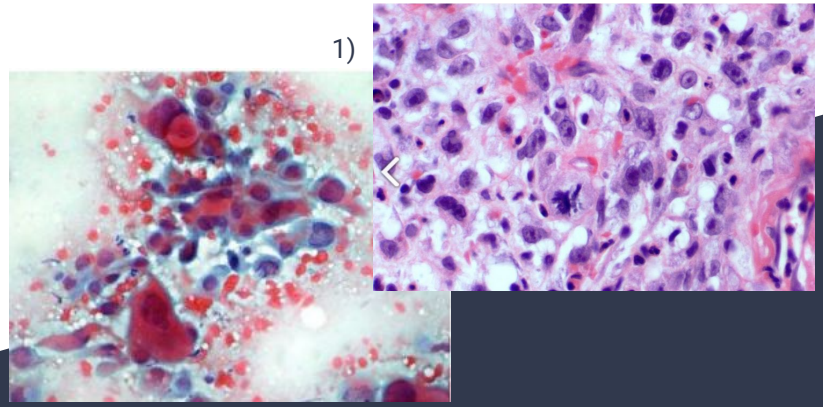
Transcriptomic and epigenomic biomarkers to differentiate between smokers with lung or laryngeal cancer

Final presentation

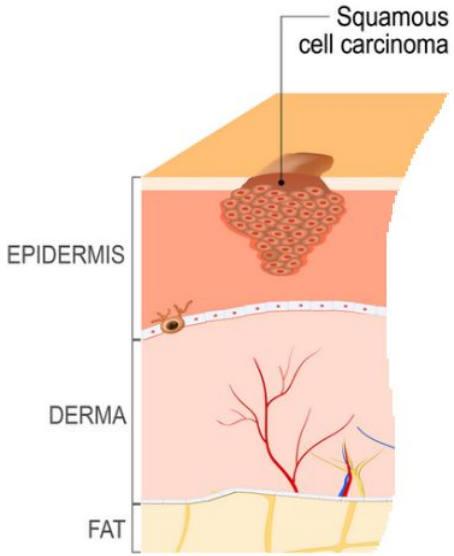
1)



Group 5
Mia Anscheit
Matanat Mammadli
Friederike Wohlfarth

1) Squamous Cell Carcinoma of the lung (left) and of the larynx (right). Images from were obtained from https://www.pathologyoutlines.com/topic/larynxcarcinomageneral.html, https://www.cellnetpathology.com/, last accessed 2024/07/07.

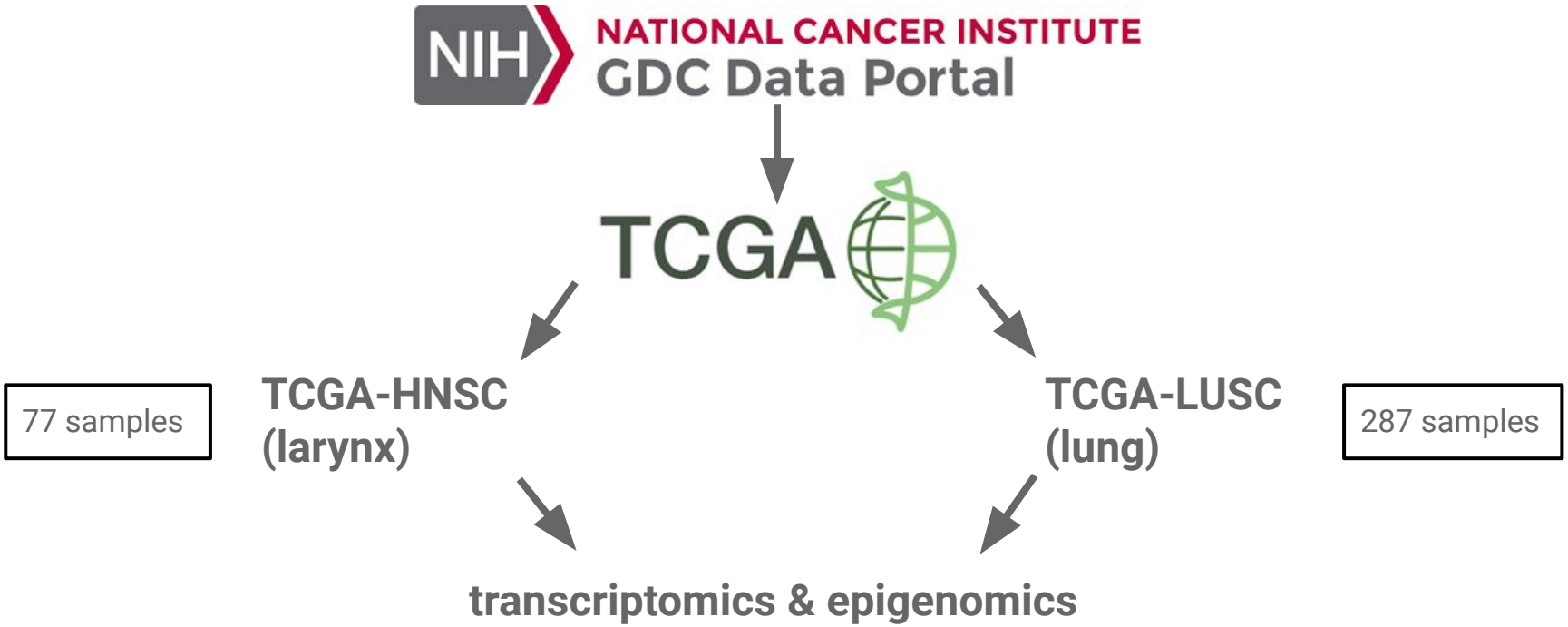
Project introduction



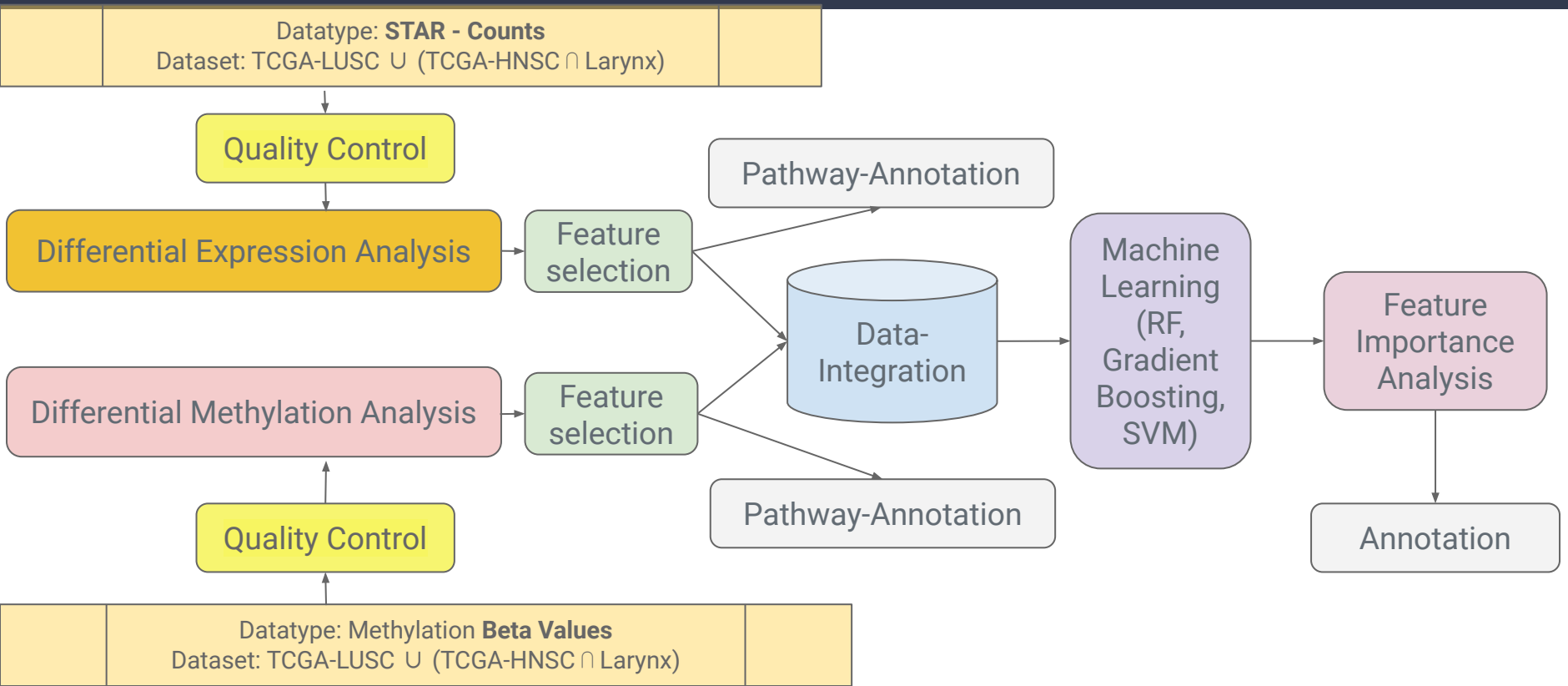
→ in respiratory organs mostly caused by smoking

Which transcriptomic and epigenomic biomarkers are useful for the differentiation between squamous cell carcinoma in the lung and in the larynx?

Data


NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
TCGA
77 samples
TCGA-HNSC
(larynx)
TCGA-LUSC
(lung)
287 samples
transcriptomics & epigenomics

Overview


Datatype: **STAR - Counts**
Dataset: TCGA-LUSC ∪ (TCGA-HNSC ∩ Larynx)
Quality Control
Differential Expression Analysis
Feature selection
Pathway-Annotation
Data-Integration
Differential Methylation Analysis
Feature selection
Pathway-Annotation
Quality Control
Machine Learning (RF, Gradient Boosting, SVM)
Feature Importance Analysis
Annotation
Datatype: Methylation **Beta Values**
Dataset: TCGA-LUSC ∪ (TCGA-HNSC ∩ Larynx)

Quality control / Preprocessing Expression

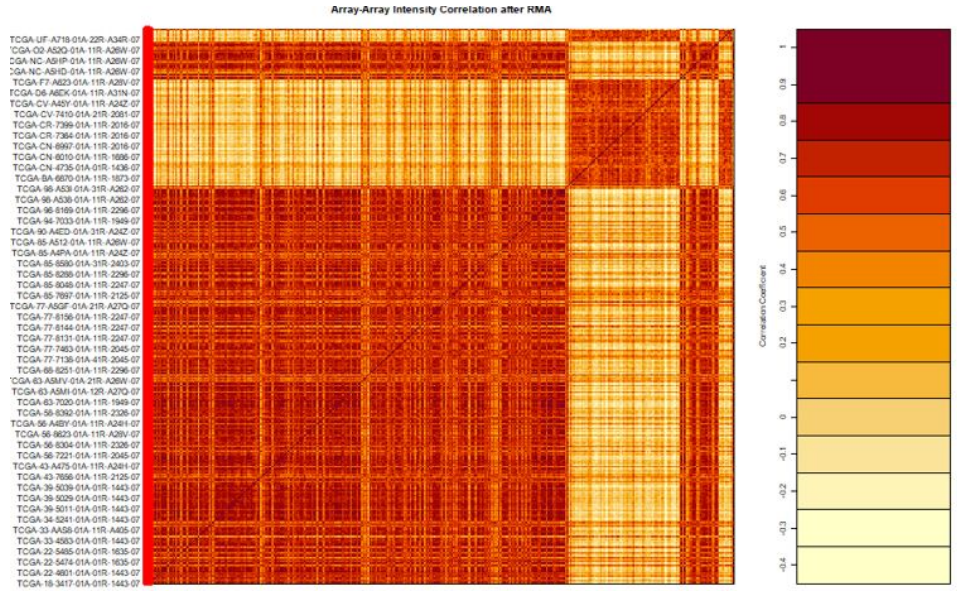
Raw Counts in 60,660 genes

Preprocessing, Normalization for GC Content

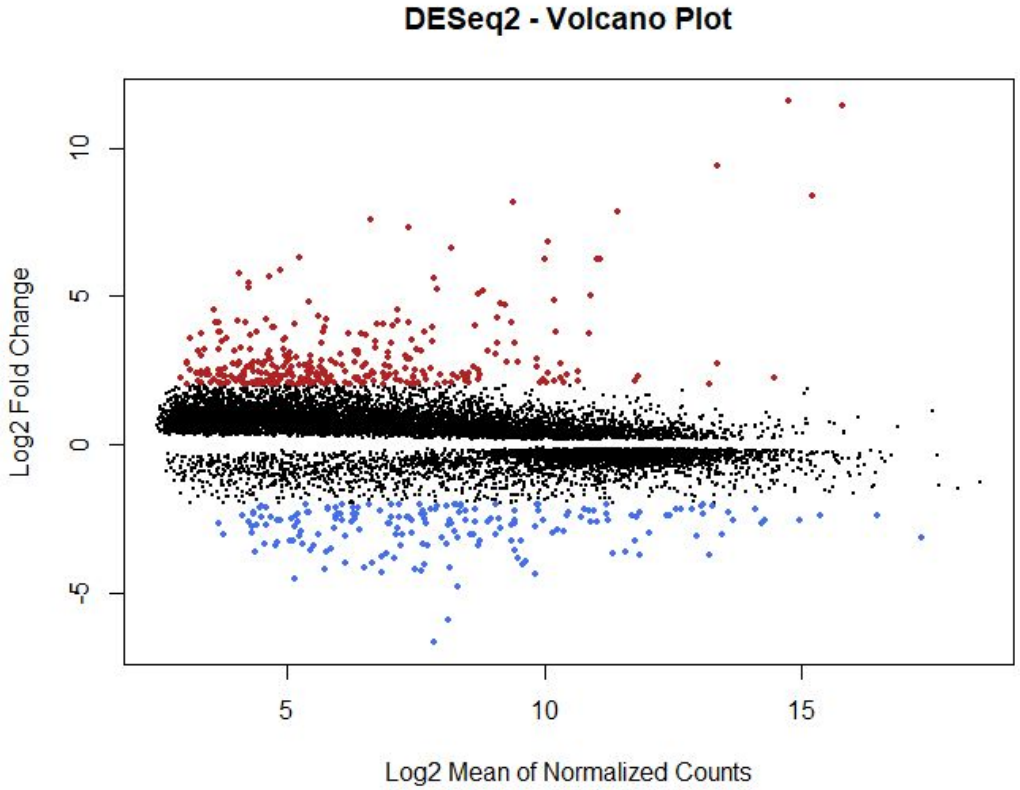
60,660 genes

0.25 Quantile Filtering

45,266 genes



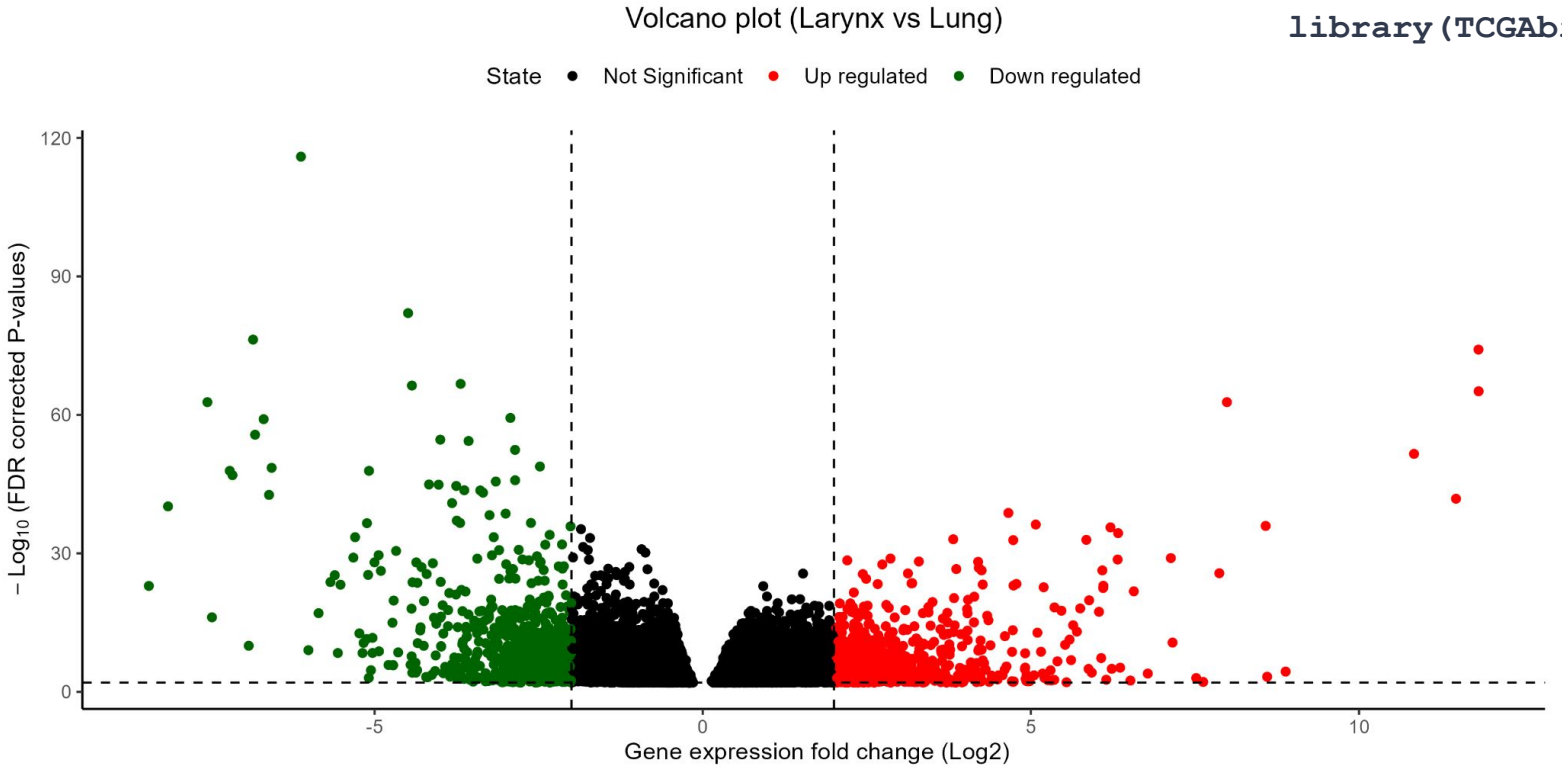
Differential gene expression analysis (DESeq2)



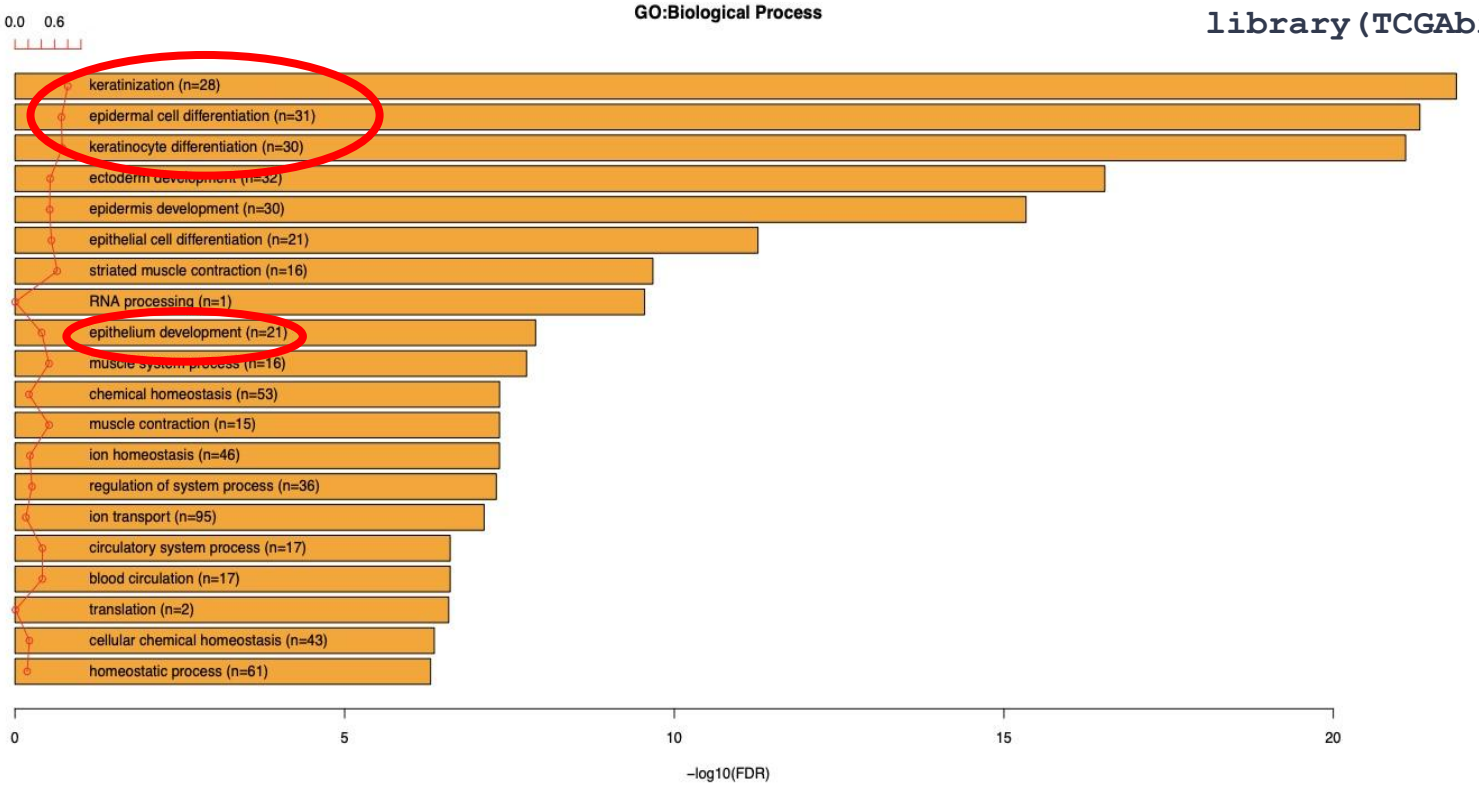
512 genes
with adjusted p < 0.01 &
abs(log2FoldChange) > 2

Differential gene expression analysis

```
library(TCGAbiolinks)
```

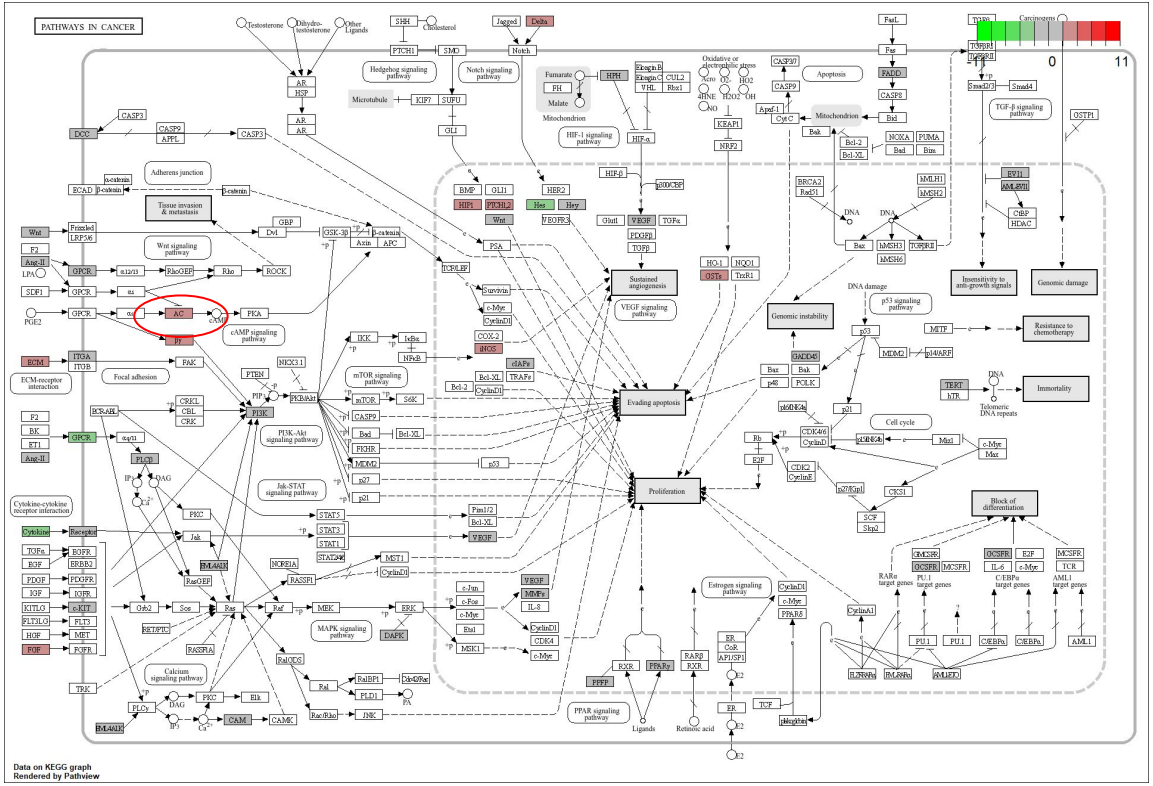


Annotation



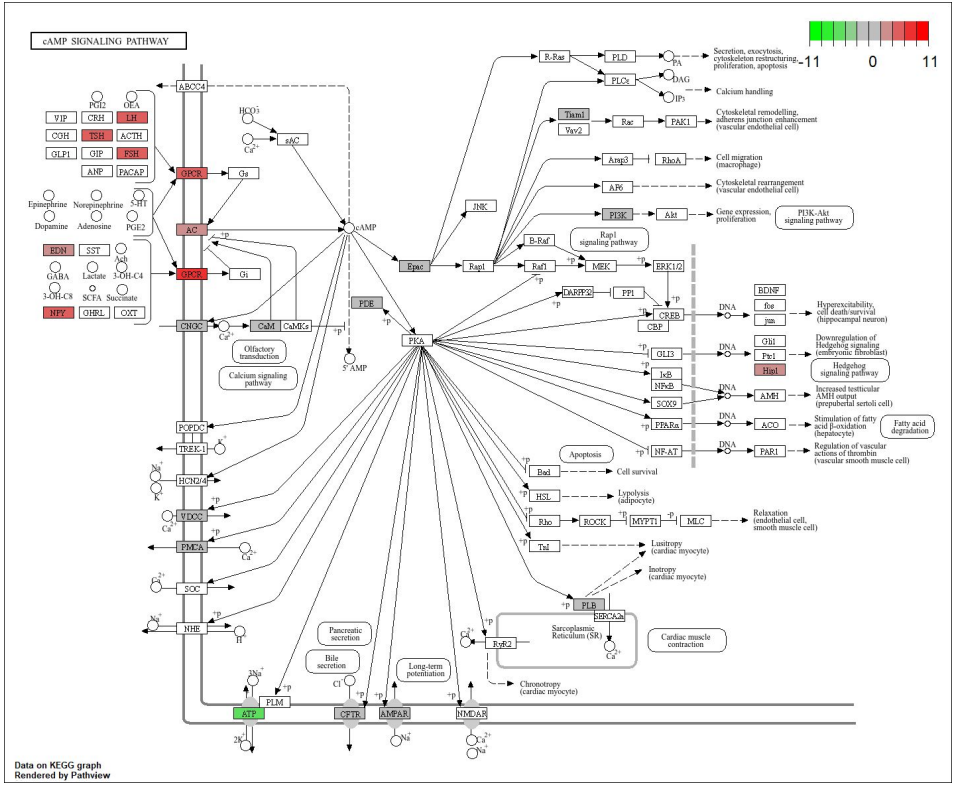

```
library(TCGAbiolinks)
```

Annotation



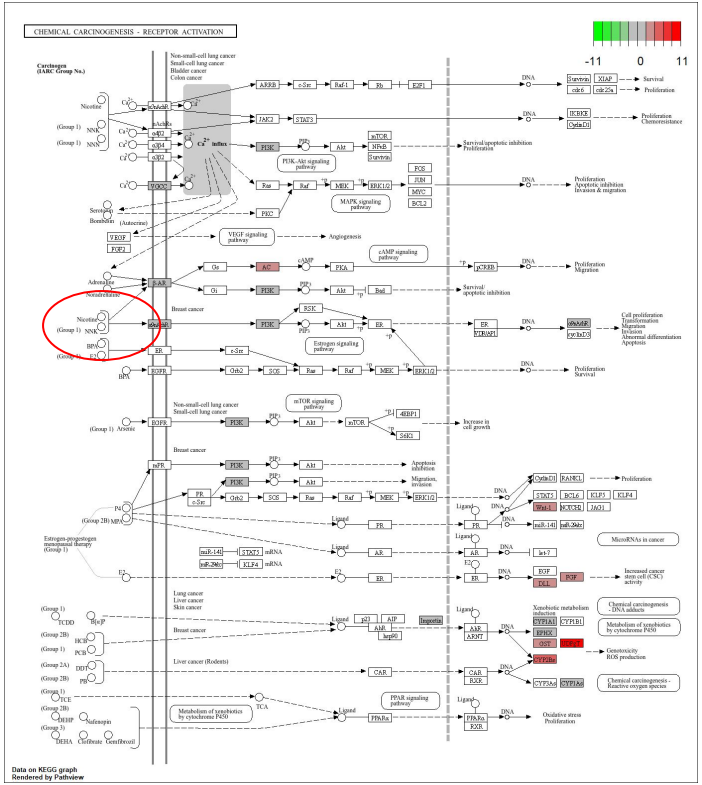

```
library(TCGAbiolinks)
```

Annotation




```
library(pathview)
```

Annotation




```
library(pathview)
```

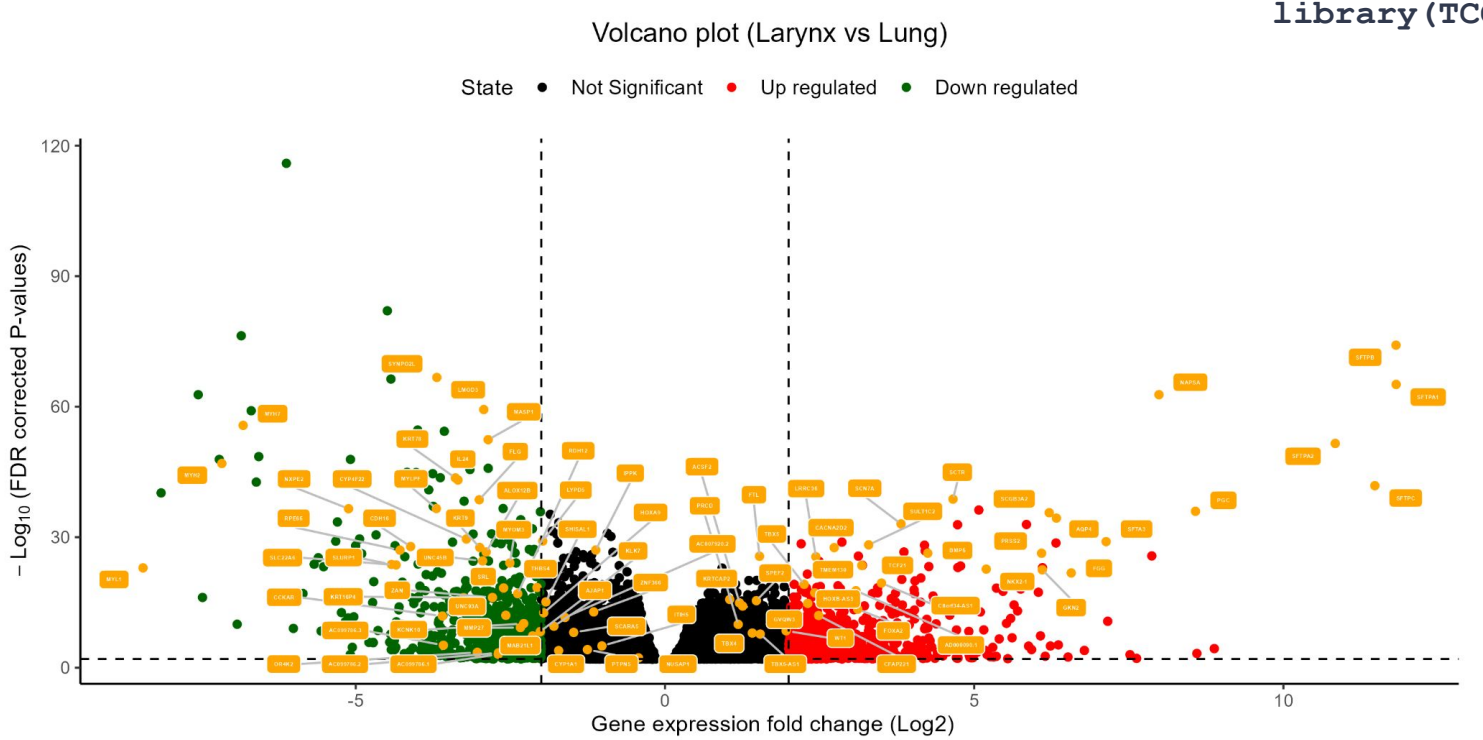
DEG Analysis – Feature selection

AC007920.2
AC099786.1
AC099786.2
AC099786.3
ACSF2
AD000090.1
AJAP1
ALOX12B
AQP4
BMP5
C8orf34-AS1
CACNA2D2
CCKAR
CDH16
CFAP221
CYP1A1
CYP4F22
FGG
FLG
FOXA2
FTL
GKN2
GVQW3
HOXA9
HOXB-AS3
IL24
IPPK
ITIH5
KCNK10
KLK7
KRT16P4
KRT78
KRT9
KRTCAP2
LMOD3
LRRC36
LYPD5
MAB21L1
MASP1
MMP27
MYH2
MYH7
MYL1
MYLPF
MYOM3
NAPSA
NKX2-1
NUSAP1
NXPE2
OR4K2
PGC
PRCD
PRSS2
PTPN5
RDH12
RPE65
SCARA5
SCGB3A2
SCN7A
SCTR
SFTA3
SFTPA1
SFTPA2
SFTPB
SFTPC
SHISAL1
SLC22A6
SLURP1
SPEF2
SRL
SULT1C2
SYNPO2L
TBX4
TBX5
TBX5-AS1
TCF21
THBS4
TMEM130
UNC45B
UNC93A
WT1
ZAN
ZNF366

```
library(Caret)
```

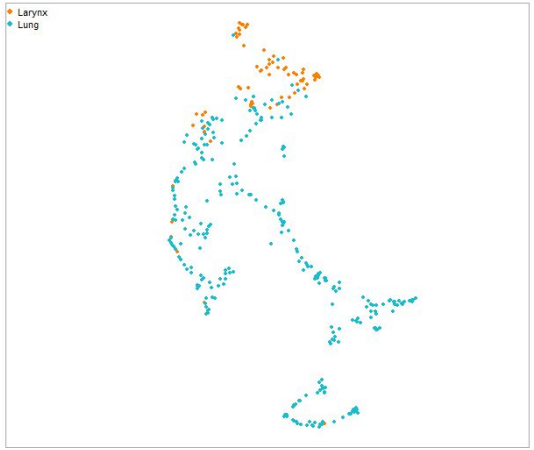
DEA – Feature selection

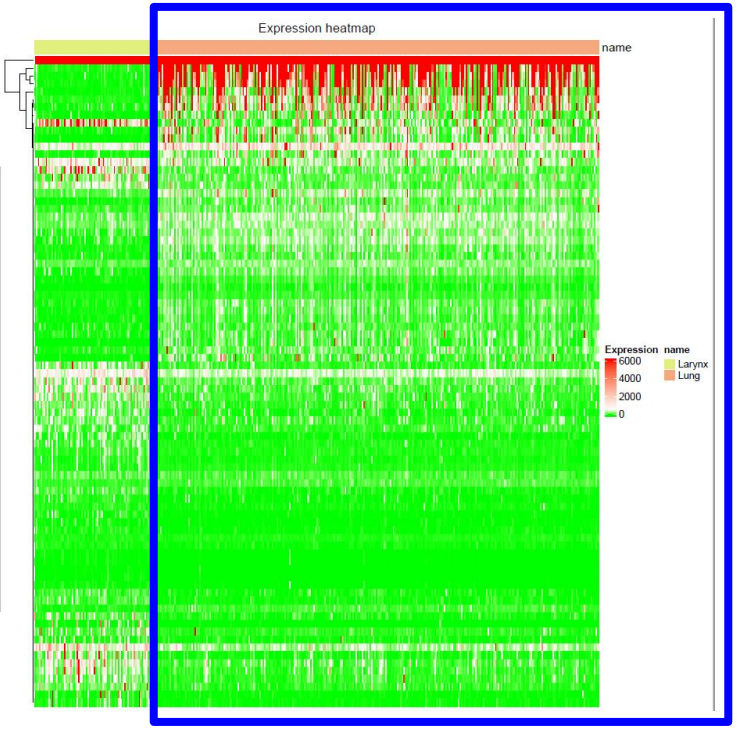
```
library(TCGAbiolinks)
```



DEA – Feature selection

- Before feature selection: 60,660 genes
- After feature selection: 83 genes

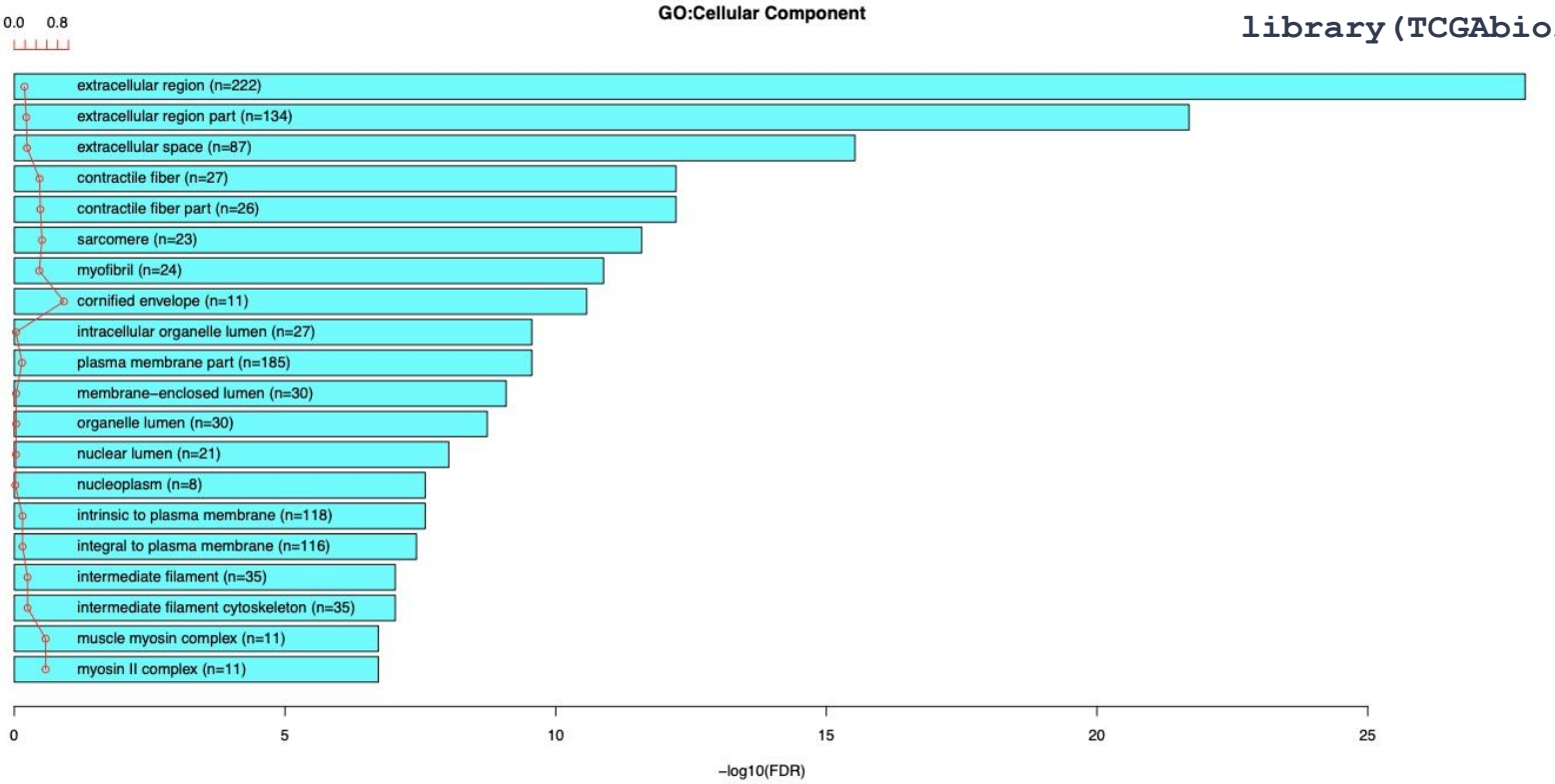





```
library(TCGAbiolinks)
        library(umap)
```

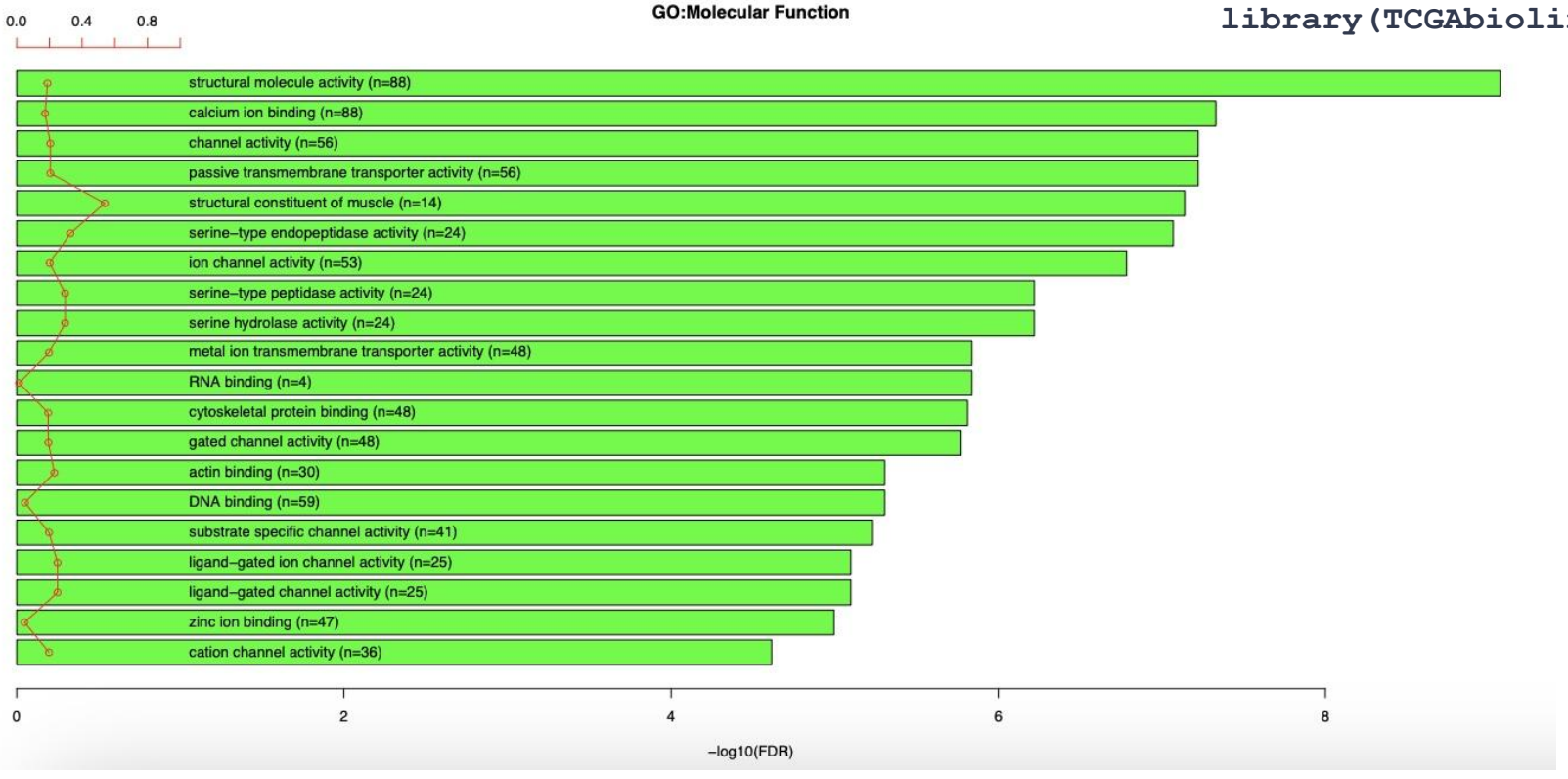
Annotation

```
library(TCGAbiolinks)
```



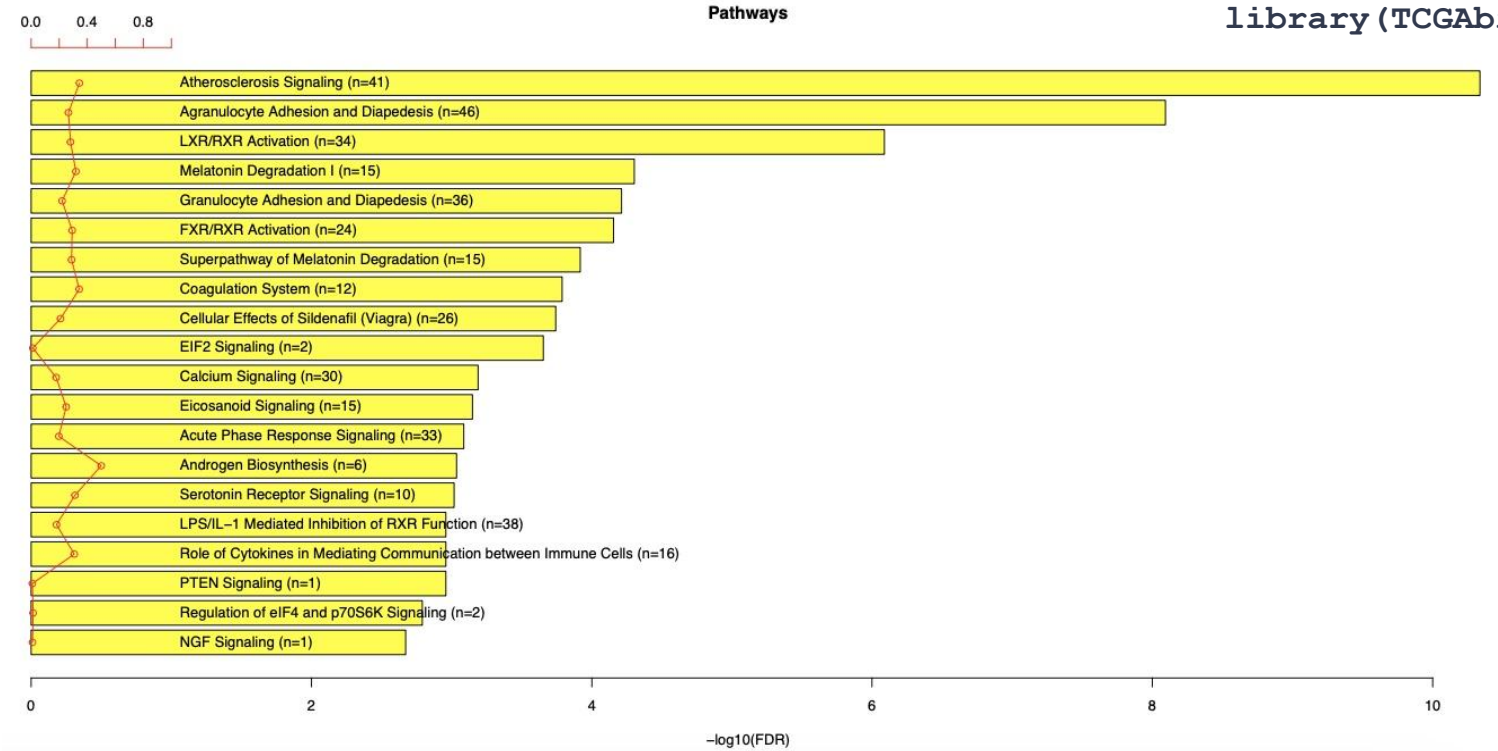
Annotation

```
library(TCGAbiolinks)
```



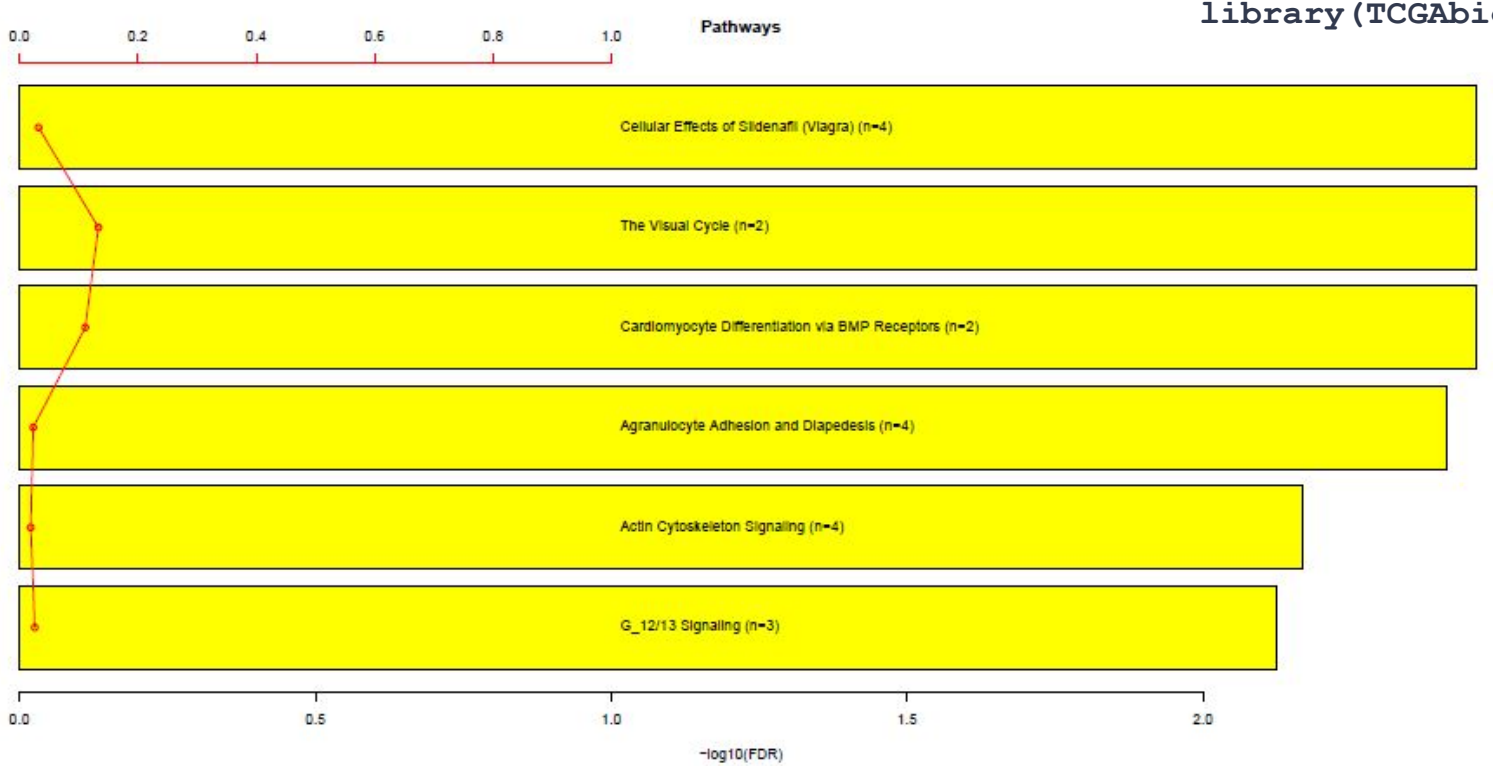
Annotation

```
library(TCGAbiolinks)
```

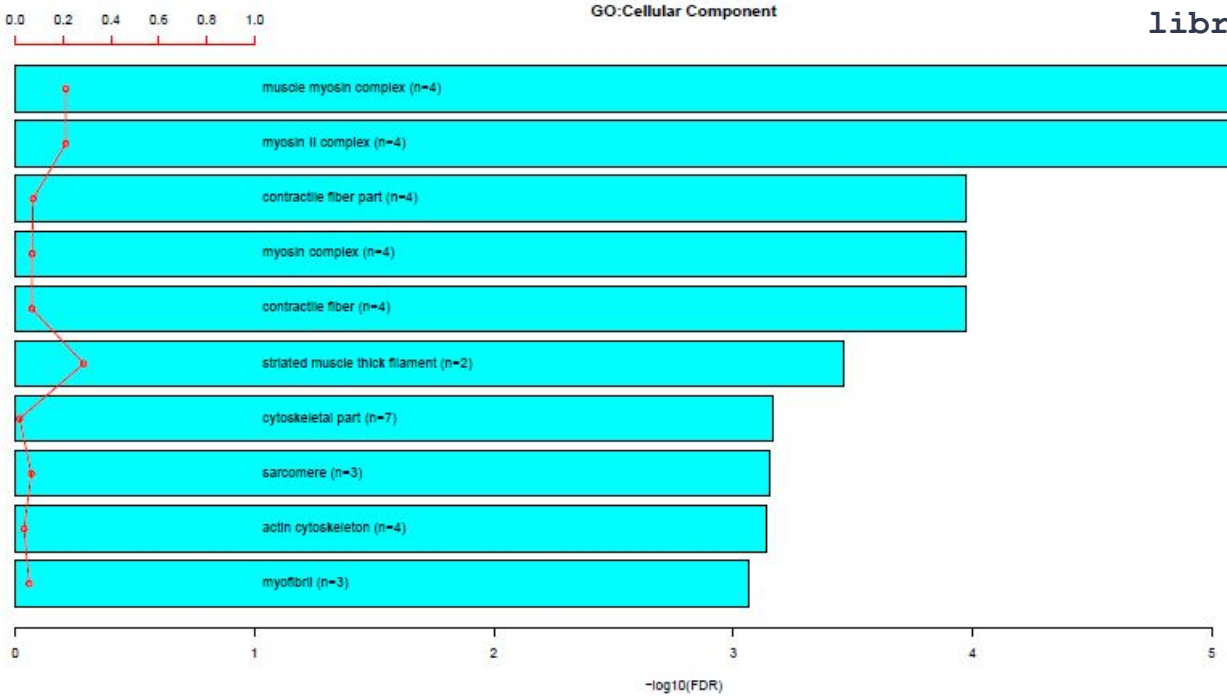


DEG Analysis - Feature selection

```
library(TCGAbiolinks)
```



DEG Analysis - Feature selection

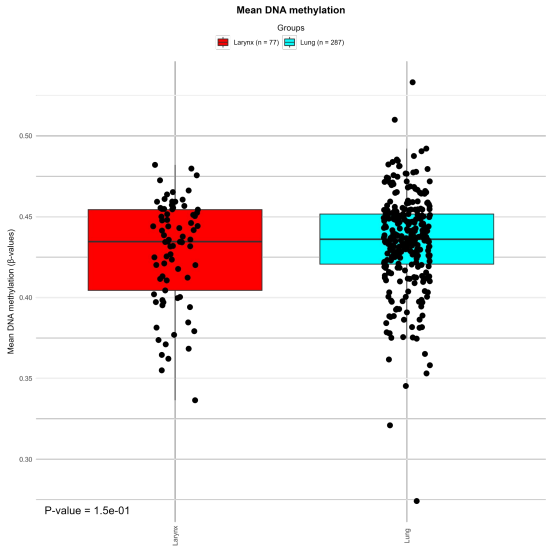


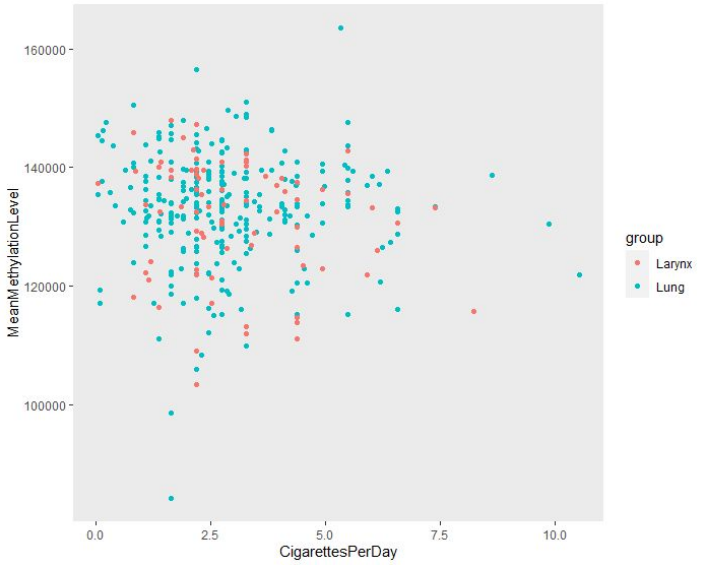

```
library(TCGAbiolinks)
```

Quality control / Preprocessing Methylation

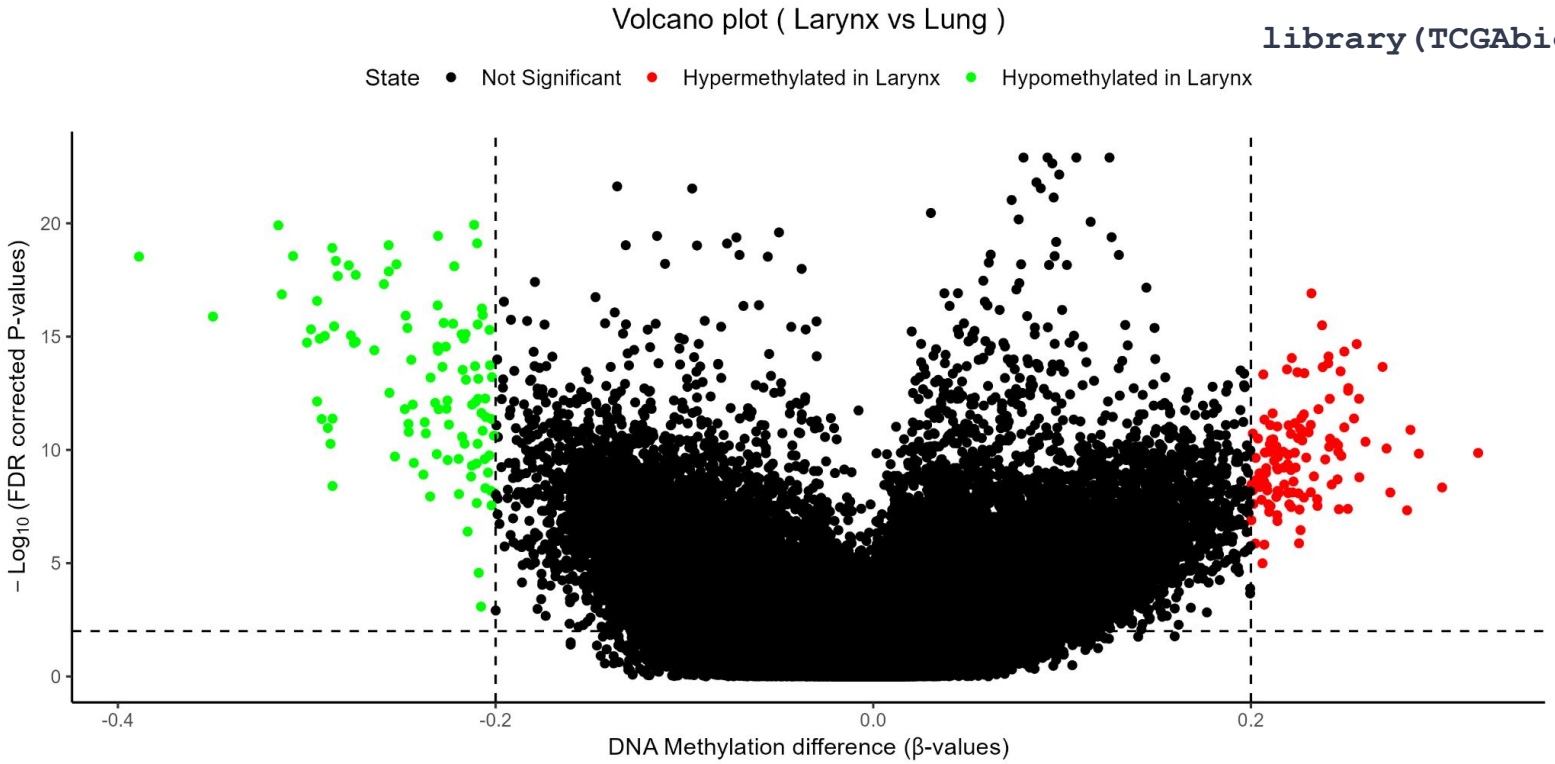
```
library(TCGAbiolinks)
```

- 485,577 cpg islands in 35,557 genes, beta value > 0 ⟿ 312,864 cpg islands
- excluded X, Y and unknown chromosomes ⟿ 306,903 cpg islands





Differential methylation analysis




```
library(TCGAbiolinks)
```

DM Analysis – Feature selection

Cells. 2020 Jul; 9(7): 1613.
Published online 2020 Jul 3. doi: 10.3390/cells9071613

Methylation in *HOX* Clusters and Its Applications in Cancer Therapy

Ana Paço,[1] Simone Aparecida de Bessa Garcia,[2] and Renata Freitas[2,3,*]

CLASP1
COLEC11
CTB-49A3.4
CTB-57H20.1
CTD-2555C10.3
CYR61
DHCR7
DNHD1
EPDR1
ESR2
FLOT1
FOXC1
FOXQ1

HOXA-AS2
HOXA-AS3
HOXA2
HOXA3
HOXA4
HOXA5
HOXA7
HOXB-AS3
HOXB3
HOXB4
HOXB5
HOXB6
IER3
INPP5D

MIR6080
MRO
NEURL1B
NRG2
PACS1
PCDHA1
PCDHA10
PCDHA11
PCDHA12
PCDHA13
PCDHA2
PCDHA3
PCDHA4

PCDHA5
PCDHA6
PCDHA7
PCDHA8
PCDHA9
PCDHAC1
PLEKHM1P
PPP1R9A
RANBP17
RBM20
RP1-170O19.22
RP1-170O19.24
RP11-399K21.11
RP11-660L16.2

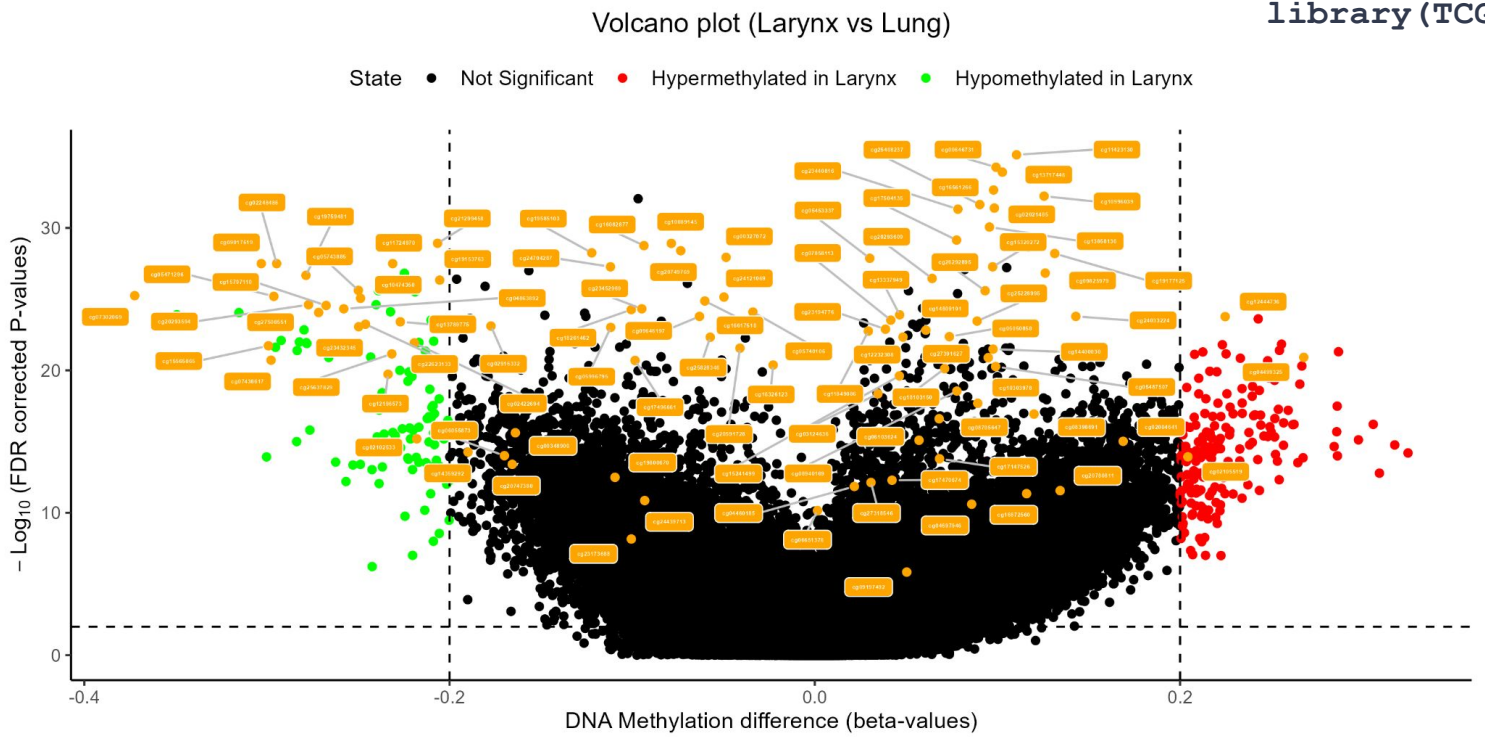
SCARF2
SFRP4
SKAP1
SMAD6
SMYD4
ST14
SYNE2
TARID
TIMM10B
TMCO3
TMEM201
ZFPM1
ZNF467

Clustered protocadherins methylation alterations in cancer

Ana Florencia Vega-Benedetti,[#1] Eleonora Loi,[#1] Loredana Moi,[1] Sylvain Blois,[1] Antonio Fadda,[1] M
Antonella Arcella,[3] Manuela Badiali,[4] Felice Giangaspero,[2,3] Isabella Morra,[5] Amedeo Columbano,[6]
Luigi Zorcolo,[7] Viviana Gismondi,[8] Liliana Varesco,[8] Sara Erika Bellomo,[9] Silvia Giordano,[9,10] Matte

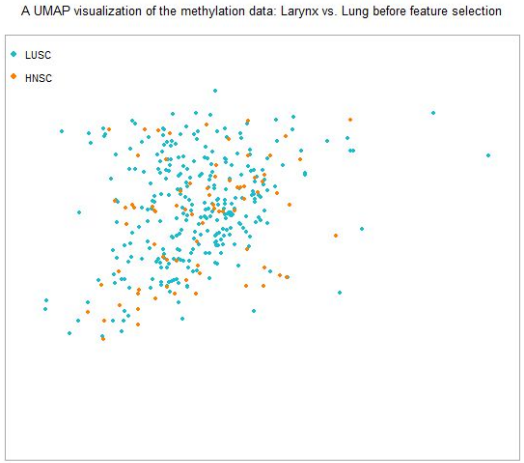
DM Analysis - Feature selection

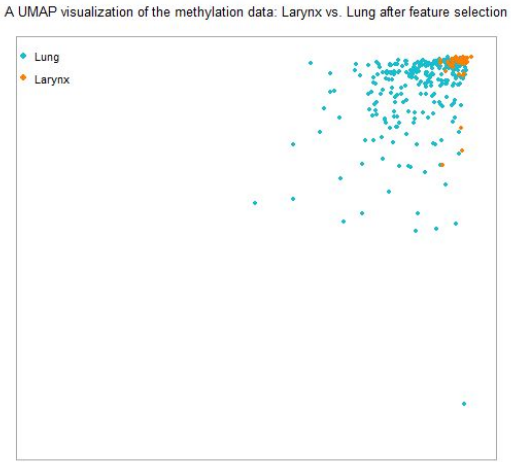
```
library(TCGAbiolinks)
```

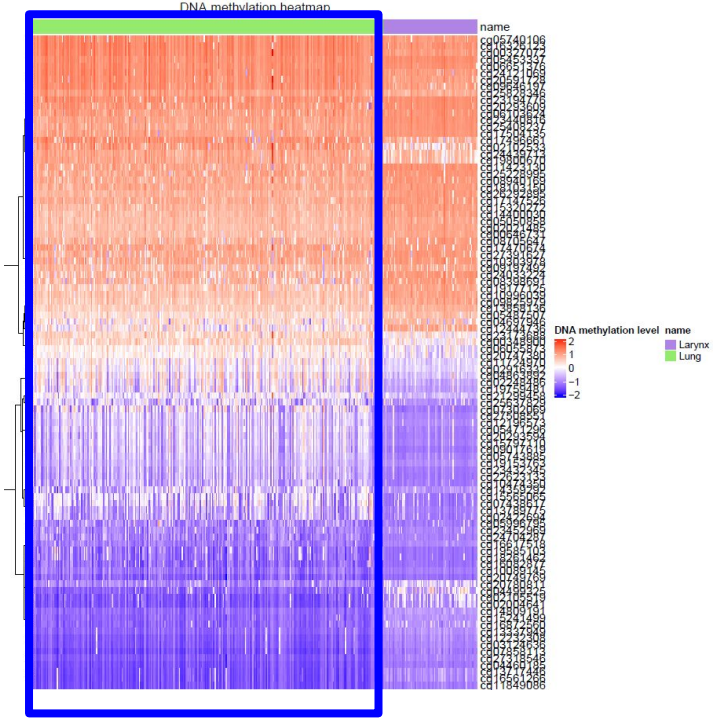


DM Analysis - Feature selection

- Before feature selection: 306,903 cpg islands
- After feature selection: 97 cpg islands





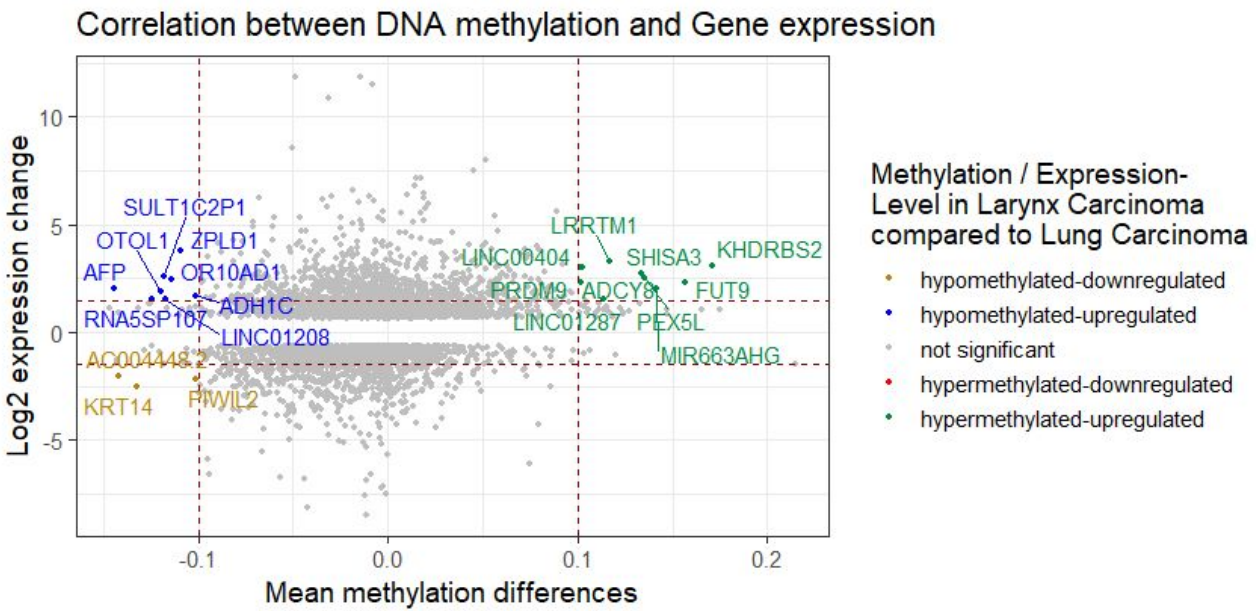



```
library(TCGAbiolinks)
library(umap)
```

Data integration Connecting Expression and Methylation

```
library(TCGAbiolinks)
```

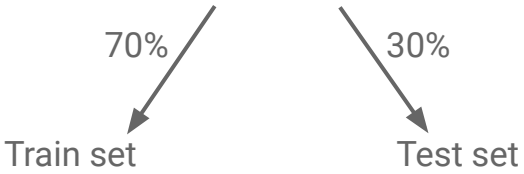
- aggregated mean beta values for methylation and counts for expression for each gene



KR
ZPLD1
ADCY8
PIWIL2
KHDRBS2
SHISA3
PEX5L
FUT9
SULT1C2P1
OR10AD1
MIR663AHG
PRDM9
AFP
ADH1C
OTOL1

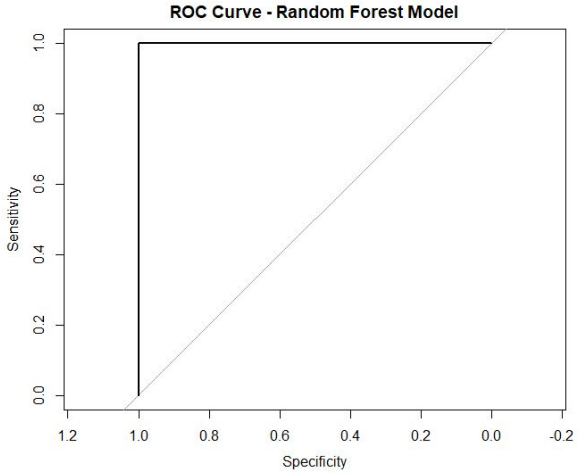
Machine Learning (with caret)

364 samples in total (77 larynx / 287 lung) with **180 features** in total (83 transcriptomic / 97 epigenomic)

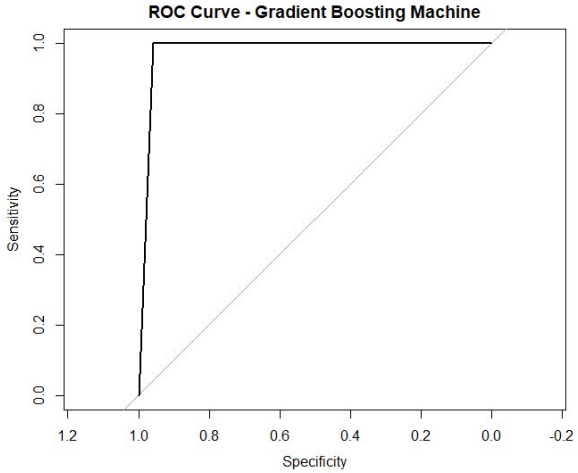


- Cross validation with 10 folds

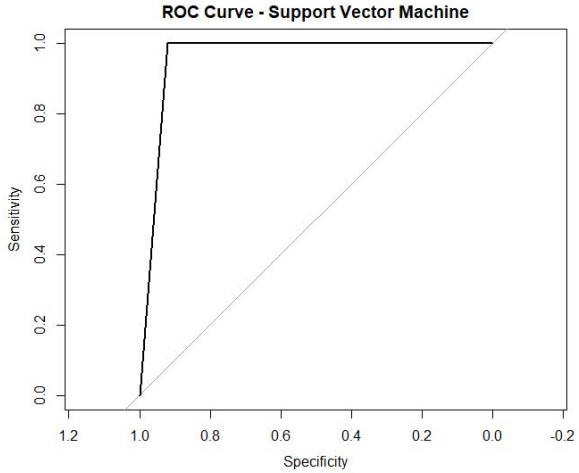
Machine Learning (with caret)




```
          Reference
Prediction larynx lung
    larynx     23    0
    lung        0   86

               Accuracy : 1
                 95% CI : (0.9667, 1)
    No Information Rate : 0.789
    P-Value [Acc > NIR] : 6.037e-12
```



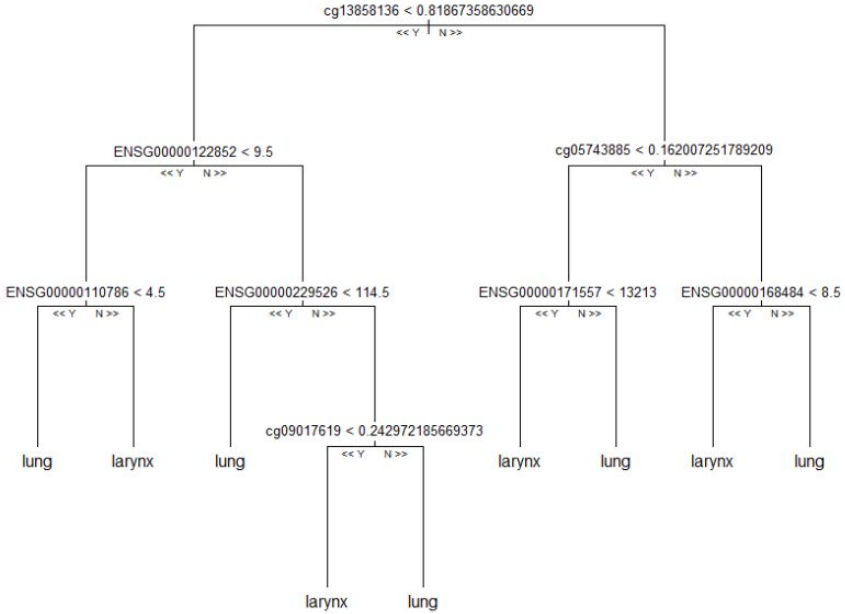

```
          Reference
Prediction larynx lung
    larynx     23    1
    lung        0   85

               Accuracy : 0.9908
                 95% CI : (0.9499, 0.9998)
    No Information Rate : 0.789
    P-Value [Acc > NIR] : 1.82e-10
```



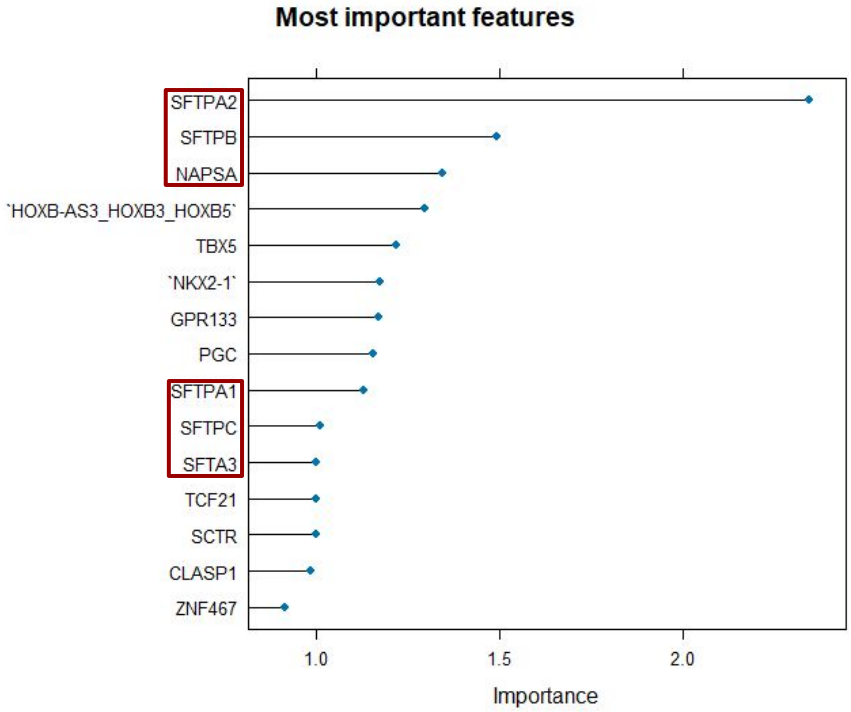

```
          Reference
Prediction larynx lung
    larynx     23    2
    lung        0   84

               Accuracy : 0.9817
                 95% CI : (0.9353, 0.9978)
    No Information Rate : 0.789
    P-Value [Acc > NIR] : 2.724e-09
```

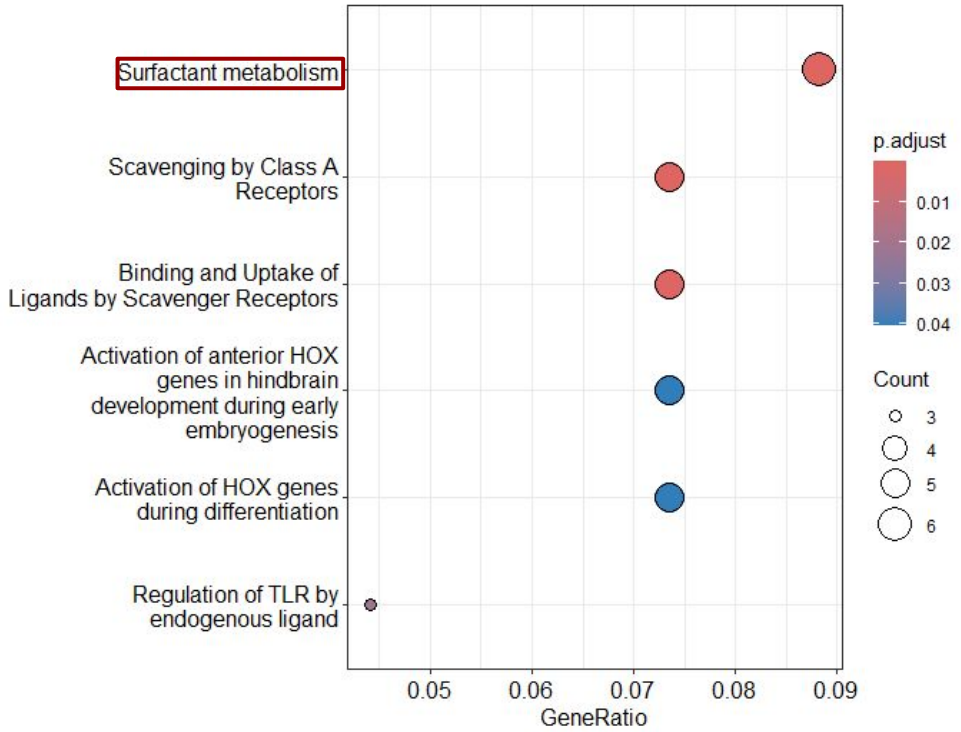
Example of the Random Forest decision trees

```
library(reprtree)
```



Feature importance (with caret)





Summary & Conclusion

- we trained three models to predict the site of squamous cell carcinoma
- differences in expression and methylation level may be characterized by tissue of origin
- the application could be diagnostic of patients with Cancer of unknown primary (CUP)
- Here, the cancer spreads within lymph nodes in the and the primary cannot be found, the ML can be used to predict were the tumor originates from
- the model could be applied to predict the primary site of the tumor

References

- **Maximilian Leitheiser et al.** “Machine learning models predict the primary sites of head and neck squamous cell carcinoma metastases based on DNA methylation”. en. In: The Journal of Pathology 256.4 (2022). Eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.5845, pp. 378–387. issn:1096-9896. (visited on 07/06/2024).
- **Philipp Jurmeister et al.** “Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases”. In: Science Translational Medicine 11.509 (Sept. 2019). Publisher: American Association for the Advancement of Science, eaaw8513. url: https://www.science.org/doi/10.1126/scitranslmed.aaw8513 (visited on 07/06/2024).
- **Anil Vachani et al.** “A 10-Gene Classifier for Distinguishing Head and Neck Squamous Cell Carcinoma and Lung Squamous Cell Carcinoma”. In: Clinical Cancer Research 13.10 (May 2007), pp. 2905–2915. issn: 1078-0432. url: https://doi.org/10.1158/1078-0432.CCR-06-1670 (visited on 07/06/2024).
- **Yawei Li et al.** “Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis”. In: Genomics Proteomics Bioinformatics 20.5 (2022). issn: 850-866. doi: 10.1016/j.gpb.2022.11.003.
- **Yin Li et al.** “Transcriptomic and functional network features of lung squamous cell carcinoma through integrative analysis of GEO and TCGA data”. In: Sci Rep 8.1 (2018). doi: 10.1038/s41598-018-34160-w.
- **Alana Sorgini et al.** “Analysis of the TCGA Dataset Reveals that Subsites of Laryngeal Squamous Cell Carcinoma are Molecularly Distinct”. In: Cancers 13.1 (Dec. 31, 2020), p. 105. issn: 2072-6694. doi: 10.3390/cancers13010105.
- **Joshua D. Campbell et al.** “Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas”. In: Cell Reports 23.1 (Apr. 3, 2018), 194–212.e6. issn: 2211-1247. Doi: 10.1016/j.celrep.2018.03.063.
- **Katherine A. Hoadley et al.** “Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer”. In: Cell 173.2 (Apr. 5, 2018), 291–304.e6. issn: 1097-4172. doi: 10.1016/j.cell.2018.03.022