Ruprecht-Karls-University Heidelberg
Faculty for Life Sciences
Molecular Biotechnology

# Thyorid cancer: Comparison of linar model and neuronal network (xxx)

## 3-Sätze-Zusammenfassung

## sfsf

Data Science Project SoSe 2022

Autoren Anna Lange, David Matuschek, Jakob Then, Maren Schneider
Geburtsort Heidelberg
Abgabetermin 20.07.2022

# Contents


1 Introduction 6
1.1 Hallmarks of cancer . . . 6
1.1.1 Histological tumor types . . . 6
1.2 Thyroid carcinoma . . . 7
1.3 Computational tools . . . 8
1.3.1 Our Analysis . . . 10
1.3.2 Linear regression analysis . . . 10

2 Material and Methods 12
2.1 Our data sets . . . 12
2.2 Metabolic pathway selection . . . 12
2.3 Preprocessing . . . 13
2.3.1 Deleting Not Available Values (NA's) . . . 13
2.3.2 Low-Variance Filtering . . . 13
2.3.3 Biotype Filtering . . . 13
2.4 Descriptive analysis . . . 14
2.4.1 Mean-variance plot . . . 14
2.4.2 Violin plot . . . 14
2.4.3 Jaccard-Index . . . 15
2.4.4 Volcano plot . . . 15
2.5 Data Reduction and Pathway Activities . . . 16
2.5.1 PCA . . . 16
2.5.2 UMAP . . . 16
2.5.3 GSEA . . . 16
2.5.4 GSVA . . . 17
2.5.5 Figure X . . . 17
2.5.6 Linear Regression . . . 18
2.5.7 Neuronal Network . . . 18

3 Results 19
3.1 Preprocessing . . . 19
3.2 Descriptive analysis . . . 20
3.3 Pan cancer analysis . . . 21
3.4 Focused analysis . . . 24
3.5 Regression analysis . . . 26

4 Discussion 27

5 References 28

6 Appendix 29
6.1 Plots . . . 29
6.2 Code . . . 29

Thank You

Thank You

# 1 Introduction

In 2019, 230,000 humans died from cancer in Germany xxx

$[KrebsrateundKrebs-SterberateinDeutschland(krebsinformationsdienst.de)](https://www.krebsin$

. To detect and fight tumors, the development of new treatment and detection methods is essential. Therefore it is inevitable to find a tumors mutational cause. Therefore transcriptomic profiling methods like RNA-seq can be used.

The provided data in the following analysis originates from transcriptomic profiling methods like RNA-seq. Transcriptomic profiling sequences all the RNA that has been generated by transcription of a cells DNA. The difference to sequencing of DNA is, that it only sequences those genes, that are going to be expressed in that cell (Alberts and Walter, 2015).

## 1.1 Hallmarks of cancer

The Hallmarks of Cancer are properties of tumors, that can be detected in each tumor. Among others resisting cell death, inducing angiogenesis, enabling replicative immortality, activating invasion and metastasis evading growth suppressors were the first detected hallmarks (Hanahan and Weinberg, 2011).

### 1.1.1 Histological tumor types

The observed tumors can be classified into different histological types. Carcinomas contain adenocarcinomas, Squamous cell carcinoma, transitional cell carcinoma and carcinomas in general, which include all of the mixed carcinomas. Carciomas derive from epithelial cells. Melanoma is a tumor of the skin, a sarcoma derives from connective or supportive tissue cells. A glioblastoma is a tumor in the brain and leukemia affects the blood (Alberts and Walter, 2015).

#### 1.1.1.1 RNA-sequencing xxx

RNA-sequencing (RNA-seq) is performed by cleaning of RNA, fragmentation, translation of RNA to cDNA, sequencing of cDNA and comparing with the reference genome. The advantage of RNA-seq is that it includes information about gene expression that is especially important in the analysis of tumors such as epigenetic changes (e.g. epigenetic gene silencing) or fusion proteins (Alberts and Walter, 2015).

The results from RNA-seq used for the analysis originate from the cancer genome atlas (TCGA).

## 1.2 Thyroid carcinoma

Thyroid carcinoma (THCA) incidence increased dramatically over the past few years (Cabanillas *et al.*, 2016). To enlarge the understanding of THCA and thereby hopefully improve patients prognosis, this project focuses on finding genes that have a significantly different expression in THCA compared to other cancers and especially to normal tissue.\ The main tasks of the thyroid gland are synthesizing hormones and regulating body temperate and metabolism (Tsibulnikov *et al.*, 2020). A lack of thyroid hormones can cause symptoms like headaches, nausea and depression. Most THCAs derive from thyroid cells and thereby the thyroid gland loses their function, resulting in a lack of thyroid hormones.\ Thyroid cancer can occur in two different types, differentiated and undifferentiated thyroid cancer. Those two types again have histological subtypes. Papillary thyroid cancer (PTC), the most common THCA, follicular thyroid cancer (FTC) and tall cell variant cancer (TCV) are subtypes of differentiated thyroid cancer (DTC) while medullary and anaplastic thyroid cancer are subtypes of undifferentiated thyroid cancer (UTC). Prevalence of DTCs is clearly higher than of UTCs (Prete *et al.*, 2020).\ Regarding the presented DTCs, PTCs have the best clinical prognosis (Lin, 2007), while TCV cancers have the worst clinical outcome (Coca-Pelaz *et al.*, 2020). Therefore, the detection of the tumor type would be important and for more specific therapy options. Even though, all thyroid cancers are treated with thyroidectomy and radioactive iodine, the additional therapy differs for each histological type (Kant *et al.*, 2020).

RODRIGUES_DCC_TARGETS_UP is a gene set containing oncogenes found in colon cancer. Those genes are upregulated in colon cancer cells and thereby promote metastasis, tumor cell survival and invasion. By identifying the upregulation of those genes in a cancer, additional therapies can be started to suppress tumor survival, tumor growth and invasion.

xxx (Opposing roles of netrin-1 and the dependence receptor DCC in cancer cell invasion, tumor growth and metastasis - PubMed (nih.gov)) If this upregulation can also be detected in THCA, therapies, that are used in colon cancer may be useful for therapy of THCA.

#### 1.2.0.1 Integrin

xxx

## 1.3 Computational tools

#### 1.3.0.1 Gene Set Enrichment Analysis

To analyse and compare the activity of pathways of gene expression data, a Gene Set Enrichment Analysis (GSEA) is performed. The aim of the GSEA is to analyse and to identify highly expressed pathways **GSEAxxx**. For this, two conditions with replicates are compared, so a reference of normal expression data is needed.\ First, a gene list is defined. Then the statistically enriched pathways are identified and lastly, the results are visualized.\ GSEA is performed with the package 'fgsea' ref(xxx).

Korotkevich, Gennady, Vladimir Sukhov, and Alexey Sergushichev. 2019. "Fast Gene Set Enrichment Analysis." Journal Article. bioRxiv, 060012. https://doi.org/10.1101/060012 xxx

#### 1.3.0.2 Gene Set Variation Analysis

The Gene Set Variation Analysis (GSVA) is performed with the same intention as the GSEA, so to analyse the pathway activities from gene expression data. Like the GSEA, the approach helps to reduce noise, to further reduce dimensions and to improve the interpretation process (**GSVA?**). The difference to the GSEA is that there no reference expression data is to perfom the GSVA.

GSVA is performed with the package xxx.

#### 1.3.0.3 Uniform Manifold Approximation and Projection for Dimension Reduction

The Uniform manifold approximation and projection for dimension reduction (UMAP) is a method to reduce the dimension of a multidimensional data set. In comparison to the PCA, UMAP can reduce dimensions where the data is not linear (Sharma *et al.*, 2021 xxx). Thereby, the high dimensional structure of the data is maintained. In further visualization, the structure can be represented in clusters that would not be visible using PCA. Therefore, the identification of the clusters is a lot easier. is Thereby the UMAP keeps the overall structure of the data set, therefore clusters are easier.\ The problem of the UMAP is, that although the overall structure is conserved, the distance between the individual points is not proportional to the real distance in the data set. This arises from the non-linear dimensional reduction.

UMAP is performed with the package xxx.

#### 1.3.0.4 Principal component analysis xxx QUELLE

A Principal component analysis (PCA) is used to reduce the dimension of a given data set. The dimensions are summarized in principal components (PCs) which do not correlate. Because the PCs summarize the dimensions, the first PCs explain most of the variance of the data set and thereby can be selected to explain the data. Still, one has to keep in mind, that by reducing the dimensions, not all of the variance is explained and some of the information is lost in the process. The ideal number of PCs can be determined with an elbow-plot. In our analysis we use a PCA as a foundation for the UMAP, because the UMAP can not work with correlated dimensions. Furthermore it is used to detect the most important pathways, which explain most of the first PCs.

In the analysis, a PCA is performed for the pan cancer analysis on the TCGA gene expression data, to find similarities and differences in pathway activity for each tumor type. Furthermore a PCA is performed for the focused analysis of THCA and normal tissue.

PCA is performed with xxx.

#### 1.3.0.5 Jaccard index

The Jaccard index is the intersection, divided by the union of two sets. Therefore, it can be used to identify the similarity of the sets.

### 1.3.1 Our Analysis

In the following, two analyses are performed: a pan cancer analysis and a focused analysis about THCA.

#### 1.3.1.1 Pan Cancer Analysis

For the pan cancer analysis 3 data sets are provided. One containing expression data of 60,000 genes in 10,000 tumor patients, another one with clinical annotations concerning those patients and one with hallmark pathways and their included genes. In the following analysis this data is cleaned by removing NAs, biotype filtering and low-variance filtering. After that a descriptive analysis is performed. Those two steps lead to the actual analysis, a gene set variation analysis to detect significantly altered pathways compared to the other pathways in tumor tissue. In the end a linear regression analysis is performed to predict pathway activity based on other pathways??? xxx Furthermore a neuronal network is built to improve prediction.

#### 1.3.1.2 Focused analysis on THCA patients

Furthermore a analysis of THCA patients is performed. For this analysis a data set containing the gene expression of 60 patients in tumor an normal tissue and their clinical annotations. First the data is cleaned and described like the pan cancer data, to prepare the data for the gene set variation analysis, which is also performed for the THCA data in the bigger pan cancer data set, to confirm results from the smaller data set. In this analysis a linear regression analysis is performed to predict the activity of thyroxine biosynthesis. The results are also improved with a neuronal network.

### 1.3.2 Linear regression analysis

A linear regression analysis is performed to predict the activity of xxx based on the activity of the other pathways.

Firstly, the correlation of the pathways for predicting is checked, only pathways with a low correlation were kept. In the next step, the variance is checked, 80% of the genes with low variance were omitted.

For the regression analysis only 20% of the pathways were used, to only use significant pathways.

The regression analysis was tested by

# 2 Material and Methods

## 2.1 Our data sets

For the analysis four data sets were provided.

The first data set is a Gene expression data frame. The Gene expression data frame contains 60,000 genes and and their expression in 10,000 patients. It is derived from The Cancer Genome Atlas (TCGA). The expression of the genes was obtained by RNA-seq.

The second data frame contains 37 clinical annotations like Tumor type, age, gender, etc. for each of the 10,000 patients from the Gene expression data frame.

The third object is a list that contains five lists for the focused analysis, one list for each tumor type (BRCA, KIRC, LUAD, PRAD, THCA). For the focused analysis, the THCA data (only DTC) was used. The THCA data contains 3 data frames, each one with information about the same 60 patients. The first data is a gene expression matrix from THCA tissue, the second data contains the gene expression from normal tissue and the third data frame contains the clinical annotations like age and gender. Gene expression data was obtained by RNA-seq.

The fourth object contains 46 pathways involved in phenotypes partly included in the hallmarks of cancer and the genes involved in those pathways.

SIND DIE DATEN NORMALISIERT –> Normalisiert glaub ich (Anna) ODER ALS COUNTS?

## 2.2 Metabolic pathway selection

From the Molecular Signature Database (MSigDB) **xxx** metabolic pathways were selected. First, they were compared to the given Hallmark-Pathways in order to select pathways that differ to the Hallmark-Pathways. The goal was to identify more pathways, that are important for the development of cancer. Therefore it was important that as many

genes from the selected pathways as possible are also included in the provided Hallmark pathways. To identify the relevant pathways, the intersection of genes was calculated and the genes with an intersection of at least 99% were maintained for further analysis.

xxx??????????????????????

To avoid duplicates in between the metabolic pathways and between the Hallmark pathways and the metabolic pathways, the pathways were checked for duplicates with the Jaccard index. Pathways with a sum of Jaccard indices beyond the 1-sigma range were discarded.

## 2.3 Preprocessing

### 2.3.1 Deleting Not Available Values (NA's)

Deleting of NA's was done with the R-function na.omit(x).

### 2.3.2 Low-Variance Filtering

Low variance filtering is performed to delete genes with a low variance in gene expression from the data set. It is performed to delete genes that are expressed the same in all cancer types (pancancer analysis) or the same in normal cells. To calculate the variance of the gene expression of a gene, the r-function var(x) is used and genes with a lower variance than a certain threshold value are removed.\ For the focused analysis the variance of the gene expression for each gene in tumor tissue was calculated. Genes with a variance beneath a certain threshold were deleted in the data sets of tumor and normal tissue.

### 2.3.3 Biotype Filtering

The biotype filtering was conducted for the pancancer data and the focussed analysis data. The biotype of each gene was determined (protein coding, RNA,...) and compared with the biotypes of pathways. To allow an appropriate comparison of the expression data and further reduce the data, only biotypes were kept that are available in the pathways. The biotype can be determined with the R-function checkbiotypes(x) from the package biomaRt **??**.

#### 2.3.3.1 Selection of metabolic pathways

(da eine hohe jaccard summe eine hohe überschneidung mit anderen pathways bedeutet. In einer heatmap sind hohe Jacccard indices weiß bis rot gefährbt. Ein niedriger Jaccard index ist blau gefärbt.)

To test for duplicate pathways in the selected metabolic pathways compared to the hallmark pathways and the compared tp the metabolic pathways themselves, the Jaccard index between to pathways were calculated.\ There were a few duplicates between the metabolic and Hallmark pathways. Those metabolic pathways with a high Jaccard index were discarded. The success of the cleaning was checked by again calculating the Jaccard index between the metabolic and the hallmark pathways. The values of the Jaccard index were then illustrated in a heatmap **??**. It can be assumed, that the selection of relevant pathways was successful because the pathways differ between each other. The number of metabolic pathways could be reduced from xxx to 600.

## 2.4 Descriptive analysis

### 2.4.1 Mean-variance plot

In a mean-variance plot the variance is plotted over the mean of expression values of the single genes across all patients. Thus, the variance and mean were calculated by the R-functions var(x) and mean(x). This is done to determine genes, which differ a lot in their expression levels across all patients. The plot is created with the package **??**

### 2.4.2 Violin plot

To check the distribution of a data set and compare it with other data sets violin plots are used. Based on how similar the violin plots are, it can be implied that the data is normalized. Violin plots are tilted and mirrored density plots of gene expression values. The y-axis shows the gene expression value and the x-axis shows the amount of genes with a certain gene expression value.

### 2.4.3 Jaccard-Index

The Jaccard-Index is a method to describe the similarity between two quantities. It is computed via dividing the union by the intersection. This is used to determine the degree in which metabolic pathways are similar to each other.

### 2.4.4 Volcano plot

A volcano plot is used to identify significantly differentially expressed genes. This is done to determine genes or pathways, which are up- or down- regulated in tumor tissue vs. normal tissue. The mean of each gene is calculated for normal and THCA tissue and used for the calculation of the Log2-Foldchange (Log2FC) in the following way, since the provided expression data is already log2 data:

$$log2FC = mean(normaltissue) - mean(tumortissue)$$

In the next step, a two-sided t-test was performed to determine the significance of a difference in expression.

To avoid the accumulation of type 1 errors, a bonferroni correction was performed. n is the number of genes in the cleaned data set for focused analysis:

$$\alpha = \frac{0.025}{n}$$

In the volcano plot the -log10 of the calculated p values is plotted against die Log2FC. Genes with a a lower p-value than the corrected alpha-value are significantly differently expressed. If the Log2FC is additionally higher than 0.1, the genes are significantly over expressed in tumor tissue, if the Log2FC is higher lower than -0.1, the genes are significantly under expressed in tumor tissue.

## 2.5 Data Reduction and Pathway Activities

### 2.5.1 PCA

The package xxx is used to perform the PCA. Therefore the data obtained from the GSEA was used. After performing the PCA, the results were plotted to visualize the different clusters.

The PCA was performed for pathway and gene activity. For analysis of the gen activity the package xxx was used. Dazu wurde noch analysiert, wie die Pathways auf die PCs verteilt sind.

### 2.5.2 UMAP

Like PCA, UMAP is a technique to reduce dimensions and to understand and visualize high dimensional data sets. Compared to PCA, UMAP better preserves the global structure and is much faster than other comparable techniques (for example t-SNE **xxx**).\ The algorithm starts by setting up a high-dimensional graph representation of the data. From each data point, a radius is extended and when two radii come into contact the points are connected. The radius is chosen individually for each point based on the distance to the nearest neighbor. The algorithm does not stop before every point is not connected at least to its closest neighbor.\ The resulting clustered high-dimensional graph is then optimized for a visualization in low-dimensions.\ Using this technique, the pan-cancer data is visualized.

### 2.5.3 GSEA

The GSEA is used to identify enriched pathways in tumor tissue. Next to the tumor tissue data, the THCA data includes also a normal tissue gene expression data frame which is used as a reference for activity comparison.

First, the log2FC is calculated for every gene of each each patient and is then ranked in a vector. This vector begins with the highest log2FC and ends with the lowest. A high log2FC implies that the this gene is higher expressed in tumor tissue compared to normal tissue in this particular patient.

Using the ranked log2FC vectors, the activity of each pathway for the patient is calculated. By iterating over every gene of the ranked vector, it was checked if it lies or does not lie in

a particular pathway. If a gene lies in the pathway, the log2FC value is summed up to a running sum. If the gene does not lie in the pathway, the log2FC value is subtracted from the running sum. Therefore, when a pathway is highly expressed compared to normal tissue, the the running sum scores a high value in the beginning and decreases to the end of the iteration. This results in a cumulative function that has a peak at a certain place. At this index of the ranked vector, the expression value of the corresponding gene is taken as the enrichment score of the analysed pathway and the patient belonging to the used vector. This process is then repeated for each pathway and each patient.

### 2.5.4 GSVA

Next to the GSEA, the GSVA is an approach to identify the pathway activities from gene expression data. Differently to the GSEA, it does not need a reference data frame to compare to.\ Hence, there was no expression data provided for comparison in the TCGA analysis, GSVA was used. There are various solutions to perform GSVA, one of them is performed by Hänzelmann et al xxx by following those five steps.\ For performing a GSVA, firstly the cumulative density distribution of a gene over all samples is estimated. Then the expression statistic of a gene in a sample based on the cumulative density distribution is calculated to bring all of the expression values to the same level. The third step is to rank the genes based on the expression statistic and to normalize the ranks with z-transformation. The last step is to compute the enrichment score based on the obtained ranked list. Therefore the Kolmogorov-Smirnov-like rank statistic is calculated for each gene set. That is used to calculate the enrichment score for each pathway in each patient, which is shown a heatmap. (Hänzelmann, Castelo, and Guinney 2013) xxx Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. 2013. "GSVA: Gene Set Variation Analysis for Microarray and Rna-Seq Data." Journal Article. BMC Bioinformatics 14 (1): 7. https://doi.org/10.1186/1471-2105-14-7.

### 2.5.5 Figure X

To identify pathways with the highest p-Value, obtained from GSVA and t-testing, a figure x is generated.

For generating figure x, the data from generating a volcano plot is used to identify the pathways, that are significantly over- or underexpressed based on the p-value. Pathways with a p-value smaller than 0.025 and a log2FC bigger than zero are significantly overexpressed, if the log2FC is smaller than zero, the pathways are significantly underexpressed.

In the next step, the pathways are ranked based on their p-value and the -log10(p-value) of each pathways is plotted against its rank. One plot is generated for overexpressed pathways and the other one for under expressed pathways.

### 2.5.6 Linear Regression

### 2.5.7 Neuronal Network

# 3 Results

## 3.1 Preprocessing

#### 3.1.0.1 Data cleaning

All dataframes were checked for NAs, which were subsequently deleted. Genes with a variance lower than 0.1 were removed to reduce dimensionality, as they contribute very little to the overall variance of the data set and are most likely house-keeping genes. Doing so, the number of genes in the pan-caner data set was reduced from 60,000 to approximately 19,000 genes.

The low-variance filtering of the THCA data set was done in a similar way. Genes with a lower variance than 0.06 were deleted in the tumor tissue and the normal tissue data. This resulted in a reduction from approximately 20,000 genes to 15,000 genes in both data frames.

#### 3.1.0.2 Biotype filtering

To reduce dimensionality further, we determined biotype of the hallmark pathway genes, which was almost exclusively protein coding. To match this, only protein coding pathways were kept in all expression data sets for further analysis.

#### 3.1.0.3 Pathway selection

The pathways from the MSigDB database were first aligned with the genes in our expression data. Only pathways with a coverage of over 99% were kept. To avoid biases during enrichment analysis jaccard indices between hallmark pathways and MSigDB pathways were computed and pathways with a high similarity were removed.

## 3.2 Descriptive analysis

#### 3.2.0.1 Mean-variance plot of TCGA expression data shows highly variant genes

To determine the genes from the TCGA expression data with a high variance, the variance was plotted over the mean (Figure @ref(fig:showmeanvariance)). Additionally those genes with a variance higher than 33 were labeled with their EnsembleID. The distribution of genes in this plot shows that the highly variant genes are around a log2 mean expression level of 0. The plot also shows, that very few genes are at a low mean expression level or at a very high mean expression level. Most genes are expressed across all patients at a log2 mean expression level of approximately 0. With this plot we were able to determine which genes differ significantly in their expression level across all cancer patients.

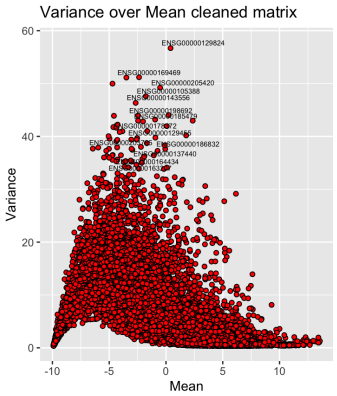


**Figure 3.1:** Mean-variance plot of cleaned TCGA expression data. Y-axis shows variance of a genes expression, x-axis shows mean of a genes expression

#### 3.2.0.2 Significantly up- and down regulated genes in THCA obtained from volcano plots

To determine those genes that are up- or down-regulated in THCA, the expression data from tumor tissue was compared to the data from normal tissue by mean log2 fold change. Associated p-Values were comuted with a Wilcoxon rank sum test. (Figure @ref(fig:showvolcanoplot)). The significance level adjusted to 1.755e-06 with a Bonferrioni adjustment. (!!!!! WICHTIG Welche sind up-regulated, welche down-regulated???)xxx

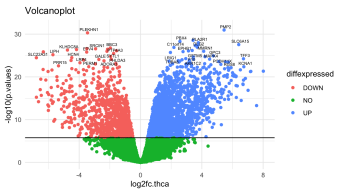


**Figure 3.2:** Volcano plot of THCA expression data

## 3.3 Pan cancer analysis

#### 3.3.0.1 GSVA of TCGA expression data reveals four clusters of cancer types

To find general clusters a heatmap with the mean expression of each gene in each tumor type was generated and clustered hierarchically. Figure @ref(fig:meanexp)

The tumor types were clustered based on their mean pathway activity and formed four clusters correlating with their histological type. The first cluster contains mainly adenocarcinamas, while the second one contains predominately glioblastomas. Leukemias are only found in the third cluster and the last cluster is enriched with sarcomas and carcinomas. Melanomas appear in the second and fourth cluster.

Furthermore, three observations were made regarding specific information about pathway activity.

Pathways, which are important for nucleus import and export like Nasopharygeal carcinoma (NPC) and Ran shuttle pathways, as well as pathways for transcription regulaturs in embryonic stem cells are down-regulated in glioblastoma and adenocarcinoma. However, these pathways are up-regulated in all other histological types. This seperation into two clusters is in line with the research of Ben-Porath et al., that shows an embryonic stem cell-like gene expression only in poorly differentiated tumors, such as leukemia An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors - PubMed (nih.gov). In that way it could be conluded that the differentation stage of a tumor correlates with pathway activty specific to certain histological types @ref(fig:meanexp).

Another observation is the clustering of glioblastoma. Pathways initiating neurogenesis https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8330525/ xxx and pathways linked to differentiation of the neural crest are up-regulated only in glioblastoma. Two other pathways, that are up-regulated in glioblastoma cells are pathways linked to the activity of

tyrosine kinases. The up-regulation of tyrosine kinases promote cell growth and proliferation. (Quelle: the cell) xxx. Taken together these two observations are in line with the expected high proliferation rate commonly found in glioblastoma.

The third cluster is mainly related to adenocarcinomas, more specifically liver hepatocellular carcinoma (LIHC), kidney renal papillary cell carcinoma (KICH) and kidney renal clear cell carcinoma (KIRC). The up-regulated pathways are involved in metabolism of carbohydrates, synthesis of lipids, synthesis of amino acids and detoxification. An up-regulation of all of these pathways may lead to cell growth and proliferation, due to higher metabolic activity, providing more biomass and energy.

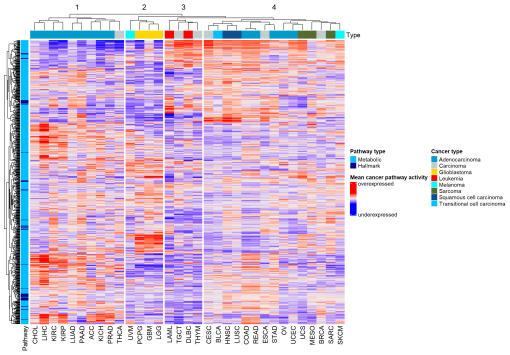


**Figure 3.3:** Mean expression of each pathway in each tumot type, annotated with pathway type, histological cancer type and clusters.

#### 3.3.0.2 PCA

#### 3.3.0.3 Dimension reduction of GSVA pan-cancer data reveals clusters in pathway activity

PCA was performed on GSVA pan-cacner data to provide uncorrelted varibales for better UMAP analysis. No apparent clustering was observed only in PCA data (compare Figure xxx supplementray materail). Subsequent UMAP analysis however, showed clear clusters for most cancer types. @ref(fig:UMAPPanType) @ref(fig:UMAPPanForm). This complemetns the results obtained from our heatmap and reassures, that the tumor types have characteristic pathway activities. However, some cancers cluster better with their histological type rather than tumor type. This was observed mainly for carcinomas like squamous cell carcinoma and transitional cell carcinoma, as well as sarcoma, lung adenocarcinoma and ovarian cancer. These are the same histological types that proofed difficult to cluster in the mean GSVA of TCGA expression. The UMAP confirmed the assumption, that the histological type of a tumor has a major impact on the patients gene expression profile.

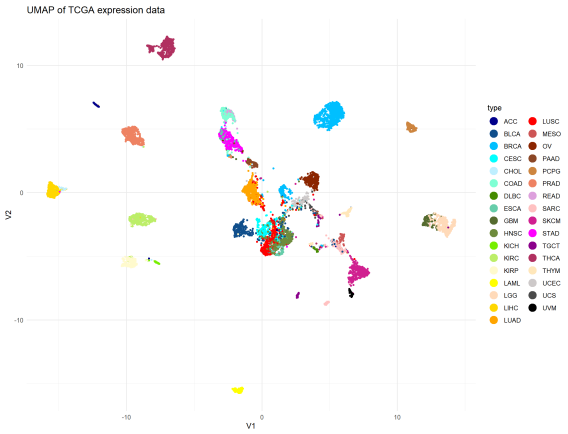


**Figure 3.4:** UMAP of TCGA expression data, colored by tumor type

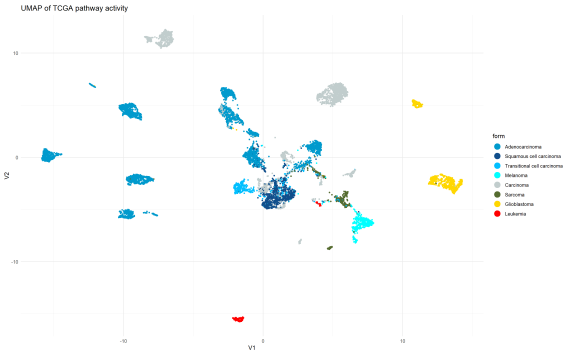


**Figure 3.5:** UMAP of TCGA expression data, colored by form of the tumor

The same analysis was performed for gene expression activity instead of pathway activity to check for reliability of the results. Similar clusters were observed, which confirms our results (see Fig. appendix)xxx.

## 3.4 Focused analysis

#### 3.4.0.1 GSVA on THCA expression data reveals pathways driving thyroid carcinogenesis.

To grasp a general overview of the differences in pathway activity between THCA and homeostatic thyroid tissue, GSVA was performed for the THCA expression data. Then, changes in pathway activity were computed by log2 fold change and the respective p-values were computed by a Wilcoxon rank-sum test. The most significantly altered pathways were then characterized. Most prominently among them were pathways linked to proliferative signaling such as upregulation of p53 inhibitory proteins and hedgehog pathway activating Gli proteins. Further, the alpha6beta4 integrin signaling pathway and associated pathways such as IL-36 signaling and Typ I hemidesmosome synthesis were significantly enhanced in THCA. These findings are consistent with previous studies that linked alpha6beta4 signaling to the development of aggressive forms of thyroid cancer. Quelle: Effect of beta 4 Integrin Knockdown by RNA Interference in Anaplastic Thyroid Carcinoma, Queen D, Ediriweera C and Liu L (2019) Function and Regulation of IL-36 Signaling in Inflammatory Diseases and Cancer Development. Front. Cell Dev. Biol. 7:317. doi: 10.3389/fcell.2019.00317, Msigdb Also, oncogenic signaling pathways commonly associated with different cancer types were significantly upregulated in THCAs. Among them, we observed ERBB2 QUELLE MSigDB and MST1 signaling commonly found in breast cancer. A role for MSP/Ron in breast cancer has recently been elucidated, wherein this pathway regulates tumor growth, angiogenesis, and metastasis. Kretschmann KL, Eyob H, Buys SS, Welm AL. The macrophage stimulating protein/Ron pathway as a potential therapeutic target to impede multiple mechanisms involved in breast cancer progression. Curr Drug Targets. 2010 Sep;11(9):1157-68. doi: 10.2174/138945010792006825. PMID: 20545605. Further, signaling through the EWSR1/FL1-fusion protein was significantly upregulated in THCA and previously shown to promote the rapid development of myeloid/erythroid leukemia in mice Quelle Msigdb. Lastly, THCAs showed downregulation of non-histone protein methylation. This process was identified as an import modulator of intracellular signaling by the MAPK, WNT, BMP, Hippo, and JAK/STAT pathways and might play an important role as a driver of carcinogenesis in THCA. Biggar, K., Li, SC. Non-histone protein methylation as a regulator of cellular signalling and function. Nat Rev Mol Cell

Biol 16, 5–17 (2015). https://doi.org/10.1038/nrm3915 Together these findings give a general overview of mechanisms driving carcinogenesis in THCA. However, no information about possible THCA subtypes or differences in patway activity between patients can be obtained from this data.

#### 3.4.0.2 Pan-cancer data GSVA reveals three subtypes of THCA altering in proliferative signaling.

To investigate potential subtypes of THCA, the respective samples were taken from the pan-cancer GSVA data. The optimal number of clusters was determined by an elbow plot and subsequent K-means clustering revealed a total of three subtypes in THCA. This is consistent with the three clusters of THCA observed in the full pan-cancer GSVA data. The follicular histological type was enriched in cluster B, with no tall cell types present in this cluster. Judging from histological type alone no difference in clusters A and C was observed. Most significant changes in pathway activity were observed in pathways concerning proliferative signaling. In comparison with all other tumor types, cluster A displayed high activity of RAS, JAK/STAT and EWSR1/FL1-fusion mediated signaling as well as elevated signatures associated with carcinogenesis driven by alpha6beta4 activity. In contrast, these pathways were downregulated in cluster B, with it showing elevated activity in mTOR, MAPK, PI3K, and EGFR signaling cascades. Cluster C was found to upregulate all the aforementioned forms of proliferative signaling. All clusters showed a homogenous upregulation of hedgehog, ERBB2, and MST1 pathway activity. Regarding immune response, cluster C showed no significant alterations in the respective hallmark pathways, however, these pathways were downregulated in both clusters A and B. With this data, we can identify two seemingly different forms of proliferative signaling driving carcinogenesis in THCA. These forms can either occur separately as in the case of clusters A and B or combined as for cluster C.

#### 3.4.0.3 THCA subtypes do not differ in their metabolism.

To investigate how the identified subtypes compare to homeostatic thyroid tissue, GSEA was performed for the THCA data. Consistent with the pan-cancer analysis of THCA data, k-means clustering obtained three different clusters in pathway activity – verified as the optimal number of clusters via an elbow plot. All clusters showed a similar change in metabolism. Katabolic pathways are downregulated whereas anabolic pathways e.g., fatty acid synthesis show increased activity in comparison with normal tissue. These changes in metabolic activity are in line with the Warburg effect. Further, the results

seem consistent with the proliferative signaling activities found previously. Alpha6beta4, RAS, JAK/STAT, and EWSR1/FL1-fusion mediated signaling is upregulated in clusters one and three with low expression in cluster two. However, the expected upregulation of mTOR, MAPK, PI3K, and EGFR signaling in clusters two and three was observed only in some samples. Regarding, immune response the expression profiles are again consistent with differences observed in the GSVA pan-cancer data: Both clusters one and two show a lower immune response compared to cluster three. From these GSEA results, we can conclude that the three subtypes of THCA differ in carcinogenesis and associated immune response but share a similar metabolism consistent with the Warburg effect.

## 3.5 Regression analysis

# 4 Discussion

In der PCA war nicht so ein gutes Ergebnis zu erkennen, weil nur die ersten 2 PCs verwendet wurden und dadurch nicht die gesamte Varianz erklärt wurde, durch die Verwendung von mehr PCs oder von anderen PCs (zB 2 und 3) könnte man evtl. eine bessere Darstellung erhalten. Bei verwendung aller PCs in der UMAP waren eindeutige Cluster zu erkennen allerdings ist UMAP auch nicht die optimale Methode, weil da die Abstände innerhalb der Cluster nicht proportional zu den tatsächlichen Abständen sind weil die mehrdimensionalen Daten irgendwie auf 2 Dimensionen runtergebrochen werden mussten.

Die Ergebnisse der UMAPs entsprechen der Erwartungen. Carcinomas haben alle ähnliche Expressionsmuster, da sie alle epithelialen Zellen entspringen und deshalb ähnliche genetische Mechanismen brauchen, um zu einer Tumorzelle zu werden. Die expressionsmuster zu anderen histological tumor types unterscheiden sich. Das liegt daran, dass verschiedene genetische Mechanismen zur Tumorentstehung führen, da sie anderen Zellen entspringen. Dadurch ist auch die unterschiedliche Genexpression zwischen den einzelnen Tumortypen, die zwar den gleichen histological type haben, aber sich in ihrem Tumortyp unterscheiden, wie zB die Adenocarcionma.

Die Ergebnisse der GSVA ...

GSVA mittlerer expressionswerte: dass hallmark pathways alle weiß sind entspricht unseren erwartungen

Outlook:

Epigeentic profiles = auch epigenetische veränderungen werden in die Expressionsdate mit inebezogen, das wäre sehr sinnvol für die ANalyse, wird hier aber nicht beachetet

Vergleich zwischen GSVA und GSEA –> weil einmal Vergleich mit tumor und normal

# 5 References

Alberts, J, B., and Walter, P (2015). Molecular biology of the cell, New York: Garland science.

Cabanillas, ME, McFadden, DG, and Durante, C (2016). Thyroid cancer. Lancet 388, 2783–2795.

Coca-Pelaz, A et al. (2020). Papillary thyroid cancer-aggressive variants and impact on management: A narrative review. Adv Ther 37, 3112–3128.

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

Kant, R, Davis, A, and Verma, V (2020). Thyroid nodules: Advances in evaluation and management. Am Fam Physician 102, 298–304.

Lin, JD (2007). Papillary thyroid carcinoma with lymph node metastases. Growth Factors 25, 41–49.

Prete, A, Borges de Souza, P, Censi, S, Muzza, M, Nucci, N, and Sponziello, M (2020). Update on fundamental mechanisms of thyroid cancer. Front Endocrinol (Lausanne) 11, 102.

Sharma, S, Quinn, D, Melenhorst, JJ, and Pruteanu-Malinici, I (2021). High-dimensional immune monitoring for chimeric antigen receptor t cell therapies. Current Hematologic Malignancy Reports 16, 112–116.

Tsibulnikov, S, Maslov, L, Voronkov, N, and Oeltgen, P (2020). Thyroid hormones and the mechanisms of adaptation to cold. Hormones (Athens) 19, 329–339.

# 6 Appendix

## 6.1 Plots

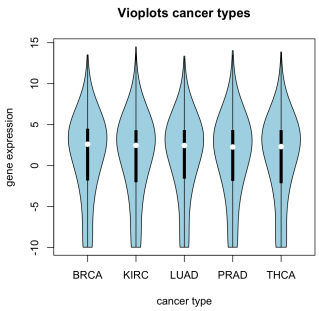


**Figure 6.1:** Mean-variance plot of cleaned TCGA expression data

## 6.2 Code

world

```
#createn einer liste mit allen patienten in dfs sortiert nach krebs
cancers = list();cancers = vector('list',length(table(tcga_anno$cancer_type_abbreviatio
names(cancers) = names(table(tcga_anno$cancer_type_abbreviation))
i=1
for (i in 1:length(cancers)){
  cancers[[i]] = tcga_exp_cleaned[,tcga_anno$cancer_type_abbreviation == names(cancers)
}
#function die einen krebstypen df und genesets als input nimmt und ein df mit pvalues a
enrichment = function(expressiondata, genesets = genesets_ids){
  ESmatrix = sapply(genesets, FUN = function(x){
    ins = na.omit(match(x,rownames(expressiondata)))#indices der gene im aktuellen set
    outs = -ins#indices der gene nicht im aktuellen set
```

```
    #gibt einen vektor der für jeden patienten den pval für das aktuelle gene enthält
    res = NULL
    for (i in 1:ncol(expressiondata)){#testet für jeden patienten
      res[i] = wilcox.test(expressiondata[ins,i],expressiondata[outs,i],'two.sided')$p.
    }
    return(res)
  })
  row.names(ESmatrix) = colnames(expressiondata); return(ESmatrix)
}
pvalueslist = lapply(cancers, enrichment)#für die tests für jeden krebstypen durch
```

```
get_top10pathways_from_pvalues = function(df_p_values, length_genesets) {

  require(ggplot2)

  results <- list()

  df_p_values_log10 <- -log10(as.data.frame(df_p_values))

  mean_pathway <- as.data.frame(apply(df_p_values_log10, 1, mean))
  rownames(mean_pathway) <- rownames(df_p_values_log10)

  ordered_score <- mean_pathway[order(-mean_pathway[ ,1]), 1]
  top_10 <- data.frame(ordered_score[1:10])
  colnames(top_10) <- "mean_pathway"

  ordered_names <- order(-mean_pathway[ ,1])
  top_10_names <- ordered_names[1:10]
  top_10$pathway_names <- row.names(mean_pathway)[top_10_names]

  results[[1]] <- top_10

  results[[2]] <- ggplot(data = top_10, aes(x = mean_pathway, y = reorder(pathway_names
    geom_bar(stat = "identity")+
    coord_cartesian(xlim =c(3, 3.75))+
    labs(title = names(df_p_values),
         x = "mean p-value pathway",
```

```
        y = "pathway name")

  pathway_size <- order(-mean_pathway[ ,1])
  top_10_size <- pathway_size[1:10]
  top_10$pathway_size <- length_genesets[top_10_size]

  results[[3]] <- ggplot(data = top_10, aes(x = mean_pathway, y = reorder(pathway_names
                                                                          mean_pathway)
    geom_point(aes(size = pathway_size))+
    labs(title = names(df_p_values),
         x = "mean p-value pathway",
         y = "pathway name")

  return(results)
}
```