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

Abhinav Mishra, Kristin Köhler, Sanket Gosavi, Utkarsha Kandale

Introduction

Motivation To better predict the prognosis and outcomes of a disease it is necessary to integrate multiple biomarkers, and look at how they interact and affect each other.

Goal Are there potential epigenomic and transcriptomic biomarkers associated with the overall survival rate of patients with hepatocellular carcinomas?

Source https://portal.gdc.cancer.gov/projects/TCGA-LIHC (phs000178.v11.p8)

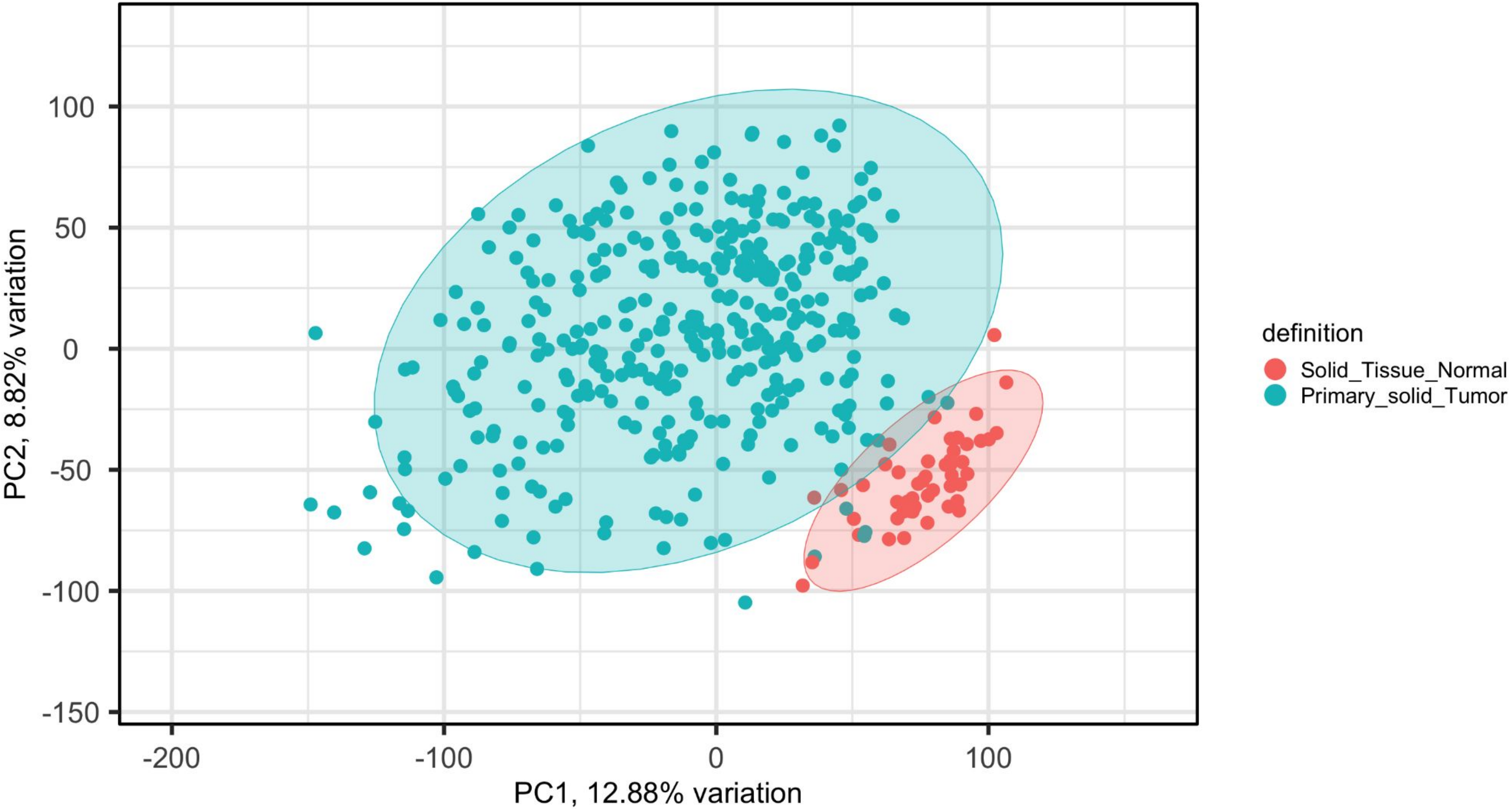
Disease Liver Hepatocellular Carcinoma (Adenomas and Adenocarcinomas)

Data Type

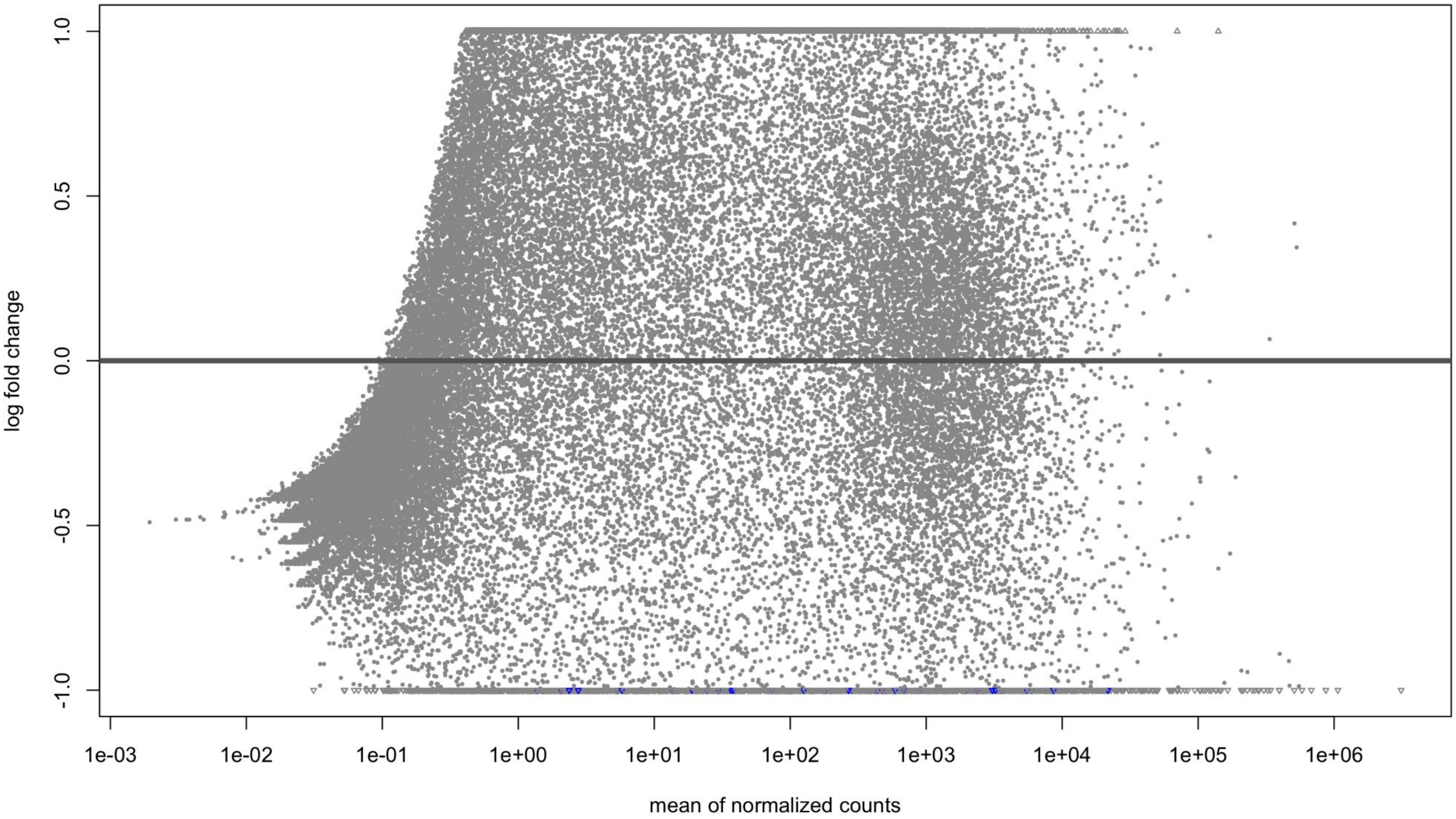
1. **Transcriptome profiling - Gene Expression Quantification using RNA-Seq** in *tsv* format (371 cases)
2. **DNA methylation - Methylation beta values using methylation array** in *txt* format (377 cases)

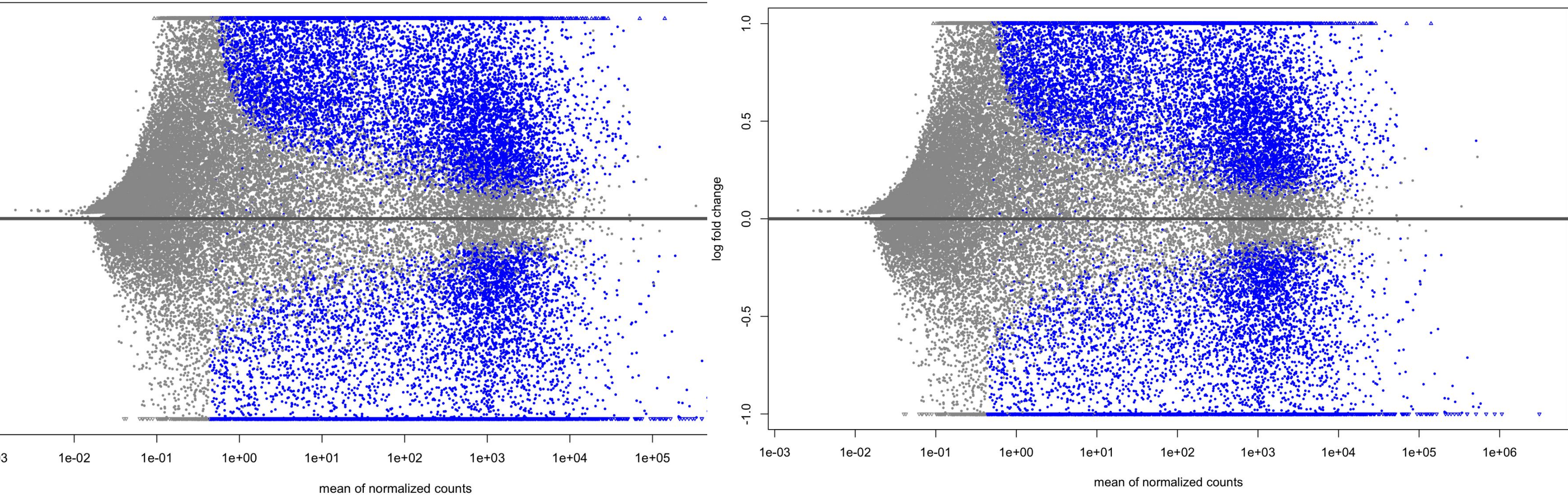
Differential Gene Expression Analysis using DESeq2

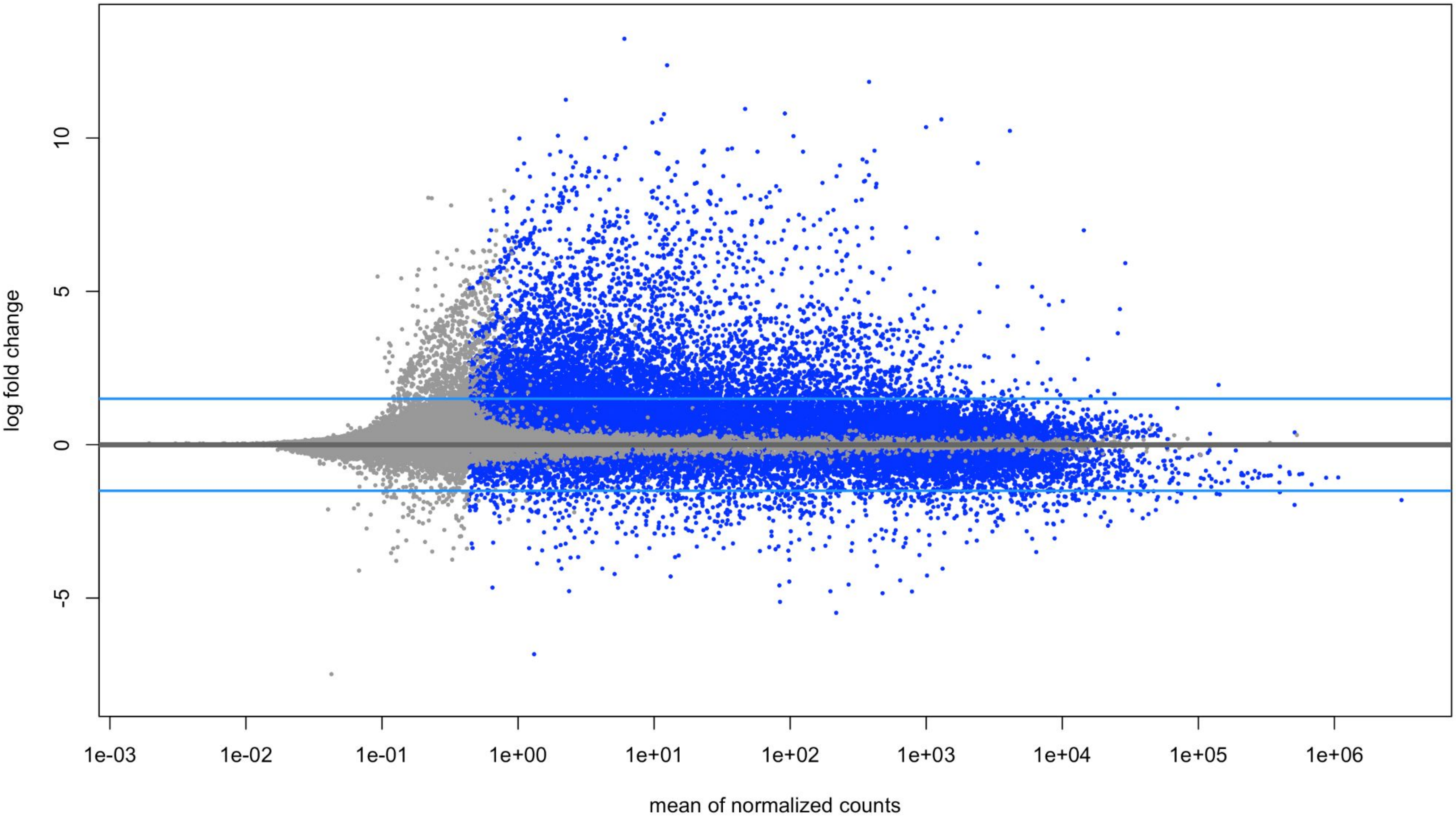
1. Querying the data in *GDC Data Portal*
2. Downloading the expression data
3. Data transformation, quality control and normalization
4. Hierarchical Clustering
5. Differential Gene Expression Analysis
6. Plot *Gene Count*: top 6 genes by *p-value*
7. ID Conversion: *ENSEMBL* to *HGNC*
8. Up/Down Regulated Genes
9. Diagnostic Plots
10. Shrinkage detection, and *MA plots* with outlier detection
11. Volcano plots by *p-adjusted value* and *logFC*
12. Accessing results for top "*x*" upregulated or downregulated genes

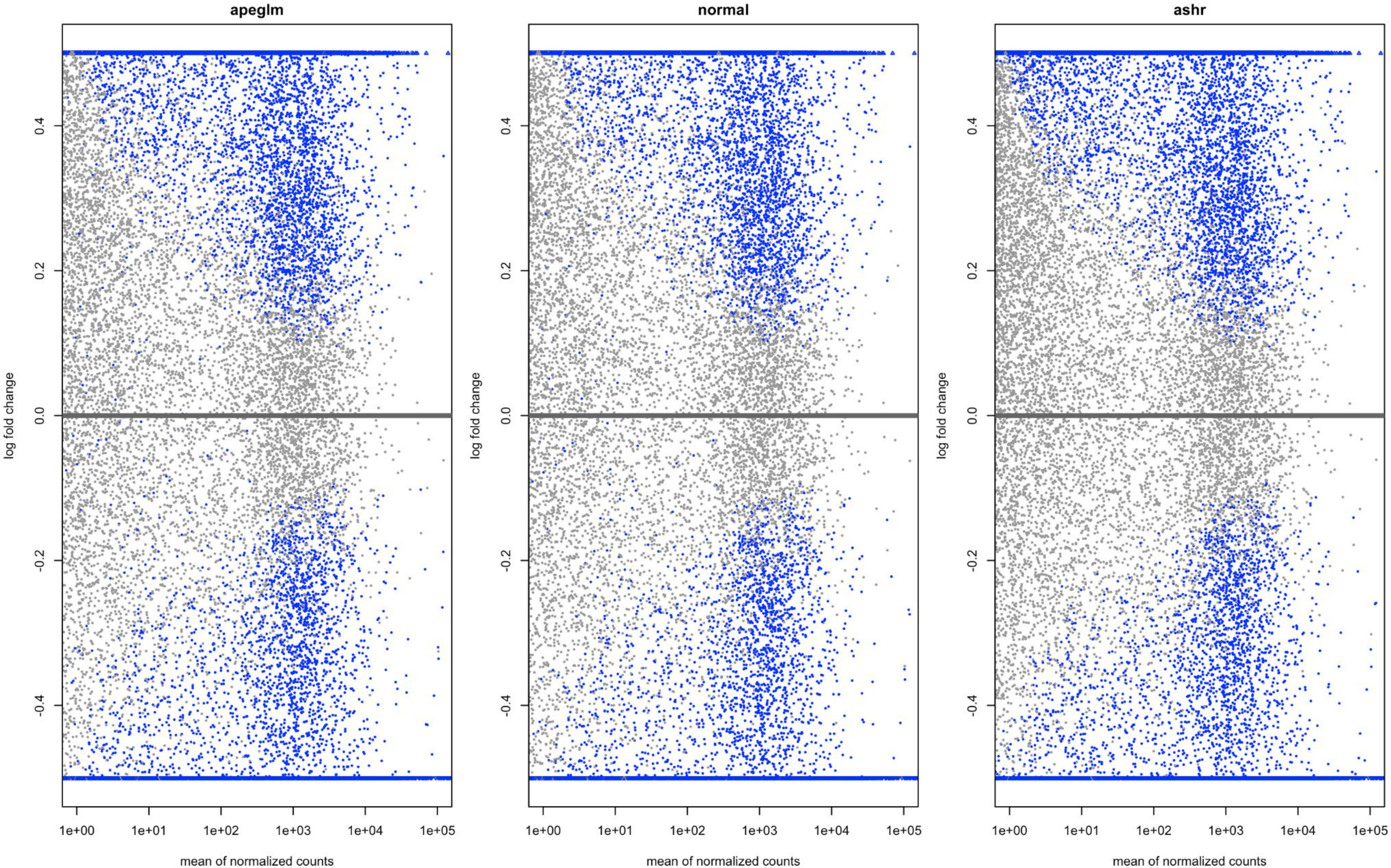


definition
Solid_Tissue_Normal
Primary_solid_Tumor
PC2, 8.82% variation
PC1, 12.88% variation
100
50
0
-50
-100
-150
-200
-100
0
100

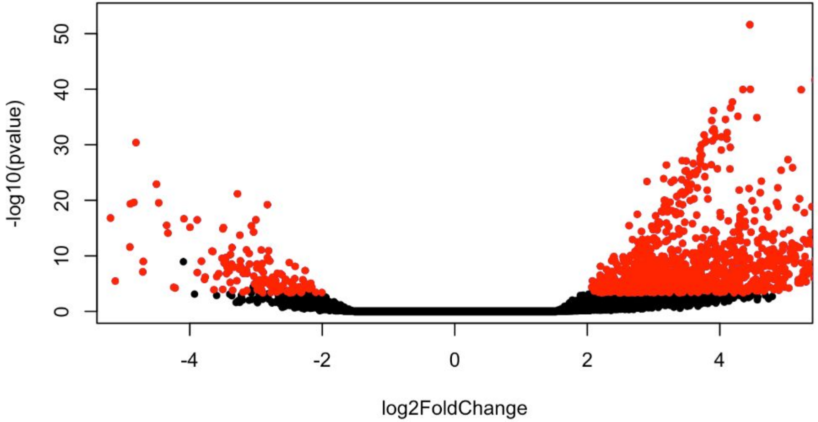

1.0
0.5
0.0
-0.5
-1.0
log fold change
1e-03
1e-02
1e-01
1e+00
1e+01
1e+02
1e+03
1e+04
1e+05
1e+06
mean of normalized counts


3
1e-02
1e-01
1e+00
1e+01
1e+02
1e+03
1e+04
1e+05
mean of normalized counts
1.0
0.5
0.0
-0.5
-1.0
log fold change
1e-03
1e-02
1e-01
1e+00
1e+01
1e+02
1e+03
1e+04
1e+05
1e+06
mean of normalized counts

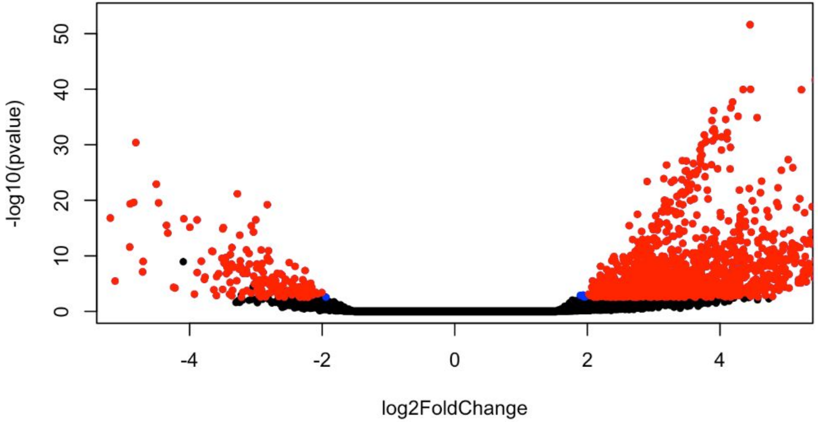

log fold change
10
5
0
-5
1e-03
1e-02
1e-01
1e+00
1e+01
1e+02
1e+03
1e+04
1e+05
1e+06
mean of normalized counts


apeglm
normal
ashr
log fold change
0.4
0.2
0.0
-0.2
-0.4
1e+00
1e+01
1e+02
1e+03
1e+04
1e+05
mean of normalized counts

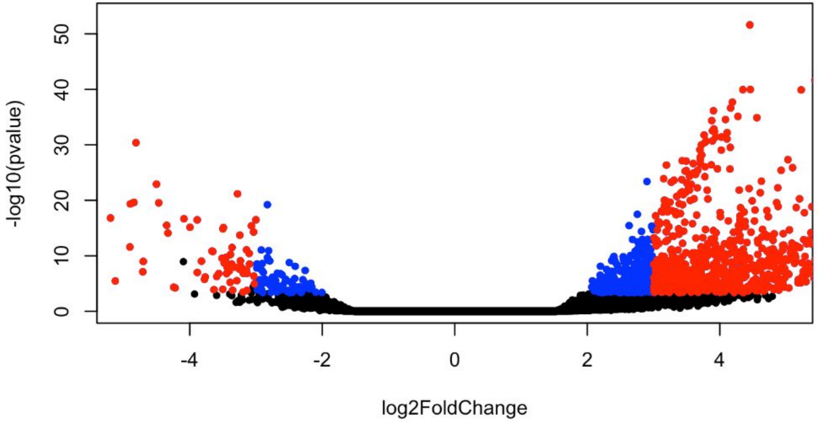
Volcano plot (p < 0.01 & logFC > 2)



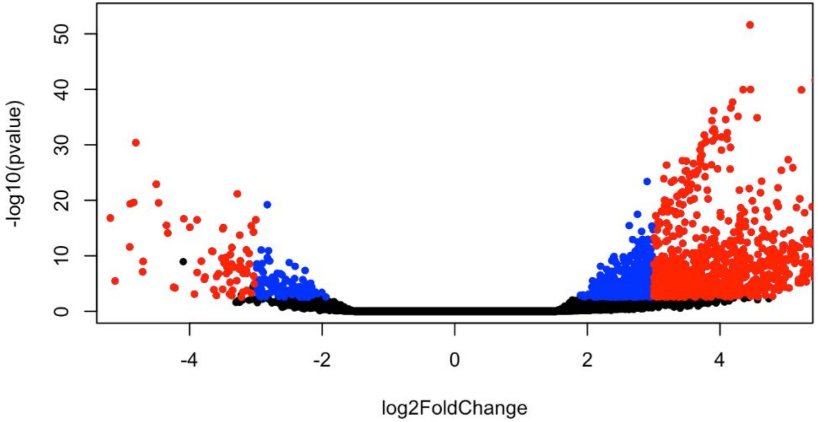
Volcano plot (p < 0.05 & logFC > 2)



Volcano plot (p < 0.01 & logFC > 3)



Volcano plot (p < 0.05 & logFC > 3)



Methylation Data and Preprocessing

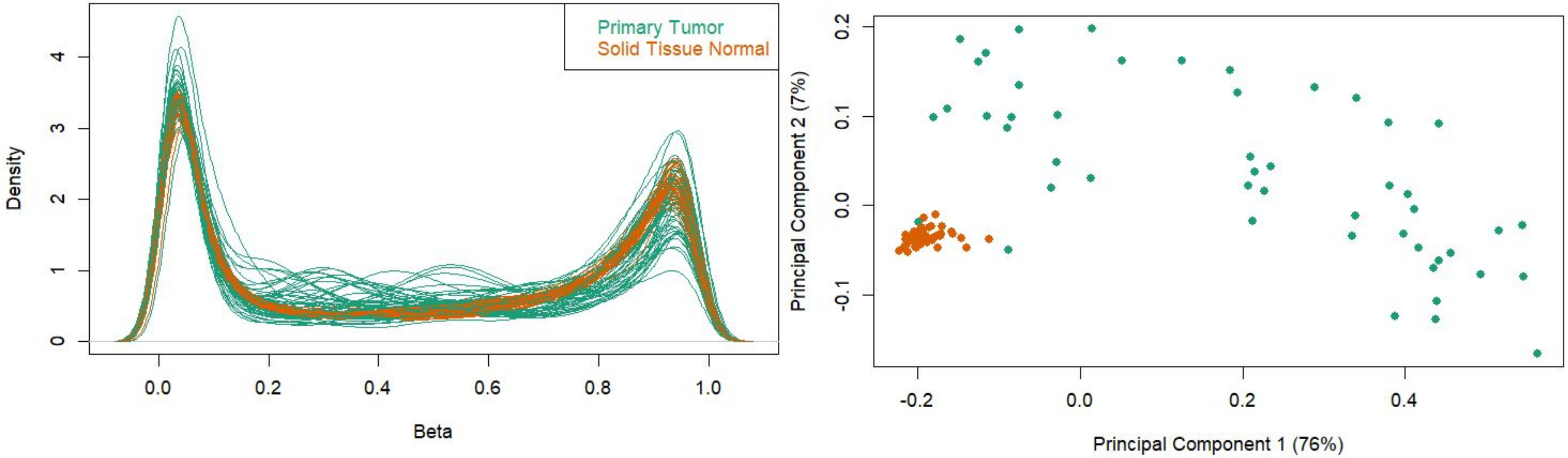
Data:

- 50 matched samples from tumor and normal tissue from TCGA
- Illumina methylation array 450k processed with SeSAMe
- files include beta-values of about 480.000 CpG loci

Preprocessing:

- exclusion of CpGs with missing beta-values
- exclusion of CpGs mapping to X or Y chromosome
- exclusion of CpGs overlapping with SNPs with MAF > 0.01

Quality Control and EDA



Differential methylation analyses

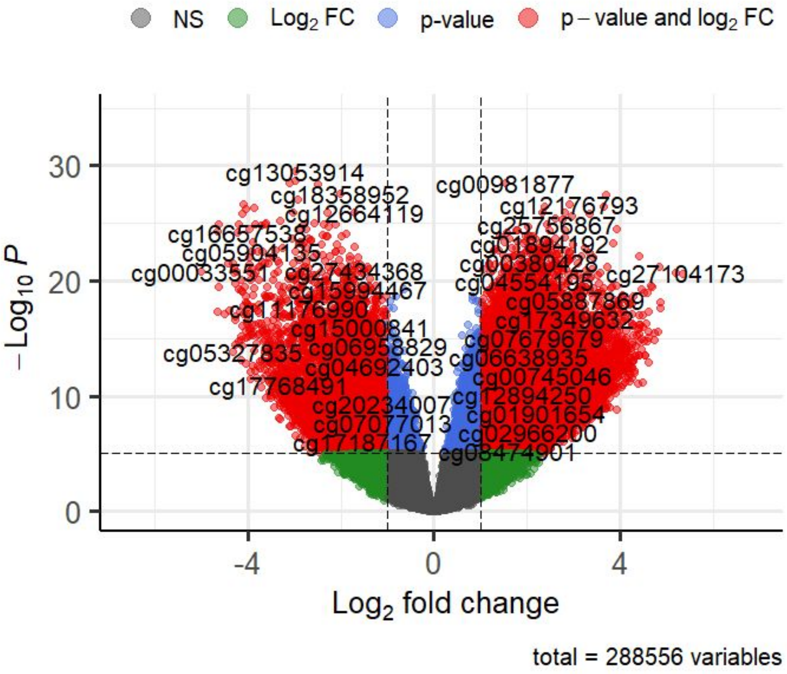
Limma

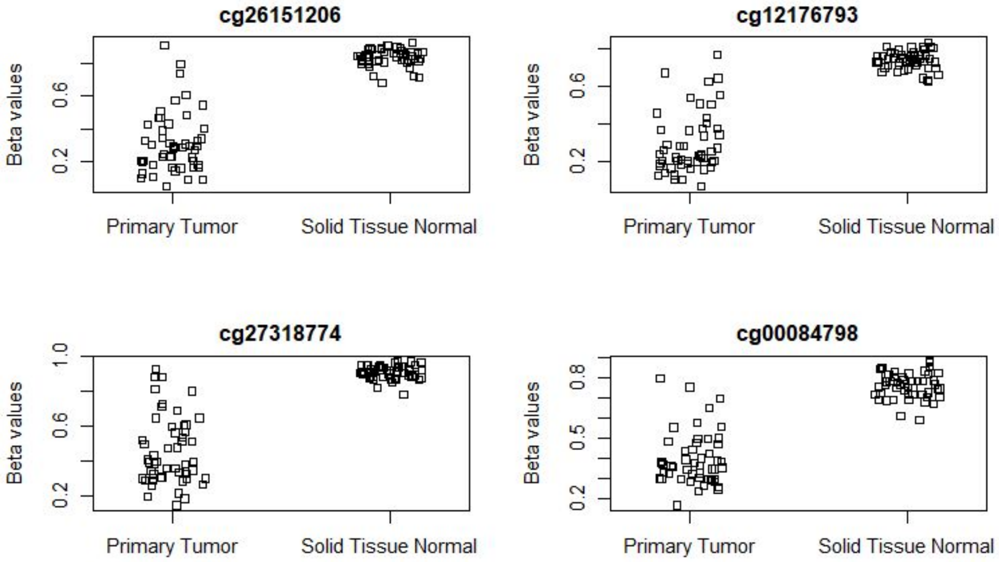
- extracted differential methylated loci (DML) between cancer and healthy tissues
- filtered the loci by adjusted p-value (<0.005) and mean beta value difference between groups (>0.2): 24457 DMLs

DMRcate:

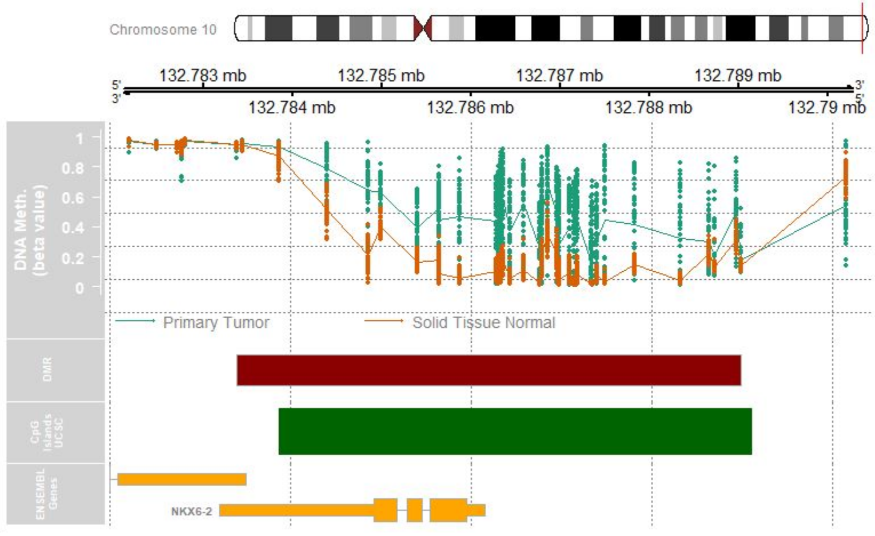
- combined CpG loci to differential methylated regions (DMRs): 4003 DMRs

Differential methylated loci

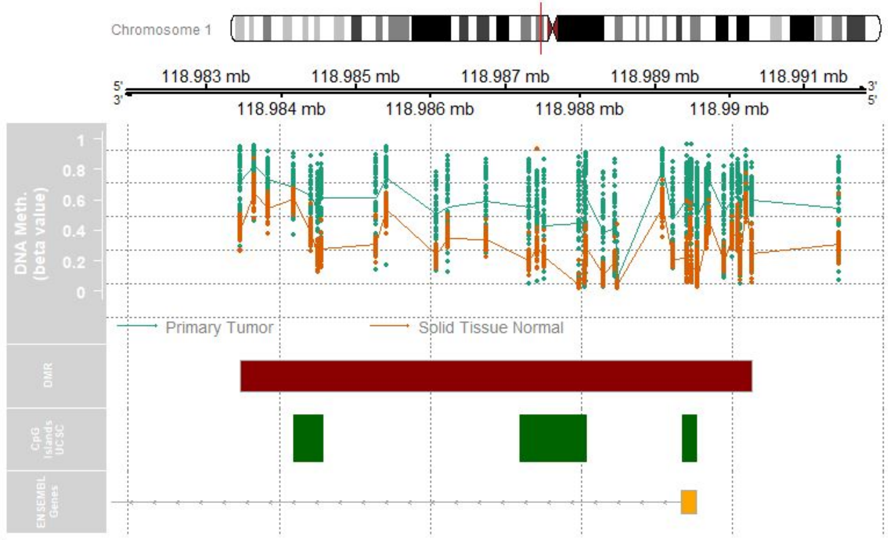




Differential methylated regions

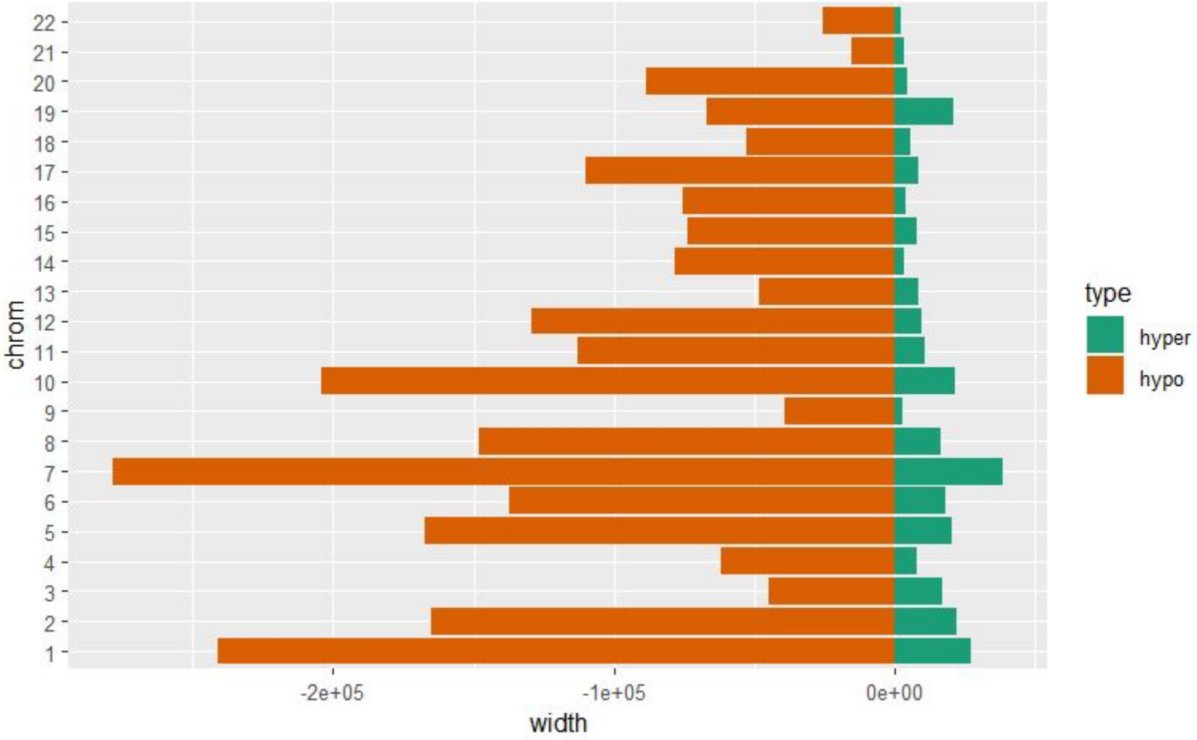


NKX6-2: predicted to be involved in cell differentiation; regulation of myelination; and regulation of transcription

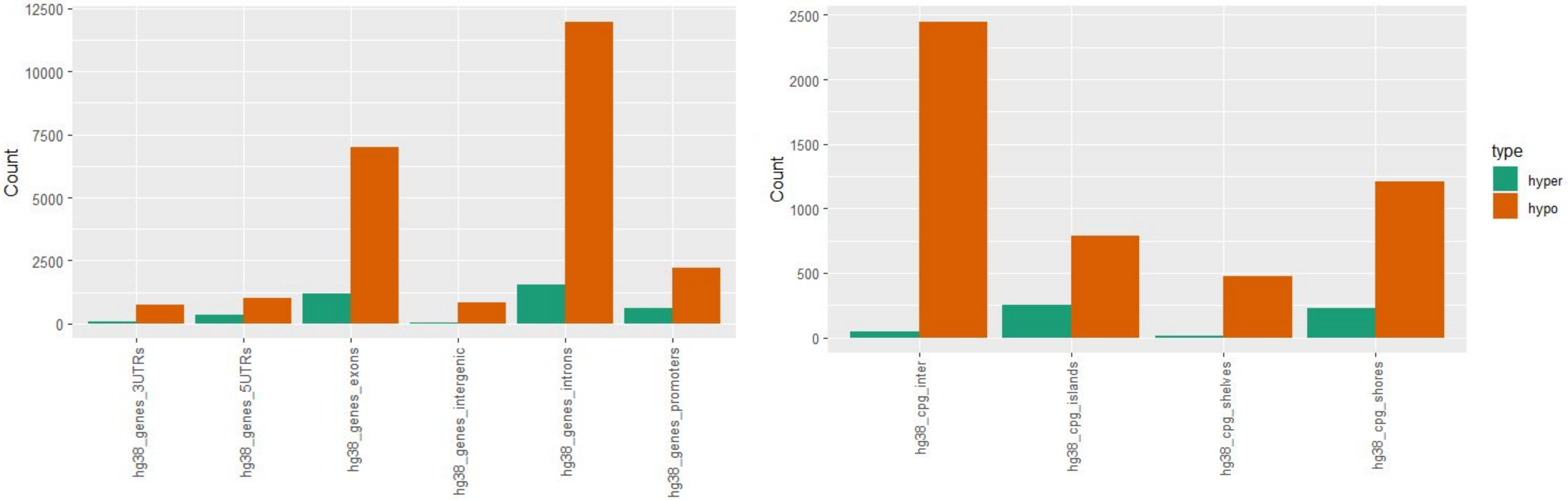


TBX15: regulate a variety of developmental processes, known to be hypermethylated in liver cancer

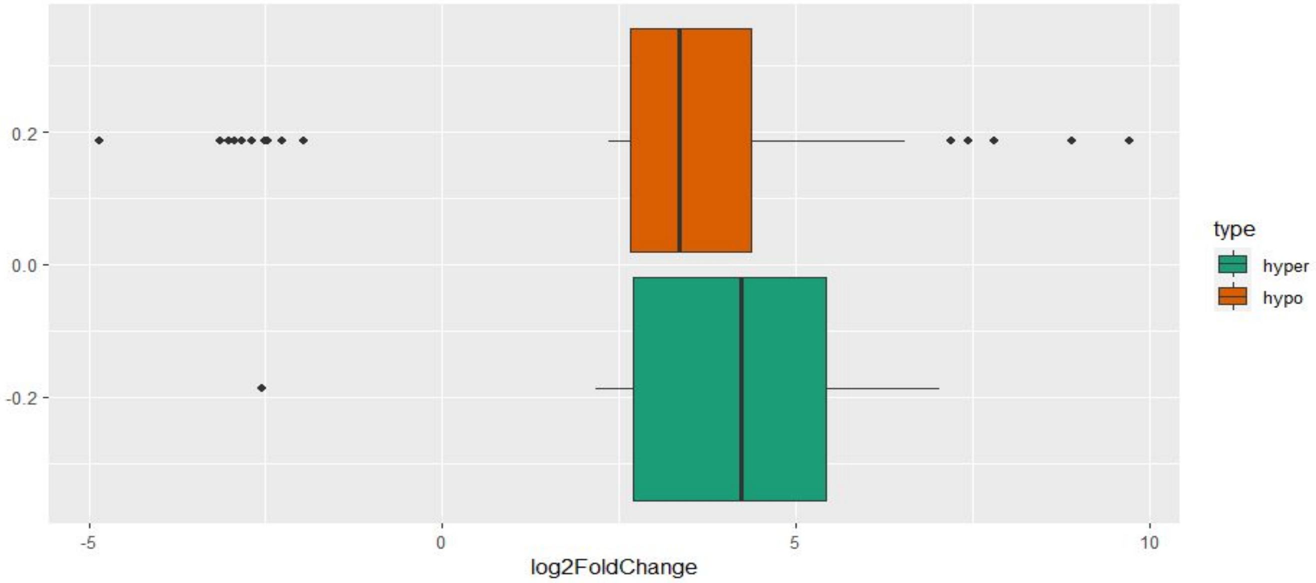
DMR annotation



DMR annotation



Promoter methylation and gene expression

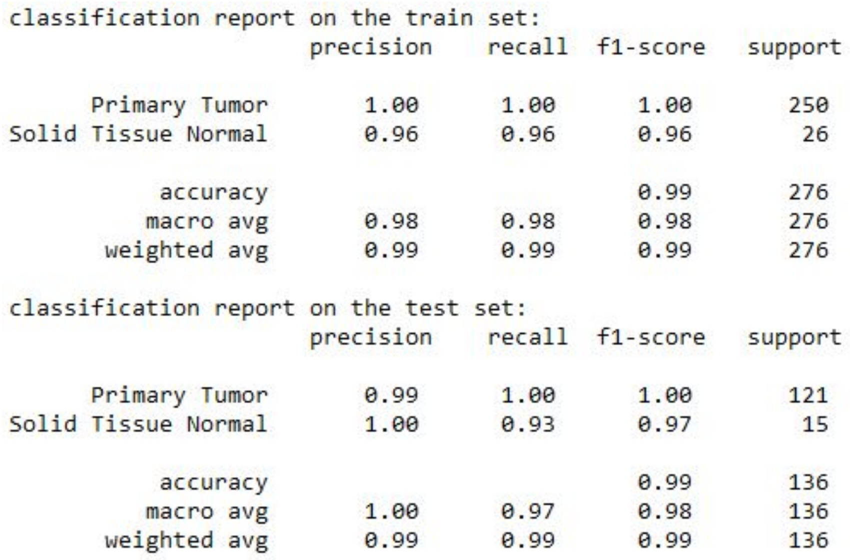


underexpressed

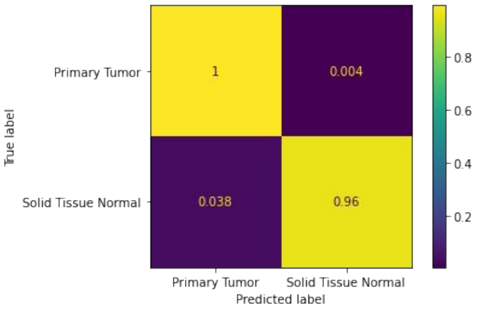
overexpressed

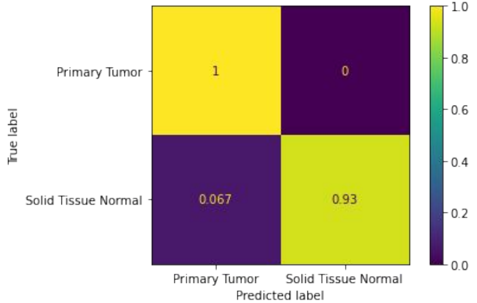
Random forest

predicts tissue type from DEGs (200 top up- and downregulated) and DMLs (top 500) between healthy (n=41) and tumor (n=371)



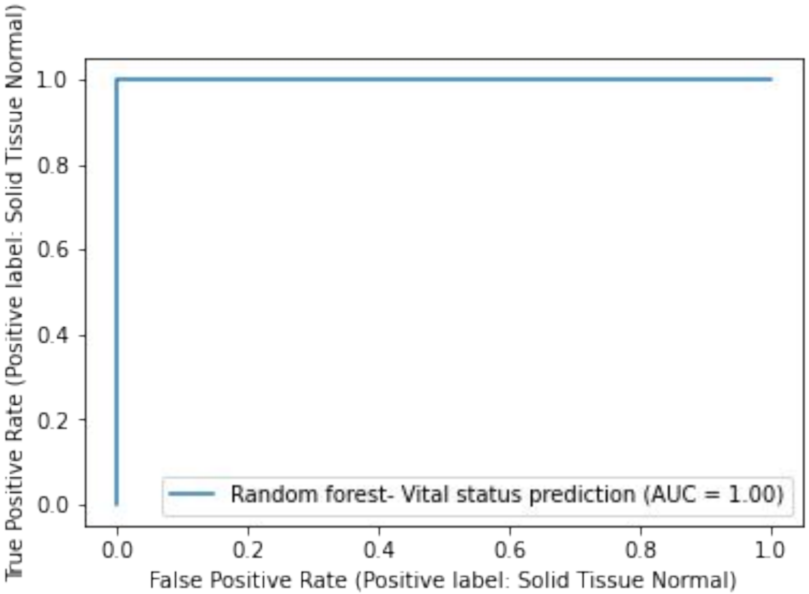
```
classification report on the train set:
                      precision    recall  f1-score   support

       Primary Tumor       1.00      1.00      1.00       250
  Solid Tissue Normal       0.96      0.96      0.96        26

            accuracy                           0.99       276
           macro avg       0.98      0.98      0.98       276
        weighted avg       0.99      0.99      0.99       276

classification report on the test set:
                      precision    recall  f1-score   support

       Primary Tumor       0.99      1.00      1.00       121
  Solid Tissue Normal       1.00      0.93      0.97        15

            accuracy                           0.99       136
           macro avg       1.00      0.97      0.98       136
        weighted avg       0.99      0.99      0.99       136
```





Random forest

predicts tissue type from DEGs (200 top up- and downregulated) and DMLs (top 500) between healthy (n=41) and tumor (n=371)

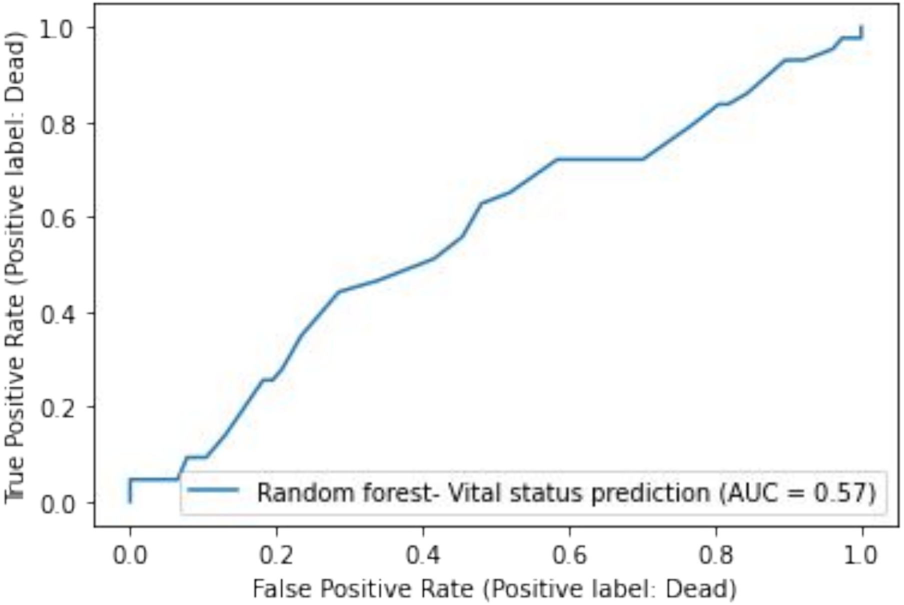


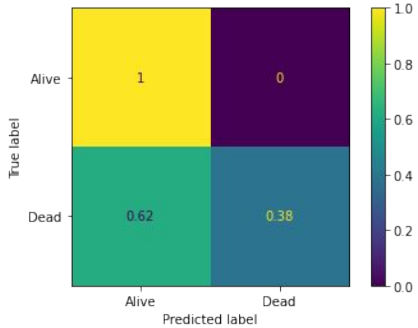
most important features:

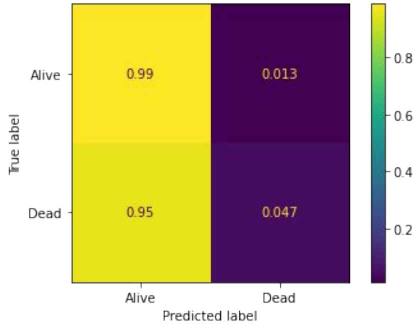
CLEC4M (known marker for hepatocellular carcinoma)
CRHBP, FCN2 …

Random forest

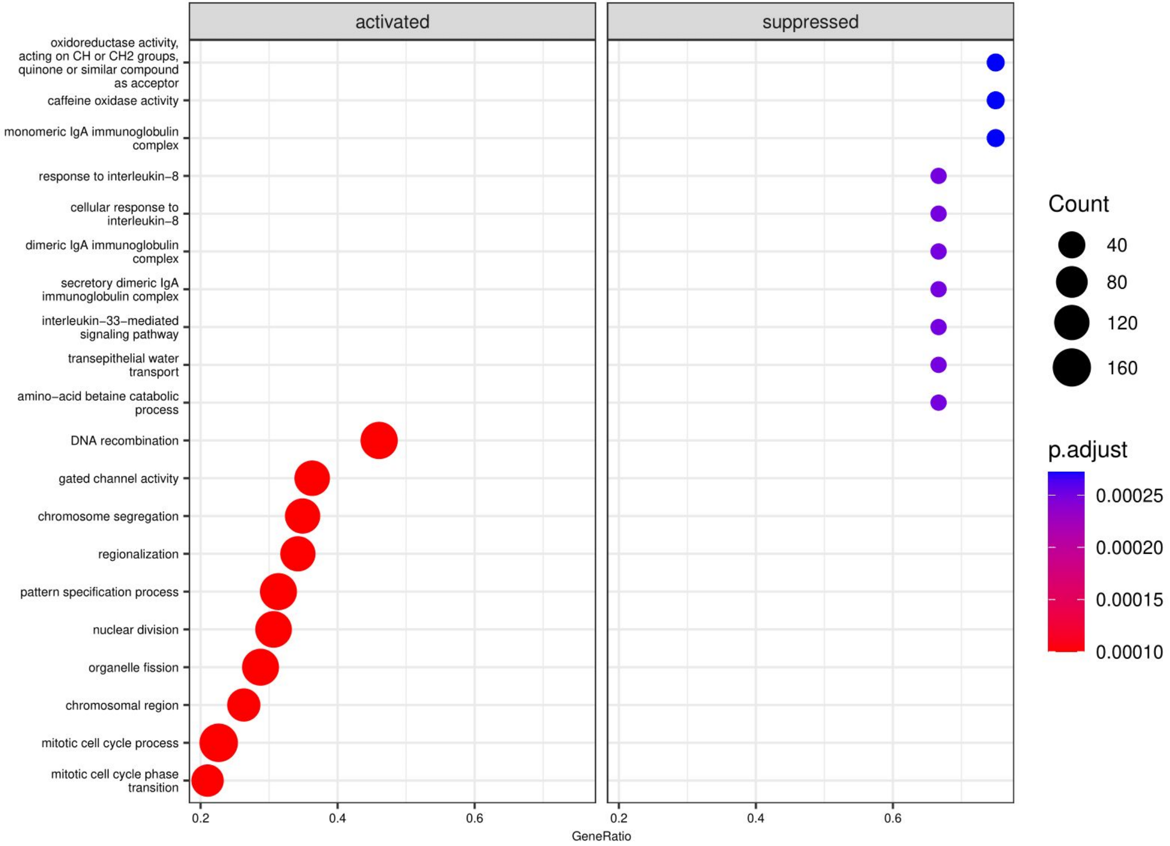
predicting vital status from DEGs, DMLs and CNVs between alive and dead patients

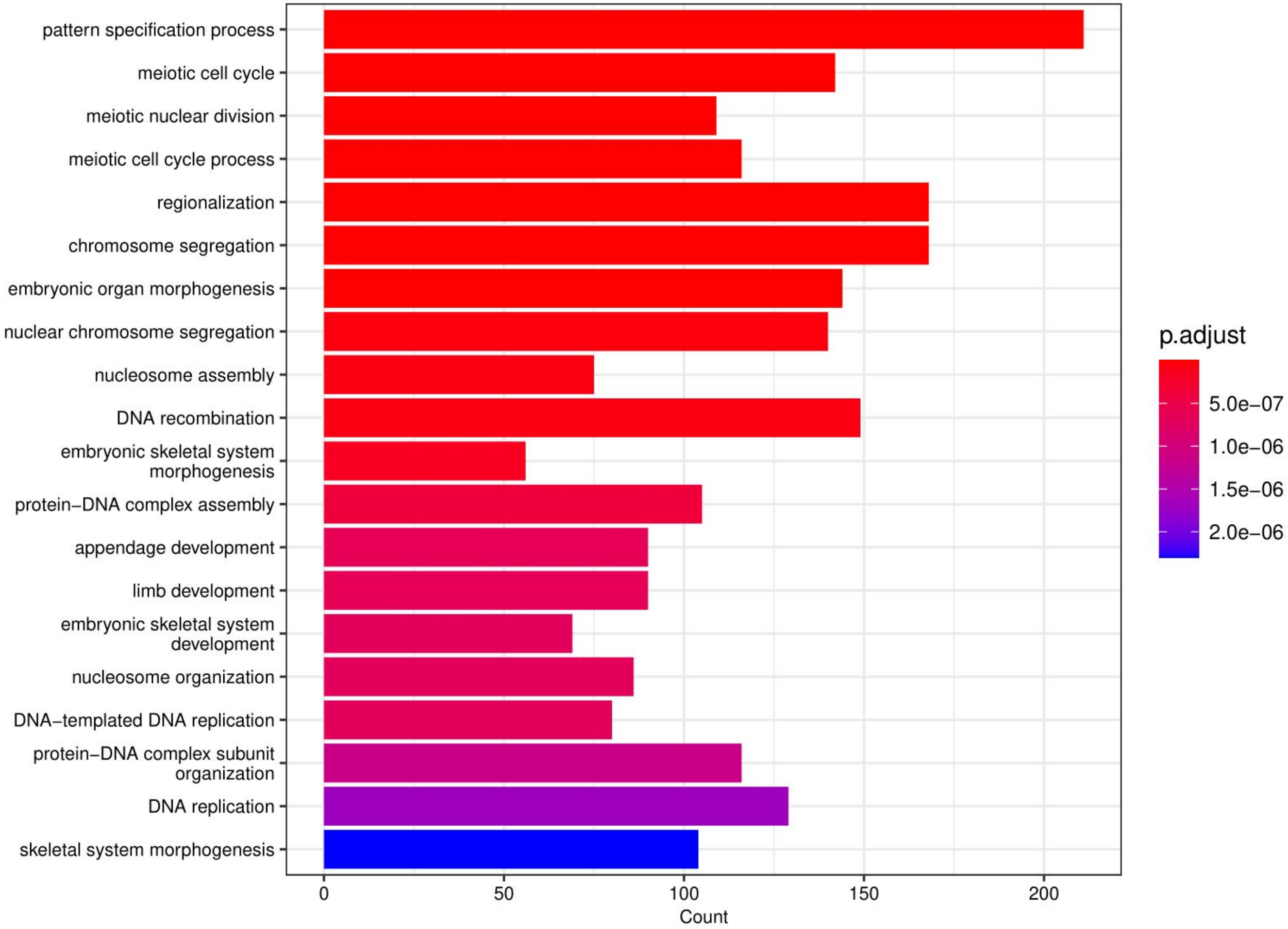






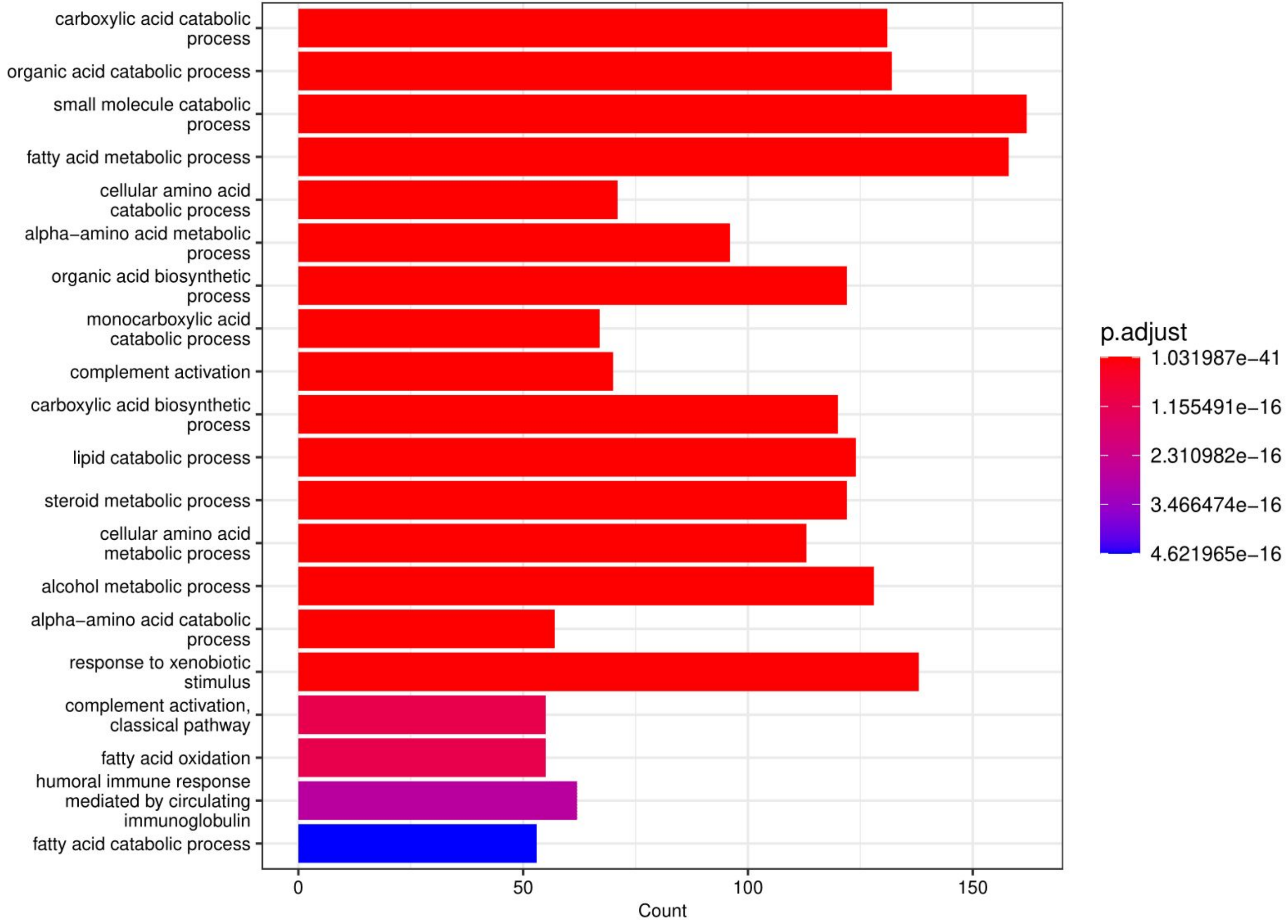
DEG Annotation





Upregulated:

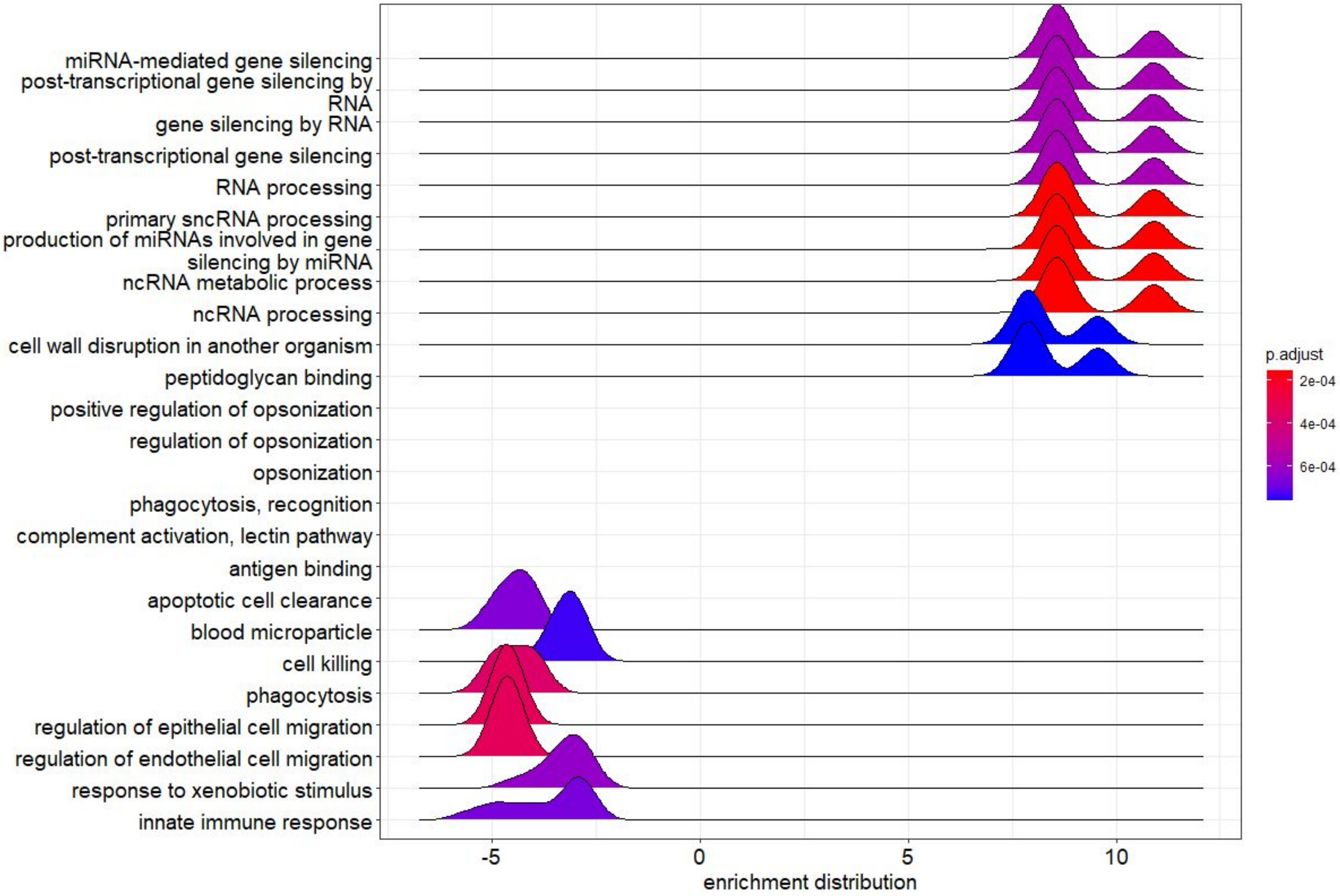
- Shows an increase in pathways that you would associate with cancer.
- Possibly shows dedifferentiation of the cells.

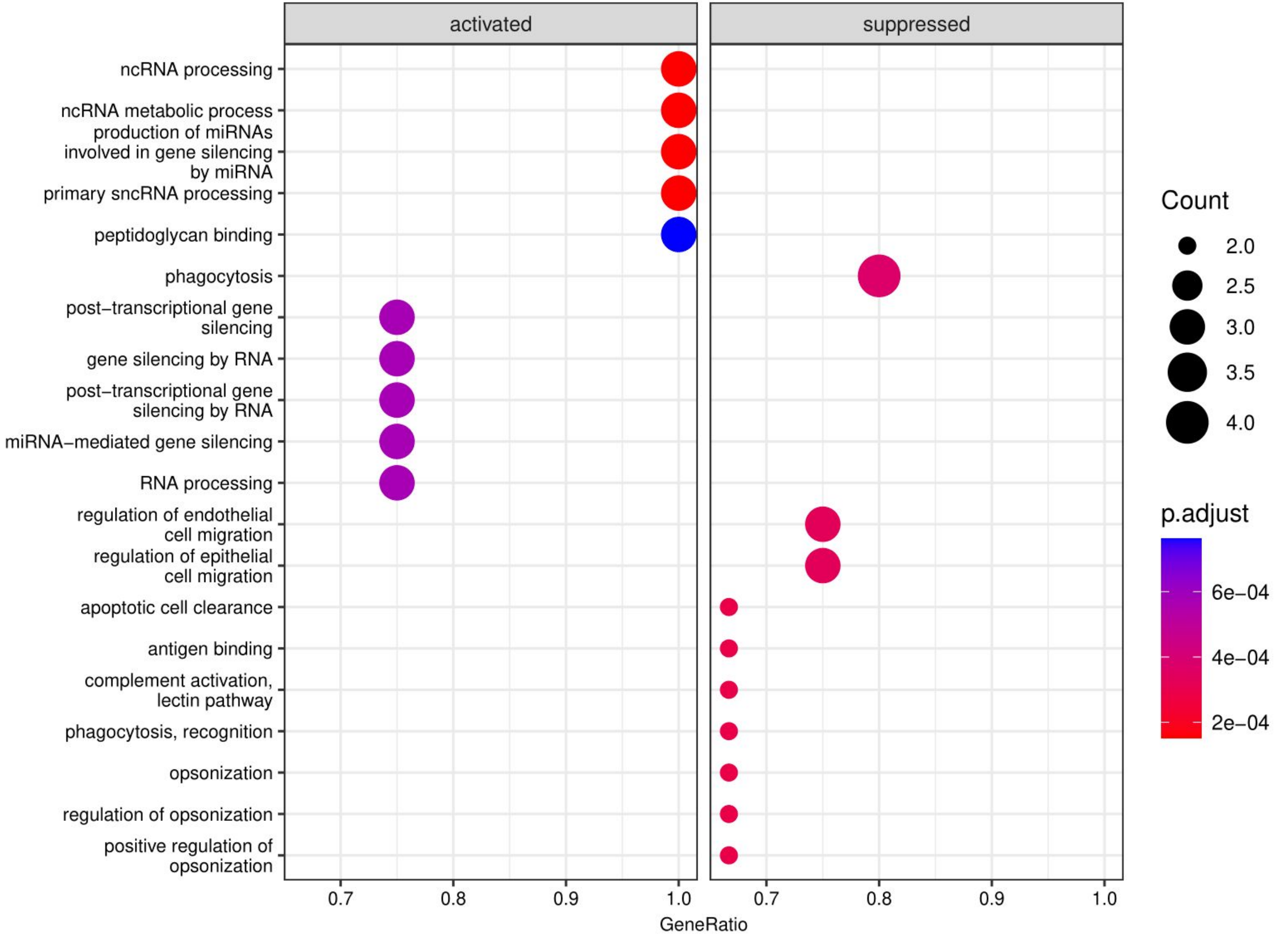


Downregulated

- Mostly associated with metabolic processes.
- Interestingly, catabolic processes are also downregulated

ML Features Annotation





ML Features

- Mostly associated gene regulation/ processing
- Also shows downregulation of immunogenic pathways

Future prospects

- Cancer evolution: One can carry out analysis of data from cancer patients suffering from various different stages of cancer (or ideally analysis of the same patient but over different time points).
- Cancer therapy: RNA metabolism and regulation seems to play an important role in determining the mortality of cancer patients. A deeper dive into these processes, especially the ones involving miRNAs, may help in discovering potential targets for therapeutic interventions.

Questions?