Project report: Transcriptomic and epigenomic biomarkers to differentiate between smokers with lung and laryngeal cancer

Group 5: Mia Anscheit and Friederike Wohlfarth

Contributing authors: miaa99@zedat.fu-berlin.de;
friederike.wohlfarth@fu-berlin.de;

Abstract

Squamous cell carcinoma is one of the most common types of cancer in the respiratory organs and in most cases is caused by smoking. This type of cancer is at risk of metastasising to other organs, so when treating the tumours, it is important to differentiate between distal metastases and second primary organs in order to provide the best possible therapy.
The aim of this project is to find biomarkers that help to differentiate between squamous cell carcinoma of the lung and squamous cell carcinoma of the larynx. We worked with both, transcriptomic and epigenomic data, on squamous cell carcinoma from smoking patients with either lung or laryngeal cancer. We performed differential gene expression and differential methylation analyses and used selected features for training machine learning models. Then, we annotated the genes and CpG islands that differed most in expression or methylation between the two cancer groups.
The tested machine learning classifiers all showed high accuracy and high AUC and can therefore be considered as good predictors. Most of the genes found to be differentially expressed or methylated in lung and larynx indicate different molecular mechanisms in lung and larynx. We found three genes that were differentially methylated and differentially expressed and were interesting in relation to cancer and / or smoking, namely SHISA3, SULT1C2 and MIR663AHG.
As stated by others, we conclude that those machine learning classifiers can facilitate diagnostics of patients were the primary tumor can not be found (CUP), which has already been applied in clinical pathology.

Keywords: Multi-Omics, Laryngeal Cancer, Lung Cancer, CUP

1 Background

1.1 Correlation between smoking and cancer

It is a well-known fact that the carcinogens in cigarette smoke can cause cancer. They form DNA adducts that prevent the polymerases in DNA replication from working properly, inducing mutations. Carcinogens can directly damage the DNA. Cigarette smoke can also cause epigenetic changes, namely methylation changes in CpG islands that have the potential to affect expression patterns in the tissue. Typically, the mechanism in cancer is as follows: promotors close to tumor suppressor genes are typically hypermethylated when the expression is downregulated. Promotors close to oncogenes are typically hypomethylated when the expression of oncogenes is upregulated [1].

1.1.1 Squamous cell carcinoma

Squamous cells form the surface of the skin as well as the interior surface of the most organs. Mutated squamous cells, squamous cell carcinoma, are exclusively located in the epidermis of the skin. This type of skin cancer is the most common one. While the main cause of the formation of carcinomas on the skin is excessive sunlight [2], the main reason for the formation of squamous cell carcinoma in the respiratory organs is smoking [3].

1.1.2 Lung cancer

Lung cancer, which is one of the three most common cancer types in male and female patients, is strongly correlated with smoking status. More men than women get diagnosed with lung cancer, which is proven to be correlated with the smoking habit. Nevertheless it remains unclear, if men are also more prone to the carcinogens in cigarette smoke, and develop cancer more likely when adjusting for smoking habit [1]. Among the three histological sub-types, squamous cell carcinoma is the one that is most often correlated with smoking habit, because it is located in the central parts of the lung. Smoking is the most common risk factor for lung cancer [4].

1.1.3 Laryngeal cancer

Another cancer type that is more prevalent in smokers than in nonsmokers and affects also the outer layer of the affected organ, namely the epithelium, is the Head and Neck Squamous carcinoma. Besides the oral cavity and the pharynx, it can also affect the larynx and, in severe cases, lead to the extraction of the larynx, which means that the patients cannot speak normally from this point onwards.

1.1.4 Cancer of unknown Primary

Sometimes, a tumor from the head and neck region produces cervical metastases. Also distant metastases in the lung can appear. In this latter case, it is important to distinguish between distal metastasis (DM) and second primary tumor (SPT) in respect to the severity of the disease and the chance of cure (if the cancer has already

spread, it is more difficult to treat the disease) [5]. This is done histologically. Sometimes, the pathologist states that the tumor is not primary, but the primary tumor cannot be found. This is a case of Cancer of unknown Primary (CUP), which has a bad prognosis.

1.2 Related work

Recently, with the advent of multi-omics studies that combine bioinformatics tools with experiment data from different sources, for example methylation and expression data, it became feasible to classify cancer types based on expression and methylation patterns. A machine learning model trained on this data can be tailored to predict the primary tumor in patients with CUP-syndrome in the Head and Neck region [6, 7].
There are several studies that have already explored the differences between lung and laryngeal cancer in gene expression and/or methylation, but have used different data or methods. We hereby focus on three.
The study "Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases" by Jurmeister et al. [6] dealed with differentiating between head and neck squamous cell carcinoma and lung squamous cell carcinomas to recognize primary and spreaded tumor cells. The authors used DNA methylation profiling of 408 HNSCC/LSCC patients for training the models, and a cohort of 279 HNSCC/LSCC patients for the validation process. They trained three different machine learning methods: an artificial neural network with 96.4 % accuracy and 99.3 % AUC, a support vector machine with 95.7 % accuracy and 99.2 % AUC and a random forest model with 78.8 % accuracy and 97.1 % AUC. They selected the 2000 most variable CpG site for model training and performed a gene set enrichment analysis on them resulting primarly in GO terms related to tissue differentiation. The subsequent study by Leitheiser et al. performed a similar examination and was even able to predict the primary site of the HNSCC tumors ('oral cavity', 'oropharynx', 'hypopharynx or larynx') [7].
Thirumani et al., the authors of the paper "The Molecular Landscape of Lung Metastasis in Primary Head and Neck Squamous Cell Carcinomas" [8], compared tumorous and normal gene expression profiles from paired patient samples either with head and neck squamous cell carcinoma (HNSCC) or with lung squamous cell carcinoma (LSCC) to discover differential expressed genes and pathways. The data was taken from the GEO database and includes 22 HNSCC patients and 5 LSCC patients. They identified 145 overlapping DEGs in both head and neck squamous cell carcinoma and lung squamous cell carcinoma which are potential targets for personalized therapy. They determined the enrichment of genes in the biological processes of extracellular matrix organization as well as cell-substrate adhesion. Some identified DEGs are part of the collagen-containing extracellular matrix and cell-cell junctions. As affected pathways they identified the peptidase regulator and inhibitor activites, the ECM-receptor interaction and the cell cycle.

2 Goal

In this study, we aim at distinguish the primary site of the dissected tumor cells by the help of machine learning models trained on both, expression and methylation data at the same time. We used samples from the TCGA database of the projects LSCC for lung cancer and HNSCC for head and neck cancer, specifically in the larynx and integrated their data in one data matrix. This matrix we used for training. We restricted ourselves to only use patients that were smokers, because we were particularly interested in the effects of smoking on methylation and the effect of methylation on expression. Our second goal was to find the most important features that distinguish larnyx from lung cancer in smokers. We are expecting that this will give a hint to the underlying cell mechanisms that are diferentially activated in these cancer types.

3 Data and Preprocessing

3.1 Data and samples

The data stems from the NCI database GDC Data Portal [9], which contains various cancer genome programs with many different types of data such as transcriptomic, epigenomic, proteomic and genomic data. We sourced our data from the TCGA program of the two projects TCGA-LUSC, which focuses on squamous cell carcinomas of the lung, and TCGA-HNSC, which focuses on squamous cell carcinomas of the head and neck, here filtering for laryngeal tissue samples. We then removed all samples for which either transcriptomic or epigenomic data were not available, samples that did not contain information about the individual's smoking behavior (we only included smokers), and samples that were from healthy tissue. There were too few healthy tissue samples to consider a differentiated gene expression or methylation analysis between healthy and cancerous tissue. In the end, we had the transcriptomic data in the form of RNAseq counts and the epigenomic data in the form of DNA methylation beta values for CpG islands from the platform "Illumina Human Methylation 450" from a total of 77 laryngeal samples and 287 lung samples.

3.2 Pre-processing steps

The data from the TCGA database is broadly pre-processed, ensuring they are ready for downstream analysis. Nevertheless we conducted the following pre-processing steps: For the expression data we first searched for outliers up to a correlation coefficient of 0.6. There were no outliers. We then normalized the expression data for the GC content and performed quantile filtering up to a 0.25 quantile to remove lowly expressed genes. For the methylation data we excluded the CpG islands that were unmethylated in all samples. We further excluded X and Y chromosomes from both, expression and methylation datasets in order to make the data from male and female patients comparable and to account for the sex imbalances within and between both datasets. Because the data was obtained from two different experiment pipelines, we further had to merge the metadata columns from the lung and the laryngeal cancer datasets.

4 Methods

After downloading the corresponding data into R and pre-processing it using the package 'TCGAbiolinks' [10–12], we performed different steps of statistics, machine learning and annotation as depicted in Figure 1.

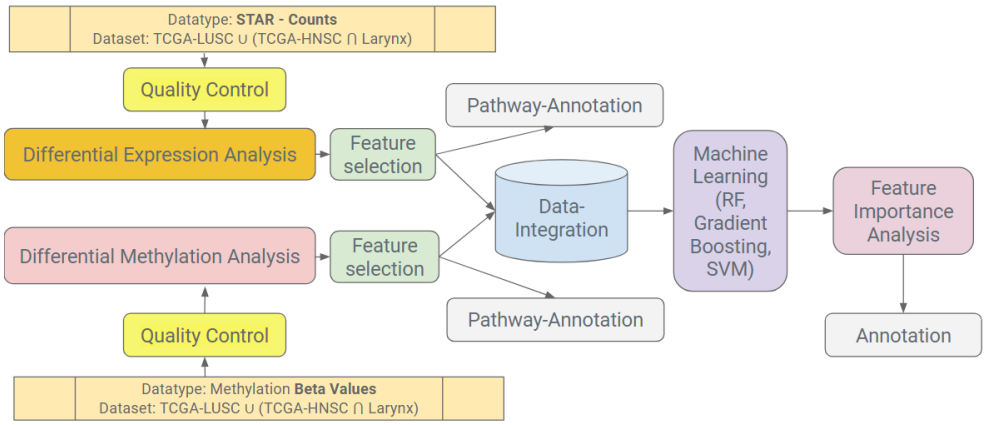


Fig. 1: Workflow of the project.

4.1 Differential gene expression and methylation analysis

At first, differential gene expression analysis (DEA) was done on the count matrix, once with the package 'DESeq2' [13], and once with the package 'TCGAbiolinks' with a corrected p-value of 0.01 and a difference of 2 for the log fold change values. The differential methylation analysis (DMA) was done on the beta value matrix using 'TCGAbiolinks'-package and a p-threshold of 0.01 and a threshold of 0.2 for the difference in mean beta values. We visualized our results in volcano plots for both analyses. We further performed enrichment analyses on the differential expressed genes using the 'TCGAbiolinks' and the 'clusterProfiler' [14, 15] package for the GO terms and KEGG pathways.
As we were interested in the effect of methylation through smoking on the expression of genes, we integrated the results from the DEA and the DMA and analysed the genes that were both, hypo- or hypermethylated as well as up- or downregulated. For that purpose, we computed the mean beta value for each gene from all CpG islands that were in the promotor regions of this gene. The gene names were already annotated in the dataset (CpG islands 1,500 bp upstream of the transcription start site to the end of the gene body were used). We visualized the result in a so-called starburst plot. Instead of a delta value of 0.2 for the beta values, we used a delta value of 0.1 in this case, because there were no significant results for a delta value of 0.2.

4.2 Machine Learning

For the machine learning part, we first integrated expression and methylation data into one big matrix. We scaled the values from the expression dataset to make them comparable to the methylation data. We then separated the feature matrix into train and test datasets using a ratio 70:30. To obtain our features for the machine learning part, we performed feature selection on the train dataset. We first subsetted the train matrix to columns that were significant and had a delta value greater than the chosen threshold in DEA and DMA. We then performed feature selection using the 'caret' [16] package, first separately on the expression and methylation data, second on the integrated dataset, and compared the results. The selected CpG islands were annotated with the affected gene names. We also checked publication databases considering the selected features and the terms "cancer" and "smoking", to find cancer-related and smoking-related features. We then performed importance analysis on the selected features for the random forest model only, because the other models do not provide this functionality. The most important features were again enriched for the affected GO terms and KEGG pathways.

4.2.1 Classifiers and hyperparameter tuning

We then subsetted our train matrix to include only the selected features which we used for the training of the machine learning models using the package 'caret'. We trained three different classifiers using a 10-fold cross-validation with five repeats: random forest (RF), gradient boosting machine (GBM) and support vector machine (SVM) (once with a linear, once with a radial kernel). For the Random Forest, we performed a grid search to find the optimal hyperparameter for the number of trees in the random forest. The train function in 'caret' automatically chooses the optimal hyperparameter mtry, the number of variables considered in each split. We further used the 'reprtree' [17] package to plot a most representative tree from the random forest for visualization purposes. We further adjusted for the class imbalance between larynx and lung by defining a weights function.

4.2.2 Model testing

We tested the performance of our models on the test matrix. To visualize the effect of feature selection, we computed UMAP on the whole integrated data matrix (training and testing with all variables), and once for the data matrix with the selected features. We also plotted heatmaps for the selected features of the expression data and the methylation data and clustered for the affected tissue.

5 Results

The original expression dataset consisted of 60.660 genes. After quantile filtering we obtained a total of 45.266 genes. For the methylation dataset, we started with 485.577 CpG islands. After exclusion of unmethylated regions, 312.864 CpG islands remained. 306.903 CpG islands remained after exclusion of X- and Y-chromosomes. Considering the clinical data, all of our samples belonged to patients that were smokers.

Male patients were overrepresented in both samples, the proportion of female to male patients was 1/5 for lung and 1/3 for laryngeal cancer patients. The patients were 40-85 years old and age was normally distributed. The mean methylation level was equally normally distributed between larynx and lung cancer samples and did not correlate with the number of cigarettes smoked per day.

5.1 Differential Expression and Methylation Analysis

DEA resulted in 12.539 significant genes. 728 genes were upregulated in lung compared to larynx and 804 genes were downregulated in lung compared to larynx. They are depicted as red and green points in Figure 2. Most of the selected features after feature selection had a low p-value and a high log2 fold change (yellow annotations in Figure 2).

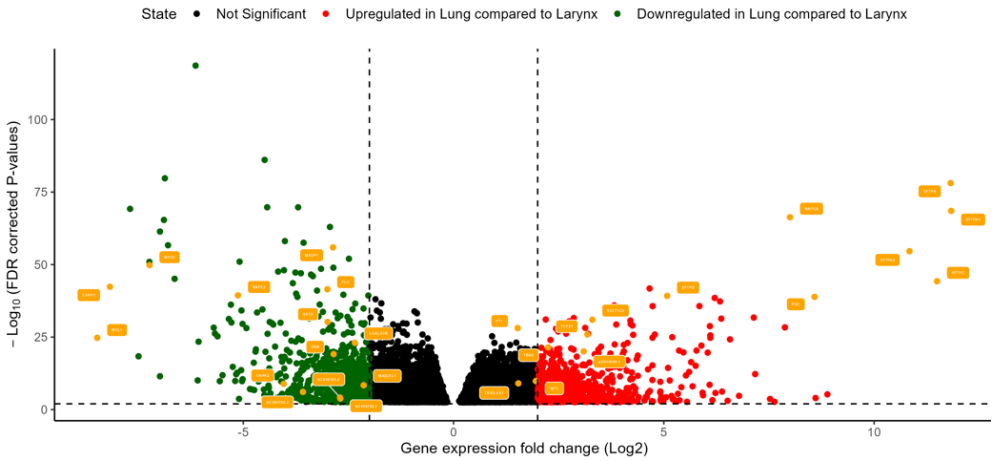


Fig. 2: Volcano Plot - Differential Gene Expression Analysis.

We performed DMA on 306.903 CpG islands. 53.597 CpG islands were significant, 122 of them were hypomethylated in lung compared to larynx and 105 were hypermethylated in lung compared to larynx using a threshold of 0.2 for the difference in mean of the beta values as depicted in Figure 3 (again, the result of the feature selection is annotated in the yellow).

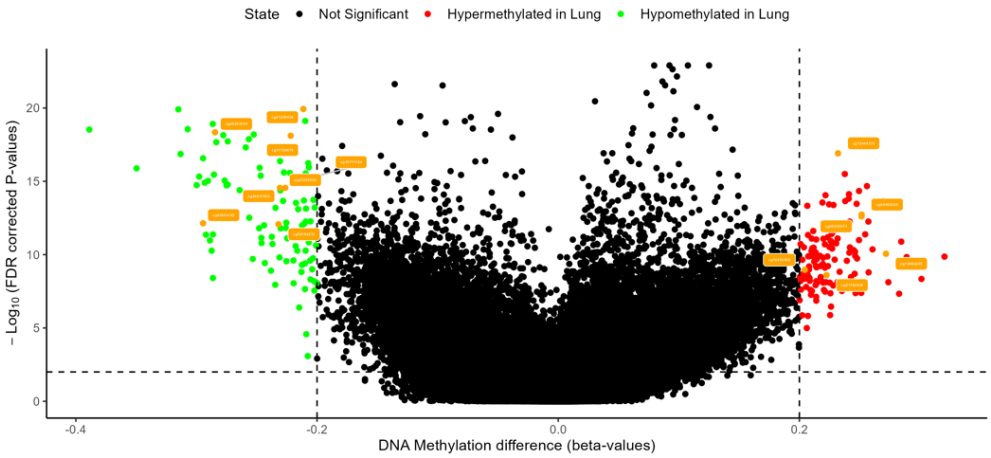


Fig. 3: Volcano Plot - Differential Methylation Analysis.

5.1.1 Prominent Genes in DEA and DMA

After summarizing the beta values for each CpG island into bins for each gene and merging the outcome with the result of the DEA, 17 genomic locations were hyper- or hypomethylated and over- or underexpressed at the same time, indicating a relation between methylation and expression level. Three of them were hypermethylated and downregulated using the mentioned thresholds as in tumor suppressor genes, six were hypermethylated and upregulated and eight were hypomethylated and upregulated as in oncogenes. 13 of these regions were protein coding genes, three were lncRNAs and one was a pseudogene. The combined visualization of beta and expression levels is depicted in the starburst plot in Figure 4, and the significant genes are listed in Table 1. As a side effect, we found that the mean number of CpG islands per gene in our samples was 17.

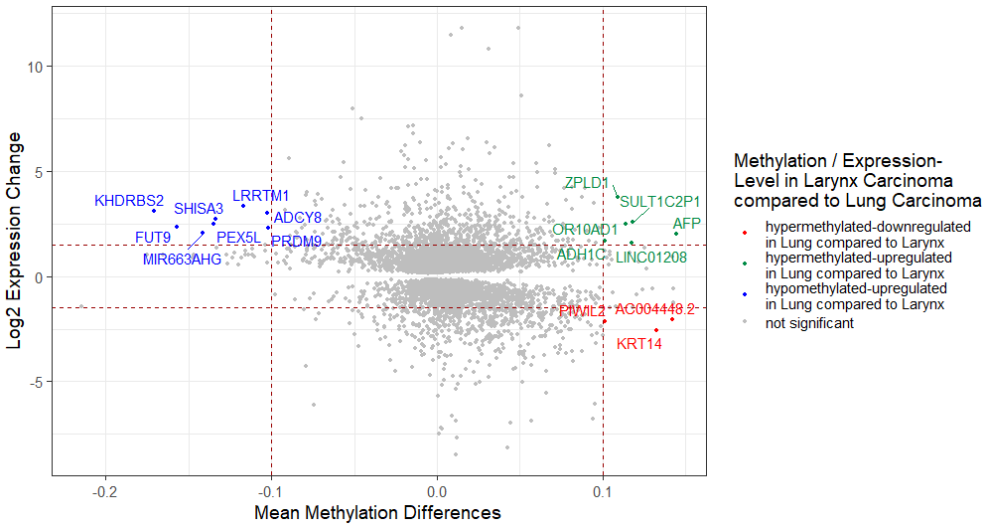


Fig. 4: Starburst Plot - Expression fold change vs. methylation differences.

Gene Name	Gene Type	expression and methylation status
AFP	protein coding	hypermethylated-upregulated
OR10AD1	protein coding	hypermethylated-upregulated
ADH1C	protein coding	hypermethylated-upregulated
SULT1C2P1	transcribed unprocessed pseudogene	hypermethylated-upregulated
LINC01208	lncRNA	hypermethylated-upregulated
ZPLD1	protein coding	hypermethylated-upregulated
PIWIL2	protein coding	hypermethylated-downregulated
AC004448.2	lncRNA	hypermethylated-downregulated
KRT14	protein coding	hypermethylated-downregulated
KHDRBS2	protein coding	hypomethylated-upregulated
PEX5L	protein coding	hypomethylated-upregulated
ADCY8	protein coding	hypomethylated-upregulated
LRRTM1	protein coding	hypomethylated-upregulated
PRDM9	protein coding	hypomethylated-upregulated
FUT9	protein coding	hypomethylated-upregulated
SHISA3	protein coding	hypomethylated-upregulated
MIR663AHG	lncRNA	hypomethylated-upregulated

Table 1: Annotated genes in the starburst plot.

We checked those 17 genes considering their role in smoking associated cancer and for for tumor suppressor genes and oncogenes.

SULT1C2 is hypermethylated in lung samples and upregulated in comparison to larynx. The gene is known to be hypermethylated and downregulated in smokers with a different type of lung cancer, lung adenocarcinoma. Normally, the gene is upregulated in lung and helps to deal with carcinogens in cigarette smoke. Mehthylation through smoke represses that process. Methylation in the SULT1C2 promoter region suppresses activation of the SULT1C2 detoxification enzyme and the cell cannot respond to cigarette smoke exposure as before anymore. Also, low expression of SULT1C2 is associated with low survival rates [18]. PRDM9 has also been identified as cancer gene and appears to be frequently mutated in head and neck squamous cell carcinoma [19]. PIWIL2 was reported as cancer-gene considering many cancer types ans also lung cancer [20].

Additionally we found evidence for two tumor suppressor genes: SHISA3 and MIR663AHG. SHISA3 was upregulated and hypomethylated in lung compared to larynx in our samples. The gene was recently described as a tumor suppressor gene in lung cancer [21]. The protein encoded by SHISA3 acts as a tumor suppressor by accelerating beta-catenin degradation [22]. SHISA3 is hypomethylated in the lung. Its missing methylation of the CpG islands can lead to gene activation preferentially in promoter regions of oncogenes. MIR663AHG was described as tumor suppressor gene only in colon cancer [23]. It was also upregulated and hypomethylated in lung compared to larynx.

5.2 Feature Selection

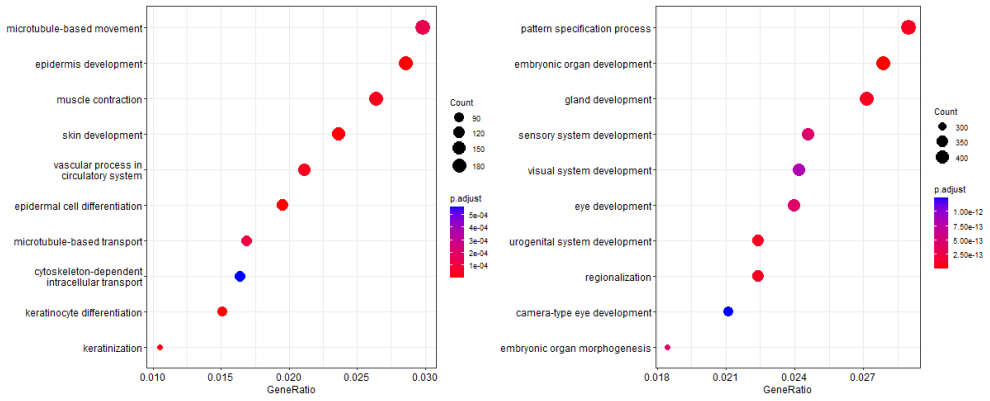
Feature selection separately the expression dataset resulted in 22 genes, chromosomes 1, 2 and 10 were overrepresented. Three SFT-pulmonary surface proteins were found and they appeared to be upregulated in the lung samples. SFT-pulmonary surface proteins take part in the surfactant metabolism, a metabolism that occurs in the lung epithelia and is responsible for the production, function and regulation of the pulmonary surfactant which lines the inner surfaces to prevent an alveolar collaps at the end of every air expiration [24].

When we performed feature selection separately on the methylation data, we ended up with 18 CpG islands in 16 genes, seven of them were close to the HOX genes either on chromosome 7 oder 17. There was no intersection with the genes from the feature selection on the expression data. For none of the CpG islands we found evidence for smoking-related cancer in the literature. Nevertheless, for seven CpG islands we found evidence considering specific profiles for LSCC or HNSCC: cg09017619, cg27508551, cg10474350, all located in promotors that affect HOXA7. All of them were all hypermethylated in lung compared to larynx and we found studies that link them to lung cancer or esophageal adenocarcinoma [25, 26]. One CpG island, cg21546671, in the promotor region of the HOX gene HOXB4, was found to be specific for lung adenocarcinoma [27]. cg13914083 is located in GRIA2 and is hypomethylated in HNSCC and in LSCC [25]. In our samples, the CpG island was hypermethylated in laryngeal cancer compared to lung cancer. cg25365934, close to zinc finger protein 503, was reported as one specific CpG island to distinguish between lung adenocarcinoma and LSCC [28]. Lastly, cg16711124, which we could not assign to a particular gene, was reported as a biomarker for cancer in general [29].

Feature selection on the integrated dataset with both, methylation and expression data at the same time, ended up in 43 variables, 29 genes and 14 CpG islands. The selected features were a good mixture of up- and downregulated, as well as hypo- and hypermethylated regions. 17 of the selected genes overlapped with the features from the feature selection on the expression dataset, ten with the features of the methylation dataset. Again, of the affected genes in the CpG islands in the selected features, there were no overlaps with the genes (the selected features were not hypo- or hypermethylated and up- or downregulated at the same time). HOX genes again popped up in this list. Methylation events affecting HOX gene expression play crucial roles in tumorgenesis specific methylation profiles in the HOX genes are recognized as potential biomarkers in several cancers [30]. The genes and CpG sites from the feature selection on the integrated data matrix are depicted in Appendix A, Tables 3 and 4 and are further analysed in section 5.7.

5.3 Enrichment Analysis

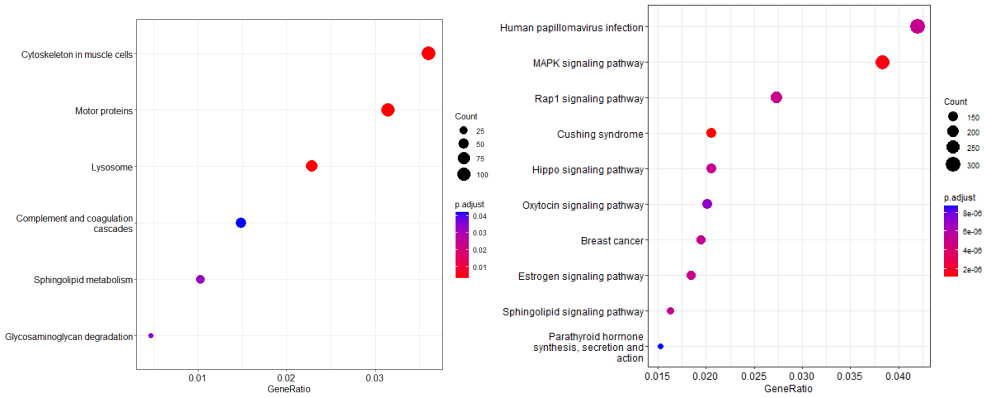
GO enrichment on the differentially expressed genes returned biological processes that describe the epidermis / skin, cytoskeleton and that include keratinization or keratinocytes as depicted in Figure 5. Keratins form intermediate filaments in epithelial cells that are located in the cytoskeleton. When we check the data from the DEA, we find 74 keratins that are differentially expressed. Only 25 of them are overexpressed in the lung, more than half of the keratins are overexpressed in larynx.



(a) DEGs. (b) DMGs.

Fig. 5: GO enrichment for biological processes.

In the pathway enrichment, again a cytosceleton-pathway pops up, as well as the motor protein pathway the HPV infection pathway, indicating that one of the sample groups was significantly more often affected by HPV. Figure 6.



(a) DEGs (b) DMGs

Fig. 6: KEGG pathway enrichment on DEGs and DMGs.

5.4 Machine Learning

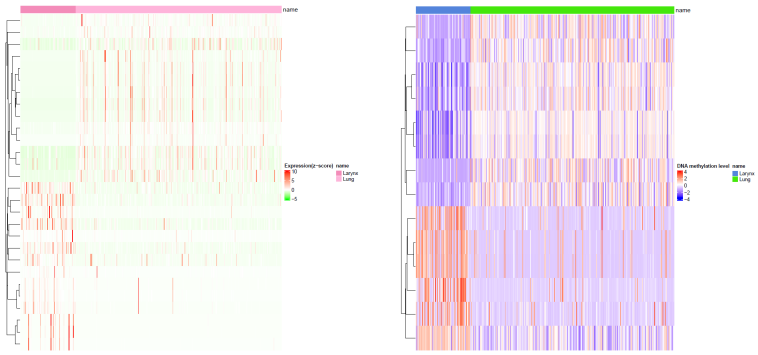
As input for our machine learning classifiers, we took the 43 features (Tables 3 and 4) and the integrated datasets from the feature selection and scaled the expression values.

5.5 Visualization of the selected features

We inspected the results of the feature selection visually after performing nonlinear dimensionality reduction (Figures 7a and 7b) and in heatmaps for the methylation and expression data (Figures 8a and 8b) and were confident about the clustering of the cancer types after feature selection.

(a) UMAP before feature selection. (b) UMAP after feature selection.

Fig. 7: UMAP Plots.



(a) Expression. (b) Methylation.

Fig. 8: Heatmaps on the selected features. Rows indicate patients, columns the selected features.

5.6 Model building and performance

We trained our classifiers on the selected features. We started with the random forest model. Grid search resulted in an optimal number of trees in the random forest model of 100 with an accuracy on the training set of 96 %. Even with five trees, we still obtained an accuracy of 88 %. After applying our model on the test-set, we obtained a sensitivity of 100 %, a specificity of 91 % and an accuracy of 93%. The area under the curve was 100 %. The performance of the Gradient Boosting Machine and the SVM were similarly good as depicted in Table 2 and in the receiver operating characteristic curve in Figure 11. Choosing a radial kernel in the SVM did not change the performance. From a distance metric for trees, using the package 'reprtree', the most representative tree out of the 100 trees is depicted in the appendix B in Figure 11. Our model of choice is the random forest model, as the decision trees as in Figure 11 are easy interpretable by medical doctors that seek to understand the biological rules.

Model	Accuracy	Sensitivity	Specificity	AUC
RF	96 %	100 %	91 %	100 %
GBM	89 %	100 %	86 %	100 %
SVM	91 %	100 %	99 %	100 %

Table 2: Accuracy, sensitivity, specificity and area under the curve (AUC) of the classifiers.

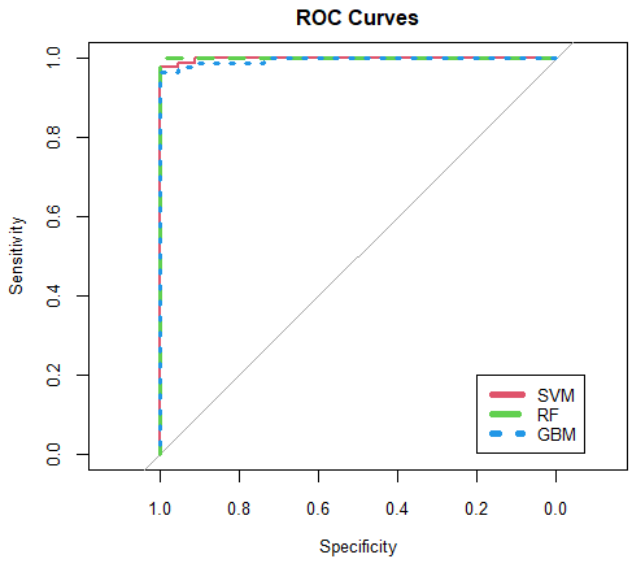


Fig. 9: ROCs from the performed machine learning models.

5.7 Feature importance analysis

The result of the feature importance analysis for the random forest classifier is depicted in Figure 10. The feature importance analysis picked 20 out of the 43 as most important features. Again, SFT-pulmonary surface proteins, SFTPA2, SFTPA1, SFTPB, SFTPC and SFTPD, were under the selected features. And again, all of them are all upregulated in the lung samples. Similar to our findings from the combined DEA and DMA results, the gene SULT1C2 popped up in our analysis pipeline [31]. Also, PGC was significant, which is mostly important in the process of digestion in the stomach, but it also occurs in the alveolar cells in the lung [32]. It is upregulated in lungs in comparison to larynx in our study. Its expression is upregulated in different types of cancer. NAPSA is also upregulated in the lung. It is enriched in alveolar cell types and already has been described as a marker for a different subtype of lung cancer, namely lung adenocarcinoma and renal cell carcinoma. [33] FLG is responsible for keratininzation and is upregulated in the larynx. KRT 9, Keratin 9, is also upregulated in the larynx. Although it does not translate into proteins, expression of KRT9 is typical for HNSCC [34]. MYH2 and MYL1 are motor proteins that are downregulated in larynx tissues in our samples. Motor proteins already appeared in the KEGG enrichment analyses. MYL1 is also known as a poor prognosis marker in HNSCC patients, where a relatively higher level of MYL1 indicates a bad outcome [35]. Again, we found HOX genes, namely genes from the HOXA-family.
We also checked for tumor suppressor and oncogenes again and found three tumor suppressor genes: WT1 is an upregulated gene in the lungs. It is known as a cancer-related gene as it is involved in tumour formation as a tumor suppressor [36]. Three

of the most important features were CpG islands located in promotor regions from tumor suppressor genes: The PDCHA1 gene, a protocadherin, is hypermethylated in the larynx. 29 CpG islands affect this gene, 28 are hypermethylated in larynx and the gene is downregulated in larynx. The protocadherin gene family is known as a potential tumor suppressor and its increased methylation can is reported as a useful marker for the early detection of cancer [37]. Two of the 18 CpG islands in the aforementioned tumor suppressor SHISA3 were selected as most important features, both hypomethylated in the lung. The other CpG islands in that region were also hypomethylated in lung except for four of them. Two of the genes in our most important features were additionally also selected by our manual analysis considering the starburst plot, the aforementioned SHISA3, and SULT1C2.

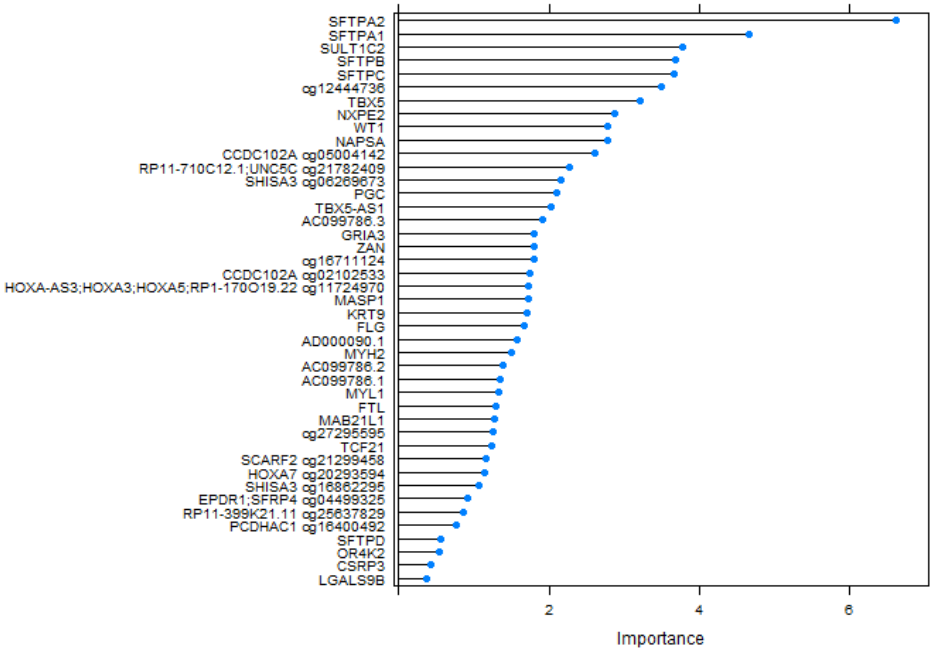


Fig. 10: Feature importance of selected features with annotated gene names.

6 Discussion

Although we did not find any evidence for smoking associated regions in the CpG islands from our feature selection, we found three genes that were both differentially methylated and differentially expressed and that were interesting in respect to cancer: SHISA3, MIR663AHG and SULT1C2. The first two were deregulated tumor suppressors in laryngeal carcinoma, the latter was downregulated in laryngeal carcinoma, indicating that the cells have a worse defence mechanism against smoking. The

other differentially expressed/ methylated regions we attribute to differences from the underlying tissue. In the lung we find ciliated epithelium, in the larynx stratified squamous epithelium. In lung cancer, specifically in LSCC, the ciliated epithelium transforms into squamous epithelium. Stratified squamous epithelium can keratinize, forming precancerous cells that might turn into cancer. Keratinization is one of the biological processes that popped up in our enrichment analysis. Together with the upregulation of almost all keratins in our larynx samples, we conclude, that this process plays a major role in laryngeal cancer. The reason and the exact biological background are not clear to us at this point. Maybe cigarette smoke first affects the upper parts of the respiratory pathway, so that the level of keratinization in larynx is elevated. Another reason could be, that the difference can be accounted to the normal tissue that might be present in the tumor sample (we were not handling single-cell RNA-Seq data). In that context the underlying reason could be, that the proportion of squamous cells in a laryngeal cancer sample is bigger compared to a lung cancer sample, and so is the keratinization level.
We found evidence for three tumor suppressors that were all downregulated in laryngeal compared to lung cancer: SHISA3, MIR663AHG and WT1. For the first two, we found evidence that they were also hypermethylated. SULT1C2, a gene that transcribed into a detoxification enzyme, is hypermethylated and underexpressed in lung compared to larynx. The hypermethylation can be an effect of smoking.
In summary, the differentially methylated and expressed genes mainly elucidate the different molecular mechanisms in laryngeal and lung carcinoma, which we conclude from our enrichment analyses. In the machine learning models on the integrated dataset, the most important features from methylation and expression datasets did not overlap. This indicates, that the distinction of the two sample types is possible by looking at methylation and expression data separately, and that it is not clearly attributable to a direct link between methylation (from smoking) and expression.

7 Conclusion and Limitations

We could not find evidence of a direct link between methylation on CpG site due to smoking and the expression of genes. Nevertheless, we found genes that were both, differentially expressed and methylated, indicating a difference in methylation between laryngeal and lung cancer at these sites that influence gene expression levels. Most of the differences we attribute to the different molecular mechanisms in larynx and lung cancer samples, some of them might not be specific for cancer.
Our random forest model resulted in a similarly high performance compared to the original study by Jurmeister et al. The three different models performed equally good but our dataset was relatively small. Applying our classifiers on new datasets (f.e. from the GEO database) would probably result in new insights and a worse performance. Also studying the HPV-status would be interesting, as it is a common cause of laryngeal cancer.
We conclude, that relatively simple machine learning classifiers may facilitate the differentiation between distant metastases and primary tumor. Random forests are the

models of choice in this case, because they perform well and are easy to interpret by medical staff.

8 Author contributions

Mia participated in the selection of the data and filtered the samples. She performed differential gene expression analysis with DESeq2 and performed data integration of differentially expressed genes and differentially methylated regions. She selected and trained the machine learning models and performed a feature importance analysis. Using the resulting genes, she analysed the pathways enrichments and put them into a biological context.
Friederike performed data preprocessing steps and the differential expression and methylation analysis (including visualization) using the TCGAbiolinks package. She summarized and visualized the selected features (UMAP, heatmaps and starburst plot). Also, she performed GO, KEGG enrichment. She contributed in the discussion of the results and the research concerning the literature databases and medical background.

A Appendix

ID	delta	pval	annot	affected_gene
ENSG00000122852	11.83167	1.718369494378 31e-65	Upregulated in Lung	SFTPA1
ENSG00000168878	11.81926	9.14695621477125e-75	Upregulated in Lung	SFTPB
ENSG00000168484	11.49531	7.82905101151806e-42	Upregulated in Lung	SFTPC
ENSG00000185303	10.84037	6.48431210318239e-52	Upregulated in Lung	SFTPA2
ENSG00000096088	8.58387	1.41001174344012e-36	Upregulated in Lung	PGC
ENSG00000131400	7.99971	2.38557880054046e-63	Upregulated in Lung	NAPSA
ENSG00000133661	5.07899	6.76117322483884e-37	Upregulated in Lung	SFTPD
ENSG00000198203	3.30098	6.05616025755168e-29	Upregulated in Lung	SULT1C2
ENSG00000118526	3.18256	4.12849583075751e-24	Upregulated in Lung	TCF21
ENSG00000283907	3.0945	1.61829834686735e-18	Upregulated in Lung	AD000090.1
ENSG00000089225	2.24977	8.09653674595423e-20	Upregulated in Lung	TBX5
ENSG00000184937	1.95531	3.68734939737716e-9	Upregulated in Lung	WT1
ENSG00000255399	1.54051	1.9585742210081e-8	Upregulated in Lung	TBX5-AS1
ENSG00000087086	1.52522	3.27481803217989e-26	Upregulated in Lung	FTL
ENSG00000125675	-2.11646	8.25421298361855e-28	Downregulated in Lung	GRIA3
ENSG00000180660	-2.13671	7.44748727670029e-8	Downregulated in Lung	MAB21L1
ENSG00000170298	-2.3476	2.58104060801723e-21	Downregulated in Lung	LGALS9B
ENSG00000224127	-2.69202	0.000818696199612317	Downregulated in Lung	AC099786.1
ENSG00000224149	-2.69734	0.000428360863787589	Downregulated in Lung	AC099786.2
ENSG00000146839	-2.8513	1.126743275772 94e-17	Downregulated in Lung	ZAN
ENSG00000127241	-2.86591	3.30637091316813e-53	Downregulated in Lung	MASP1
ENSG00000171403	-2.99511	3.3089967471522e-28	Downregulated in Lung	KRT9
ENSG00000143631	-2.99935	4.22458579965728e-39	Downregulated in Lung	FLG
ENSG00000261213	-3.58092	0.000008219890553427 36	Downregulated in Lung	AC099786.3
ENSG00000165762	-4.03905	2.78937656509318e-8	Downregulated in Lung	OR4K2
ENSG00000204361	-5.12687	4.7066487931381e-37	Downregulated in Lung	NXPE2
ENSG00000125414	-7.22694	3.19852039438047e-47	Downregulated in Lung	MYH2
ENSG00000129170	-8.16953	5.96898060556931e-40	Downregulated in Lung	CSRP3
ENSG00000168530	-8.47133	5.45993411071864e-23	Downregulated in Lung	MYL1

Table 3: Genes after features selection. Positive delta values indicate upregulation in lung, negative delta values indicate downregulation in lung compared to larynx. Color code: Yellow = downregulated in lung compared to larynx, Green = upregulated in lung compared to larynx.

ID	delta	pval	annot	affected_gene
cg16862295	0.27182	8.72244982230305e-11	Hypomethylated in Lung	SHISA3
cg04499325	0.25162	1.86558332649985e-13	Hypomethylated in Lung	EPDR1;SFRP4
cg06269673	0.25151	2.45255760201449e-13	Hypomethylated in Lung	SHISA3
cg12444736	0.23202	1.22988761769238e-17	Hypomethylated in Lung	
cg21782409	0.22251	2.49471576245535e-9	Hypomethylated in Lung	RP11-710C12.1;UNC5C
cg16400492	0.20448	1.08141307605671e-9	Hypomethylated in Lung	PCDHA1;PCDHA10;PCDHA11;PCDHA12;PCDHA13
cg16711124	-0.2032	5.07274640303486e-16	Hypermethylated in Lung	
cg21299458	-0.21137	1.16961637833739e-20	Hypermethylated in Lung	SCARF2
cg11724970	-0.22188	7.83507690481821e-19	Hypermethylated in Lung	HOXA-AS3;HOXA3;HOXA5;RP1-170O19.22
cg27295595	-0.22635	2.79373657113536e-15	Hypermethylated in Lung	

Table 4: CpG islands from feature selection. Color code: Yellow = hypomethylated in lung compared to larynx, Green = hypermethylated in lung compared to larynx.

B Appendix

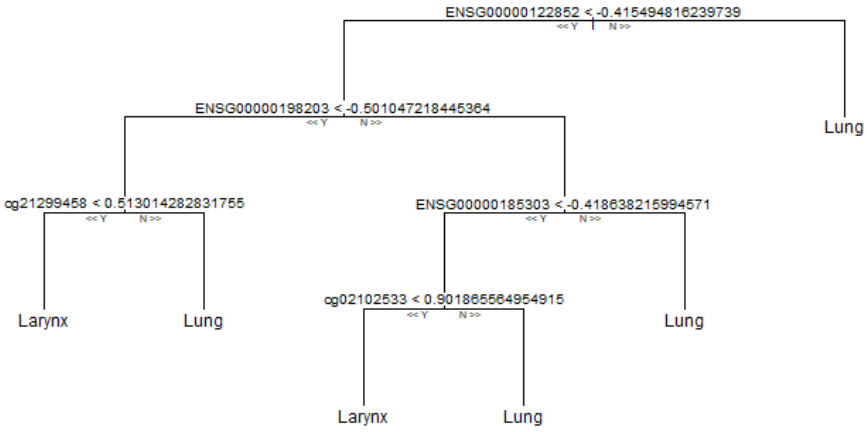


Fig. 11: Most representative tree of the random forest using 100 trees.

References

[1] Khuder, S.A.: Effect of cigarette smoking on major histological types of lung cancer: a meta-analysis **31**(2), 139–148 https://doi.org/10.1016/S0169-5002(00)

00181-1 . Accessed 2024-07-23

[2] Howell, J.Y., Hadian, Y., Ramsey, M.L.: Squamous Cell Skin Cancer. https://www.ncbi.nlm.nih.gov/books/NBK441939/. Updated 2024 Mar 27. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. (2024)

[3] Sabbula, B.R., Gasalberti, D.P., Mukkamalla, S.K.R., et al.: Squamous Cell Lung Cancer. https://www.ncbi.nlm.nih.gov/books/NBK564510/. Updated 2024 Feb 14. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. (2024)

[4] Stapelfeld, C., Dammann, C., Maser, E.: Sex-specificity in lung cancer risk **146**(9), 2376–2382 https://doi.org/10.1002/ijc.32716 . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.32716. Accessed 2024-07-23

[5] Geurts, T.W., Velthuysen, M.L.F., Broekman, F., Huysduynen, T.H., Brekel, M.W.M., Zandwijk, N., Tinteren, H., Nederlof, P., Balm, A.J.M., Brakenhoff, R.H.: Differential diagnosis of pulmonary carcinoma following head and neck cancer by genetic analysis. Clinical Cancer Research **15**(3), 980–985 (2009) https://doi.org/10.1158/1078-0432.CCR-08-1968

[6] Jurmeister, P., Bockmayr, M., Seegerer, P., Bockmayr, T., Treue, D., Montavon, G., Vollbrecht, C., Arnold, A., Teichmann, D., Bressem, K., Schüller, U., Laffert, M., Müller, K.-R., Capper, D., Klauschen, F.: Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. Science Translational Medicine **11**(509), 8513 (2019) https://doi.org/10.1126/scitranslmed.aaw8513 . Publisher: American Association for the Advancement of Science. Accessed 2024-07-06

[7] Leitheiser, M., Capper, D., Seegerer, P., Lehmann, A., Schüller, U., Müller, K.-R., Klauschen, F., Jurmeister, P., Bockmayr, M.: Machine learning models predict the primary sites of head and neck squamous cell carcinoma metastases based on DNA methylation. The Journal of Pathology **256**(4), 378–387 (2022) https://doi.org/10.1002/path.5845 . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/path.5845. Accessed 2024-07-06

[8] Thirumani, L., Helan, M., S, V., Jamal Mohamed, U., Vimal, S., Madar, I.H.: The molecular landscape of lung metastasis in primary head and neck squamous cell carcinomas. Cureus **16**(4), 57497 (2024) https://doi.org/10.7759/cureus.57497

[9] Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., Staudt, L.M.: Toward a shared vision for cancer genomic data. New England Journal of Medicine **375**(12), 1109–1112 (2016) https://doi.org/10.1056/NEJMp1607591

[10] Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D.,

Sabedot, T., Malta, T.M., Pagnotta, S.M., Castiglioni, I., Ceccarelli, M., Bontempi, G., Noushmehr, H.: Tcgabiolinks: An r/bioconductor package for integrative analysis of tcga data. Nucleic Acids Research (2015) https://doi.org/10.1093/nar/gkv1507

[11] Silva, T.C., Colaprico, A., Olsen, C., D'Angelo, F., Bontempi, G., Ceccarelli, M., Noushmehr, H.: Tcga workflow: Analyze cancer genomics and epigenomics data using bioconductor packages. F1000Research **5** (2016)

[12] Mounir, M., Lucchetta, M., Silva, T.C., Olsen, C., Bontempi, G., Chen, X., Noushmehr, H., Colaprico, A., Papaleo, E.: New functionalities in the tcgabiolinks package for the study and integration of cancer data from gdc and gtex. PLOS Computational Biology **15**(3), 1006701 (2019) https://doi.org/10.1371/journal.pcbi.1006701

[13] Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology **15**, 550 (2014) https://doi.org/10.1186/s13059-014-0550-8

[14] Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., Yu, G.: clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation **2**(3), 100141 (2021) https://doi.org/10.1016/j.xinn.2021.100141

[15] Yu, G., Wang, L.-G., Han, Y., He, Q.-Y.: clusterprofiler: an r package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology **16**(5), 284–287 (2012) https://doi.org/10.1089/omi.2011.0118

[16] Kuhn, Max: Building predictive models in r using the caret package. Journal of Statistical Software **28**(5), 1–26 (2008) https://doi.org/10.18637/jss.v028.i05

[17] Banerjee, M., Ding, Y., Noone, A.M.: Identifying representative trees from ensembles. Statistics in Medicine (2012)

[18] Johnson, C., Mullen, D.J., Selamat, S.A., Campan, M., Offringa, I.A., Marconett, C.N.: The sulfotransferase SULT1c2 is epigenetically activated and transcriptionally induced by tobacco exposure and is associated with patient outcome in lung adenocarcinoma **19**(1), 416 https://doi.org/10.3390/ijerph19010416 . Accessed 2024-07-25

[19] Allali-Hassani, A., Szewczyk, M.M., Ivanochko, D., Organ, S.L., Bok, J., Ho, J.S.Y., Gay, F.P.H., Li, F., Blazer, L., Eram, M.S., Halabelian, L., Dilworth, D., Luciani, G.M., Lima-Fernandes, E., Wu, Q., Loppnau, P., Palmer, N., Talib, S.Z.A., Brown, P.J., Schapira, M., Kaldis, P., O'Hagan, R.C., Guccione, E., Barsyte-Lovejoy, D., Arrowsmith, C.H., Sanders, J.M., Kattar, S.D., Bennett, D.J., Nicholson, B., Vedadi, M.: Discovery of a chemical probe for PRDM9 **10**(1), 5759 https://doi.org/10.1038/s41467-019-13652-x . Publisher: Nature Publishing

Group. Accessed 2024-07-26

[20] PIWIL2 Promotes Progression of Non-small Cell Lung Cancer by Inducing CDK2 and Cyclin A Expression. https://www.springermedizin.de/piwil2-promotes-progression-of-non-small-cell-lung-cancer-by-ind/9767188 Accessed 2024-07-26

[21] Chen, C.-C., Chen, H.-Y., Su, K.-Y., Hong, Q.-S., Yan, B.-S., Chen, C.-H., Pan, S.-H., Chang, Y.-L., Wang, C.-J., Hung, P.-F., Yuan, S., Chang, G.-C., Chen, J.J.W., Yang, P.-C., Yang, Y.-C., Yu, S.-L.: Shisa3 is associated with prolonged survival through promoting -catenin degradation in lung cancer **190**(4), 433–444 https://doi.org/10.1164/rccm.201312-2256OC . Publisher: American Thoracic Society - AJRCCM. Accessed 2024-07-23

[22] The Human Protein Atlas: SHISA3 Summary. Accessed: 2024-07-23. https://www.proteinatlas.org/ENSG00000178343-SHISA3

[23] Yuan, H., Ren, Q., Du, Y., Ma, Y., Gu, L., Zhou, J., Tian, W., Deng, D.: LncRNA miR663ahg represses the development of colon cancer in a miR663a-dependent manner **9**, 220 https://doi.org/10.1038/s41420-023-01510-1 . Accessed 2024-07-23

[24] Agassandian, M., Mallampalli, R.K.: Surfactant phospholipid metabolism. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids **1831**(3), 612–625 (2013) https://doi.org/10.1016/j.bbalip.2012.09.010 . Epub 2012 Sep 29

[25] Kumar, D., Cinghu, S., Oldfield, A.J., Yang, P., Jothi, R.: Decoding the function of bivalent chromatin in development and cancer **31**(12), 2170–2184 https://doi.org/10.1101/gr.275736.121 . Accessed 2024-07-25

[26] Peng, W., Tu, G., Zhao, Z., He, B., Cai, Q., Zhang, P., Peng, X., Shi, S., Wang, X.: DNA methylome and transcriptome analysis established a model of four differentially methylated positions (DMPs) as a diagnostic marker in esophageal adenocarcinoma early detection **9**, 11355 https://doi.org/10.7717/peerj.11355 . Publisher: PeerJ Inc. Accessed 2024-07-25

[27] Shen, N., Du, J., Zhou, H., Chen, N., Pan, Y., Hoheisel, J.D., Jiang, Z., Xiao, L., Tao, Y., Mo, X.: A diagnostic panel of DNA methylation biomarkers for lung adenocarcinoma **9**, 1281 https://doi.org/10.3389/fonc.2019.01281 . Accessed 2024-07-25

[28] Jiang, J.-H., Gao, J., Chen, C.-Y., Ao, Y.-Q., Li, J., Lu, Y., Fang, W., Wang, H.-K., Castro, D.G., Santarpia, M., Hashimoto, M., Yuan, Y.-F., Ding, J.-Y.: Circulating tumor cell methylation profiles reveal the classification and evolution of non-small cell lung cancer **11**(2), 224–237 https://doi.org/10.21037/tlcr-22-50 . Accessed 2024-07-25

[29] , , : Method and System for Determining Cancer Status. TW202011416A. https://patents.google.com/patent/TW202011416A/en Accessed 2024-07-25

[30] Paço, A., Bessa Garcia, S.A., Freitas, R.: Methylation in HOX clusters and its applications in cancer therapy **9**(7), 1613 https://doi.org/10.3390/cells9071613 . Accessed 2024-07-25

[31] The Human Protein Atlas: SULT1C2 Summary. Accessed: 2024-07-23. https://www.proteinatlas.org/ENSG00000198203-SULT1C2

[32] The Human Protein Atlas: PGC Summary. Accessed: 2024-07-23. https://www.proteinatlas.org/ENSG00000096088-PGC

[33] The Human Protein Atlas: NAPSA Summary. Accessed: 2024-07-23. https://www.proteinatlas.org/ENSG00000131400-NAPSA

[34] Expression of KRT9 in Cancer - Summary - The Human Protein Atlas. https://www.proteinatlas.org/ENSG00000171403-KRT9/pathology Accessed 2024-07-23

[35] Li, C., Guan, R., Li, W., Wei, D., Cao, S., Chang, F., Wei, Q., Wei, R., Chen, L., Xu, C., Wu, K., Lei, D.: Analysis of myosin genes in HNSCC and identify MYL1 as a specific poor prognostic biomarker, promotes tumor metastasis and correlates with tumor immune infiltration in HNSCC **23**, 840 https://doi.org/10.1186/s12885-023-11349-5 . Accessed 2024-07-23

[36] The Human Protein Atlas: WT1 Summary. Accessed: 2024-07-23. https://www.proteinatlas.org/ENSG00000184937-WT1

[37] Dutra, T., Bezerra, T., Luna, E., Carvalho, F., Chaves, F., Barros Silva, P., Costa, F., Pereira, K.: Do protocadherins show prognostic value in the carcinogenesis of human malignant neoplasms? systematic review and meta-analysis. Asian Pacific Journal of Cancer Prevention **21**(12), 3677–3688 (2020) https://doi.org/10.31557/APJCP.2020.21.12.3677