Ruprecht-Karls-Universität Heidelberg
Fakultät für Biowissenschaften
Bachelorstudiengang Molekulare Biotechnologie

Cancer Hallmark and Metabolic Pathways in Cancer
Topic 02 Team 03
Exploration of Biomarkers in Lung Adenocarcinoma (LUAD)

Data Science Project SoSe 2022

Authors Paul Brunner, Marie Kleinert, Felipe Stünkel, Chloé Weiler
Submitted on 18.07.2022

Abstract

Lung adenocarcinoma (LUAD) is the most common type of lung cancer. Its poor early diagnosis rate and high mortality create a great demand to identify biomarkers improving diagnosis and prognosis. This study aims to determine driver genes suitable to predict LUAD based on a patient's gene expression pattern.

To this end, two RNA-seq datasets from The Cancer Genome Atlas, one containing molecular data from 33 different cancer types and one containing data from LUAD tissue and from corresponding healthy tissue were analysed based on gene expression and pathway enrichment. In an effort to characterize LUAD-specific deregulations, Gene Set Enrichment Analysis and Gene Set Variation Analysis were performed using the geneset list provided by the Molecular Signatures Database.

Patients suffering from LUAD exhibit a general downregulation of the renal system but upregulation of HIF1 alpha. In comparison to other cancer types, LUAD expresses a more immunogenic and pro-inflammatory molecular signature. Moreover, genes involved in alveolal development show a notably higher expression rate. Further analysis revealed two subclasses of LUAD differing mainly in their blood clotting capabilities.

With the help of these findings a regression model was developed that helps predict LUAD and non-LUAD patients with 98.6 percent accuracy.

Our study provides a diverse analysis of driver genes and pathways underlying LUAD progression. Identifying differences between LUAD and healthy tissue, LUAD to itself and between LUAD and other cancer types suggests related diagnostic, prognostic and therapeutical value.

Contents

Abbreviations

ADGRF1	Adhesion G protein-coupled receptor F1	LGG	Low grade glioma
AGTR1	Angiotensin II receptor 1	LIHC	Liver hepatocellular carcinoma
AML	Acute myeloid leukemia	LUAD	Lung adenocarcinoma
AUC	Area under the curve	MSigDB	Molecular Signature Database
BRCA	Breast cancer	ORC	Origin Recognition Complex
CALCA	Calcitonin Related Polypeptide Alpha	OV	Ovarian serous cystadenocarcinoma
CD36	Cluster of differentiation 36	PAAD	Pancreatic adenocarcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	PC	Principal component
CYP	Cytochrome p450	PCA	Principal Component Analysis
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	PCPG	Pheochromocytoma and Paraganglioma
EDNRB	Endothelin receptor type B	PPAR gamma	Peroxisome proliferator-activated receptor gamma
EMT	Epithelial to mesenchymal transition	PRAD	Prostate adenocarcinoma
ESCA	Esophageal carcinoma	RAS	Renin-angiotensin system
FGA	Fibrinogen alpha chain	RNA-seq	RNA sequencing
GBM	Glioblastoma multiformae	ROC	Receiver operating characteristic
GSEA	Gene Set Enrichment Analysis	SARC	Sarcoma
GSVA	Gene Set Variation Analysis	SKCM	Skin Cutaneous Melanoma
HIF1 alpha	hypoxia inducible factor 1 subunit alpha	TCGA	The Cancer Genome Atlas
HNSC	Head and neck squamous cell carcinoma	Th17	IL-17–producing CD4 helper T cells
INSL4	Gene encoding insulin-like 4 protein	THCA	Thyroid carcinoma
KICH	Kidney chromophobe	TPM	Transcripts per million
KIRC	Kidney renal clear-cell carcinoma	UCEC	Uterine Corpus Endometrial Carcinoma
KIRP	Kidney renal papillary cell carcinoma	UMAP	Uniform Manifold Approximation and Projection
LAML	Acute Myeloid Leukemia	UVM	Uveal melanoma
LDL	low-density lipoprotein	VEGF	Vascular endothelial growth factor

1 Introduction

1.1 Background

To this day, lung cancer is the leading cause of cancer death worldwide (Zhang et al., 2020). Lung adenocarcinoma (LUAD) is a form of non-small cell lung cancer which accounts for approximately 40% of lung cancer cases and which is characterized by a remarkably low 5-year overall survival rate of merely 18% (Li and Lu, 2022; Wang et al., 2020). In theory, every cell is capable of developing into a cancer cell through acquisition of so-called hallmark capabilities that drive tumor formation due to numerous genetic mutations (Hanahan and Weinberg, 2011; Peng et al., 2018). In order to gain an insight into which mutations drive cancer development we revert back to the pan cancer project. The Cancer Genome Atlas (TCGA) is a publicly available collection of datasets that store the most important cancer-causing genomic alterations in order to create an 'atlas' of cancer genomic profiles (Tomczak et al., 2015). In the Pan Cancer project, data is collected from thousands of cancer patients and subsequently analysed and interpreted in an attempt to gain a deeper understanding of the genomic changes that drive a normal cell to become cancerous. The TCGA project provides a vast amount of RNA sequencing (RNA-seq) data which makes it possible to go beyond static genome analysis and to gain an insight into the transcriptional landscape of a cell. Studying gene expression profiles of a given cell through monitoring RNA synthesis enables researchers to gain a deeper understanding of how gene expression is regulated in cells and its impact on the cell's phenotype (Marguerat and Bähler, 2010).

1.2 Computational Tools

1.2.1 Uniform Manifold Approximation and Projection (UMAP)

Uniform manifold approximation and projection (UMAP) is a k-neighbor graph based algorithm that is used for non-linear dimension reduction (Smets et al., 2019) (McInnes et al., 2018). After data normalization, the Euclidean distances between points in a two-dimensional space of the graph are calculated and a local radius is determined (Vermeulen et al., 2021). In general the closer two points are to each other, the more similar they are. UMAP makes a density estimation to find the right local radius. A large number of k-nearest neighbors will ensure that more global structure is preserved whereas a smaller number of neighbors will ensure the preservation of more local structure (McInnes et al., 2018). An alternative method for dimensionality reduction is principal component analysis (PCA) which is explained in detail in the appendix. UMAP is a newer method than PCA and it is generally believed to facilitate interpretation and grouping of data (Milošević et al., 2022).

1.2.2 Gene Set Variation Analysis (GSVA)

Gene set variation analysis (GSVA) is an unsupervised method to estimate pathway activities based on gene expression data. GSVA quantifies enrichment in a sample-wise manner independently of phentoypes which makes GSVA the most adequate method to perform on the `tcga_exp`. GSVA estimates a cumulative distribution for each gene over all samples. The gene expression values are then converted according to these estimated cumulative distributions into scaled values. Based on these new values, the genes are ranked in each sample. Next, the genes are classified into two distributions and a Komogorow-Smirnow statistic is calculated to judge how similar the distributions are to each other and to obtain an enrichment score (ES). A highly positive or negative ES indicates that the studied geneset is positively or negatively enriched compared to the genes not in the geneset, respectively. If the ES for a given geneset is close to zero, then the geneset is probably not differentially expressed compared to the genes not in the geneset (Hänzelmann et al., 2013a). An alternative method to evaluate gene scoring is gene set enrichment analysis (GSEA) which is explained in detail in the appendix.

1.3 Objective

Knowing the devastating impact of LUAD on all those affected by the disease, we set out to learn more about its underlying genetic mutations. By comparing gene expression patterns in tumorous tissue to that in healthy tissue within one patient as well as to other cancer types, we hope to gain a better understanding of which metabolic deregulations are the root cause of LUAD and what makes LUAD unique. Identifying the precise genes that are up- or downregulated in LUAD tumor cells reveals the pathways most involved in tumor development and opens up new doors for cancer diagnostics. Not only do we hope to find a way to predict LUAD based on a cell's gene expression pattern but by revealing the driver genes of the disease, whole genome sequencing could be replaced by more efficient panel sequencing methods.

2 Methods

An overview of all used packages can be found in the appendix (Appendix, **Tab. 7.1**).

2.1 Our Data and Geneset Extraction

We were given four datasets, two of which contained RNA-seq data, one of which contained clinical annotations pertaining to one of the RNA-seq data frames and one of which contained a list of genesets for cancer hallmark analysis. The first RNA-seq dataset containing data from almost 10,000 TCGA cancer patients of 33 different tumor types was used to perform pan cancer analysis and to create a logistic regression model. The second RNA-seq dataset is a smaller data frame containing the TCGA expression data of tumor tissue and the corresponding healthy tissue for five different cancer types. It was used to perform a focused analysis. All expression data was already log2(TPM) transformed to reduce skewness in data and to make it more conform to a normal distribution. Here, TPM stands for 'Transcripts per million' and refers to a method of RNA-seq normalization in which one first accounts for gene length before adjusting for sequencing depth which helps to reduce type I and type II errors (Yuen In and Pincket, 2022).
The Molecular Signature Database (MSigDB) is a database offering a variety of annotated genesets publicly available for analysis. The import of genesets from MSigDB into RStudio can easily be performed using the R package "msigdbr" (Dolgalev, 2022) which allows for the extraction of species-specific genesets of the category of interest. The aim of this was to extract curated (C2) and ontology (C5 BP) genesets which were used for focused analysis as well as pan cancer analysis. The curated genesets that regulate the metabolism of cells were also used for comparison with given hallmark pathways that are often deregulated in cancer cells.

2.2 Data cleanup on TCGA datasets

After checking for NAs in the big TCGA dataset, all genes that showed a standard deviation lower than the 50% quantile of the standard deviation distribution were removed. Using the BiomaRt package (Durinck et al., 2005) the biotypes of all genes were compared to biotypes of the given geneset list, as well as to the geneset lists retrieved from MSigDB. All genes linked to biotypes not part of the biotypes of said genesets and possible pseudogenes were removed. Moreover, the small TCGA dataset for LUAD patients was cleaned using the same workflow as explained before. However, this time only zero variance genes were deleted.
To uncover clusters in the data, a combination of a PCA and UMAP was conducted on the cleaned TCGA expression dataset. Aiming to minimize artefacts caused by correlating variables, a PCA using the `RunPCA` command from the Seurat package (Satija et al., 2022) was conducted (McInnes et al., 2018). Afterwards, the UMAP analysis was done on the produced principle components using the uwot package (Melville, 2021) and visualized with ggplot2 (Wickham, 2016). After splitting the dataset into 33 data frames, one for each cancer type, the aforementioned workflow was then used on each of the cancer type subsets. To analyse the created plots, clusters k-means clustering (R Core Team,

2022a) was performed aiming to assign each of the patients to their corresponding cluster. The ideal number of clusters was evaluated using the silhouette method with function from the cluster package (Maechler et al., 2021). To create a volcano plot with the EnhancedVolcano package (Blighe et al., 2021), foldchanges between each cluster and the rest of the patients were calculated using the foldchange function (Warnes et al., 2022a) and a two sided Wilcoxon test was conducted on each cluster and the rest of the data points.

2.3 Differential Expression Analysis

A GSVA was conducted with the homonymous package to evaluate the enrichment of genes that play a role in the aforementioned pathways (Hänzelmann et al., 2013b). This package enables multi-core calculations for reduction of calculation time. Two different geneset lists retrieved from MSigDB were used: gene ontology genesets and curated genesets, each containing roughly 3,000 genesets. Genesets that showed an overlap below 95% with the genes in the chosen dataset were removed. Afterwards, the output matrix was cleaned from pathways with a low standard deviation and then split into the different tumor types. Calculating the means per pathway over all patients of one tumor type resulted in a tumor type pathway enrichment matrix which was visualized using the ComplexHeatmap package (Gu et al., 2016). The same workflow was repeated for each geneset in the large dataset as well as for the small dataset.

To assess how strongly pathway activity differs between tumor and normal tissue, a GSEA was performed using the fgsea package (Korotkevich et al., 2019). A previously created signed-rank list of decreasing p-values and a combination of metabolism and hallmark genesets served as input for the pathway list. An enrichment score and a leading edge containing the genes that contribute most to the enrichment score were calculated for each pathway. To sort the pathways by expression rate, the mean expression of their leading edges was calculated and visualized in a barplot using the ggplot package (Warnes et al., 2022b).

2.4 Regression

In order to identify LUAD risk patients from RNA-seq data, a logistic regression model was trained through binary classification between LUAD and non-LUAD patients. First of all, the cleaned dataset was split by a 70 to 30 percent ratio into a training and testing dataset respectively. In an effort to find genes that could be used as explaining variables, gene foldchanges (Warnes et al., 2022a) between LUAD and other cancer types were calculated using the large cleaned dataset. The genes were tested for correlation and in the case of highly correlating genes, one of them was removed. The 10 most over- and 10 most underexpressed genes were chosen for further testing. As a quality control, PCA (Satija et al., 2022) and UMAP (Melville, 2021) were conducted on all patients for these 20 chosen genes. A first rough model was trained using all 20 of the chosen genes and the glm function (R Core Team, 2022a). The blorr package was used to determine the best composition of the given 20 genes (Hebbali, 2020). With the best configuration, the final model was trained on the training dataset. Using the native predict function (R Core Team, 2022a), the model was evaluated by predicting whether the patients of the testing dataset were LUAD patients. The resulting probabilities were transformed into predictions for LUAD or non-LUAD using a cutoff value of 50%. Next a confusion table was used to estimate the false-positive and false-negative rates using the known tumor type of each patient and comparing that to the prediction of the model. As a final evaluation step the package ROCR (Sing et al., 2005) was used to create a ROC curve.

3 Results

3.1 Data cleaning

The analysis of the expression dataset showed no NAs in any of the patients, which meant none of the patients had to be removed in order to continue working with the dataset. After the further clean-up process described in part 3.4 the expression dataset could be reduced from over 60.000 to roughly 17.000 genes while keeping all 9741 patients. The tumor vs. normal dataset for LUAD patients was reduced to around 17.000 genes for 58 patients.

3.2 Cancer hallmark pathways

The similarity between chosen pathways was measured using the Jaccard index and afterwards visualized in a heatmap (Appendix, **Fig. 6.1**). Similarity is generally low and the Jaccard index does not exceed 0.2. The highest amount of shared genes is found in the alanine, aspartate and glutamate metabolism with glutamine metabolism.

3.3 Focused analysis

Overall, as shown in **Fig. 6.2** more upregulated than downregulated pathways were observed. Among the upregulated, Meta_HIF1 and the ascorbate and alderate metabolism pathways presented with the highest mean as well as some pathways associated with nucleotide, amino acid, and sugar metabolism. Downregulated pathways are mainly linked to the immune response but also to the metabolism of specific amino acids and fatty acids. The two most significantly downregulated pathways regulate linolenic acid and nitrogen metabolism.

The heatmap (**Fig. 3.1**) shows two recognizable groups with different expression patterns. The left half shows only tumor samples while the right half shows almost exclusively healthy samples, with a few tumor samples among them framed in black.
Strong differences can be seen in the pathways marked in violet. Many of the pathways that are strongly upregulated in tumor cells are associated with DNA replication or chromosome distribution during mitosis.The pathways marked in blue are connected to immune response, for example that of T-helper 17 cells. In this case, some tumor samples show upregulation while most show downregulation. In normal tissue the expression of most pathways is clearly upregulated with a few exceptions near the right margin. The volcano plot (**Fig. 6.3**) confirms these findings.

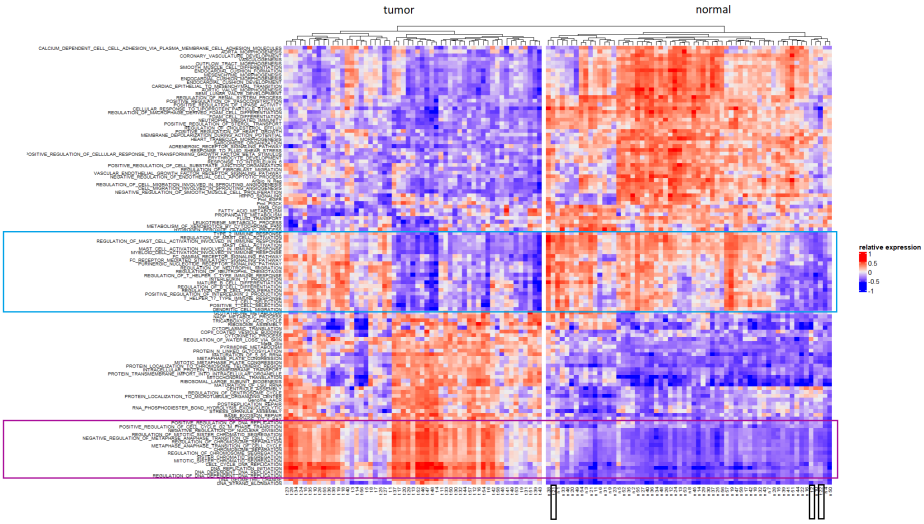

Figure 3.1: Heatmap for comparing pathway expression in normal and tumor tissue The heatmap shows the GSVA results for the small TCGA dataset. On the x axis one can see the normal and tumor samples, while the y axis shows the pathways with the highest standard deviation. The expression of each pathway is color-coded, from high expression (red), to low expression (blue).

A total of four pathways with particularly significant p values and high absolute foldchanges were selected.In (**Appendix Fig. 6.4**), genes of these pathways are highlighted in black and some with particularly low p values are labeled. Pathways involved in DNA replication and cell cycle progression were globally overexpressed (**Appendix Fig. 6.4**). Two down regulated pathways, namely renal system regulation and cellular response to lipoprotein particle stimulus are seen in **Appendix Fig. 6.4**(C, D). PPAR gamma or CD36 seem to be significantly down-regulated in the latter pathway.

3.4 Pan cancer analysis

3.4.1 Visualization of TCGA patients of different tumor types

The first step to pan cancer comparison was to evaluate potential clusters in our data. Dimension reduction of the cleaned data conducted by performing PCA and UMAP results in the plot shown in (**Fig. 3.2**). Data point were colored according to the cancer types they belong to in order to find clusters. Based on the 33 different types occurring in the dataset, the reduced data results in approximately 16 clusters. Notably, BRCA, LIHC, KIRP, SKCM, UVM, THCA, PCPG and PAAD exhibit well defined clustering. Additionally, LGG and GBM form a united cluster. Patients suffering from LUAD show a similar gene expression indicated by the isolated, turquoise cluster in the right, bottom corner.

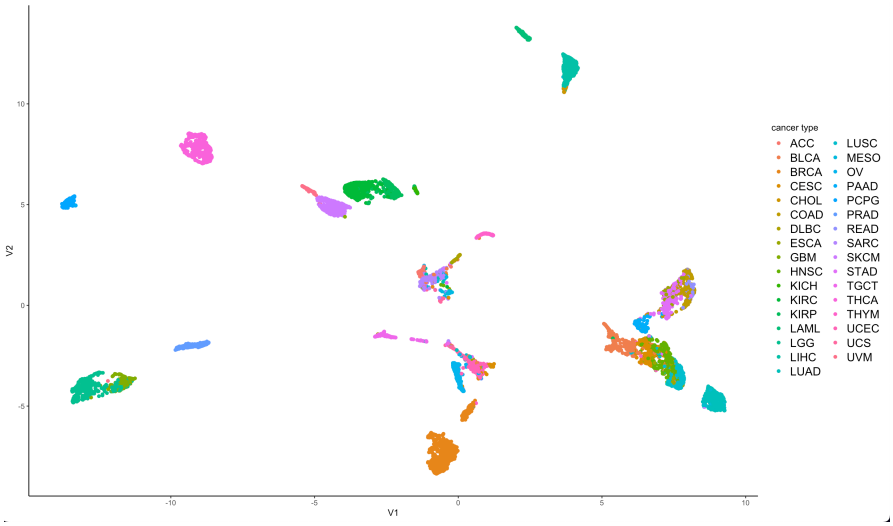

Figure 3.2: UMAP plot on TCGA expression dataset The x-axis is defined by the first umap component, whereas the y-axis assigned to the second component. The data plots are colored by the patients cancer type

3.4.2 Pathway enrichment

The diagnosis of the cancer type a patient suffers from is not only based on the cancer's location in the body but by the molecular signature it exhibits. The deregulation of pathways is characteristic for each cancer type. Aiming to identify differences in pathway activities based on the cancer type, two geneset list were extracted from MSigDB, as described in section 3.3. One list contained curated genesets whereas the other list contained ontology genesets. GSVA was performed twice on the TCGA expression dataset; once using the curated geneset list and once with the ontology geneset list for quality control. By utilizing the genesets separately, the better fitting geneset for the analysed dataset could be selected (Appendix, **Fig. 6.6**; Appendix, **Fig. 6.7**). The selected geneset list contains only ontology genesets that overlap with the genes from the expression data by more than 95%. Cancer types that result in an isolated and well defined cluster after GSVA are LIHC, KIRK, THCA, PRAD, PCPG, LGG and LAML.

Based on the geneset enrichment matrix created with GSVA, a pathway enrichment heatmap was created (**Fig. 3.3**). Performing k-means clustering, three clusters of cancer types were identified. The cancer types allocated to the first cluster can be categorized into kidney carcinomata, gliomata, carcinomata of the sexual organ as well as thyroid and liver carcinoma. The other two clusters exhibit no specific subcategories explaining similar pathway deregulation patterns. Cancer types belonging to cluster one show a general strong down regulation of pathways in comparison to the other cancer types. Cluster 2 contains cancer types with distinct deregulation of pathways. The third cluster exhibits a strong deregulation of pathways relatively to the other cancer types, some being upregulated while others are severely downregulated. The first cluster includes pathways that regulate the cell cycle, DNA replication and chromatid segregation.Cancer types included in the first cluster inhibit immune activation.

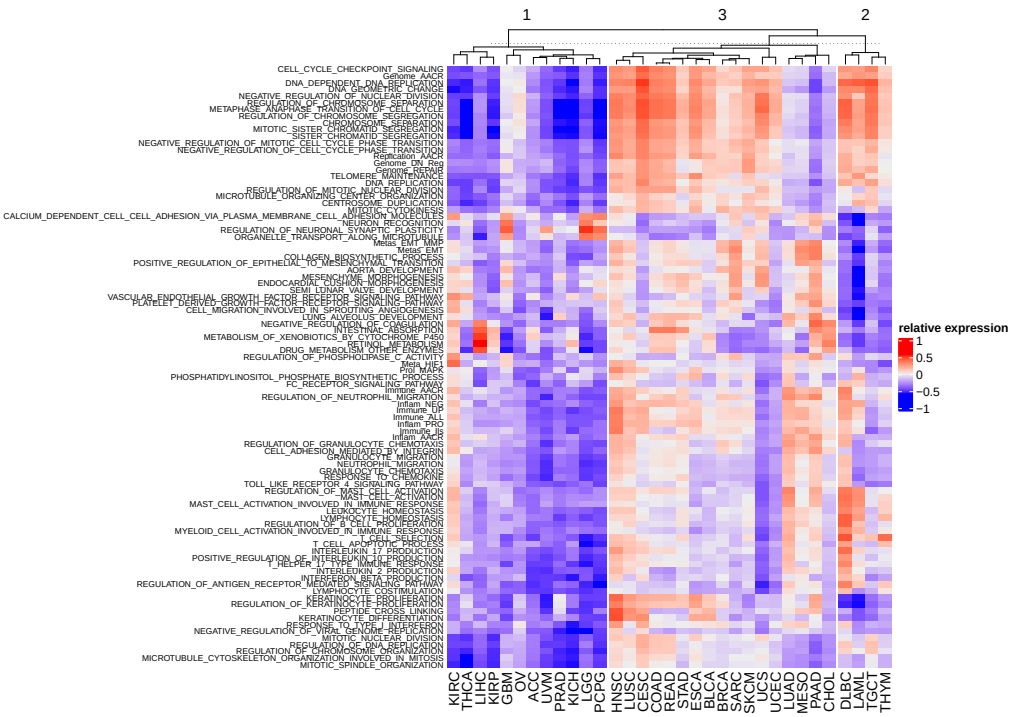

Figure 3.3: Pathway enrichment heatmap. The clustering of cancer types was conducted using kmeans.

3.4.3 Geneset enrichment comparison between LUAD and other cancer types

The identification of marker pathways for LUAD and the comparison of geneset enrichment is a central part of this project. The resulting heatmap and volcano plot could then be used to gather information of the different molecular signatures of LUAD and of other cancer types and gave critical information for the following regression analysis. The volcano plot (**Fig. 3.4**) helps with analysing the exact pathways that differ in activity between LUAD patients and other cancer patients.

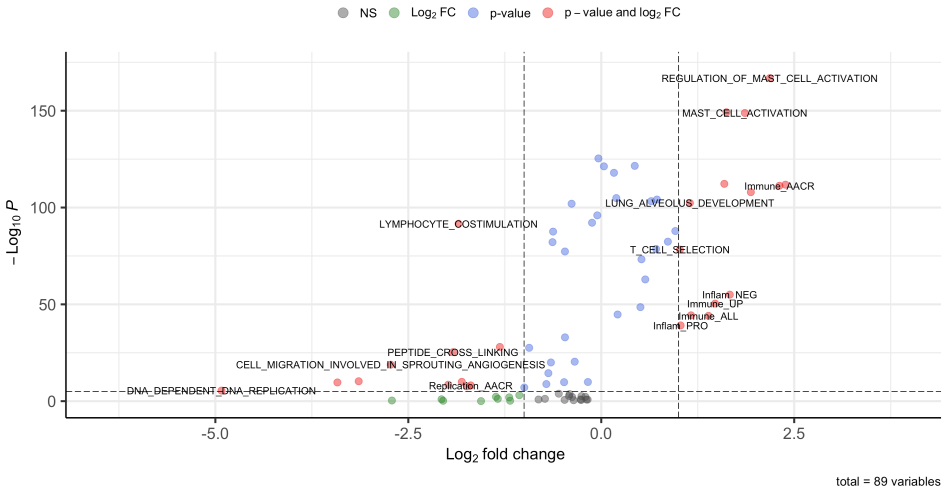

Figure 3.4: Volcano plot of geneset enrichment in LUAD compared to all other cancer types of the TCGA dataset. log2 foldchange values between LUAD and non-LUAD patients as well as the -log10 values of the corresponding p-values from the two sided Wilcoxon test are shown. Dotted lines indicate the alpha value and foldchange values of -1 and 1

The volcano plot shows several differentially expressed pathways in LUAD. The majority of them are upregulated. Most notably, a group of pathways related to inflammation and immune activation show a significant increase in activity in LUAD. Additionally, a pathway regulating mast cell activation is upregulated. Pathways concerning DNA replication and RNA translation seem to be downregulated in LUAD, as well as genesets concerning angiogenesis. Several genesets show a significant difference in expression, however the absolute value difference between the two groups does not meet our criteria of being at least 1.

3.4.4 Comparison of Clusters Within LUAD

After running separate PCA and UMAP analysis on patients for each tumor type the question arose how the patients within one tumor type differ from each other. The UMAP plots for three of the most clearly clustering tumor types can be seen in Figure (Appendix **Fig. 6.8**).
LUAD also clustered into two clusters, however these were not as clear as the ones shown before (Appendix, **Fig. 6.5**). To further our understanding of LUAD the gene activity of the clusters was compared using a volcano plot (Appendix, **Fig. 6.9**).

3.5 Regression

The logistic model was trained on the TCGA expression dataset. The model's goal was to predict whether a cancer patient suffers from LUAD or not. In order to be used reliably, the model has to be precise enough. Testing of our model revealed the following characteristics: The model predicts 136 LUAD patients correctly, as well as 2752 non-LUAD cases. Also shown in the confusion table (**Fig. 3.5** (A)) are the 27 false-negative occurrences and 7 false-positive occurrences. This results in an accuracy of 0.986.

For further evaluation a ROC plot was produced which enables an estimation of model performance in relation to the false-positive rate (**Fig. 3.5** (B)). The ideal estimator would have an area under the curve (AUC) of 1 and would fill out the top left corner. The trained regression model exhibits an AUC of 0.9159 and a nearly linear increase.

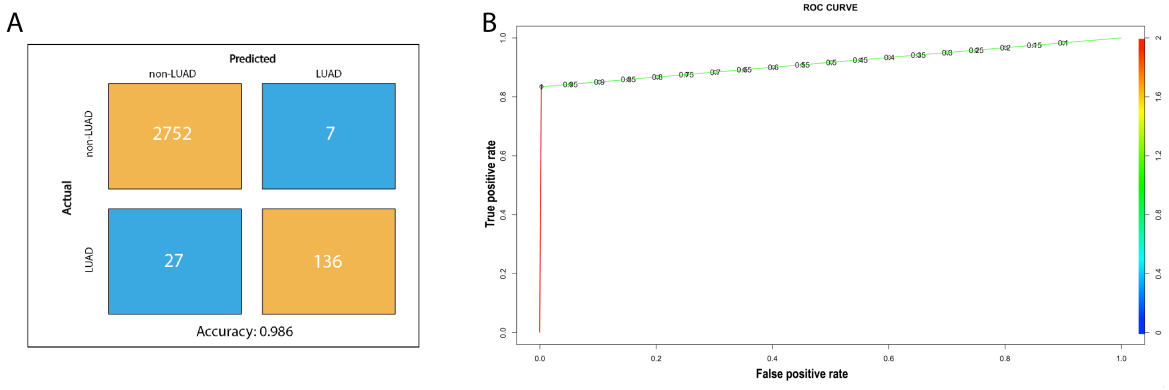

Figure 3.5: A) Confusion matrix and **B) ROC plot** of the trained logistic model with an AUC value of **0.9159**.

4 Discussion

4.1 Focused Analysis

The mean expression per pathway is shown in **Fig. 6.2**. Of all pathways, HIF1 mean expression was the most elevated which is plausible as it plays an important role in tumor progression and metastasis (Ren et al., 2013). Pathways linked to nucleotide, amino acid and sugar metabolism like ascorbate and aldarate metabolism are shown to be generally enriched in tumor tissue which is confirmed by previous research (Araujo et al., 2018). Gamma linolenic acid suppresses HIF1 alpha induced proliferation and invasion of non-small cell lung cancer cells (Wang et al., 2020). Therefore it is entirely plausible that the linoleic acid pathway has the lowest mean expression of all observed pathways. As indicated in our results (**Fig. 6.3**), pathways linked to DNA replication and cell cycle regulation are strongly overexpressed in tumor tissue compared to normal tissue. Immune response mechanisms like T-helper 17 type immune response were partly up and partly downregulated in tumor tissue. Th17 cells play a crucial role in promoting chronic tissue inflammation which has often been linked to the development of cancer (Chang et al., 2014). However Th17 function may vary accordingly to stage of disease (Wilke et al., 2011). This suggests that the expression rate of the T-helper 17 type immune response can vary between samples which would lead to inconclusive results during gene expression analysis. Analysis of **Fig. 6.4** yielded that the renal system pathway shows a particularly low expression rate. The renin-angiotensin system (RAS) plays an important role in suppression of lung cancer development, therefore our findings confirm our theoretical expectations (Xiong et al., 2021). Overall our findings for the most differentially expressed pathways could be backed by previous research.

4.2 Pan Cancer Analysis

4.2.1 Identification of Clusters in Gene Expression Data

Visualization of the data after dimension reduction reveals strong clustering based on the cancer types. This indicates that patients diagnosed with certain cancer types, particularly LUAD, BRCA, LIHC, KIRP, and UVM developed a unique gene expression pattern. Due to the fact that LGG and GBM cluster together, a similar pattern of genetic mutations can be concluded. The latter observation is unsurprising as both tumor types are glioma.

4.2.2 Pathway Enrichment

To minimize the information loss, we performed GSVA with two different geneset lists and chose the ontology geneset list which retained the most information. By visualizing the pathway enrichment relative to the cancer type the formation of three cancer type clusters can be observed (**Fig. 3.3**).This can be explained by the fact that several cancer types can be assigned to a joint cancer class like glioma and kidney carcinomata. No higher category could be assigned to the second cluster as it showed no distinct deregulation pattern. The third cluster contains cancer types related to cells of internal organs.The three pathway clusters that can be seen in the heatmap correspond to cell cycle and genome regulation, regulation of different phases of metastasis, and activation of the immune response. Cell cycle deregulation are cancer hallmarks and therefore at the base for tumor progression, as they can promote unregulated growth (Bruce, 1983). It is important to highlight that our analysis is relative to different cancer types and does not have any informative value concerning the relation to healthy cells. Accordingly, cancer types that exhibit downregulation of cell cycle pathways only show aggressive spread than other cancer types. The first cluster of cancer types, particularly kidney cancer and thyroid adenocarcinoma, result in a relative downregulation of these pathways. While our findings in THCA are according to our expectations (Coca-Pelaz et al., 2020), the strong downregulation

in kidney cancer is not supported by other studies. A cancer's ability to metastasize depends on its location and molecular signature (Bruce, 1983). Budczies et al. found melanoma, breast cancer and lung cancer to feature a high metastatic potential, while cancer cells deriving from the liver and sexual organs show the lowest rates of metastasis (Budczies et al., 2015). Our findings confirm this research. PAAD resulted in the most severe upregulation of metastasis related pathways which is also supported by studies (Ayres Pereira and Chio, 2019). Our analysis further revealed downregulation of metastatic pathways in liver and reproductive cancers that is supported by literature (Budczies et al., 2015). Cancer types of the first cluster show a severe downregulation of immune activation pathways. This clustering is according to our expectations as cancer types of each cluster can be categorized by immune-infiltration (Wang et al., 2020). Even though UCS is assigned to cluster two, it results in a strongly decreased immune pathway activity. This is confirmed by first studies (Ali et al., 2020). In contrast, DLBC and HNSC are examples for cancer types in which immune infiltration is severly upregulated. In this case inflammation can support the tumor mircoenvironment because it promotes tumorigenesis by supplying growth and survival factors. Tamma et al. observed the same phenomenon in DLBC (Tamma et al., 2020) and increased inflammatory pathways of HNSC were verified by the studies of He et al. (He et al., 2022). Overall, our results show great compliance with previous studies. Nevertheless, some findings do not conform to expectations as they imply a different transformation of some cancer types. Possible reasons for these discrepancies are the chosen pathways and the general loss of information during conduction of GSVA.

4.2.3 Geneset Enrichment Comparison Between LUAD and Other Cancer Types

Using the volcano plot which compared geneset enrichment of LUAD and non-LUAD patients several conclusions can be drawn.

Due to the overexpression of inflammatory and immune activity pathways, it can be deduced that LUAD is generally more immunogenic than the other cancer types. This explains the increase in tissue inflammation and T-cell selection as well as mast cell activation (**Fig. 3.3**). The upregulation of mast cell activity further supports the hypothesis that LUAD is more immunogenic as mast cells play a vital role in inflammatory and constrictory processes by secretion of cytokines (Tataroğlu et al., 2004). These findings are supported by Xu *et al.*, who claim that especially in the immunity high LUAD subtype a higher expression in immune system pathways and pro-inflammatory genes can be found. This also correlates with better response to immunotherapy (Xu et al., 2020).

Furthermore, the increased expression of alveolal developmental genes fits our expectation, as LUAD is a non-small cell lung cancer and thus growth of alveoli should be overexpression. Sainz de Aja *et al.* even suspect the affected alveolar progenitor cells to be the source of the tumor growth (Sainz de Aja et al., 2021).

The downregulation of genesets involved in replication compared to other genesets leads to the conclusion that LUAD does not exhibit the same increase in replication as other cancer types do. Furthermore, angiogenesis seems to be less advanced in LUAD as in other cancer types. Tataroğlu *et al.* suggest that the level of angiogenesis expression in LUAD patients is connected to the cancer stage the patients find themselves in. As our dataset provided patients over all stages the expression level of angiogenesis could have been skewed by patients in low angiogenesis stages (Tataroğlu et al., 2004).

In conclusion LUAD could be described as a rather immunogenic and pro-inflammatory cancer, protruding from alveolar progenitor cells. Immunotherapy is a promising therapy approach for LUAD patients, especially for the immunity high subtype (Xu et al., 2020).

4.2.4 Comparison of Clusters Within LUAD

The UMAP plots of SARC, ESCA and BRCA show perfectly clear clusters, which were also confirmed by k-means clustering. LUAD did not cluster as clearly, however further analysis of the differences between its patients was possible by using volcano plots. The volcano plot clearly shows that the two clusters differ in expression of certain genes. Most of the genes that are differentially expressed are connected to signal transmission over various pathways. For example

ADGRF1 which influences the way GPCRs behave in the two LUAD clusters and thus even influences CREB activity, which can promote anti-tumor cell programs (Abdulkareem et al., 2021). INSL4 is normally found during embryonic development as it can bind the insulin-like growth factor receptor. In LUAD it is significantly overexpressed which is unsurprising insofar that cancer progression results in reactivation of early development genes (Veitia et al., 1998). Additionally, FGA, the fibrinogen alpha chain, differs, meaning a difference in blood clotting (Freissmuth et al., 2016). It was shown that the two found clusters differ in very specific aspects of biological processes, namely CREB activation, growth by INSL4, blood clotting capabilities and calcium household by CALCA expression. We expected to find distinct clusters corresponding to the LUAD subtypes found by Qin *et al.* which are characterized by immune activity (Qin et al., 2020). However even in the most differentially expressed genes we found no significant difference in immune activity between the clusters. Qin *et al.* had access to both genomic and transcriptomic data and analysed the datasets specifically for changes in immune response which influences the results.

4.3 Regression

Since LUAD patients clustered clearly throughout the UMAP plot before, we expected to be able to built a rather robust logistic regression to differentiate between LUAD and non-LUAD patients. This expectation was further fueled by the LUAD patients also clustering during quality control using only the genes we chose as our explaining variables (**Fig. 6.10**). The confusion table shows a low amount false-positives and a high number of true-negatives. The model seems to be able to recognize clear non-LUAD patients fairly easily. However there are 27 false-negatives, which means that 16.6 % of all LUAD patients have not been labelled right. The reason for those false-negatives could be the fact that while LUAD patients show a clear cluster there are some non-LUAD patients in the same cluster. The patients that are close to these other cancer types are at risk of wrongfully being labelled as the neighbouring cancer type, as it is shown in the quality control plot (**Fig. 6.11**). Nevertheless due to the high amount of patients (2921 total patients in the testing dataset) the accuracy of 98.6 % shows a rather reliable model. The models performance is further underlined by the ROC curve that was created during analysis. In this case the ROC curve shows a steep progression at first and then inclines linearly. The area under curve of 0.9159 ranks this model as reliable, as generally AUC closer to 1 are regarded as good (Narkhede, 2018). In conclusion this model could be used to reduce the amount of genes that have to be screened by RNA-seq in order to diagnose a patient with LUAD. However there still is potential to differentiate between more cancer types by using a multinomial logistic regression. Additionally neural-networks have been shown to be a more reliable and functional alternative to logistic regression. Way *et al.* even showed this possible solution on the same TCGA dataset (Way et al., 2018).

4.4 Conclusion

Over the course of the project we were able to discern the genes that are differentially expressed in LUAD cells compared to normal cells and to assign them to their respective pathways. Furthermore, we managed to distinguish LUAD to other cancer types through analysis of their respective gene expression patterns. The vast majority of our findings could be confirmed by preexisting research. The crowning achievement of the project is a logistic regression model that allows us to predict LUAD with 98.6 % accuracy and to reduce the amount of genes that would have to be screened to diagnose a patient with LUAD.

4.5 Outlook

Performance of UMAP and PCA on our data (**Fig. 6.5**) showed that LUAD forms two distinct clusters. Further analysis may reveal the genetic differences within LUAD that lead to clustering. Different LUAD patient clusters might belong to different immune subtypes and thus respond differently to immunotherapy. A link between immune

activity in LUAD and metastasis formation could be further researched, as roughly half of LUAD patients develop brain metastases (Shih et al., 2020). It is a well-known fact that lung cancer is a smoker's disease therefore further research could be to investigate differences in smokers and non-smokers that suffering LUAD. However, in recent years studies have found that lung cancer incidence is decreasing in smokers and increasing in non-smokers. The same study states that the genomic profile of lung cancer in non-smokers differs from that in smokers (Qiu et al., 2015). Inspired by this study, a possible next step would be to subgroup the data into smokers and non-smokers and to compare the two groups in order to determine which pathways are differentially expressed and if the results of Qiu *et al.* can be replicated with our data. For an easier diagnosis during biopsy of early stage cancer cells an improved logistic regression model could be developed. The regression model we developed could be further sophisticated to additionally predict the cancer stage on the grounds of a patient's genomic profile. With enough training, this model could help determine an adequate treatment plan with reduced patient trauma. In order to really enhance analysis and modelling, not only transcriptomic data but also epigenetic data could be used to gain more insights into the cancer's behavior, as strategic DNA methylation or demethylation is a crucial driver in many cancer types (Esteller, 2008). In the end one of the most important steps following our bioinformatic analysis would be experimental validation of our findings. In particular the differences in tissue and cells could be studied using and comparing immunostained tissue slides from the TCGA project.

5 References

Abdi, H., and Williams, L.J. (2010). Principal component analysis. WIREs Computational Statistics *2*, 433–459.

Abdulkareem, N.M., Bhat, R., Qin, L., Vasaikar, S., Gopinathan, A., Mitchell, T., Shea, M.J., Nanda, S., Thangavel, H., Zhang, B., et al. (2021). A novel role of ADGRF1 (GPR110) in promoting cellular quiescence and chemoresistance in human epidermal growth factor receptor 2-positive breast cancer. The FASEB Journal *35*, e21719.

Ali, A.M.R., Tsai, J.-W., Leung, C.H., Lin, H., Ravi, V., Conley, A.P., Lazar, A.J., Wang, W.-L., and Nathenson, M.J. (2020). The immune microenvironment of uterine adenosarcomas. Clinical Sarcoma Research *10*, 1–8.

Araujo, J.M., Flores, C.J., Tirado-Hurtado, I., Rolfo, C.D., Raez, L.E., Prado, A., Saravia, C.H., and Pinto, J.A. (2018). RNA-seq data analysis to identify enriched metabolic pathways and a prognostic signature in squamous cell lung cancer. (American Society of Clinical Oncology).

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics.

Ayres Pereira, M., and Chio, I.I.C. (2019). Metastasis in pancreatic ductal adenocarcinoma: Current standing and methodologies. Genes *11*, 6.

Blighe, K., Rana, S., and Lewis, M. (2021). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.

Bruce, A. (1983). Molecular biology of the cell (Garland publishing).

Budczies, J., Winterfeld, M. von, Klauschen, F., Bockmayr, M., Lennerz, J.K., Denkert, C., Wolf, T., Warth, A., Dietel, M., Anagnostopoulos, I., et al. (2015). The landscape of metastatic progression patterns across major human cancers. Oncotarget *6*, 570.

Chang, S.H., Mirabolfathinejad, S.G., Katta, H., Cumpian, A.M., Gong, L., Caetano, M.S., Moghaddam, S.J., and Dong, C. (2014). T helper 17 cells play a critical pathogenic role in lung cancer. Proceedings of the National Academy of Sciences *111*, 5664–5669.

Coca-Pelaz, A., Shah, J.P., Hernandez-Prera, J.C., Ghossein, R.A., Rodrigo, J.P., Hartl, D.M., Olsen, K.D., Shaha, A.R., Zafereo, M., Suarez, C., et al. (2020). Papillary thyroid cancer—aggressive variants and impact on management: A narrative review. Advances in Therapy *37*, 3112–3128.

Dolgalev, I. (2022). Msigdbr: MSigDB gene sets for multiple organisms in a tidy data format.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomaRt. Nature Protocols *4*, 1184–1191.

Esteller, M. (2008). Epigenetics in cancer. New England Journal of Medicine *358*, 1148–1159.

Freissmuth, M., Offermanns, S., and Böhm, S. (2016). Pharmakologie und toxikologie: Von den molekularen grundlagen zur pharmakotherapie (Berlin ; Heidelberg: Springer).

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. Cell *144*, 646–674.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013b). GSVA: Gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics *14*, 7.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013a). GSVA: Gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 1–15.

He, L., Ren, D., Lv, G., Mao, B., Wu, L., Liu, X., Gong, L., and Liu, P. (2022). The characteristics and clinical relevance of tumor fusion burden in head and neck squamous cell carcinoma. Cancer Medicine.

Hebbali, A. (2020). Blorr: Tools for developing binary logistic regression models.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. Bull Soc Vaudoise Sci Nat *37*, 241–272.

Karakaslar, O., and Ucar, D. (2022). cinaR: A computational pipeline for bulk 'ATAC-seq' profiles.

Kassambara, A. (2020). Ggpubr: 'ggplot2' based publication ready plots.

Kolde, R. (2019). Pheatmap: Pretty heatmaps.
Korotkevich, G., Sukhov, V., and Sergushichev, A. (2019). Fast gene set enrichment analysis. bioRxiv.
Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. Journal of Statistical Software *25*, 1–18.
Li, X., and Lu, Z. (2022). Role of von willebrand factor in the angiogenesis of lung adenocarcinoma. Oncology Letters *23*, 1–7.
Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2021). Cluster: Cluster analysis basics and extensions.
Marguerat, S., and Bähler, J. (2010). RNA-seq: From technology to biology. Cellular and Molecular Life Sciences *67*, 569–579.
McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). UMAP: Uniform manifold approximation and projection. The Journal of Open Source Software *3*, 861.
McKenzie, A. (2016). Bayesbio: Miscellaneous functions for bioinformatics and bayesian statistics.
Melville, J. (2021). Uwot: The uniform manifold approximation and projection (UMAP) method for dimensionality reduction.
Milošević, D., Medeiros, A.S., Stojković Piperac, M., Cvijanović, D., Soininen, J., Milosavljević, A., and Predić, B. (2022). The application of uniform manifold approximation and projection (UMAP) for unconstrained ordination and classification of biological indicators in aquatic ecology. Science of The Total Environment *815*, 152365.
Morgan, M., Wang, J., Obenchain, V., Lang, M., Thompson, R., and Turaga, N. (2021). BiocParallel: Bioconductor facilities for parallel evaluation.
Narkhede, S. (2018). Understanding auc-roc curve. Towards Data Science *26*, 220–227.
Neuwirth, E. (2022). RColorBrewer: ColorBrewer palettes.
Peng, X., Chen, Z., Farshidfar, F., Xu, X., Lorenzi, P.L., Wang, Y., Cheng, F., Tan, L., Mojumdar, K., Du, D., et al. (2018). Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers. Cell Reports *23*, 255–269.
Qin, F., Xu, Z., Yuan, L., Chen, W., Wei, J., Sun, Y., and Li, S. (2020). Novel immune subtypes of lung adenocarcinoma identified through bioinformatic analysis. FEBS Open Bio *10*, 1921–1933.
Qiu, M., Xu, Y., Wang, J., Zhang, E., Sun, M., Zheng, Y., Li, M., Xia, W., Feng, D., Yin, R., et al. (2015). A novel lncRNA, LUADT1, promotes lung adenocarcinoma proliferation via the epigenetic suppression of p27. Cell Death & Disease *6*, e1858–e1858.
R Core Team (2022a). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
R Core Team (2022b). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
R Core Team (2022c). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
Ren, W., Mi, D., Yang, K., Cao, N., Tian, J., Li, Z., and Ma, B. (2013). The expression of hypoxia-inducible factor-1α and its clinical significance in lung cancer: A systematic review and meta-analysis. Swiss Medical Weekly.
Ringnér, M. (2008). What is principal component analysis? Nature Biotechnology *26*, 303–304.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research *43*, e47.
Sainz de Aja, J., Dost, A., and Kim, C. (2021). Alveolar progenitor cells and the origin of lung cancer. Journal of Internal Medicine *289*, 629–635.
Satija, R., Butler, A., Hoffman, P., and Stuart, T. (2022). SeuratObject: Data structures for single cell data.
Shih, D.J., Nayyar, N., Bihun, I., Dagogo-Jack, I., Gill, C.M., Aquilanti, E., Bertalan, M., Kaplan, A., D'Andrea, M.R., Chukwueke, U., et al. (2020). Genomic characterization of human brain metastases identifies drivers of metastatic lung adenocarcinoma. Nature Genetics *52*, 371–377.
Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: Visualizing classifier performance in r. Bioinformatics *21*, 7881.

Slowikowski, K. (2021). Ggrepel: Automatically position non-overlapping text labels with 'ggplot2'.
Smets, T., Verbeeck, N., Claesen, M., Asperger, A., Griffioen, G., Tousseyn, T., Waelput, W., Waelkens, E., and De Moor, B. (2019). Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data. Analytical Chemistry *91*, 5706–5714.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences *102*, 15545–15550.
Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J.P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics *23*, 3251–3253.
Tamma, R., Ranieri, G., Ingravallo, G., Annese, T., Oranger, A., Gaudio, F., Musto, P., Specchia, G., and Ribatti, D. (2020). Inflammatory cells in diffuse large b cell lymphoma. Journal of Clinical Medicine *9*, 2418.
Tataroğlu, C., Kargı, A., Özkal, S., Eşrefoğlu, N., and Akkoçlu, A. (2004). Association of macrophages, mast cells and eosinophil leukocytes with angiogenesis and tumor stage in non-small cell lung carcinomas (NSCLC). Lung Cancer *43*, 47–54.
Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia *2015*, 68–77.
Trost, N. (2022). Babyplots: Easy, fast, interactive 3D visualizations for data exploration and presentation.
Veitia, R., Laurent, A., Quintana-Murci, L., Ottolenghi, C., Fellous, M., Vidaud, M., and McElreavey, K. (1998). The INSL4 gene maps close to WI-5527 at 9p24. 1→ p23. 3 clustered with two relaxin genes and outside the critical region for the monosomy 9p syndrome. Cytogenetic and Genome Research *81*, 275–277.
Vermeulen, M., Smith, K., Eremin, K., Rayner, G., and Walton, M. (2021). Application of uniform manifold approximation and projection (UMAP) in spectral imaging of artworks. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy *252*, 119547.
Wang, Y., Shi, J., and Gong, L. (2020). Gamma linolenic acid suppresses hypoxia-induced proliferation and invasion of non-small cell lung cancer cells by inhibition of HIF1α. Genes & Genomics *42*, 927–935.
Warnes, G.R., Bolker, B., and Lumley, T. (2022c). Gtools: Various r programming tools.
Warnes, G.R., Bolker, B., and Lumley, T. (2022a). Gtools: Various r programming tools.
Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., et al. (2022b). Gplots: Various r programming tools for plotting data.
Way, G.P., Sanchez-Vega, F., La, K., Armenia, J., Chatila, W.K., Luna, A., Sander, C., Cherniack, A.D., Mina, M., Ciriello, G., et al. (2018). Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas. Cell Reports *23*, 172–180.
Wickham, H. (2016). ggplot2: Elegant graphics for data analysis (Springer-Verlag New York).
Wickham, H., and Seidel, D. (2022). Scales: Scale functions for visualization.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. Journal of Open Source Software *4*, 1686.
Wilke, C.M., Kryczek, I., Wei, S., Zhao, E., Wu, K., Wang, G., and Zou, W. (2011). Th17 cells in cancer: Help or hindrance? Carcinogenesis *32*, 643–649.
Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In Implementing Reproducible Computational Research, V. Stodden, F. Leisch, and R.D. Peng, eds. (Chapman; Hall/CRC),.
Xiong, L., Wei, Y., Zhou, X., Dai, P., Cai, Y., Zhou, X., Xu, M., Zhao, J., and Tang, H. (2021). AGTR1 inhibits the progression of lung adenocarcinoma. Cancer Management and Research *13*, 8535.
Xu, F., Chen, J., Yang, X., Hong, X., Li, Z., Lin, L., and Chen, Y. (2020). Analysis of lung adenocarcinoma subtypes based on immune signatures identifies clinical implications for cancer therapy. Molecular Therapy-Oncolytics *17*, 241–249.
Yu, G. (2022). Enrichplot: Visualization of functional enrichment result.
Yuen In, H.L., and Pincket, R. (2022). Transcripts per million ratio: A novel batch and sample control method over an established paradigm. arXiv e-Prints arXiv–2205.

Zhang, Y., Tseng, J.T.-C., Lien, I.-C., Li, F., Wu, W., and Li, H. (2020). mRNAsi index: Machine learning in mining lung adenocarcinoma stem cell biomarkers. Genes *11*, 257.

6 Appendix

All statistical analyses were done in an R-environment (R Core Team, 2022a).

Table 6.1: Used packages in alphabetical order.

Package Name	Application	**Reference**
babyplots	create interactive 3D visualizations	Trost (2022)
bayesbio	calculate Jaccard coefficients	McKenzie (2016)
BiocParallel	novel implementations of functions for parallel evaluation	(Morgan et al., 2021)
biomaRt	access to genome databases	Durinck et al. (2009)
blorr	building and validating binary logistic regression models	Hebbali (2020)
cinaR	combination of different packages	Karakaslar and Ucar (2022)
cluster	cluster analysis of data	Maechler et al. (2021)
Complex Heatmap	arrange multiple heatmaps	(Gu et al., 2016)
Enhanced Volcano	produce improved volcano plots	(Blighe et al., 2021)
enrichplot	visualization of geneset enrichment results (GSEA)	Yu (2022)
FactoMineR	perform principal component analysis (PCA)	Lê et al. (2008)
fgsea	Run GSEA on a pre-ranked list	Korotkevich et al. (2019)
ggplot2	visualization of results in dot plots, bar plots and box plots	Wickham (2016)
ggpubr	formatting of ggplot2-based graphs	Kassambara (2020)
ggrepel	creates non-overlapping text labels for ggplot2-based graphs	Slowikowski (2021)
grid	implements the primitive graphical functions that underlie the ggplot2 plotting system	R Core Team (2022b)
gridExtra	arrange multiple plots on a page	Auguie (2017)
GSVA	Run GSVA on a dataset	(Hänzelmann et al., 2013b)
gplots	plotting data	(Warnes et al., 2022b)
gtools	calculate foldchange, find NAs, logratio2foldchange	Warnes et al. (2022c)
knitr	creation of citations using write_bib	Xie (2014)
limma	“linear models for microarray data”	Ritchie et al. (2015)
msigdbr	provides the ‘Molecular Signatures Database’ (MSigDB) genesets	Dolgalev (2022)
parallel	allows for parallel computation through multi core processing	R Core Team (2022c)
pheatmap	draw clustered heatmaps	Kolde (2019)
RColorBrewer	provides color schemes for maps	Neuwirth (2022)

Package Name	Application	**Reference**
ROCR	visualizing classifier performance	Sing et al. (2005)
scales	helps in visualization: r automatically determines breaks and labels for axes and legends	Wickham and Seidel (2022)
Seurat	includes RunPCA function	Satija et al. (2022)
tidyverse	collection of R packages, including ggplot2	Wickham et al. (2019)
uwot	performs dimensionality reduction and Uniform Manifold Approximation and Projection (UMAP)	Melville (2021)

6.1 Additional Computational Tools

6.1.1 Jaccard index

The Jaccard index is a widely known measure for the similarity between finite sample sets and is defined as the intersection of the sample sets devided by their union. The restricted domain ranges from zero to one. A Jaccard index close to one indicates a high similarity of the sample sets (Jaccard, 1901).

6.1.2 Principal Component Analysis (PCA)

Principal component analysis (PCA) constitutes an additional tool for dimensionality reduction besides UMAP.
Principal component analysis (PCA) is a procedure used to perform linear dimension reduction. The goal is to reduce the dimension of a given dataset whilst losing as little information as possible by retaining a maximum of the standardized dataset's variation (Ringnér, 2008).
Principal components (PC) are a set of new orthogonal variables that are made up of a linear combination of the original variables. Principal components display the pattern of similarity of the observations and of the variables as points in maps (Abdi and Williams, 2010). By convention, the PCs are ordered in decreasing order according to the amount of variation they explain of the original data (Ringnér, 2008). It is important to note that all PCs are uncorrelated.
PCA is a useful tool for genome-wide expression studies and often serves as a first step before clustering or classification of the data. Dimension reduction is a necessary step for easy data exploration and visualization (Ringnér, 2008).

6.1.3 Gene Set Enrichment Analysis (GSEA)

Gene set enrichment analysis (GSEA) is a computational method that is used to determine whether two pathway expression states are significantly different from each other (Subramanian et al., 2007). Two datasets are compared and the genes are sorted from the most to the least differential expression between the datasets according to their p-values. This creates a ranked list (L). Referring to an *a priori* defined set of genesets (S), the goal is to locate for each pathway of (S) where its corresponding genes fall in (L) and find a discerning trend. To determine the distribution of the genes from pathway (S) in (L), an enrichment score is calculated for each pathway. For this, a running-sum statistic is calculated as the list (L) is ran through. The running-sum is increased every time a gene belonging to the pathway in question is encountered and decreased otherwise. The enrichment score is defined as the maximum deviation from zero of the running-sum. Lastly, adjustment for multiple hypothesis testing is performed by normalizing the enrichment

score for each pathway to account for its size and a normalized enrichment score is obtained.(Subramanian et al., 2005)

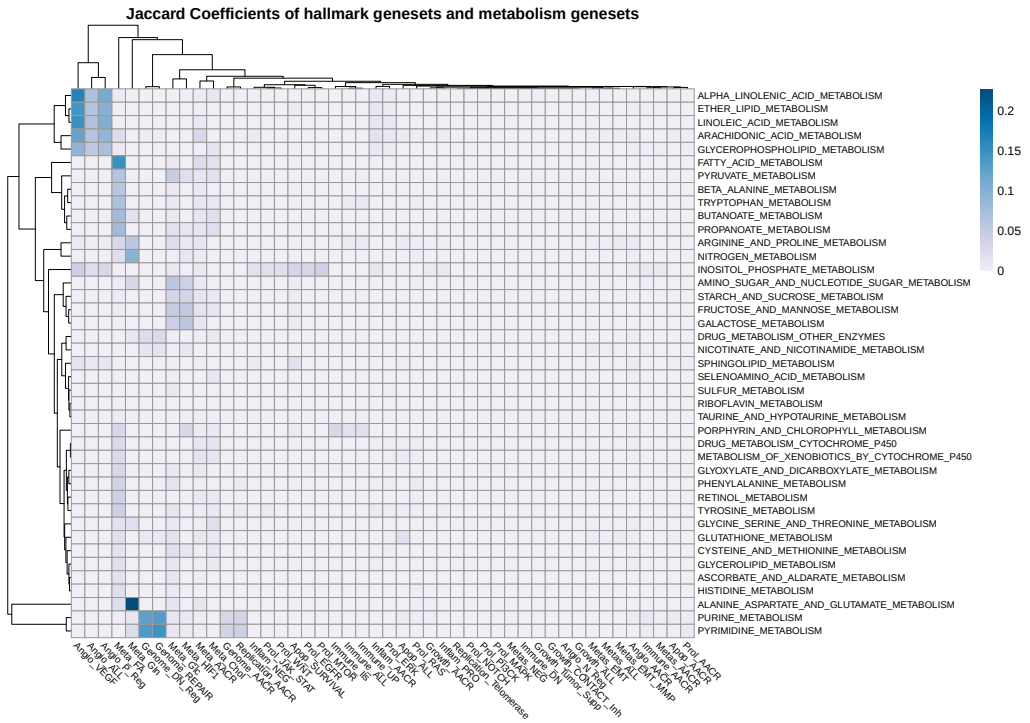

Figure 6.1: Jaccard Coefficients of hallmark genesets and metabolism genesets. The x-axis is defined by the given hallmark genesets, whereas the y-axis is assigned to the selected metabolism geneset.

Figure 6.2: Barplot for the mean regultion of hallmark gene sets and metabolism gene sets. Pathways are sorted by their mean expression. The x-achsis shows the mean expression and the y-achsis shows the pathways. Each pathway is colored according to regulation state.

Figure 6.3: volcano plot comparing pathway expression in normal and tumor tissue The -log10 of the p values are plotted against the log2(foldchange) of each pathway. The regulation is colored accordingly)).

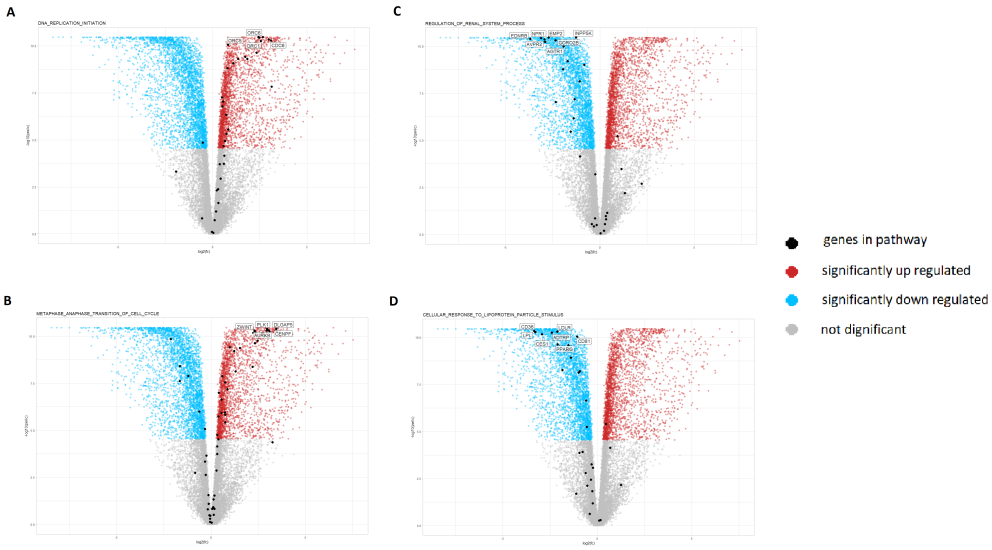

Figure 6.4: Volcano plot showing gene expression for selected pathways The plots show the number of genes in each pathway and the percentage of significantly over expressed (red), significantly under expressed (blue) and not significant differentially expressed genes (gray). The differential expression referes to the change of mean expression over all patients for each gene form normal to tumor tissue. The selected pathways were: DNA replication initiation (top left), metaphase anaphase transition of cell cycle (bottom left), regulation of renal system process (top right) and cellular response to lipoprotein particles (bottom right) The x-axis is defined by the given hallmark genesets, whereas the y-axis is assigned to the selected metabolism geneset.

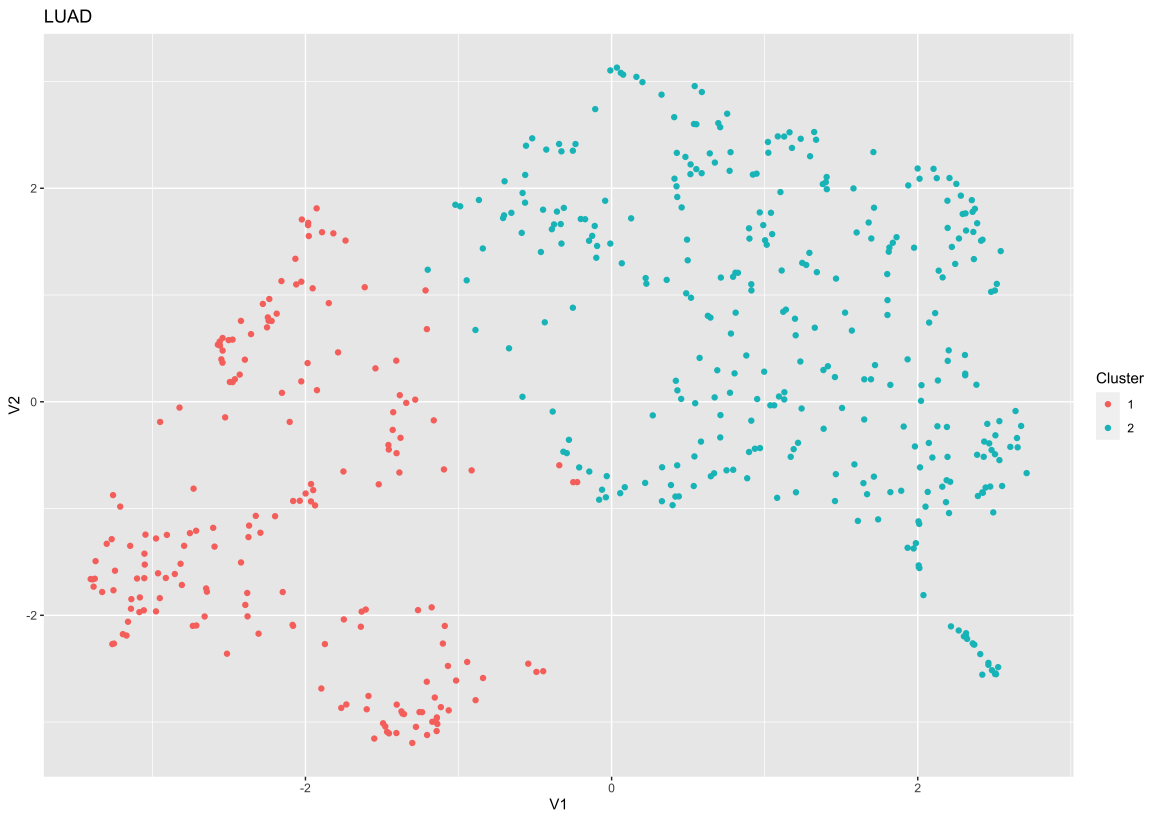

Figure 6.5: PCA UMAP plot of LUAD

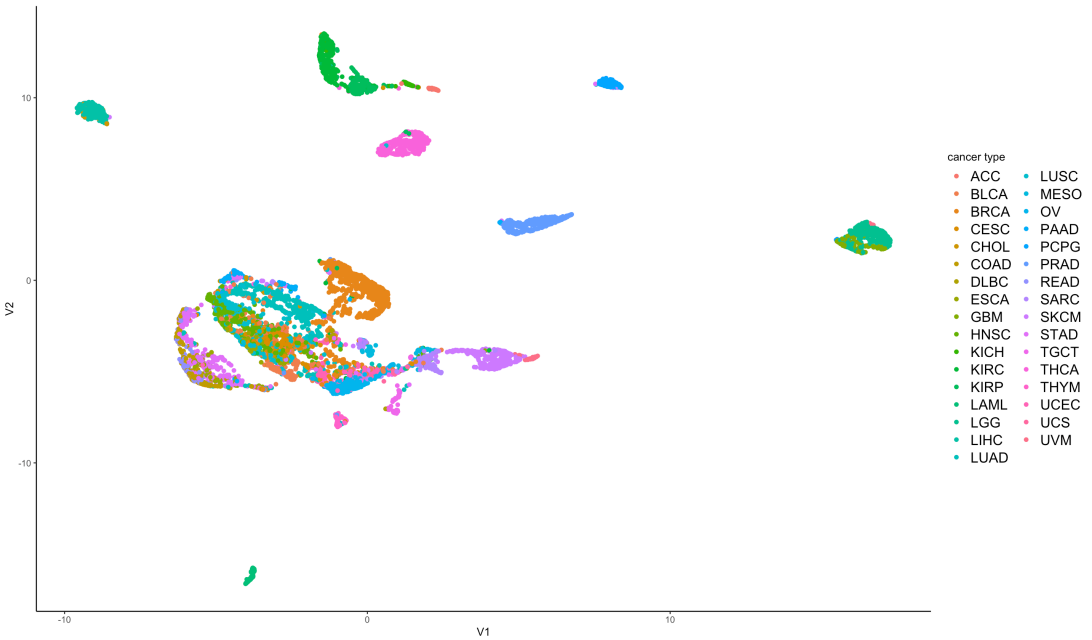

Figure 6.6: Pathway enrichment based on cancer type. The x-axis is defined by the first umap component, whereas the y-axis assigned to the second component. The data plots are colored by the patients cancer type

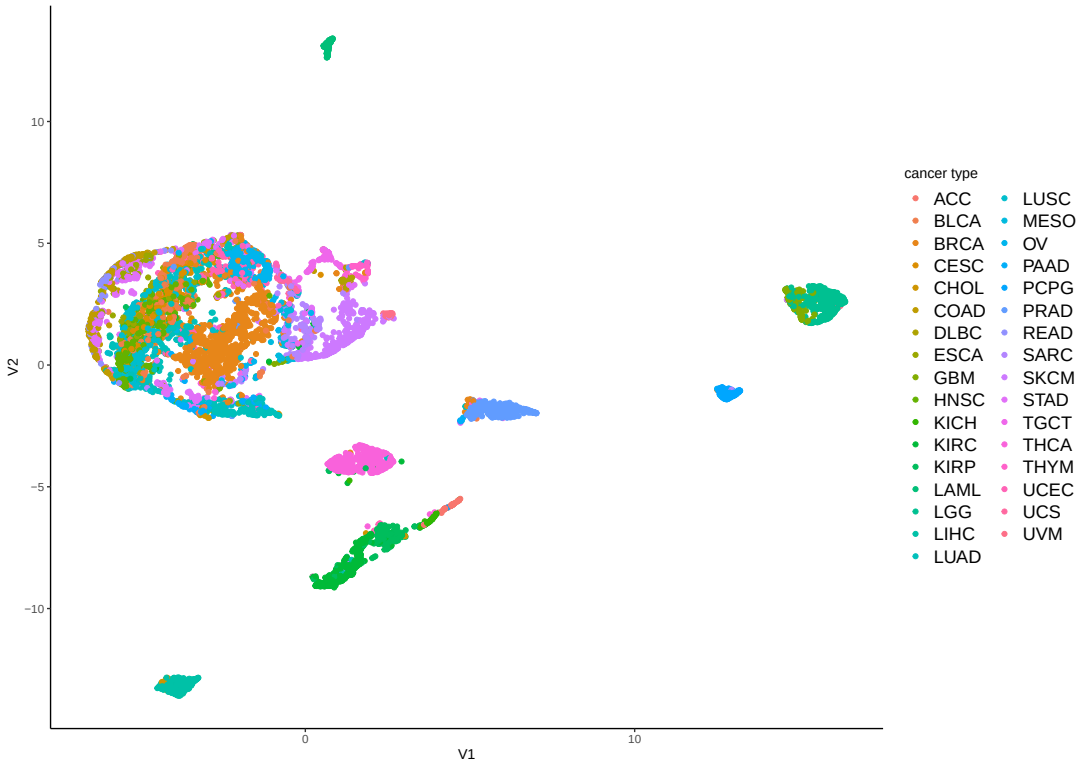

Figure 6.7: Clustering after GSVA performed with curated genesets. Colored by assigned value by the trained model to asses quality of model

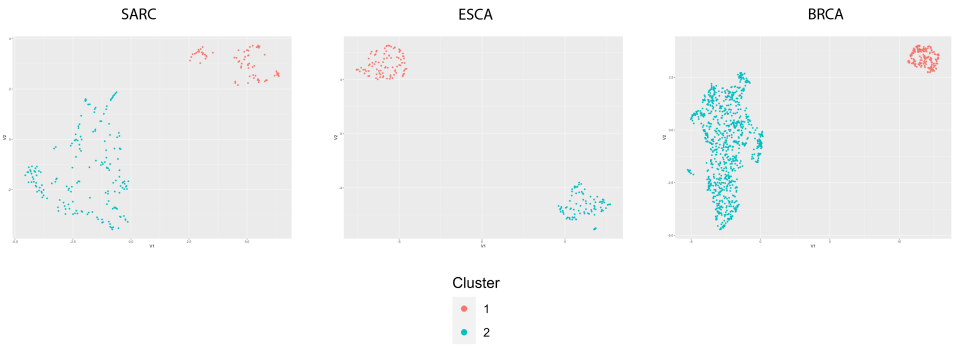

Figure 6.8: UMAP plots for SARC, ESCA and BRCA run on the corresponding subset of the gene expression dataset, colored by the cluster assigned to each datapoint by k-means clustering.

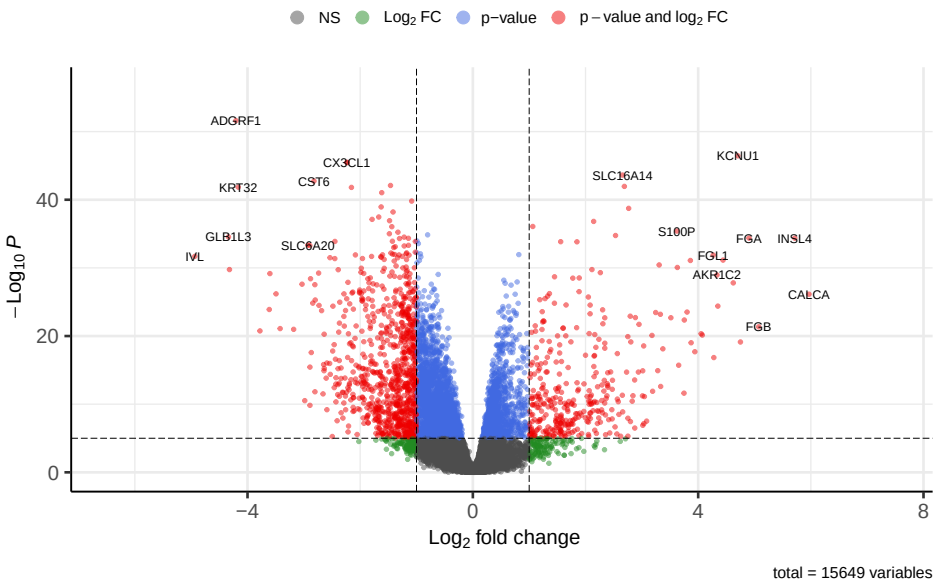

Figure 6.9: Volcano plot comparing the two observed clusters. Each datapoint is one gene, the most differentially expressed genes are marked by name.

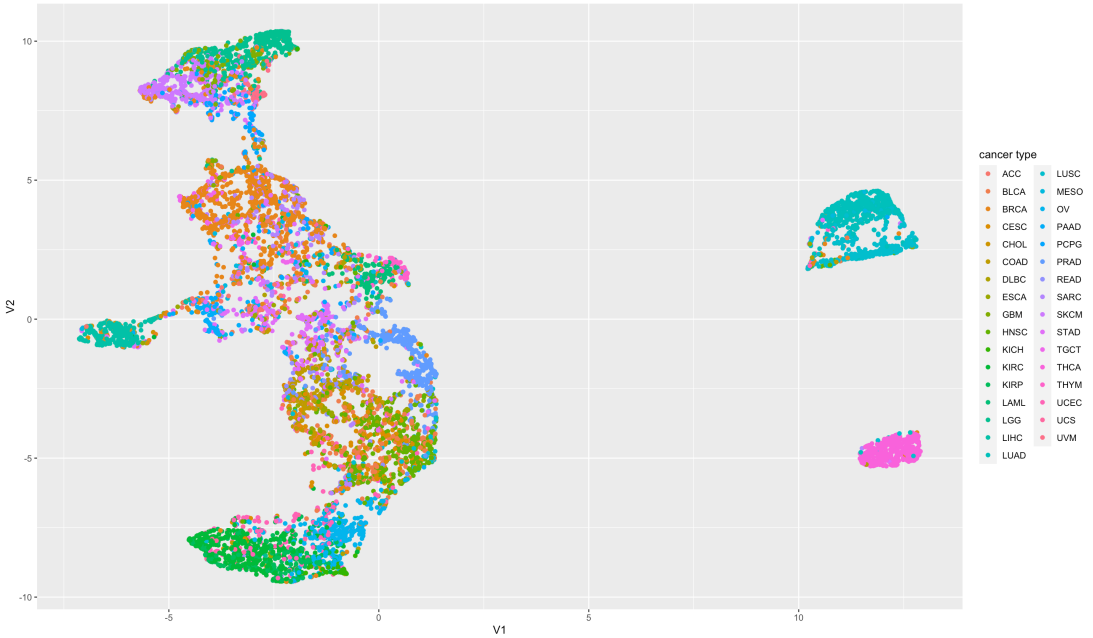

Figure 6.10: Quality control UMAP plot for regression. Shows that LUAD cluster when gene expression dataset is subset to only include explaining variables, colored by cancer type

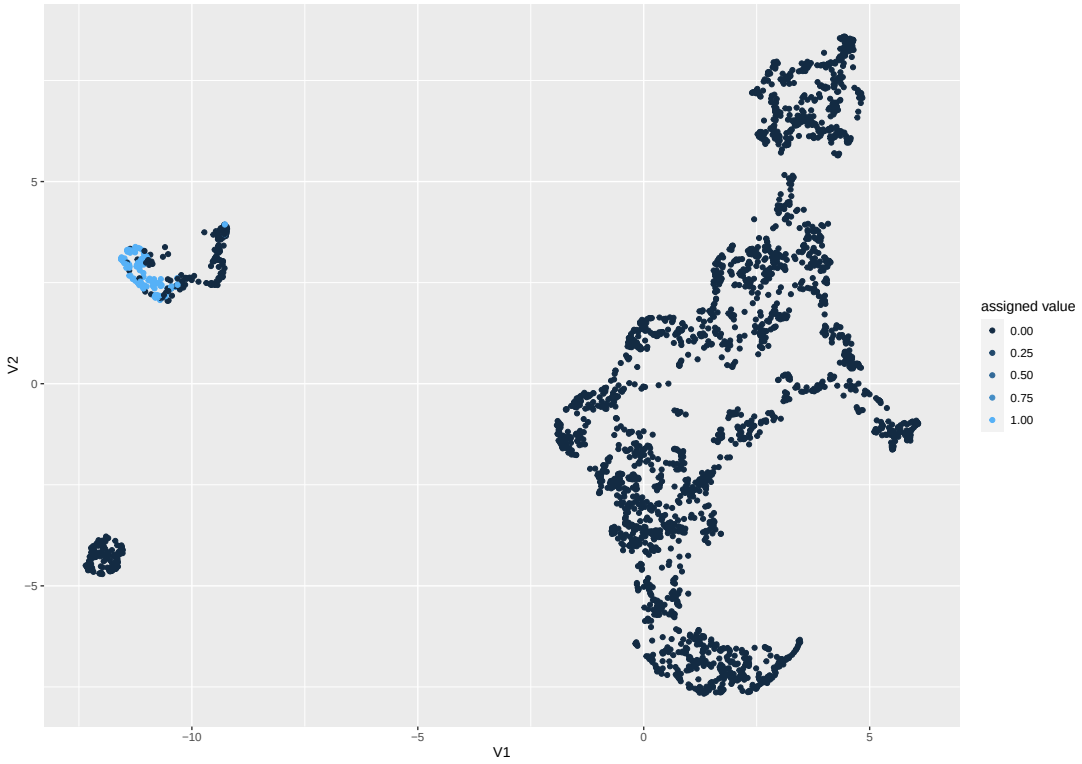

Figure 6.11: Quality control UMAP plot for regression. Colored by assigned value by the trained model to asses quality of model