Multi-Omics analysis identifies signature genes to predict bladder cancer survival

Group 2: Pham Gia Cuong, Matanat Mammadli, Samra Hamidovic

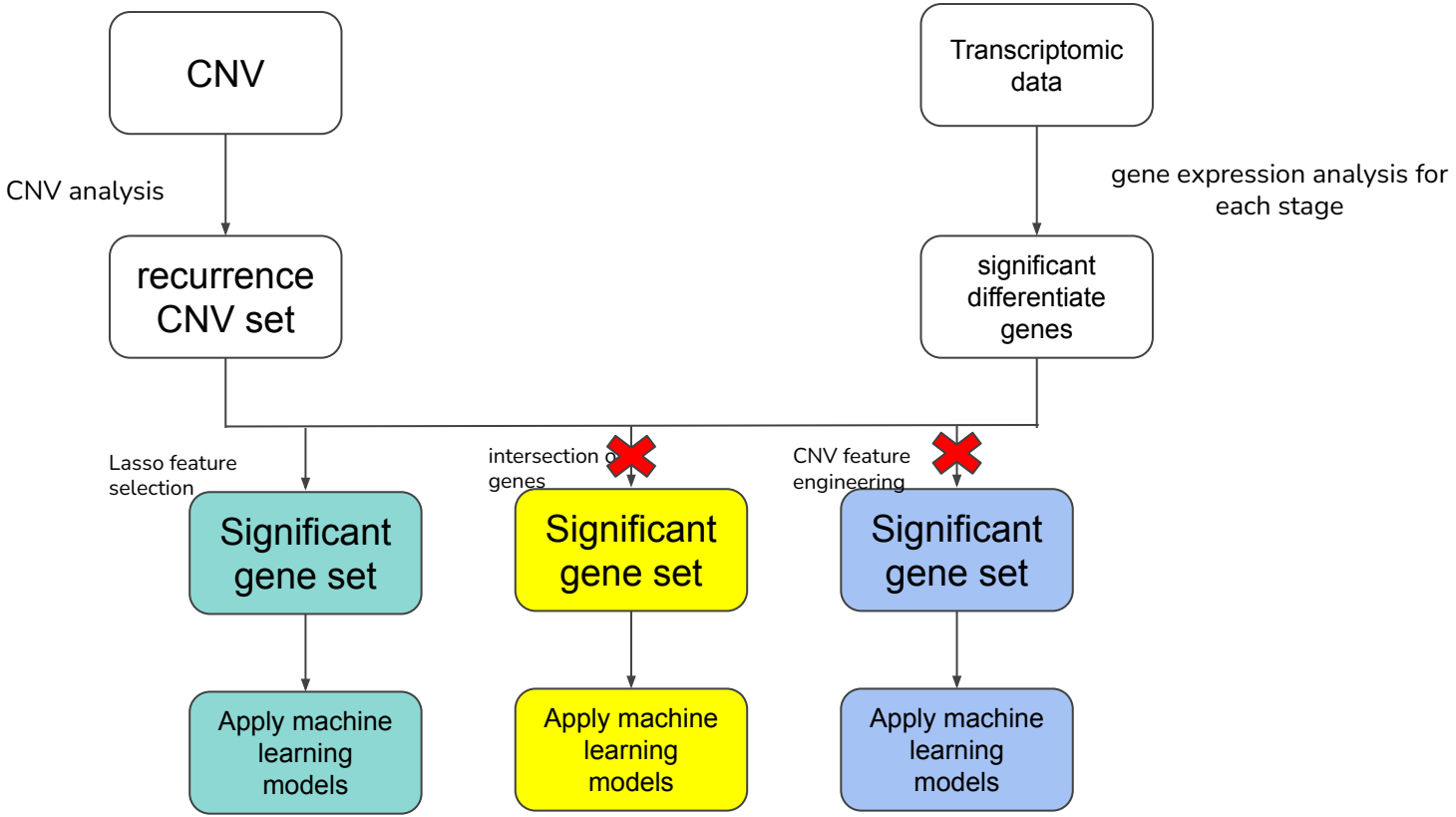
Overview

1. Remind
2. Workflow
3. CNV analysis
4. Gene expression analysis
5. Feature selection
6. Apply machine learning model
7. Survival analysis
8. GO enrichment analysis

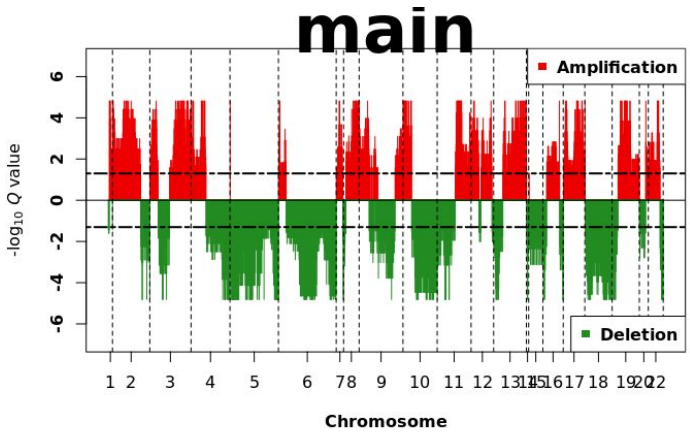
Remind:

- Aim:
  - Using CNV data and transcriptomic data to predict survival of Bladder Cancer patients
  - Key genes contain CNV and play a big role in survival of patient
  - Functional annotation
- Data:
  - TCGA Bladder cancer data
  - CNV data: CNVs from tumor and normal cells of different samples (within-sample homogeneity)
  - Transcriptomic data: 442 samples
  - Survival data: survival label and survival time.
  - Phenotype data:
    - tumor diagnose stage: 422 samples
    - lost of follow up patient: 150 samples are lost of follow up

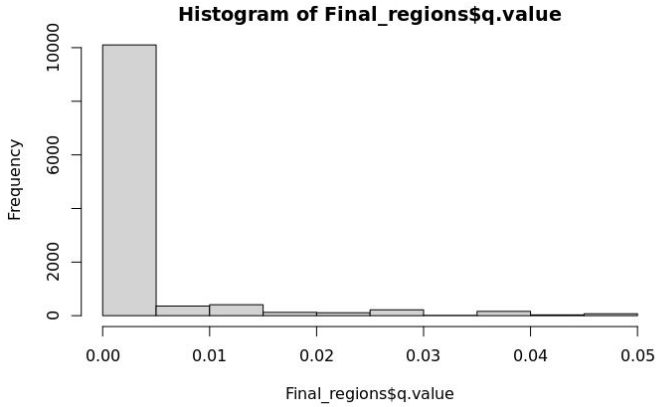
Workflow


CNV
Transcriptomic data
CNV analysis
gene expression analysis for each stage
recurrence CNV set
significant differentiate genes
Lasso feature selection
intersection of genes
CNV feature engineering
Significant gene set
Significant gene set
Significant gene set
Apply machine learning models
Apply machine learning models
Apply machine learning models

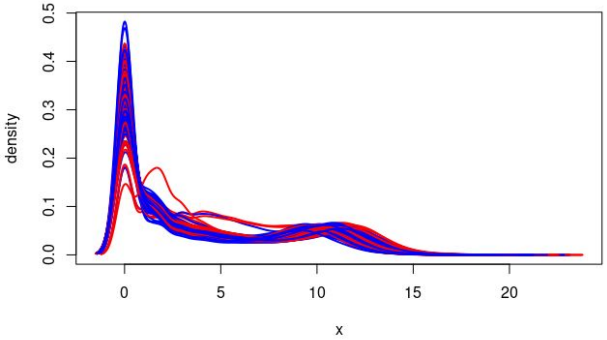
CNV analysis

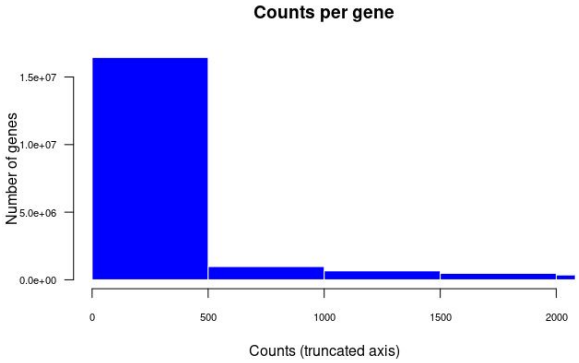


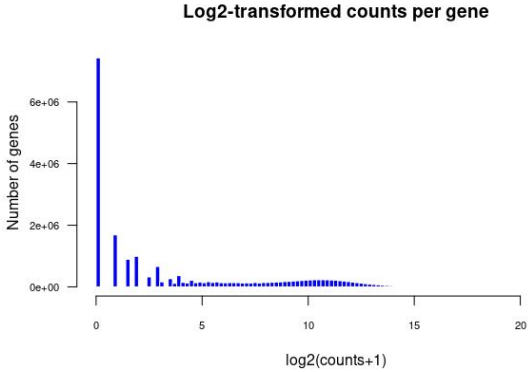
- 412 samples , each sample contains tumor and normal cell types
- CNVs: chromosome, start, end, probe and segmean
- Aim: finding independent and recurrent copy number abbreviations
    - r-package: GAIA (genomic analysis of significant chromosomal aberrations)
- Result: 3448 segments (q.value < 0.01) -> 11591 genes
- Cross checking with gene profile from gene expression data -> 9254 genes



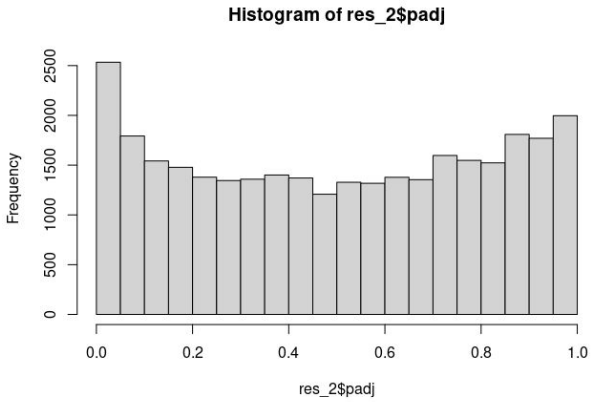
Gene expression analysis



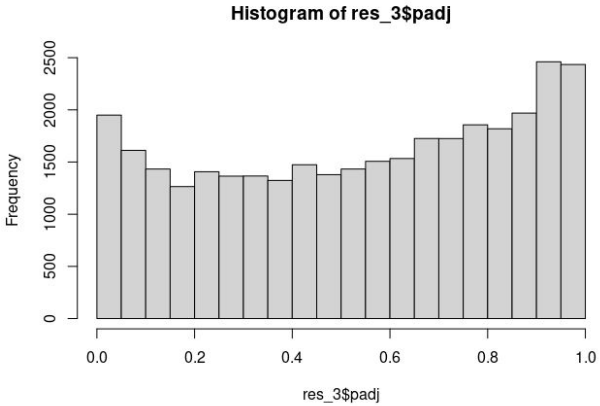




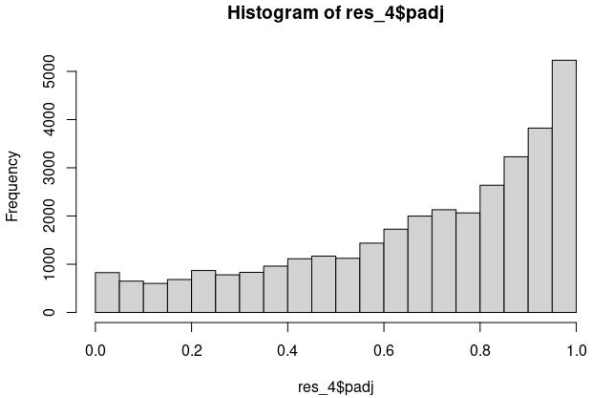
Gene expression analysis



Stage 2

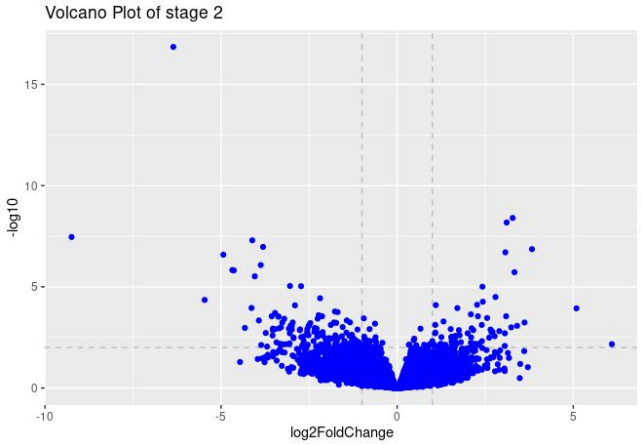


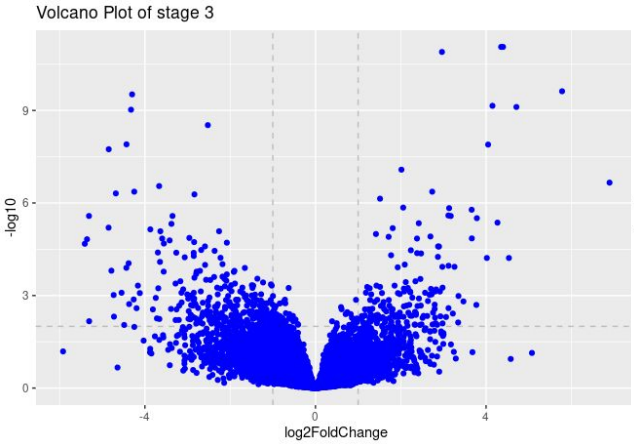
Stage 3

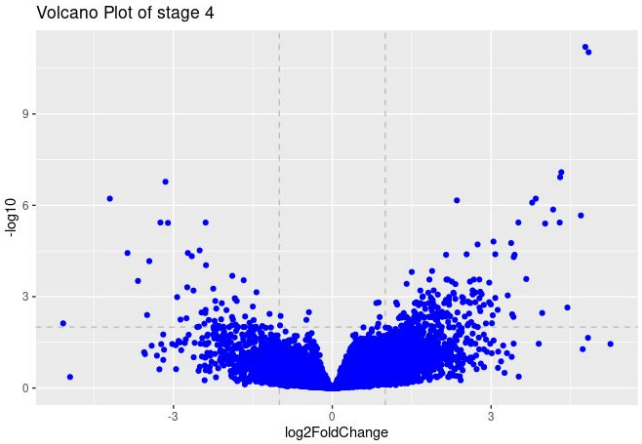


Stage 4

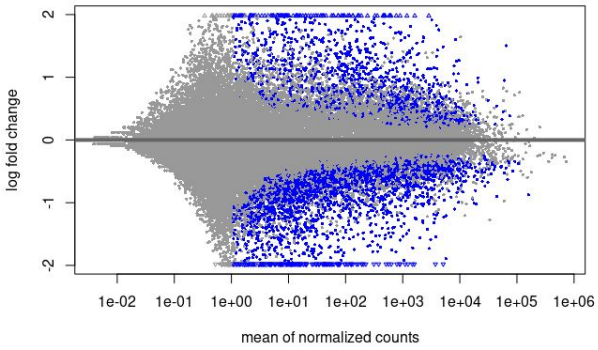
Gene expression analysis



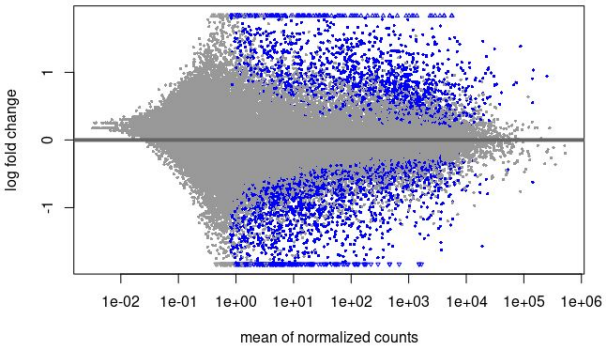




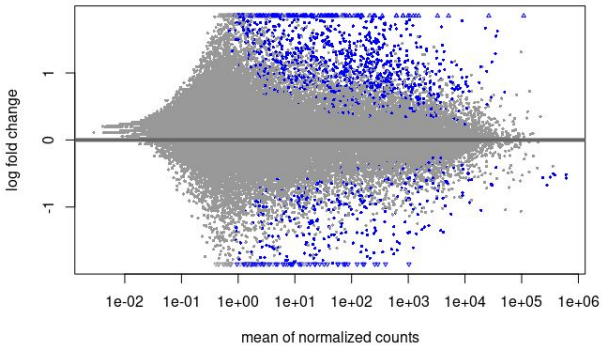
Gene expression analysis



Stage 2



Stage 3

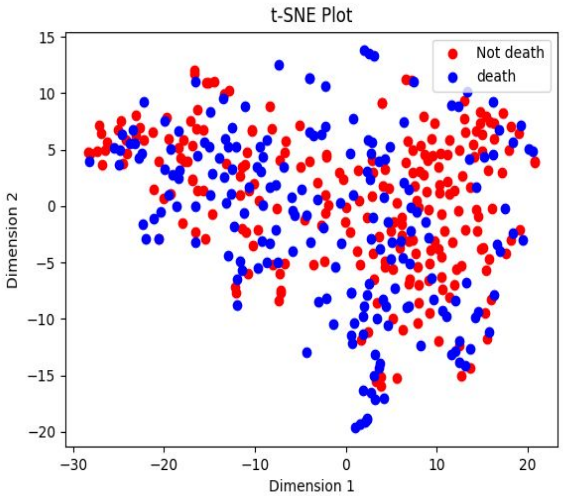


Stage 4

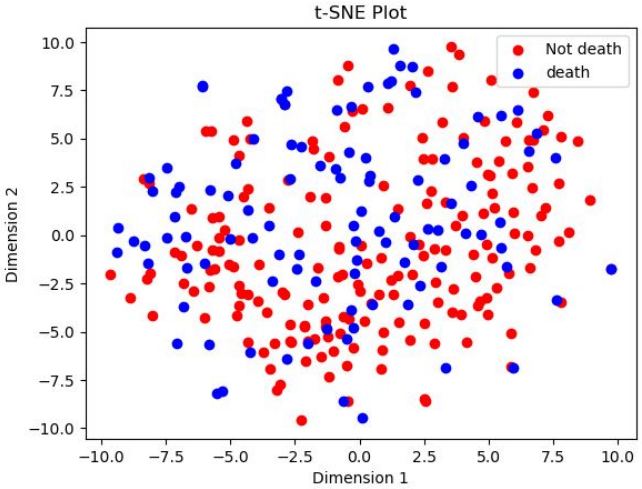
Gene expression analysis

	significant genes	samples
stage 2	619	111
stage 3	1003	97
stage 4	607	84

tSNE plot

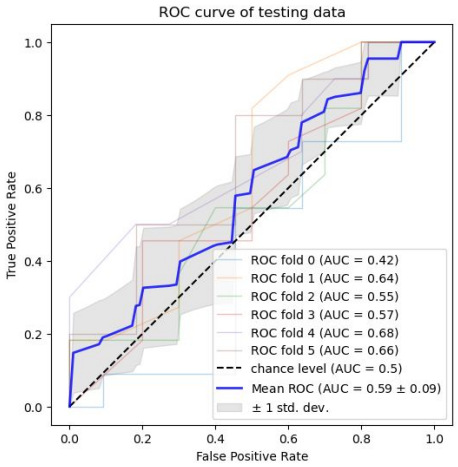


t-SNE on original raw count

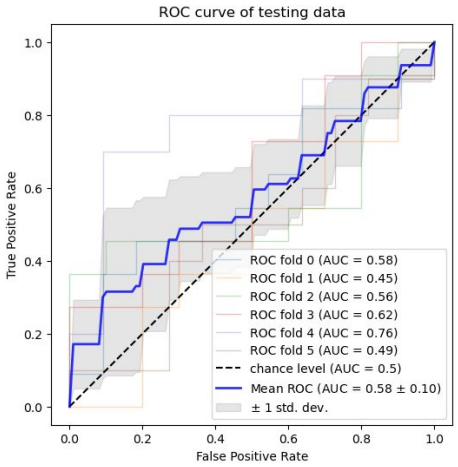


t-SNE on significant genes raw count

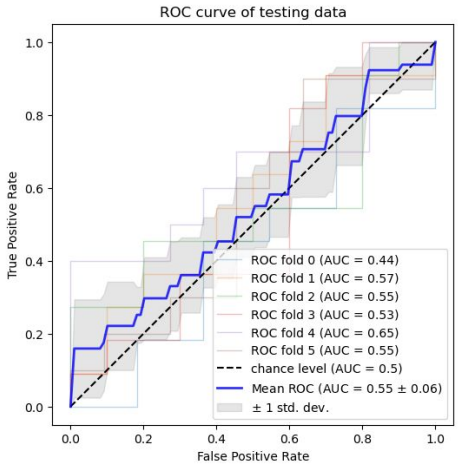
Apply machine learning models



Random forest



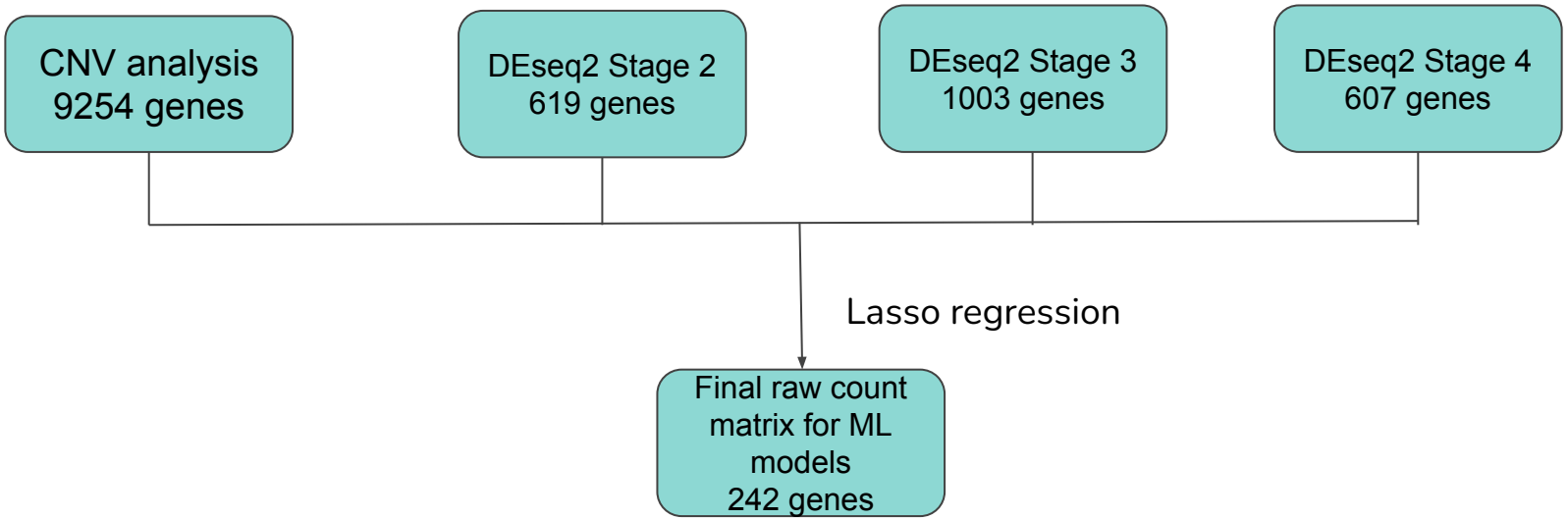
Ridge regression



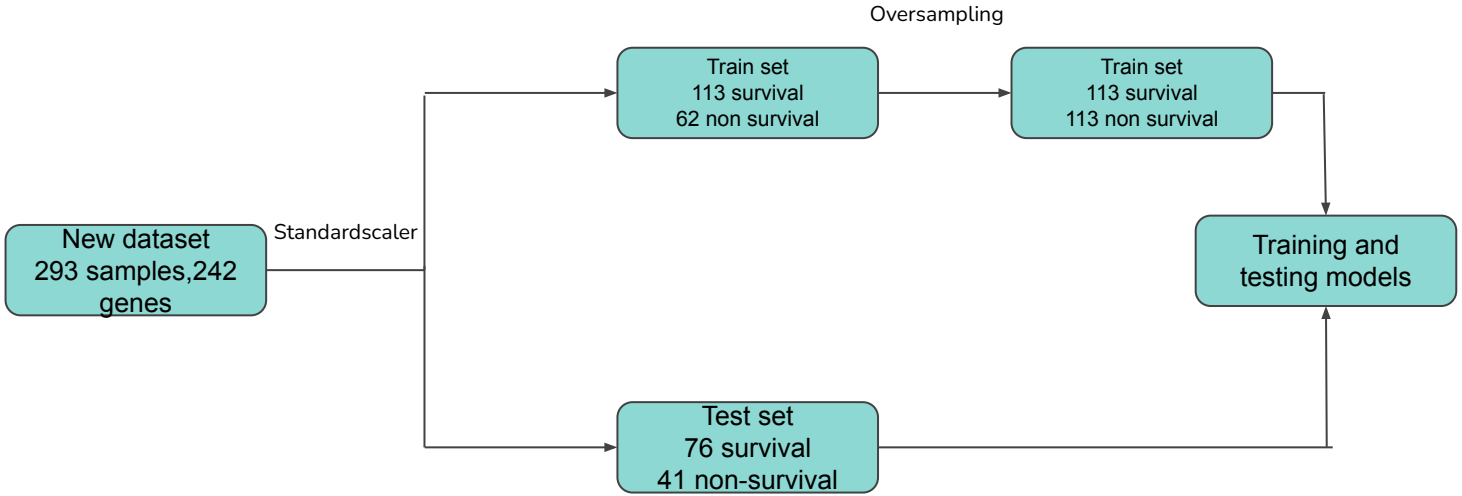
XGBoost

Result with the features from intersection of CNV gene set and DESeq2

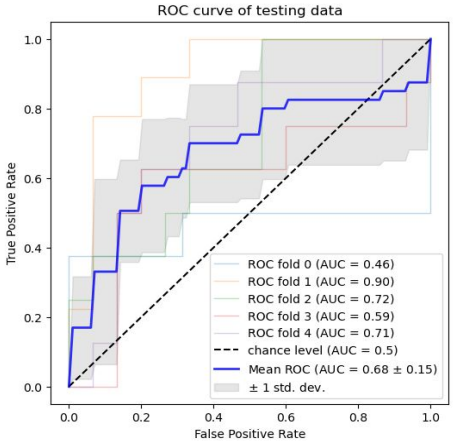
Apply machine learning model



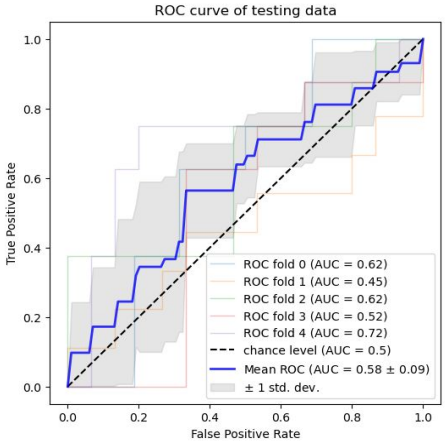
Apply machine learning model



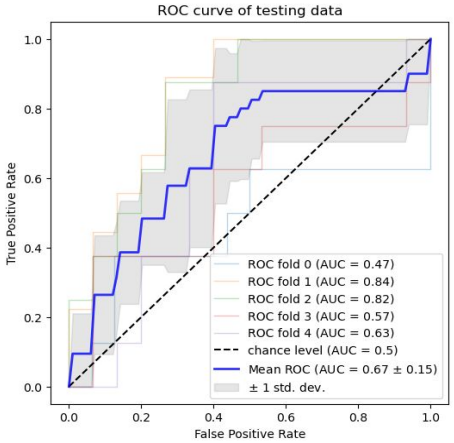
Apply machine learning model



Random forest



Ridge regression



XG Boost

Random forest

Hyperparameters:

- max features : sqrt, log2
- max depth of each tree : 10, 12, 14, 16, 20
- criterion: gini, entropy

Gridsearchcv -> max_depth=10, max_features='log2'

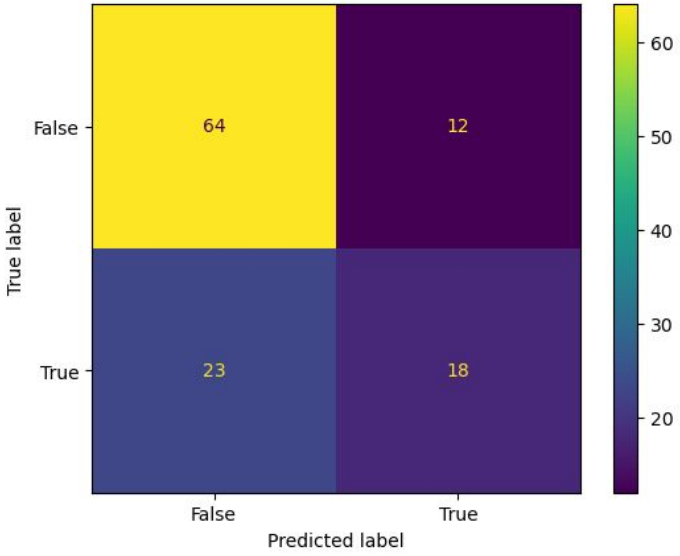
Random forest

Test set

	precision	recall	f1-score	support
0	0.74	0.84	0.79	76
1	0.60	0.44	0.51	41
accuracy			0.70	117
macro avg	0.67	0.64	0.65	117
weighted avg	0.69	0.70	0.69	117

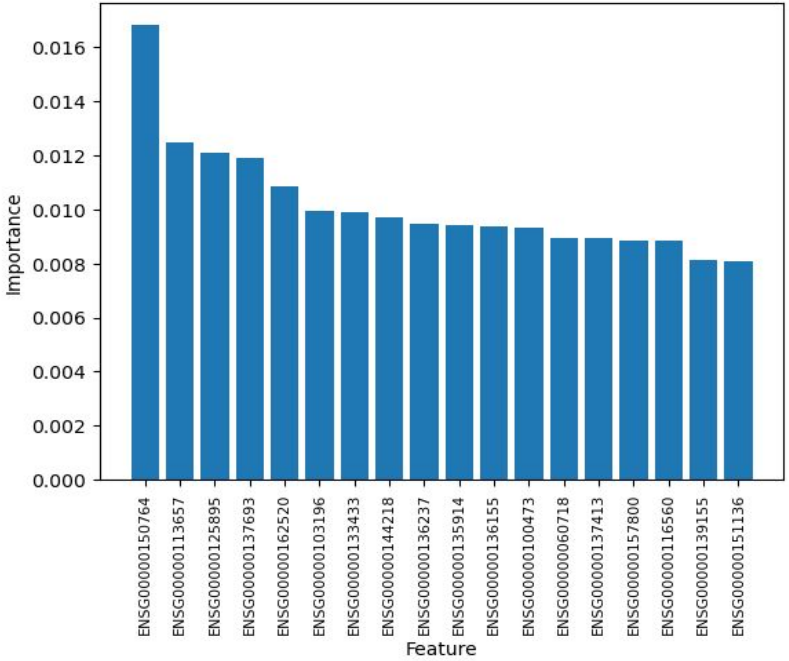
Train set

	precision	recall	f1-score	support
0	1.00	1.00	1.00	113
1	1.00	1.00	1.00	113
accuracy			1.00	226
macro avg	1.00	1.00	1.00	226
weighted avg	1.00	1.00	1.00	226



OVERFITTED :(

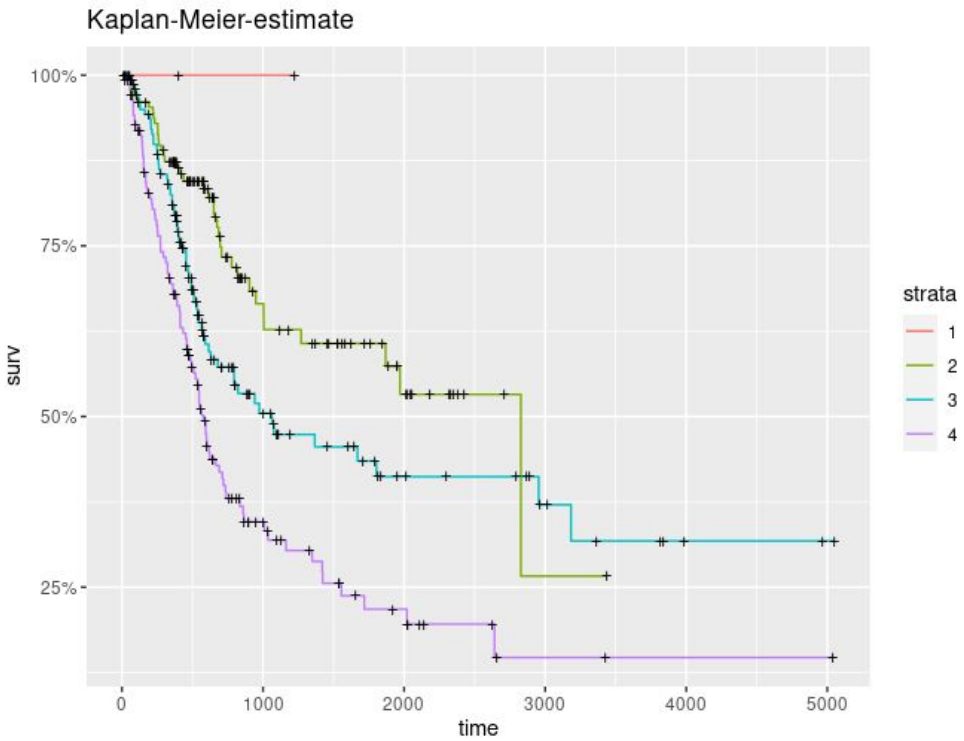
Features importance



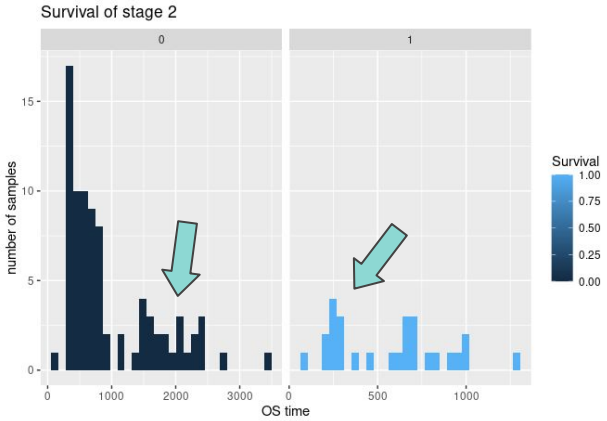
Ensemble ID	Gene name	Cancer related information
ENSG00000150764	***DIXDC1***	Participating in growing of tumor (1)
ENSG00000113657	***DPYSL3***	high DPYSL3 expression predicted a higher bladder tumor recurrence rate in patients (2)
ENSG00000125895	***TMEM74***	High expression of TMEM74 significantly shortens the surviving periods of patients in several types of cancer (3)
ENSG00000137693	***YAP 1***	Yap1 also plays an important role in the development of bladder and the deregulation of Yap1 is significantly associated with the development and metastasis of human bladder cancer (4)

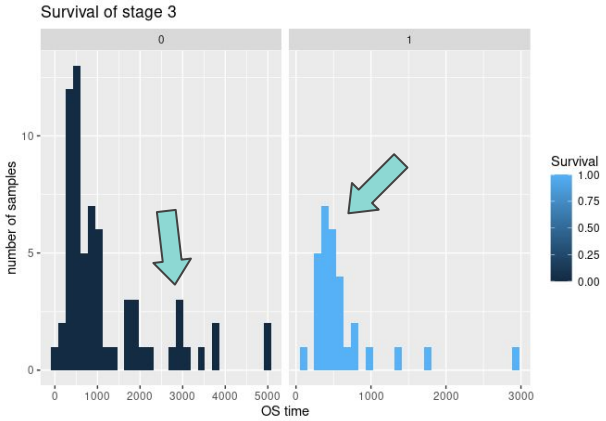
Survival analysis

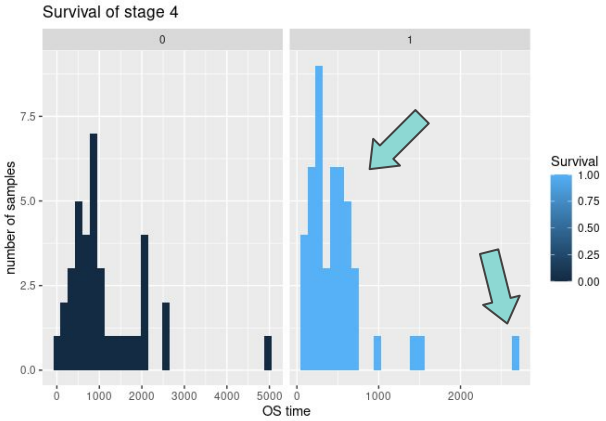
- **Reminder**:
  Kaplan-Meier-curve for each stage
- Stage = Confounder



Survival analysis

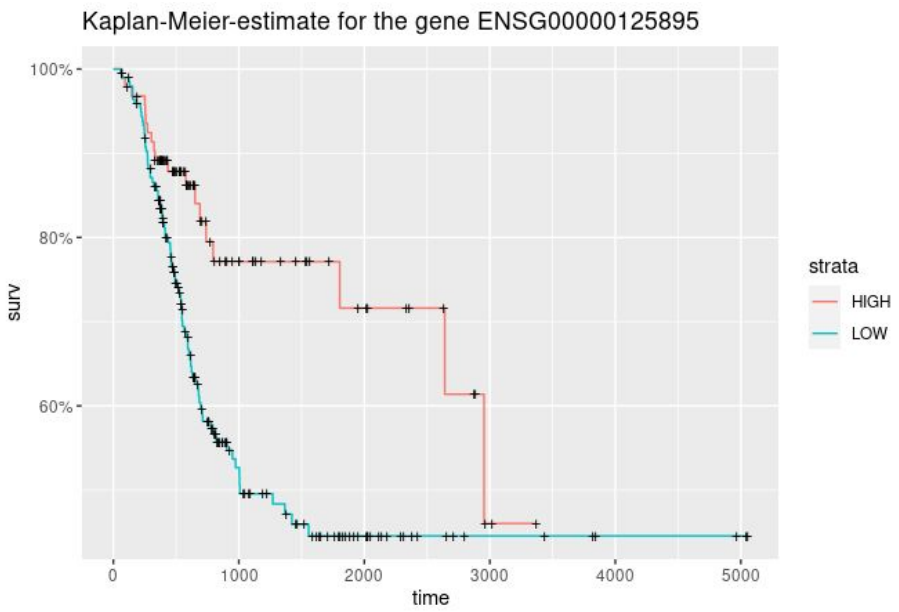




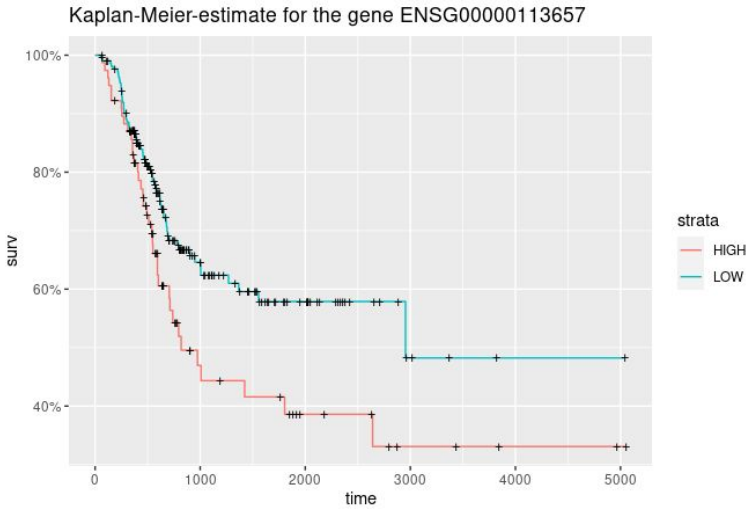


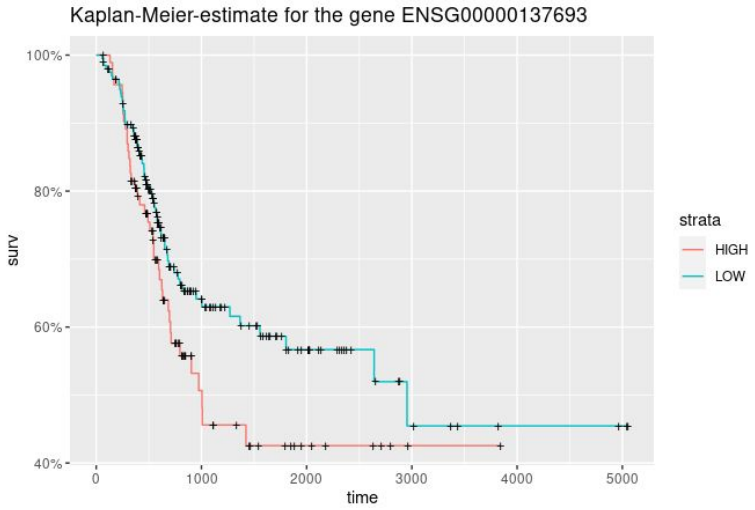
Survival analysis

- highest and lowest counts of one gene
- upregulated and downregulated
- Gene ***TMEM74***
- High expression shortens the surviving periods of patients



Survival analysis

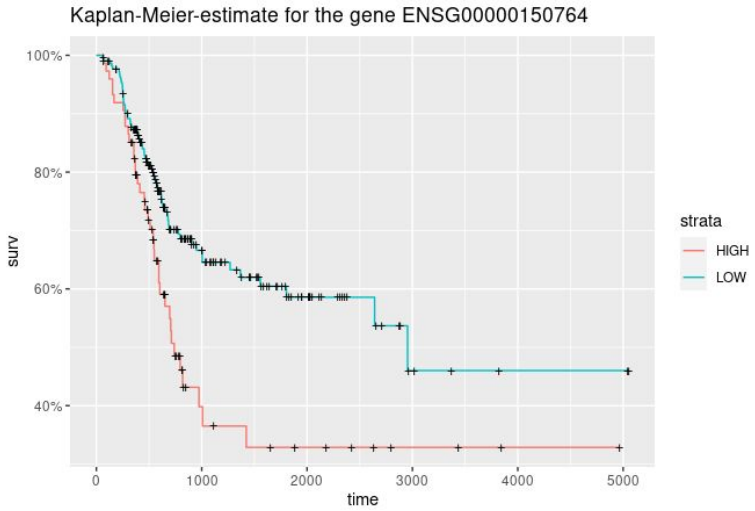




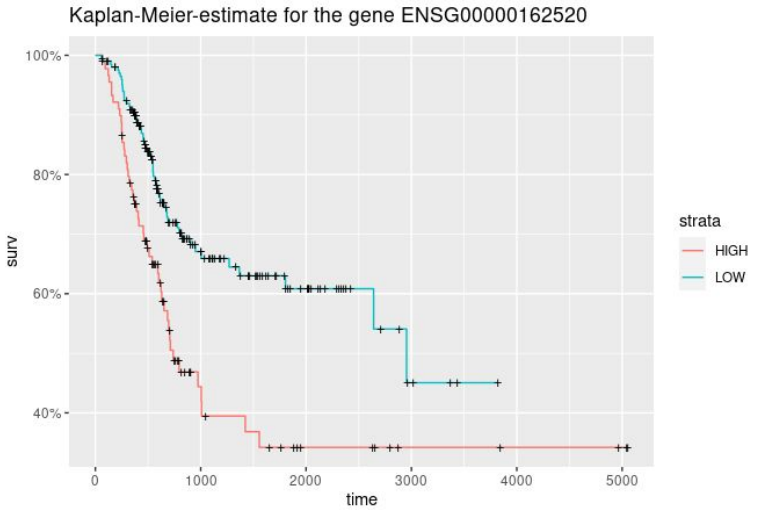
DPYSL3
high DPYSL3 expression predicted a higher bladder tumor recurrence rate

YAP 1
deregulation of Yap1 is significantly associated with the development and metastasis of BLCA

Survival analysis



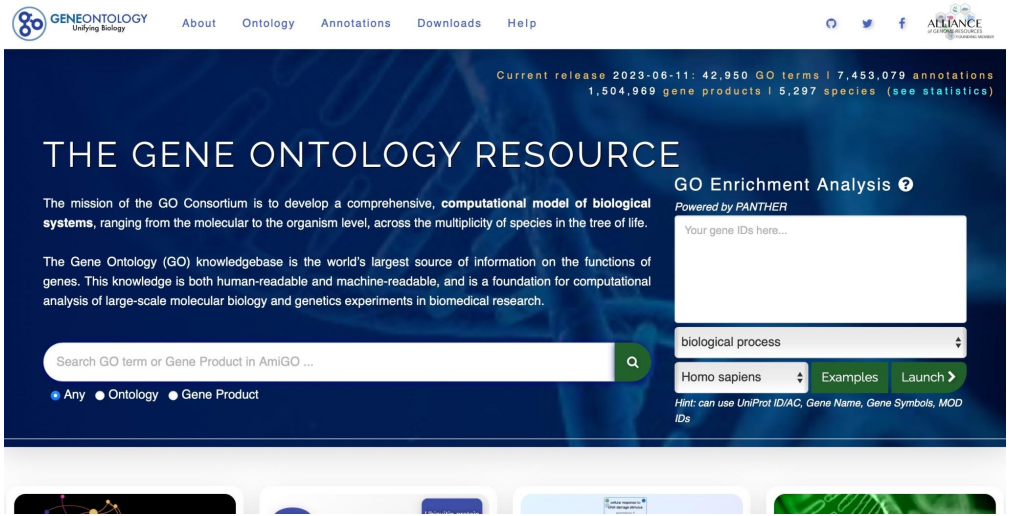
DIXDC1
high DPYSL3 expression predicted a higher bladder tumor recurrence rate



SYNC

GO Enrichment analysis

Gene Ontology (GO) enrichment analysis is used for interpreting high throughput molecular data and generating hypotheses about underlying biological phenomena of experiments.



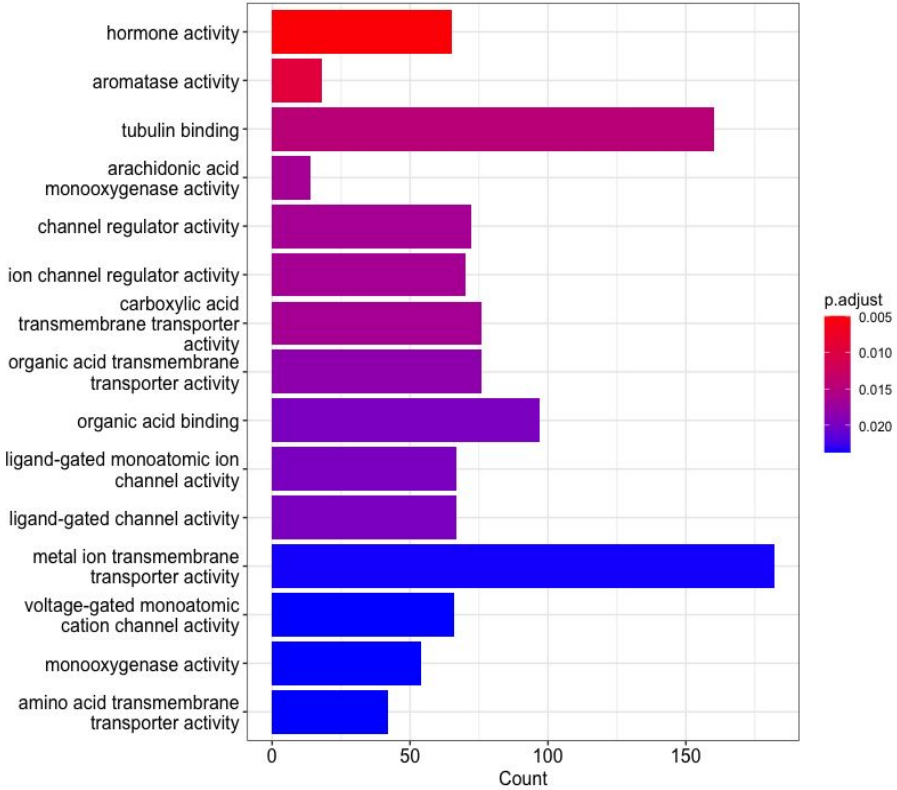
GO Enrichment Analysis

- One of the main uses of the GO (Gene Ontology) is to perform enrichment analysis on gene sets.
- For example, given a set of genes that are up-regulated under certain conditions, an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for that gene set.

```r
GO_results <- enrichGO(gene = genes_to_test,
                       OrgDb = "org.Hs.eg.db",
                       keyType = "ENSEMBL",
                       ont = "MF",
                       readable = TRUE
                       )
GO_results
```

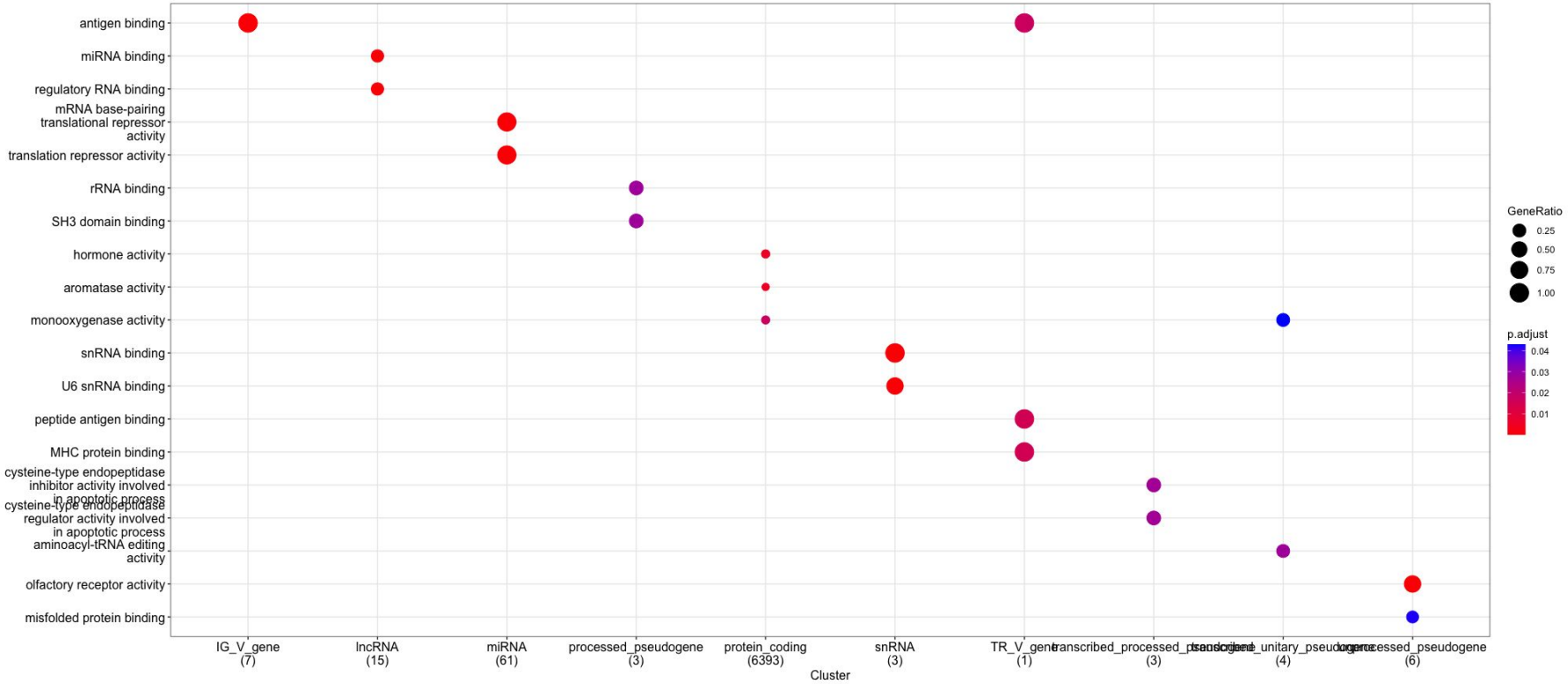
GO Enrichment analysis (Results)

Barplot of our GO results



GO Enrichment analysis (Results)

clusterProfiler::dotplot()

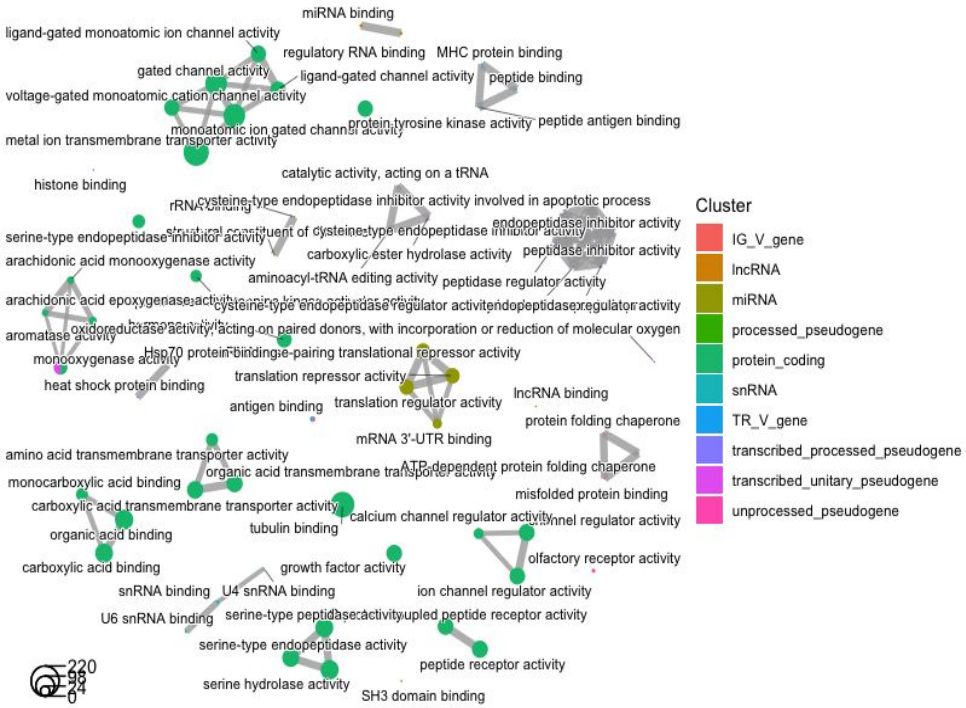


GO Enrichment analysis (Results)

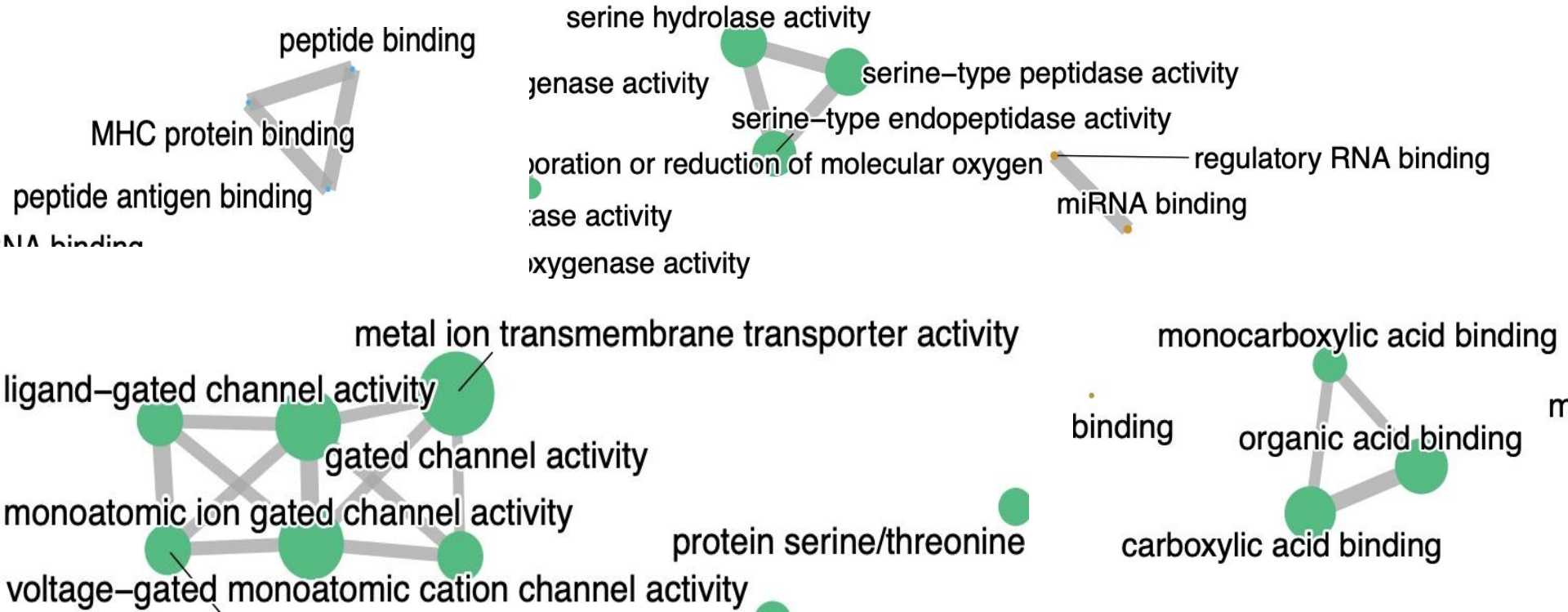
compareCluster()

enrichplot::emapplot()

cowplot::plot_grid()



Clusters close-up



Conclusion

What difficulties have we faced:

- finding dataset
- preprocessing took longer than we thought (no LIMMA, but DESeq2)
- accuracy of our data was 70% and overfitted
- confounder: stages, lost of follow up
- Install outdated R packages

What have we learned?

- CNV analysis, DESeq2 analysis
- GO enrichment analysis

What could be improved?

- Find confounders and insights of metadata before starting
- Always checking for overfitting