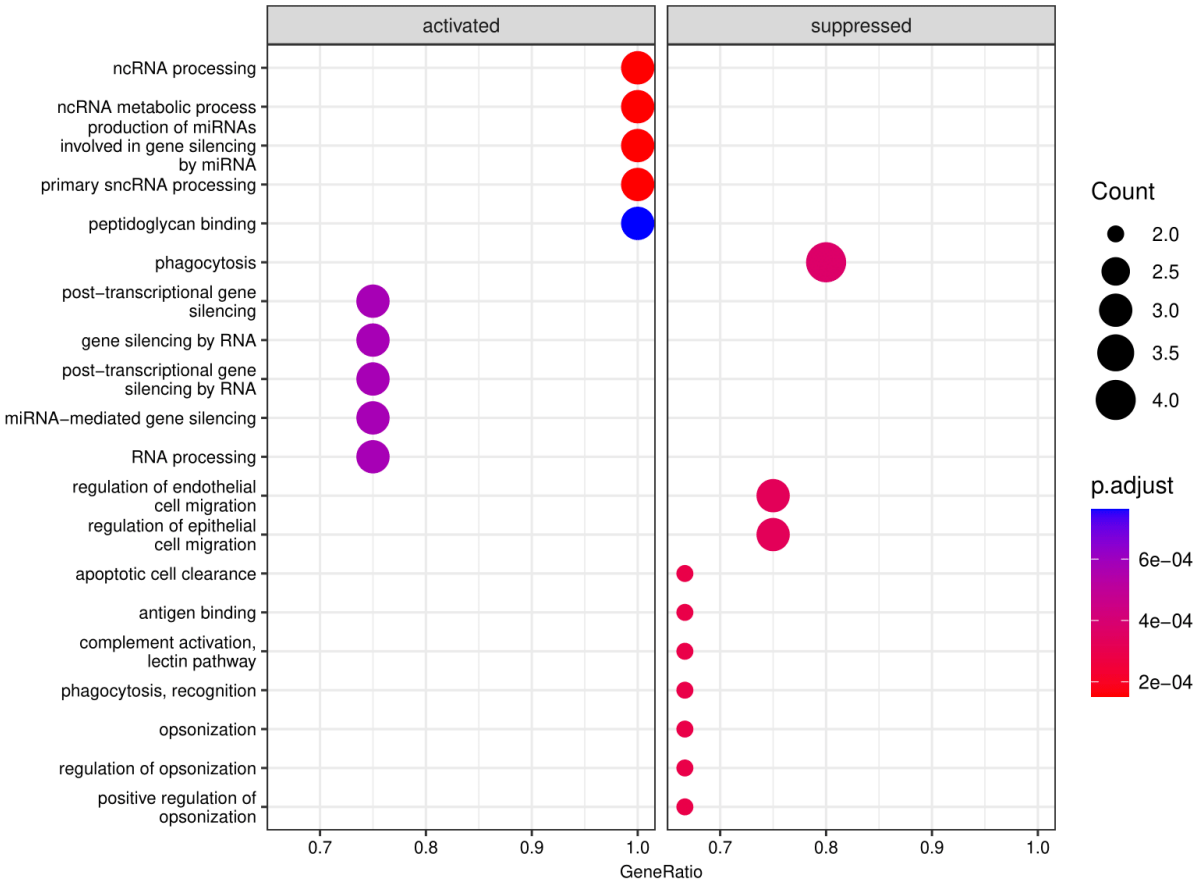


Figure 18: Features dotplot

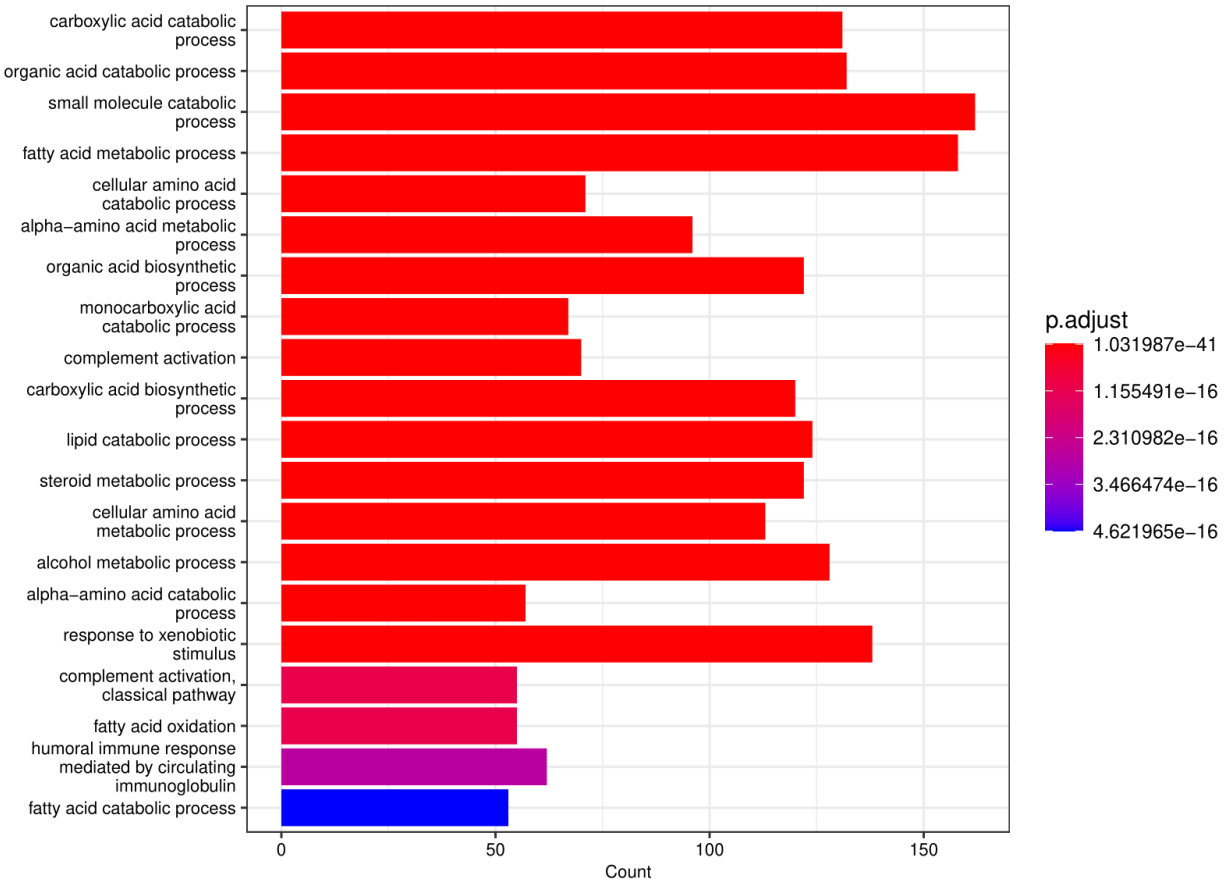


Figure 19: Downregulated DEGs

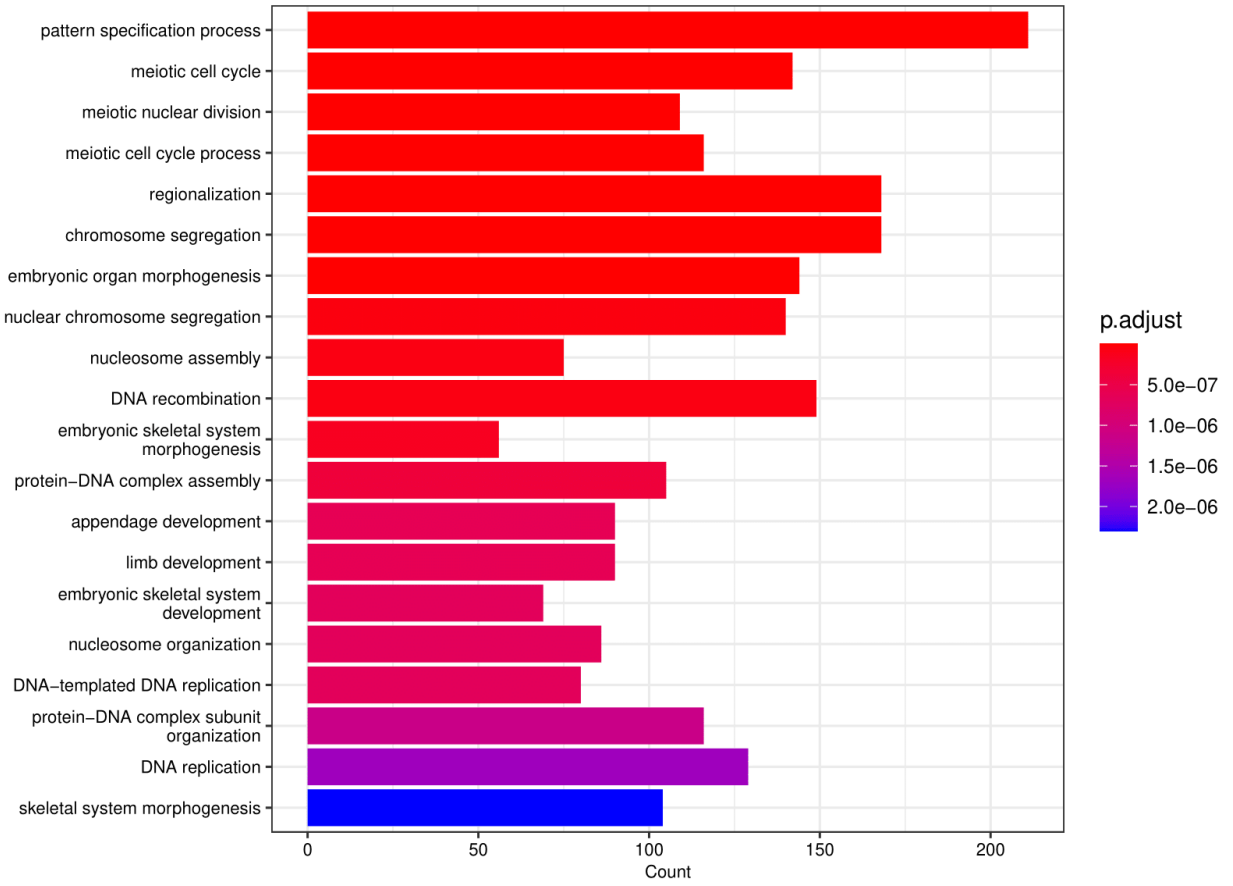


Figure 20: Upregulated DEGs