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

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

Data Science Project SoSe 2022

Autoren Paul Brunner, Marie Kleinert, Felipe Stünkel, Chloé Weiler
Abgabetermin 18.07.2022

1 Introduction

1.1 Biological Background

To this day, lung cancer is the leading cause of cancer death worldwide

@*zhang*2020

. Lung adenocarcinoma (LUAD) is a form of non-small cell lung cancer which accounts for approximately 40% of lung cancer cases

@*wang*2020

and which is characterized by a remarkably low 5-year overall survival rate of merely 18%

@*li*2022

. In theory, every cell is capable of developing into a cancer cell through acquisition of so-called hallmark capabilities that essentially cause metabolic reprogramming, immune evasion and uncontrolled proliferation due to numerous genetic mutations

@*hanahan*2011; @*peng*2018

. In order to gain an insight into which mutations drive cancer development and how to best treat different cancer types, one must start by deciphering the intricate workings of gene expression regulation within a tumor cell. In our case this feat was achieved with the help of the pan cancer project.

1.2 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 2012 TCGA Research Network launched the Pan-Cancer analysis project as a globally coordinated initiative whose main objective is to assemble coherent, consistent TCGA datasets across twelve different tumor types, one of which being lung adenocarcinoma (LUAD). Each tumor type is characterized using six different genomic, proteomic, epigenomic, and transcriptional platforms. Data collected from thousands of patients is analysed and interpreted in an attempt to gain a deeper understanding of the genomic changes that drive a normal cell to become cancerous. In the future, the aim is to analyse additional tumor types beyond the twelve original ones in the hopes that the pan cancer project will one day inform clinical decision-making and aid in the development of novel therapeutic options (Weinstein et al., 2013).

1.3 Jaccard index

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

$$J(A, B) = \frac{A \cap B}{A \cup B}$$

1.4 Principal Component Analysis (PCA)

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).

1.5 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. This variable radius is smaller in high density regions of data points and larger in low density regions. In general, the density is higher when the k-nearest neighbor is close and vice versa. The number of k-nearest neighbors controls the number of neighbors whose local topology is preserved. Precisely, a large number of 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).

UMAP is a newer method than PCA and it is generally believed to be easier to interpret and to group data than by using PCA. Furthermore, UMAP has the advantage of not requiring linear data (Milošević et al., 2022).

1.6 Gene Set Enrichment Analysis (GSEA)

Gene set enrichment analysis (GSEA) is a computational method that is used to determine whether two gene expression states are significantly different from each other or not (Subramanian et al., 2007). In this project we compared gene expression profiles between healthy and tumorous tissue of LUAD.

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. If the genes of a given pathway are randomly distributed in L then the pathway is assumed to not significantly contribute to the particular tumor's phenotype. However if the genes are primarily clustered at the top or the bottom of L then a phenotypic significance of the given pathway can be assumed.

To determine the location of the genes, 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. An enrichment score is thus calculated for each pathway. 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.

GSEA is a useful tool for interpretation of gene expression data.

(Subramanian et al., 2005)

1.7 Gene Set Variation Analysis (GSVA)

Gene set variation analysis (GSVA) is an unsupervised method to estimate pathway activities based on gene expression data. Contrarily to the aforementioned GSEA, GSVA does not rely on phenotypic characterization of the datasets into two categories but rather quantifies enrichment in a sample-wise manner which makes GSVA the better choice to perform on the tcga_exp dataset.

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 ES.

The GSVA corresponds to either the maximum deviation between both running sums or the GSVA score can be defined as the difference of the maximum deviations in the positive and in the negative direction. A highly positive or negative GSVA score indicates that the studied geneset is positively or negatively enriched compared to the genes not in the geneset, respectively. If the GSVA score 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)

2 Abbreviations

AML	Acute myeloid leukemia
AUC	Area under the curve
BRCA	Breast cancer
CALCA	Calcitonin Related Polypeptide Alpha
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
EMT	Epithelial to mesenchymal transition
ESCA	Esophageal carcinoma
FGA	Fibrinogen alpha chain
GBM	Glioblastoma multiformae
GSEA	Gene Set Enrichment Analysis
GSVA	Gene Set Variation Analysis
HIF1a	
INSL4	Gene encoding insulin-like 4 protein
KICH	Kidney chromophobe
KIRC	Kidney renal clear-cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LGG	Low grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
MSigDB	Molecular Signature Database
ORC	
PAAD	Pancreatic adenocarcinoma
PC	Principal component
PCA	Principal Component Analysis
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic

SARC	Sarcoma
TCGA	The Cancer Genome Atlas
TPM	Transcripts per million
UMAP	Uniform Manifold Approximation and Projection
UVM	Uveal melanoma

3 Methods

3.1 Our Data

At the beginning of this project 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 is a data frame containing RNA-seq data from almost 10,000 TCGA cancer patients for 33 different tumor types. The data stored within that data frame 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. A focused analysis was performed on this second dataset.

RNA sequencing (RNA-seq) data 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).

All expression data was log2(TPM) transformed. Log2 Transformation is a commonly used tool 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. A possible perk of TPM is the reduction of type I and type II errors which would otherwise falsify downstream analysis results by accounting for gene length first (Yuen In and Pincket, 2022).

3.2 Overview of used packages

Table 3.1: Tab. 1: Used packages in alphabetical order.

Package Name	Application	**Reference**
babyplots	create interactive 3D visualizations	Trost (2022)
base	basic R functions	R Core Team (2022a)
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)
ComplexHeatmap	arrange multiple heatmaps	(Gu et al., 2016)
edgeR	assess differential expression in gene expression profiles	Chen et al. (2016)
EnhancedVolcano	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)

Package Name	Application	**Reference**
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., 2022a)
gtools	calculate foldchange, find NAs, logratio2foldchange	Warnes et al. (2022b)
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)
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	visualize geneset enrichment results in dot plots	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)

3.3 Gene Set Extraction

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 the extraction of species-specific genesets of the category of interest. Afterwards, the prefix corresponding to the source of the geneset was removed so as to achieve coherent and well-arranged pathway names. The resulting output was a list containing all selected genesets with the comprising genes saved in a vector and each element named after the pathway stored within. 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 known pathways that are often deregulated in cancer cells.

3.4 Data cleanup on TCGA expression dataset

In order to enable an efficient workflow on the big TCGA dataset containing cancer patient RNA-seq expression data, the total amount of genes had to be reduced. In a first cleanup step the amount and distribution of NAs was analysed. There were no NAs in the dataset itself, which means that no patients had to be removed. Next all genes that showed a very low standard deviation were removed from the dataset. As a cutoff value the value of the 50% quantile of the standard deviation distribution was used.

To decrease the amount of genes even further and focus the analysis on important genes a biotype analysis was conducted. Using the BiomaRt package (Durinck et al., 2005) the biotypes of all genes in the dataset were retrieved and compared to biotypes of the given geneset list, as well as the geneset lists retrieved from MSigDB by using the native msigdbr package (Dolgalev, 2022). All genes linked to biotypes that were not found in the biotypes of said genesets and possible pseudogenes were removed. However genes belonging to lncRNA or siRNA were kept as they also have a significance in various biological processes (2016).

In the end the expression dataset could be reduced from over 60.000 genes to roughly 17.000 while keeping all 9741 patients.

3.5 Data cleanup on TCGA tumor vs normal dataset

In preparation for the focused analysis, the RNA-seq data from the small TCGA dataset for LUAD was extracted and cleaned.

For LUAD, the raw gene expression data included roughly 20,000 genes for 58 patients in normal and tumor tissue.

The first step was to check whether NA values were present, which was not the case. Next, all zero variance genes were removed which is prerequisite for the Shapiro-Wilk test performed in the focused analysis. In addition, the biotypes that were not present in the genesets relevant for later analysis were filtered out. For this the biomaRt package (Durinck and Huber, 2022) was used and all pseudogenes were removed. The remaining biotypes corresponded to the genesets which were predominantly protein-coding genes and few lncRNAs.

The cleaned data consisting of roughly 17,000 genes was saved to a separate file.

3.6 Pancancer analysis

3.6.1 Dimensionality reduction

The first step to pan cancer comparison was to evaluate potential clusters in our data. To uncover these, a combination of a PCA and UMAP was conducted on the cleaned tcga expression dataset. First, the PCA using the RunPCA command from the Seurat package (Satija et al., 2022) and later on the UMAP analysis was done on the produced principle components using the uwot package (Melville, 2021). Performing the PCA before the UMAP analysis was necessary to minimize artefacts caused by correlating variables in the dataset. Principle components do not correlate by nature. The UMAP results per patient were then plotted with ggplot2 (Wickham, 2016) and colored according to their corresponding tumor type in order to gain insight into cluster formation.

The dataset was subsetted into dataframes containing only patients of one cancer type. The aforementioned workflow was then used on each of the cancer type subsets. All of the created plots were then collected into one overview of intra-cancer clusters. To analyse these, 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 the help of the function from the cluster package (Maechler et al., 2021). Based on these assignments, foldchanges between each cluster and the rest of the patients were calculated using the foldchange function (Warnes et al., 2022c). Additionally, a two sided Wilcoxon test was conducted on each cluster and the rest of the data points. These two metrics were used to create a volcano plot showing the -log10 of the Wilcoxon p-values plotted against log2 foldchanges of each gene between the clusters. These plots were created using the EnhancedVolcano package (Blighe et al., 2021). From these plots the most significant and over or under expressed genes were extracted.

3.6.2 GSVA

In order to compare cancer types with each other, the genes needed to be condensed into more easily understandable metrics. For this we evaluated the enrichment of genes that play a role in the same pathways over all cancer types by conducting a GSVA. Two different geneset lists were used: gene ontology genesets and curated genesets, both retrieved from MSigDB, as mentioned above. Genesets that showed little overlap with the dataset genes, meaning below 95% of the pathways genes could be found in the dataset, were removed. Going into the GSVA, both geneset lists consisted of roughly 3,000 genesets. To conduct the GSVA a package of the same name was used, that also enables multi-core calculations. This was crucial in order to cut down on calculation time (Hänzelmann et al., 2013b). Afterwards, pathways that showed a low standard deviation were removed from the resulting pathway enrichment matrix. The patient pathway enrichment matrix was turned into a tumor type pathway enrichment matrix by subsetting it into the different tumor types and calculating the means per pathway over all patients of one tumor type. A heatmap was produced using the ComplexHeatmap package (Gu et al., 2016). To compare these two geneset lists concerning their information value, this workflow was repeated for each geneset list.

For quality control, PCA and UMAP were conducted on the newly created pathway patient enrichment matrices, to check whether the clusters that could be seen on the first UMAP plot could in part be found again. After quality control we decided to keep working with the gene ontology genesets (C5 BP).

After subsetting the pathway patient enrichment matrix into one data frame for LUAD and the other cancer types, foldchanges were calculated and a two sided Wilcoxon test was conducted between the means per pathway for all LUAD patients and all other patients. With these values another volcano plot was created and the interesting pathways were extracted.

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.

3.7 Focused analysis

3.7.1 Signed-ranked list

A Shapiro-Wilk test was applied to the cleaned data to check whether the genes were normally distributed. It turned out that over 50% of the genes were not normally distributed. Therefore,

the Wilcoxon signed-rank test was used to determine how significant the change in gene expression between normal and tumor tissue was. A p-value was assigned to each gene.

Lastly, for each gene the sign of the fold change was multiplied with the -log10 transformed p-value to get signed p-values and sort them in decreasing order. This yielded a signed-ranked list which classified the genes by significance in gene expression to show which genes were differentially expressed in tumor tissue compared to normal tissue. This list was later used as input for the GSEA.

3.7.2 GSEA

To assess how strongly pathway expression differs between tumor and normal tissue, a GSEA was performed using the fgsea package (Korotkevich et al., 2019). The previously created signed ranked list served as ranked list L and a combination of metabolism and hallmark genesets served as input for the pathway list S. 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 the expression rate, the mean expression of their leading edges was calculated and visualized in a dotplot using the packages ggplot (Warnes et al., 2022a). Therefore, the pathways were sorted from most upregulated to most downregulated.

3.7.3 GSVA

GSVA was performed using the gsva package (Hänzelmann et al., 2013b) to illustrate the pathway expression for each patient. For this the gene ontology genesets (C5 BP) comprising about 3,000 pathways were used.

The standard deviation was calculated for each pathway of the resulting pathway enrichment matrix. Those with the greatest standard deviation were retained.

The selected pathways and their expression rate per patient were then plotted in a heatmap using the ComplexHeatmap package (Gu et al., 2016).

The shortened matrix was divided into tumor and normal tissue and the foldchange of each pathway was calculated by taking the difference between the mean expression in tumor and normal tissue for each pathway. In addition, the Wilcoxon signed-rank test was applied to each pathway. This was illustrated in a volcano plot using the ggplot package (Warnes et al., 2022a) in which the -log10(p-values) were plotted against the foldchanges.

Finally, volcano plots were created for selected pathways to see which genes were over or under expressed in the individual pathways and the results were compared to literature.

3.8 Regression

To use the information gathered by the pan cancer analysis a regression model with the purpose of identifying potential LUAD patients from RNA-seq data was added to the project. In order to make the binary decision between LUAD and non-LUAD patients, a logistic regression model was trained. First of all, the cleaned dataset was split 70/30 into a training and testing dataset respectively. Additionally, every LUAD patient in these datasets was marked with a 1 and non-LUAD patients with a 0, so that the model can be evaluated later on.

In an effort to find genes that could be used as explaining variables, gene foldchanges (Warnes et al., 2022c) between LUAD and other cancer types were calculated using the cleaned dataset. The genes were additionally tested for correlation and for all highly correlating genes, one of them was removed from the dataset. The 10 most overexpressed 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 and the UMAP was plotted using ggplot2 (Wickham, 2016). The colors represented the corresponding tumor type of the patient. We chose to continue with the chosen genes.

A first rough model was trained using all 20 of the chosen genes and the glm function (R Core Team, 2022a) and specifying the model to use a binomial error distribution and a logit link. This model was then passed into the blr_step_aic_both function from the blorr package (Hebbali, 2020). This function calculates the best composition of the given 20 genes. With the best configuration the final model was trained on the training dataset.

To evaluate the model the first step was to predict whether the patients of the testing dataset were LUAD patients. For this the native predict function (R Core Team, 2022a) was used. The resulting probabilities were transformed into predictions for 1 or 0 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 predicition of the model. As a final evaluation step the package ROCR (Sing et al., 2005) was used to create a ROC curve clearly showing the performance of the model.

Contents

4 Results

4.1 Cancer hallmark pathways

By calculating the Jaccard index for each metabolism geneset to each hallmark geneset, the similarity between these pathways was measured and afterwards visualized in a heatmap (**Fig. 4.1**). This highlights that there is a general low similarity between the selected pathways and only a few genesets show a slightly higher similarity which don't exceed an index of 0.2. The most shared genes are found in the alanine, aspartate and glutamate metabolism with glutamine metabolism. Additionally, large overlaps are found in purine and pyrimidine metabolism with genome repair and down regulation as well as in lipid and fatty acid metabolism with VEGF-induced angiogenesis.

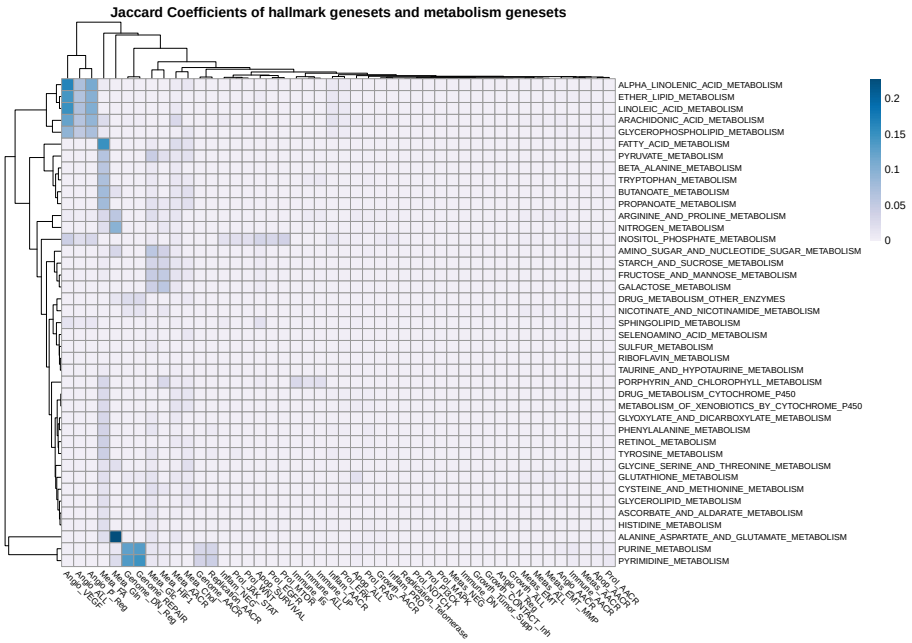

Figure 4.1: Jaccard Coefficients of hallmark genesets and metabolism genesets. The x-axis is defined by the given hallmark genesets, whereas the y-axis is assigned to the selected metabolism geneset.

4.2 Focused analysis

4.2.1 GSEA

In order to illustrate which pathways of the hallmark and metabolism genesets are most up or down regulated, a barplot was created (**Fig. 4.2**). The genes crucial to the mean expression of a pathway are the ones that are in the leading edge of the GSEA result.

Overall we observed more upregulated than downregulated pathways. Among the upregulated, Meta_HIF1 and the ascorbate and alderate metabolism pathways presented with the highest mean. In addition, some pathways associated with nucleotide, amino acid, and sugar metabolism also seem to be upregulated. Specific amino acids whose synthesis appears to be overexpressed include methionine and cysteine. Pathways pertaining to telomerase activity and cell growth were also significantly upregulated.

Downregulated pathways are mainly linked to the immune response but also to the metabolism of specific amino acids and fatty acids, such as histidine or arachidonic acid. Pathways that affect the cytochrome p450 system seem to be downregulated. The two most significantly downregulated pathways regulate linolenic acid and nitrogen metabolism.

4.2.2 GSVA

To visualize the pathway expression of each individual sample for normal and tumor tissue, a heatmap was created using the results of GSVA (**Fig. 4.3**). Two recognizable groups with different expression patterns have formed. The expression of each pathway is color coded, from high expression rate (red) to low expression rate (blue). The left half shows only tumor samples while the right half shows almost exclusively samples belonging to healthy tissue, with a few exceptions framed in black. This suggests that for some samples the tumor tissue has phenotypic manifestations that don't significantly differ from the overall expression patterns in normal tissue. Most pathways however present with a significantly different expression between tumor and normal tissue.

Strong differences can be seen in the pathways marked in green. Many of them are associated with DNA replication or chromosome distribution during mitosis. Notable pathways include "DNA dependent DNA replication", "chromosome separation", and „metaphase anaphase transition off cell cycle". Many tumor samples present with a high expression rate of these pathways, whilst normal samples present with a low expression. This suggests that the aforementioned pathways play an important role in the differentiation between normal and tumor tissue.

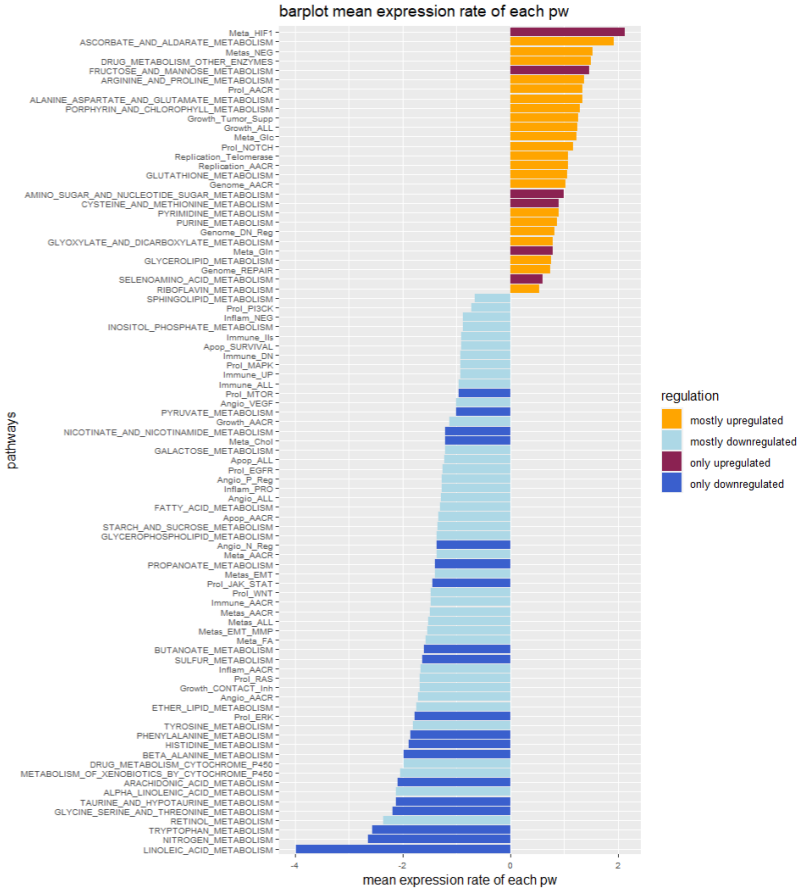

Figure 4.2: Barplot for the mean regultion of hallmark gene sets and metabolism gene sets. The pathways are sorted by their mean expression and can be seen on the y axis, while the mean expression is plotted on the x axis. Furthermore each pathway is coloured by the regultion state, meaning the genes contained in the leading edge define weather the pathway is mostly or only upregulated and vice versa.

The pathways marked in blue are connected to immune response, for example „T-helper 17 type immune response", „regulation of B-cell differentiation " or „dendritic cell migration". In this case, some tumor samples are upregulated while most are downregulated. In normal tissue the expression of most pathways is clearly upregulated with a few exceptions near the right margin.

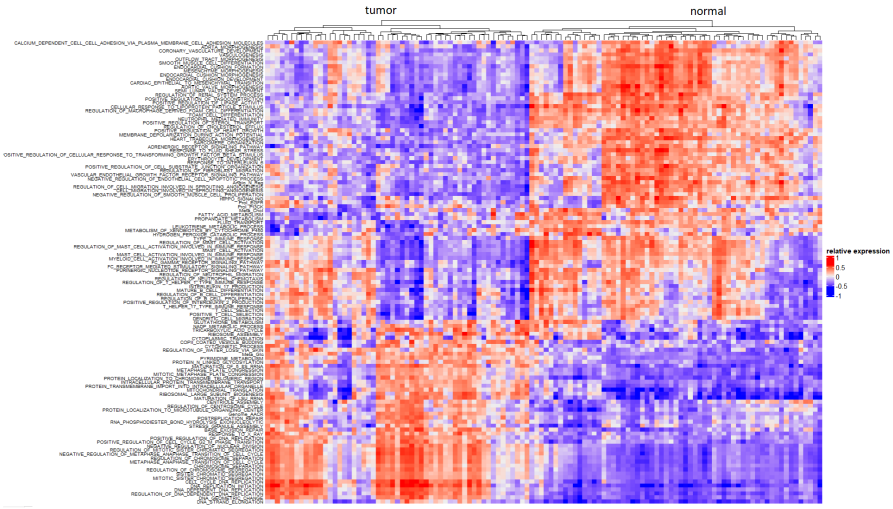

Figure 4.3: heatmap for comparing pathway expression in normal and tumor tissue The heatmap shows the GSVA results for the small TCGA dataset. On the x axis one can see the normal and tumor samples, while the y axis shows the pathways with the highest standard deviation. The expression of each pathway is color-coded, from high expression (red), to low expression (blue).

4.2.3 GSVA volcano plot

The expression of the pathways selected in the previous step is visualized in the volcano plot shown in **Fig. 4.4**. The higher the foldchange, the greater the difference in expression between tumor and normal tissue. Greater -log10(p value) corresponds to higher significance of this difference.

As already seen in the heatmap **Fig. 4.3** the pathways with the highest foldchanges and pvalues are those who relate to DNA replication and cell cycle and are therefore strongly upregulated. Pathways with less significant upregulation stand in relation to ribosome metabolism, such as ribosome assembly and ribosomal large subunit biogenesis.

The pathways involved in immune response such as mast cell activation, are significantly downregulated, while the T-helper 17 pathway is less significantly downregulated.

Other noteworthy pathways that are strongly downregulated include the cellular response to lipoprotein particle stimulus and the regulation of renal system process.

Figure 4.4: 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)).

Schreib hier evtl 1 Satz nach welchem Kriterium die PWs ausgesucht wurden (ich kenn die Fig ja nt) In **Fig. 4.5**, genes of selected pathways are highlighted in black and some with particularly high p values are labeled. The red colored genes are all genes in our cleaned TCGA tumor normal dataset that are significantly over expressed over all patients in tumor tissue compared to normal tissue. The blue colored genes imply the opposite. In gray, one can see insignificantly up or downregulated genes. All genes mentioned in this section will be discussed in more detail in the discussion.

Looking at the pathway of DNA replication initiation with 36 genes, overregulated genes like ORC and CDC6 seem to play an important role in this process. Over 60% of the genes are overexpressed and less than 2% are under expressed. Furthermore, PLK1 is significantly overexpressed in the metaphase anaphase transition of cell cycle. Out of 63 genes, over 50% are overexpressed and just under 10% under expressed. In both pathways, as already described in the section above, a clear up regulation of the pathways in tumor tissue can be seen in the gene expression patterns.

Two down regulated pathways, namely renal system regulation and cellular response to lipoprotein particle stimulus are seen on the left side of **Fig. 4.5**.

Out of the 33 genes of the renal system regulation pathway, approximately 51% are downregulated and only 3% are significantly upregulated.

A similar gene distribution can be observed in the pathway pertaining to lipoprotein particle stimulus.

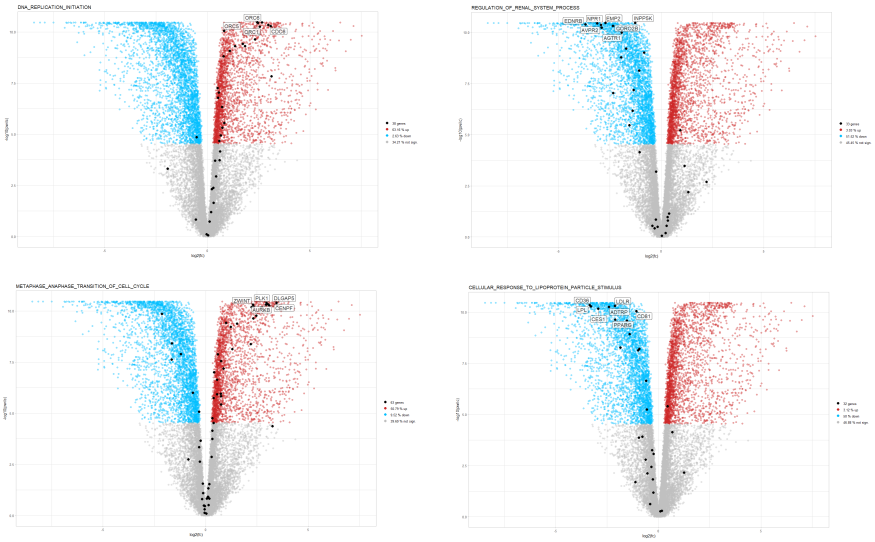

Figure 4.5: volcano plot showing gene expression for selected pathways The plots show the number of genes in each pathway and the percentage of significantly over expressed (red), significantly under expressed (blue) and not significant diferentially expressed genes (gray). The differential expression 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))).

4.3 Pan cancer analysis

4.3.1 Identification of clusters in gene expression data

Dimension reduction of the cleaned data conducted by performing PCA and UMAP results in the plot shown in (**Fig. 4.6**). To identify clustering of different cancer types, the data points of each patient was colored accordingly. Based on the 33 different types occurring in the dataset, the reduced data results in approximately 16 clusters. Notably, BRCA, LIHC, KIRP, SKCM, UVM, THCR, PCPG and PARP exhibit a 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.

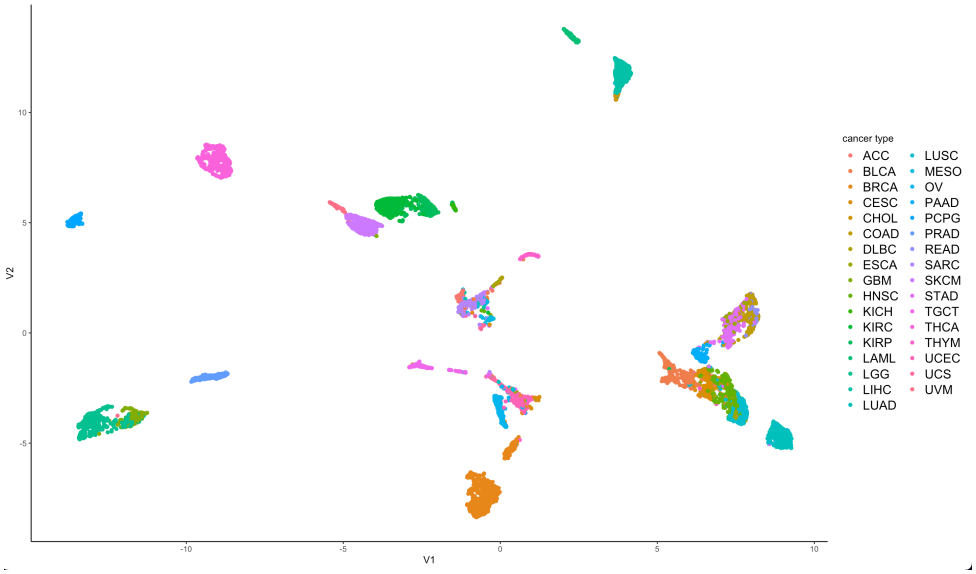

Figure 4.6: UMAP plot on TCGA expression dataset The x-axis is defined by the first umap component, whereas the y-axis assigned to the second component. The data plots are colored by the patients cancer type

4.3.2 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. Different molecular changes result in a different expression of genes and therefore an abnormal regulation of pathways. This deregulation of pathways is characteristic for each cancer type. Hence, its analysis is a crucial part of this pan cancer analysis. Aiming to identify differences in pathway activities based on the cancer type, two geneset list were extracted from MSigDB. One list contained curated genesets whereas the other list contained ontology genesets. Following, GSVA was performed twice on the TCGA expression dataset; once using the curated geneset list and one time with the ontology geneset list. By utilizing the genesets separately, the better fitting geneset for the analysed dataset can be selected. In **Fig. 4.7** the pathway enrichment of each patient is shown with highlighted cancer type. The selected geneset list contains only ontology genesets that overlap with the genes from the expression data with more than 95%. The curated geneset list was not chosen due to less clustering after conducting GSVA (Appendix, **Fig. 8.2**) . Cancer types that result in an isolated and well defined cluster are LIHC (turquoise, upper left corner), KIRK (green, left top), THCA (pink, under KIRK), PRAD (blue, center), PCPG (blue, top), LGG (green, right center) and LAML (green right bottom).

Based on the geneset enrichment matrix created with GSVA, a pathway enrichment heatmap was created (**Fig. 4.8**). Performing kmeans, three clusters of cancer types were identified. The

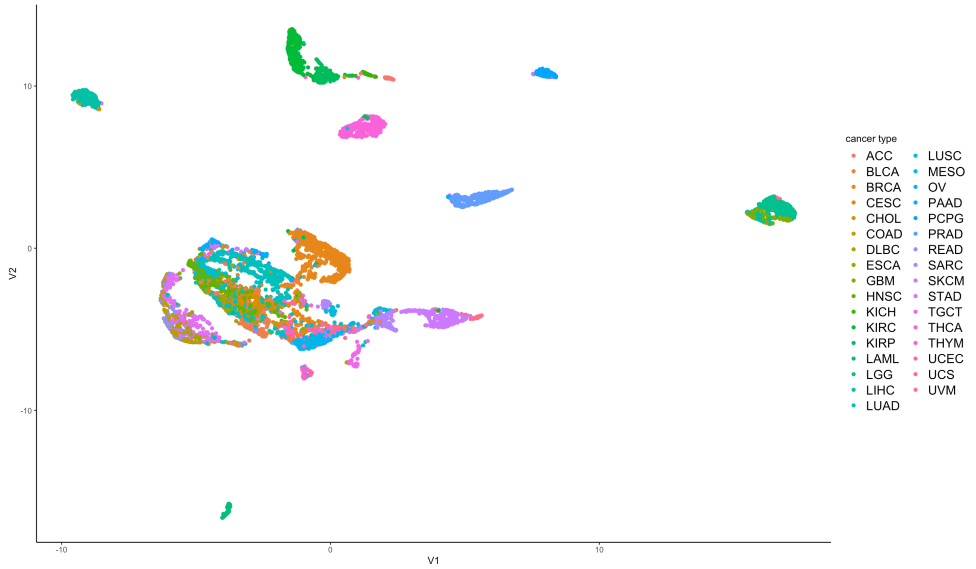

Figure 4.7: Pathway enrichment based on cancer type. The x-axis is defined by the first umap component, whereas the y-axis assigned to the second component. The data plots are colored by the patients cancer type

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 a relatively neutral enrichment 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 pathways can approximately be divided into three subsets. The first cluster includes pathways that regulate the cell cycle, DNA replication and chromatid segregation. These genesets are highly downregulated in the first cancer type cluster and moderately downregulated in 5 cancer types assigned to cluster 2. However, in the majority of cluster 2 and in cluster 3, these pathways show a higher activity. A second cluster can be found in pathways important for morphogenesis, metastasis and cell adhesion. While the cancer types from the second cluster do not drastically over- nor downregulate these genesets in comparison to the other cancer types, cluster 1 and 3 exhibit a general downregulation. The last cluster of pathways comprises pathways involved in the regulation of the immune response. On the one hand, the second and third cluster of cancer types solely exhibit a moderate deregulation of these genesets. On the other hand, cancer types included in the first cluster inhibit immune activation.

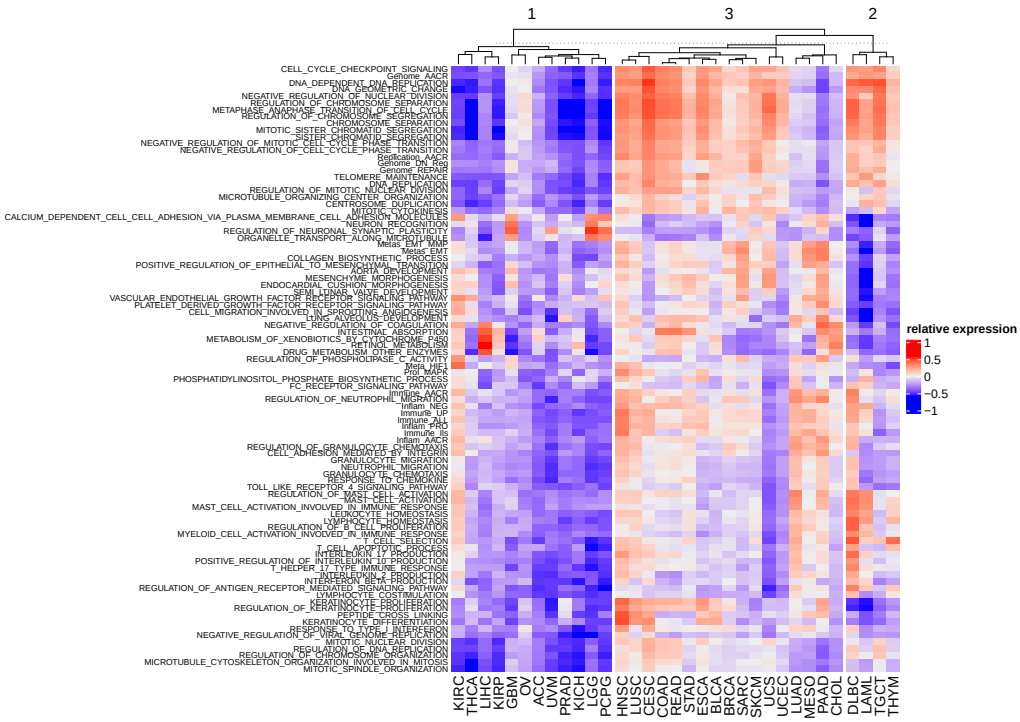

Figure 4.8: Pathway enrichment heatmap. The clustering of cancer types was conducted using kmeans.

4.3.3 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 volcano plot (**Fig. 4.9**) helps with analysing the exact pathways that differ in activity between LUAD patients and other cancer patients. It shows the log2 foldchange values between LUAD and non-LUAD patients as well as the -log10 values of the corresponding p-values from the Wilcoxon test.

The volcano plot shows several differently 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.

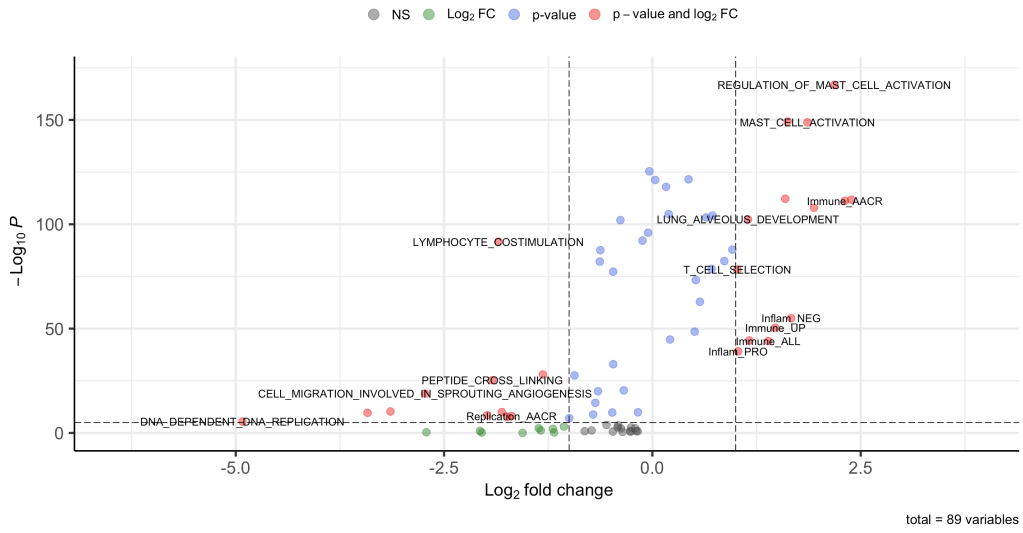

Figure 4.9: Volcano plot for geneset enrichment in LUAD compared to all other cancer types of the TCGA dataset. dotted lines indicating the alpha value and foldchange values of -1 and 1.

4.3.4 Comparison of Clusters Within LUAD

After running separate PCA and UMAP analysis on patients for each tumor type the question arose how the clusters 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 (**Fig. 4.10**).

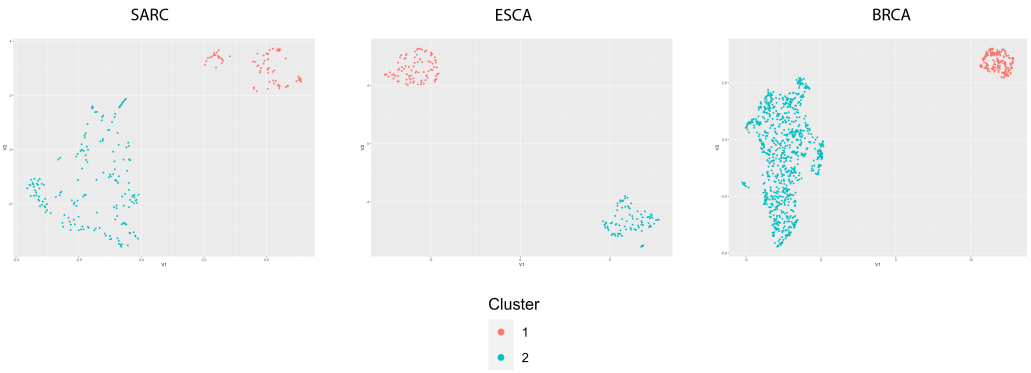

Figure 4.10: UMAP plots for SARC, ESCA and BRCA run on the corresponding subset of the gene expression datatset, colored by the cluster assigned to each datapoint by k-means clustering.

4.4 Regression

The logistic model that 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. 4.11**) are the 27 false-negative occurrences and 7 false-positive occurrences. This results in an accuracy of 0.986 %.

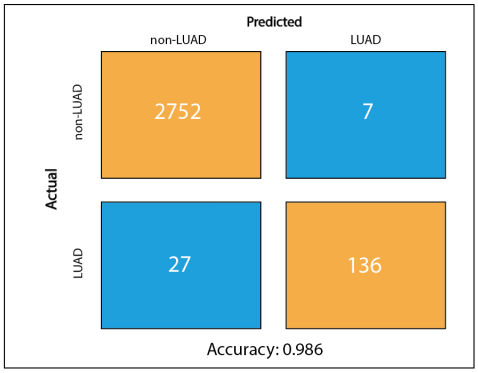

Figure 4.11: Confusion table for prediction on test datast containing 2921 patients with 163 LUAD patients.

For further evaluation a ROC plot was produced which enables an estimation of model performance in relation to the false-positive rate (**Fig. 4.12**). The ideal estimator would have an area under 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.

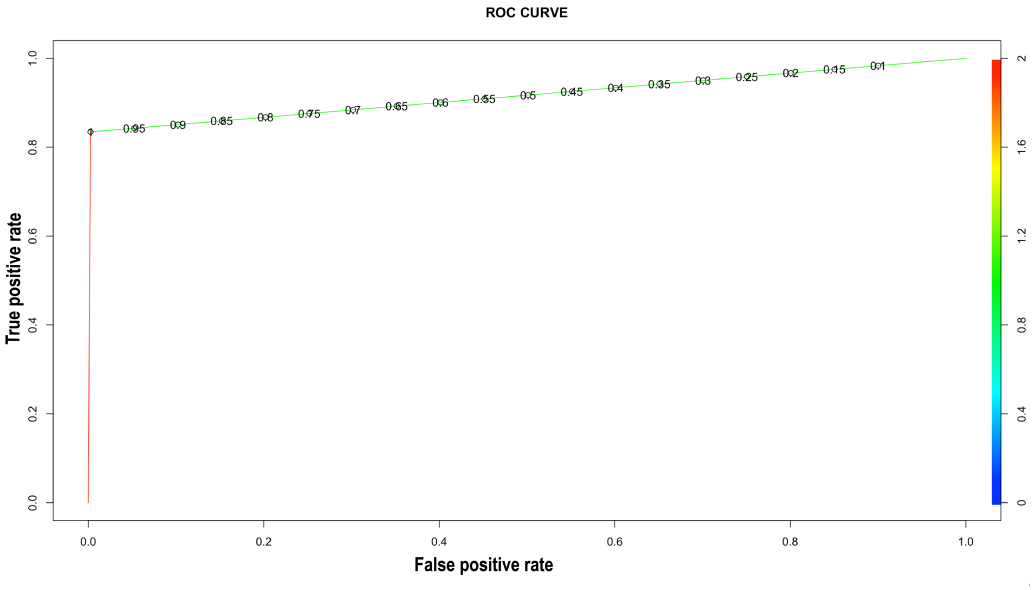

Figure 4.12: ROC plot of the trained logistic model with an AUC value of **0.9159**.

5 Discussion

5.1 Focused Analysis

5.1.1 GSEA

Literature confirms that the hypoxia-inducible factor-1 alpha (HIF1 alpha) plays an important role in tumor progression and metastasis (**Weiwei2013?**). Expression of this factor is much higher in lung cancer tissue than in normal tissue, which can also be seen in metastatic tissues (**Yang2016?**).

Pathways linked to nucleotide, amino acid and sugar metabolism like ascorbate and aldarate metabolism are generally enriched in tumor tissue (**Araujo2018?**).

Adenocarcinoma cell samples are often associated with high telomerase activity and with higher immortality (**Hiyama1995?**).

As seen in (**FigXXX**) CYP is downregulated in LUAD. This is coherent with our expectations as cytochrome p450 (CYP) is known to metabolize carcinogens , thus inactivating them (**Oyama2012?**). However, it is not mentioned if the expression rate is significantly higher or lower than in normal cells. One factor that could influence the result seen in **FigXXX** is that only 58 samples were available.

Gamma linolenic acid suppresses HIF1 alpha induced proliferation and invasion of non-small cell lung cancer cells through inhibition of HIF1 alpha (**Wang2020?**). As previously mentioned HIF1a induces proliferation therefore it downregulation of its inhibitor conforms to our expectations.

5.1.2 GSVA

As indicated in the heatmap (**FigXXX**) and the volcano plot (**FigXXX**), pathways linked to DNA replication and cell cycle regulation were strongly overexpressed in tumor tissue compared to normal tissue. ORC (Origin Recognition Complex) binds to chromatin, marking the location for replication [@ Feng2021. Furthermore ORC is linked to CDC6 thus being activated and

replication can be induced. Research has also shown, that the expression of ORC plays a significant role in the development of lung cancer (**deng2021systemic?**). This confirms the results seen in **FigXXX** (top left).

Immune response mechanisms like T-helper 17 type immune response were partly up and partly downregulated in tumor tissue, which can be seen in **Fig heatmap**. Overall, this circumstance lead to immune mechanisms to not be significantly downregulated as seen in **Fig volcano gsva**. IL-17–producing CD4 helper T cells (Th17 cells) play a crucial role in promoting chronic tissue inflammation which has often been linked to the development of cancer (**Chang2014?**). Further research showed that the Th17 function may vary according to cancer type, location, and stage of disease (**Wilke2011?**). This suggests that the expression rate of the **named pathway** can vary between samples.

Looking at the down regulated pathways, the renal system shows particularly low expression rate of AGTR1 and EDNRB **Fig** (top right). The renin-angiotensin system (RAS) plays an important role in lung cancer (**Xiong2021?**). Furthermore the angiotensin II receptor 1 (AGTR1) is part of the RAS system in the kidney and was found to be lowly expressed in most examined tumors, including in lung cancer. Another research group found, that the endothelin receptor type B (EDNRB) inhibited proliferation and migration of LUAD cells and was lowly expressed in LUAD patients (**Wei2020?**). This leads to higher proliferation and migration rates, which are cancer hallmarks.

The results indicate low cellular response to LDL particles in tumor tissue samples. As seen in **Fig**, peroxisome proliferator-activated receptor gamma (PPARgamma) is significantly downregulated. PPARgamma is proven to be involved in inhibition of development of primary tumors and metastases formation in lung cancer and may also function as a tumor suppressor (**Aravind2016?**). Moreover CD36 has different expression rates at different cancer stages. CD36 expression is invariably low in the *in situ* stage but rises when cells begin to metastasize (Bacolod et al., 2019).

5.2 Pan Cancer Analysis

5.2.1 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 cancer cells derived from 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 transformation of gene expression can be concluded. The latter observation is unsurprising as both tumor types are glioma.

5.2.2 Pathway Enrichment

Comparison of the clusters resulting from GSVA to the clusters based on the gene expression highlights that the cancer types cluster less accoring to their pathway regulation. This is due to the fact that GSVA entails a certain degree of loss of information. To minimize the information loss, we performed GSVA with two different geneset lists and chose the geneset list that retained the most information. Therefore, several cancer types with a unique gene expression also show a specific pathway enrichment pattern. Cancer types that remain clustered are LIHC, LGG and PCPG.

By visualizing the pathway enrichment relative to the cancer type the formation of three cancer type and approximately three geneset clusters can be observed. Notably, the first cluster of cancer types show a strong downregulation of pathways relative to other cancer types.

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 and last cluster contains cancer types related to cells exhibiting a high stemness like blood forming cells and germ cells (Weissman, 2000).

The three clusters in the pathways axis resemble cell cycle and genome regulation, regulation of different phases of metastasis, and activation of the immune response.

Cell cylce and genome deregulation resembles one of the hallmarks of cancer as it is the base for tumor progression (Bruce, 1983). By increasing the activity of cell cylce promoting pathways, cancer cells activate proliferation and cell growth. 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 do not have slower proliferation than normal cells but rather have less aggressive spread than other cancer types. The first cluster of cancer types, particularly kidney cancer and prostate adenocarcinoma, result in a downregulation of these pathways. While our findings in PRAD are according to expectations (Sakr and Grignon, 1997), the strong downregulation in kidney cancer is not supported by other studies. There are several subtypes of kidney cancer, for example KIRC, KIRP and KICH. KIRC indeed is a slow growing cancer but KIRP and KICH both are subtypes that show a high proliferation rate (Li et al., 2012). These aberrances to other studies may result from our selection of pathways or our data cleaning which inevitably leads to loss of

information. Cancer types that show a strong increase in cell cycle pathway activity are CESC and DLBC. Since CESC leads to acute mortality among woman (Small Jr et al., 2017), our findings are as expected and proven by further studies (Ding et al., 2020). CESC and DLBC are highly aggressive cancer types which explains the strongly increased activity in pathways regulating the genome and cell cycle (Said, 2013).

Metastasizing is a complex and inefficient progress due to numerous regulation mechanisms (Bruce, 1983). Therefore, the ability of a primary cancer depends on its location and molecular signature. As part of the analysis of metastatic landscapes, Budczies *et al.* found melanoma, breast cancer and kidney cancer to feature a high metastatic potential. However, cancer cells deriving from the liver and sexual organs show the lowest rates of metastasis (Budczies et al., 2015). Our findings confirm this research. BRCA, SKCM, and especially KIRC result in upregulated mestastasis genesets in comparison to other cancer types. In addition, the high invasion and metastasis rate of SKCM is supported by further studies (Huang et al., 2022). Secondly, our analysis revealed downregulated metastatic pathways in LIHC, PRAD, OV, CESC and UCEC which aligns with insufficient invasiveness in liver and reproductive cancers (Budczies et al., 2015). Moreover, PAAD resulted in the most severe upregulation of metastasis related pathways which is supported by several studies(Ayres Pereira and Chio, 2019). Epithelial mesenchymal transition (EMT) is an essential process of metastasis formation by enabling the cell to circulate through vessels (Kalluri et al., 2009). Therefore, the upregulation of pathways involved in EMT in PAAD matches our expectations (Rasheed et al., 2010). In contrary, LAML shows a significantly low enrichment of pathways inducing metastasis. At first, this seems paradoxical as AML is highly invasive because it is not categorized as a solid tumor (Whiteley et al., 2021). By nature myeloid cells possess a high motility and the ability to circulate through vessels. Consequently, transformed myeloid cells do not need to upregulate processes that enable cell migration or inhibit anoikis (Trendowski, 2015).

While there is no significant deregulation of immune activation in the second and third cancer type clusters, cancer types of the first cluster transform a severe downregulation. This clustering is according to our expectations as cancer types of each cluster can be categorized by immune-infiltration CpG markers (Wang et al., 2020). As a result, these cancer types inhibit immune cell infiltration hence have a low immunogenicity (Smyth et al., 2006). Notably, 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). Unfortunately, its rareness entails a small research base for analysis. As part of cluster one, glioma like GBM and LGG cause a decline in pathway activity leading to a highly immunosuppressive tumor microenvirnment (Guan et al., 2018). PCPG is a solid tumor and thus develops a tumor microenvironment that surpesses immune cell infiltration. Furthermore, transformed cells lack leukocyte infiltration enabling

tumor progression (Fishbein et al., 2017). On the other hand DLBC and HNSC are examples for cancer types upregulating immune infiltration which seems to be contradictory at first. However, inflammation can support the tumor mircoenvironment because it promotes tumorigenesis and tumor progression by supplying essential molecules like growth and survival factors. Tamma *et al.* observed this phenomenon in DLBC (Tamma et al., 2020). Additionally, HNSC increasing inflammatory pathways is 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 relative to their actual behaviour. Possible reasons for these discrepancies are the chosen pathways and the general loss of information during conduction of GSVA.

5.2.3 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. 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 increased by overexpression of the corresponding genes. 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.

5.2.4 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 a 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).

Another crucial gene for biological processes is FGA, which is overexpressed in cluster one. This codes for the fibrinogen alpha chain and thus is needed for secondary haemostasis. Patients from both clusters seem to differ in their blood clotting capabilities (Freissmuth et al., 2016). Additionally the CALCA gene, which controls the calcium household is also differentially expressed and thus further fuels the difference in blood clotting, as calcium is needed for secondary haemostasis (Singh et al., 2019).

INSL4 is normally found during embryonic development as it can bind the insulin-like growth factor receptor. In LUAD it is significantly overexpressed. Such genes were expected to be found as often cancer progression results in reactivation of early development genes (Veitia et al., 1998).

It was shown that the two clusters found do differ in very specific aspects of biological processes. We expected to find distinct clusters corresponding to the LUAD subtypes found by Qin *et al.* (Qin et al., 2020) which are characterized by immune activity. However even in the most differentially expressed genes twe 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.

5.3 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. 8.3**). The confusion table that was acquired from predicting values in the testing dataset 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. 8.4**). Nevertheless due to the high amount of patients in general (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. Generally the further the curve protrudes into the top left corner, the better the model is. 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 to 1 are regarded as good (**narkhede2018understanding?**). 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 (**way2018machine?**).

6 Outlook

6.1 Analysis of Metastasis Formation in LUAD

About a third of LUAD patients already present with brain metastases at the time of diagnosis and about half of all patients will eventually develop brain metastases (Shih et al., 2020). As a cancer's ability to metastasize dramatically impacts a patient's chances of survival, a comprehensive analysis of the genetic alterations that most often lead to metastasis formation would allow to single out patients at risk for developing metastases and to treat them accordingly.

The comparison of genomic expression profiles of cancer cells taken from the metastases with those taken from the primary tumor might reveal which mutations enable a lung cancer cell to metastasize.

6.2 Smoker vs non-smoker

It is a well-known fact that lung cancer is a smoker's disease. However, in recent years studies have found that lung cancer incidence is decreasing in smokers and increasing in non-smokers. Furthermore, 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 Qui *et al* can be replicated with our data.

6.3 Identification of Immune Subtypes

Many different research groups have already set out to subtype LUAD according to immune signature defined for instance by PD-L1 expression or immune cell infiltration. Generally, the more pronounced a cancer's immune signature is, the better the cancer will respond to immunotherapy and the better a patient's chances of survival (Xu et al., 2020). Based on the

findings of the likes of Xu *et al.*, our own data could be divided according to immune signature and determining trends within each subgroup such as survival rate could be identified.

6.4 Finding Defining Trends Between LUAD Clusters

Performance of UMAP and PCA on our data (**Fig. 8.1**) showed that LUAD forms two distinct clusters. Further analysis may reveal the genetic differences within LUAD that lead to clustering as well as the genetic similarities shared by samples belonging to the same cluster. Additionally, response to therapy might differ between clusters and thus patients belonging to one cluster or another might face vastly different chances of survival.

6.5 Prediction of Cancer Stage

Over the course of this project we trained a logistic regression model to predict whether an individual is at risk of eventually developing LUAD. This model could be further sophisticated to additionally predict the cancer stage on the grounds of a patient's genomic profile as well as other factors like age or smoking habits. With enough training, this model could ideally be used as a less invasive alternative to the current diagnostic methods and therefore help determine an adequate treatment plan with reduced patient trauma.

6.6 Epigenetics

The term epigenetics describes hereditary changes in gene expression that are not due to changes in the DNA sequence. The most common epigenetic alterations in cancer cells include global hypomethylation of repetitive DNA sequence regions and hypermethylation of tumor suppressor genes which are consequently inactivated (Esteller, 2008). Analysis of the differences in methylation patterns between tumorous and healthy tissue would allow to determine which tumor suppressor genes are most commonly inactivated in LUAD through methylation. Furthermore differences in methylation patterns between LUAD and other cancer types could further help to distinguish the process of LUAD development from that of other cancers.

6.7 Experimental Validation

The ultimate step of any analysis would be to seek empirical confirmation of novel findings. Since TCGA holds immunohistochemistry stained samples of tumor tissue it would be possible to directly study in what way different genomic profiles impact the development of tumors *in vivo*.

7 References

(2016). Non-coding RNAs in colorectal cancer (Switzerland: Springer).

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.

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.

Bacolod, M.D., Barany, F., and Fisher, P.B. (2019). Chapter seven - can CpG methylation serve as surrogate markers for immune infiltration in cancer? In Immunotherapy of Cancer, X.-Y. Wang, and P.B. Fisher, eds. (Academic Press), pp. 351–384.

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.

Chen, Y., Lun, A.A.T., and Smyth, G.K. (2016). From reads to genes to pathways: Differential expression analysis of RNA-seq experiments using rsubread and the edgeR quasi-likelihood pipeline. F1000Research *5*, 1438.

Ding, H., Xiong, X.-X., Fan, G.-L., Yi, Y.-X., Chen, Y.-R., Wang, J.-T., and Zhang, W. (2020). The new biomarker for cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) based on public database mining. BioMed Research International *2020*.

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

Durinck, S., and Huber, W. (2022). biomaRt: Interface to BioMart databases (i.e. ensembl).

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.

Fishbein, L., Leshchiner, I., Walter, V., Danilova, L., Robertson, A.G., Johnson, A.R., Lichtenberg, T.M., Murray, B.A., Ghayee, H.K., Else, T., et al. (2017). Comprehensive molecular characterization of pheochromocytoma and paraganglioma. Cancer Cell *31*, 181–193.

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.

Guan, X., Hasan, M.N., Begum, G., Kohanbash, G., Carney, K.E., Pigott, V.M., Persson, A.I., Castro, M.G., Jia, W., and Sun, D. (2018). Blockade of na/h exchanger stimulates glioma tumor immunogenicity and enhances combinatorial TMZ and anti-PD-1 therapy. Cell Death & Disease *9*, 1–16.

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.

Huang, R., Li, M., Zeng, Z., Zhang, J., Song, D., Hu, P., Yan, P., Xian, S., Zhu, X., Chang, Z., et al. (2022). The identification of prognostic and metastatic alternative splicing in skin cutaneous melanoma. Cancer Control *29*, 10732748211051554.

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.

Kalluri, R., Weinberg, R.A., et al. (2009). The basics of epithelial-mesenchymal transition. The Journal of Clinical Investigation *119*, 1420–1428.

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.-S., Yao, L., Gong, K., Yu, W., He, Q., Zhou, L.-Q., and He, Z.-S. (2012). Growth pattern of renal cell carcinoma (RCC) in patients with delayed surgical intervention. Journal of Cancer Research and Clinical Oncology *138*, 269–274.

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., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction (arXiv).

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.

Neuwirth, E. (2022). RColorBrewer: ColorBrewer palettes.

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).

Rasheed, Z.A., Yang, J., Wang, Q., Kowalski, J., Freed, I., Murter, C., Hong, S.-M., Koorstra, J.-B., Rajeshkumar, N., He, X., et al. (2010). Prognostic significance of tumorigenic cells with mesenchymal features in pancreatic adenocarcinoma. Journal of the National Cancer Institute *102*, 340–351.

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.

Said, J.W. (2013). Aggressive b-cell lymphomas: How many categories do we need? Modern Pathology *26*, S42–S56.

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.

Sakr, W.A., and Grignon, D.J. (1997). Prostate cancer: Indicators of aggressiveness. European Urology *32*, 15–23.

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.

Singh, S., Dodt, J., Volkers, P., Hethershaw, E., Philippou, H., Ivaskevicius, V., Imhof, D., Oldenburg, J., and Biswas, A. (2019). Structure functional insights into calcium binding during the activation of coagulation factor XIII a. Scientific Reports *9*, 1–18.

Slowikowski, K. (2021). Ggrepel: Automatically position non-overlapping text labels with 'ggplot2'.

Small Jr, W., Bacon, M.A., Bajaj, A., Chuang, L.T., Fisher, B.J., Harkenrider, M.M., Jhingran, A., Kitchener, H.C., Mileshkin, L.R., Viswanathan, A.N., et al. (2017). Cervical cancer: A global health crisis. Cancer *123*, 2404–2412.

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.

Smyth, M.J., Dunn, G.P., and Schreiber, R.D. (2006). Cancer immunosurveillance and immunoediting: The roles of immunity in suppressing tumor development and shaping tumor immunogenicity. Advances in Immunology *90*, 1–50.

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.

Trendowski, M. (2015). The inherent metastasis of leukaemia and its exploitation by sonodynamic therapy. Critical Reviews in Oncology/Hematology *94*, 149–163.

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. (2022b). Gtools: Various r programming tools.

Warnes, G.R., Bolker, B., and Lumley, T. (2022c). 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. (2022a). Gplots: Various r programming tools for plotting data.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The cancer genome atlas pan-cancer analysis project. Nature Genetics *45*, 1113–1120.

Weissman, I.L. (2000). Translating stem and progenitor cell biology to the clinic: Barriers and opportunities. Science *287*, 1442–1446.

Whiteley, A.E., Price, T.T., Cantelli, G., and Sipkins, D.A. (2021). Leukaemia: A model metastatic disease. Nature Reviews Cancer *21*, 461–475.

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.

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),.

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.

8 Appendix

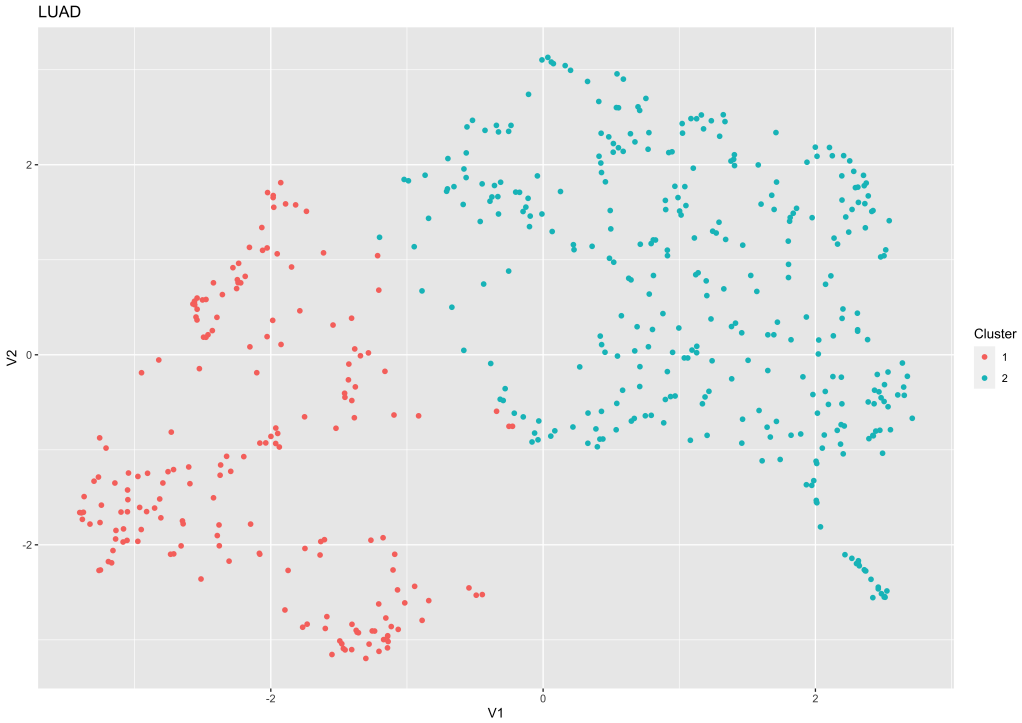

Figure 8.1: PCA UMAP plot of LUAD

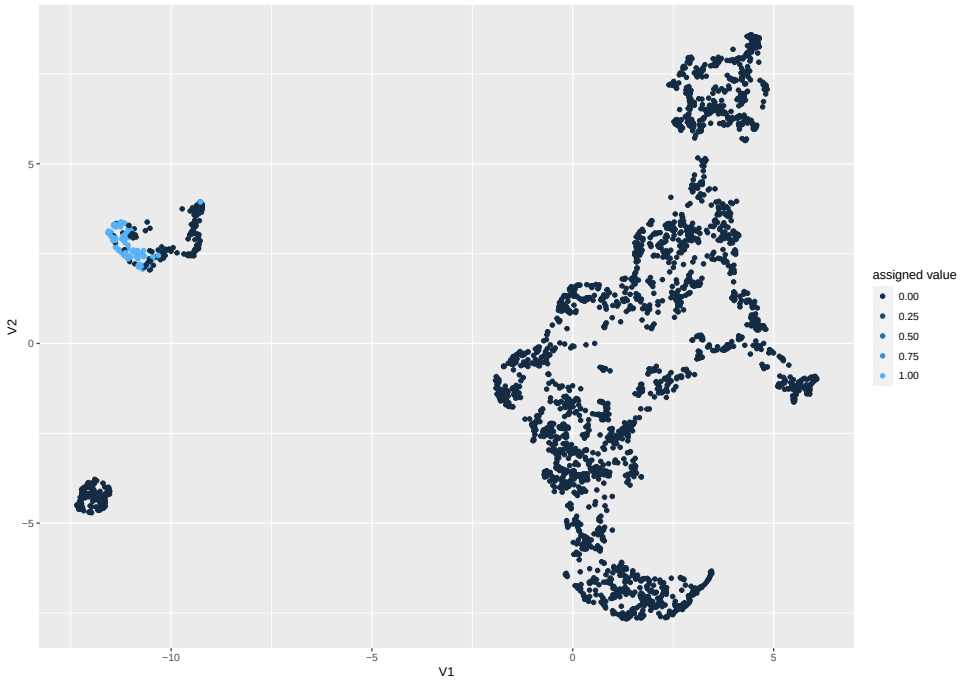

Figure 8.2: Clustering after GSVA performed with curated genesets. Colored by assigned value by the trained model to asses quality of model

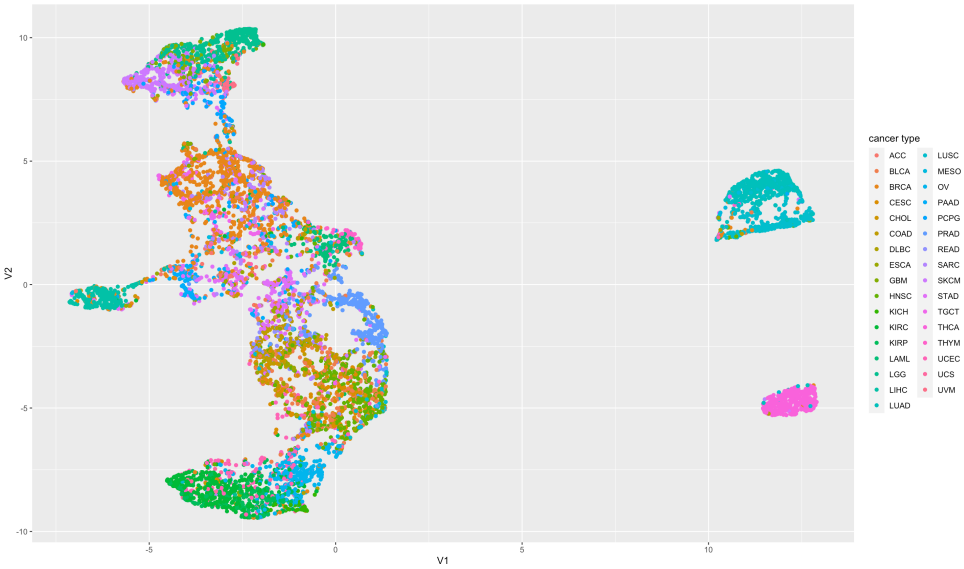

Figure 8.3: Quality control UMAP plot for regression. Shows that LUAD cluster when gene expression dataset is subset to only include explaining variables, colored by cancer type

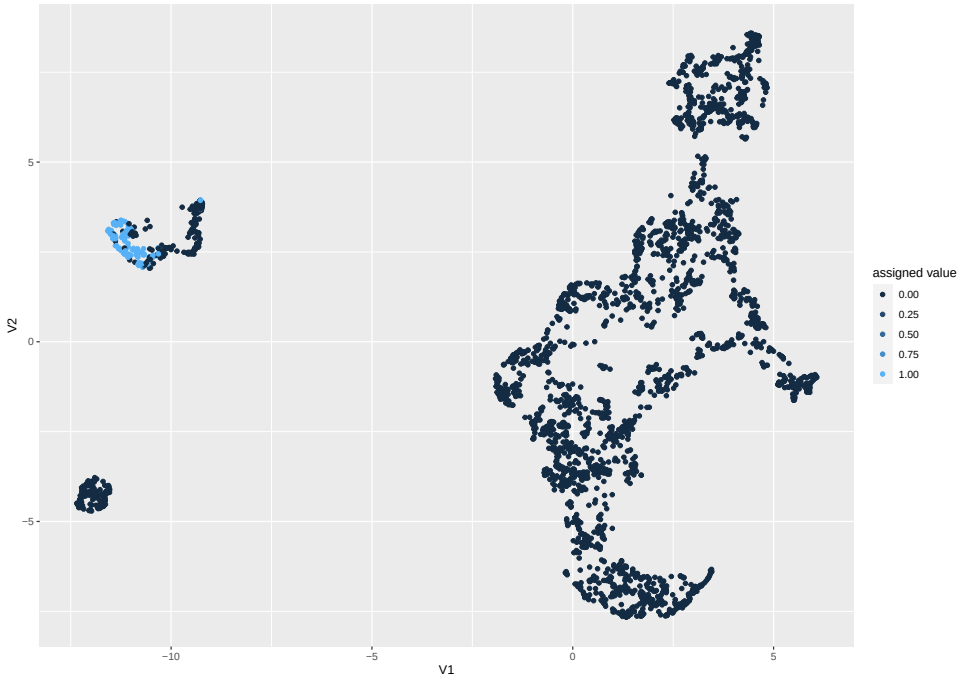

Figure 8.4: Quality control UMAP plot for regression. Colored by assigned value by the trained model to asses quality of model