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

In the recent years bioinformatic methods became a tool of utmost importance in medical research. To define specific genes and pathways in different cancer types or histological types pan-cancer analysis are done. A focused analysis is done to specify different subcategories within a certain cancer type and to identify targets for targeted therapy. The main methods in identifying up- or down-regulated pathways are GSEA and GSVA. GSVA of TCGA expression data reveals four clusters of cancer types, which are defined by different histological types like glioblastoma and adenocarcinoma. The histological types therefore seems to correlate with a specific set of pathways being especially enriched in certain cancer types. Furthermore, a GSVA of Thyroid cancer expression data shows that thyroid carcinogenisis is associated with the up-regulation of proliferative signalling pathways like the hedgehog pathway and alpha6beta4 integrin signaling pathway and associated pathways such as IL-36 signaling. It also showed the down-regulation of a pathway that is associated with an increased MAP-kinase acitivity. It is based on those proliferative signalling pathways that three subclusters form inside of the THCA patients from the pan-cancer data. One THCA subtype that could be linked to the follicular histological subtype is defined by increased mTOR and MAPK acitivity, while having low alpha6beta4 activity. In contrast another THCA subtype is defined by a low mTOR and MAPK activity, but a high alpha6beta4 activity. The third THCA subtype is linked to enhanced activity of both of these proliferative signalling pathways. These results promise better results is treatment, as a more percise diagnosis of the distinct THCA subtypse is possible To improve 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.

Thank You

1 Introduction

In 2019 230,000 cancer deaths were documented 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. For that it is beneficial to find a similarities in mutational causes across different tumors by using transcriptomic profiling methods like RNA-seq. In transcriptomic profiling all the RNA that has been generated by transcription of a cells DNA is sequenced (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 those are: resisting cell death, inducing angiogenesis, enabling replicative immortality, activating invasion and metastasis evading growth suppressors were the first detected hallmarks (Hanahan and Weinberg, 2011).

1.2 Histological tumor types

The observed tumors can be classified into different histological types. Carcinoma, which can be further subcategorized into adenocarcinomas, squamous cell carcinoma, transitional cell carcinoma. Carcioma derive from epithelial cells. Melanoma are skin tumors, sarcoma derive from connective or supportive tissue cells, glioblastoma are brain tumors and leukemia affect bloodcells (Alberts and Walter, 2015).

1.3 RNA-sequencing

RNA-sequencing (RNA-seq) is performed by cleaning of RNA, fragmentation, translation of RNA to cDNA, sequencing of cDNA and comparison with a 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 stem from data from the cancer genome atlas (TCGA).

1.4 Thyroid carcinoma

Thyroid carcinoma (THCA) incidence increased dramatically over the past few years (Cabanillas *et al.*, 2016). The main tasks of the thyroid gland are synthesizing hormones and regulating body temperate and metabolism (Tsibulnikov *et al.*, 2020). Most THCA derive from thyroid cells and result in the thyroid gland losing its function. 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 a tall cell variant (TCV) are subtypes of differentiated thyroid cancer (DTC). 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).

1.4.0.1 Integrin

Integrin is a cellular adhesion molecule, that binds to laminin in the extracellular matrix (Liberzon *et al.*, 2015). Together with other proteins they form hemidesmosomes. Thereby, integrin is essential for the integrity between cells. An important step in the development of malignant cancer is the invasion into healthy tissue. Thus, the detachment of the extracellular matrix from of the surrounding cells is essential.

1.5 Computational tools

1.5.1 Gene Set Enrichment Analysis

To analyse how the activity of a gene set differs between two sets of gene expression data, a Gene Set Enrichment Analysis (GSEA) is performed. For this, the genes in the expression data have to be ranked decreasingly by a certain metric. Such metrics can include the log2 fold change between the sample expression data and a reference set or the associated p-values for each gene. After ranking, a cumulative sum of all expression values in the ranked sample is computed. If a gene is present in the gene set to be analysed the expression value of that gene is added to the running sum. However, if the current gene does not lie in the gene set the value is subtracted. The extremum of this running sum is termed the enrichment score of the gene set. It is positive if the gene set is overexpressed in the sample compared to the reference data and negative vice versa. (Reimand *et al.*, 2019)

1.5.2 Gene Set Variation Analysis

The Gene Set Variation Analysis (GSVA) is performed with the same intention as the GSEA - to analyse the gene set activities in gene expression data. However, no reference data is required to successfully perform GSVA. There are various approaches to GSVA, one of them is performed by Hänzelmann *et al.* (2013)

1.5.3 Principal component analysis xxx QUELLE

A Principal component analysis (PCA) is used to alter the coordinates of a given dataset to its eigenvectors. This matrix rotation results in a new set of basis vectors called principal components (PCs) - the eigenvectors - that are orthogonal and show little correlation. Sorting the PCs by their associated eigenvalue, sthe PCs explaining the most variance can easily be identified, as they have the highest eigenvalue. By displaying the data set in a coordinate system span by the n most variant PCs, the dimensionalty of the dataset is reduced to $\mathbb{R}^n$ with the lowest loss in variance.

1.5.4 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. Compared to PCA, UMAP preserves the global structure of the data better and is much faster than other comparable techniques like t-SNE **Quelle: xxx 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 in the graph. The radius is chosen individually for each point based on the distance to the nearest neighbor. The algorithm goes on until k points are connected or n iterations are reached. The resulting clustered high-dimensional graph is then optimized for a visualization in low-dimensions. A disadvantage of UMAP is that although the overall structure is conserved, the distances between the individual points are not proportional to the real distance in the data set. This arises from the non-linear dimensional reduction. (Sharma *et al.*, 2021)

1.5.5 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.

1.6 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 contains clinical annotations concerning those patients and the last one contains hallmark pathways and their included genes. In the following analysis the data is cleaned by removing NAs, biotype filtering and low-variance filtering. After that a descriptive analysis is performed. After that a gene set variation analysis to detect significantly altered pathways compared to the other pathways in tumor tissue and a linear regression analysis is performed to predict pathway activity based on other pathways??? xxx Furthermore a neuronal network is built to improve prediction.

1.7 Focused analysis on THCA patients

An analysis of THCA patients is performed. This analysis is done on a data set containing the gene expression data 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. GSVA is performed on 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 other pathways based on thyroxine biosynthesis (nicht mehr!!!!) xxx. A better prediction can be achieved with a neuronal network.

1.8 Linear regression analysis

Linear regression is a statistical model that uses measurable values to predict an outcome. For this purpose, a linear function serves as basis to built the linear regression equation (Lunt, 2013). In gene expression data analysis a linear regression equation can be used to predict the activity of one gene (or pathway) by the activity of another.

1.8.1 Neural network

A neural network was performed using the package "neuralnet" (Fritsch *et al.*, 2019). In general, a deep learning network consists of an input layer, multiple hidden layers and an output layer (Riedmiller). Each one consisting of various neurons. The input layer contains as much neurons as input numbers are given for each sample. The output layer contains as much neurons as possible outputs. The number of neurons in each hidden layer and the number of hidden layers vary and will be determined for the best results. Based on the input numbers in the input layer, the number of each neuron of the next layer is determined, based on a linear regression model composed of the input data, weights, and bias.

$$\sum_{i=1}^{n} w_i x_i + bias$$

n is the number of input neurons and w the weight.

To obtain numbers in the range of 0 and 1, a min/max-scaling is performed on the input data. Furthermore, the optimal number of neurons per layer and the best random weights and biases for the first sample must be determined. This is done because some weights and biases may result in finding a local, but not global minimum of the cost function. After determining the random weights and bias, which resulted in a random output for the first sample, the cost function is calculated. To minimize the cost, resilient backpropagation with weight backtracking is used. Therefore, the gradient function of the cost function is determined. In resilient backpropagation, only the sign of the derivate is used, to avoid harmful effects of its magnitude. For minimizing the cost function, the ideal weights and biases are determined, based on the input and the expected output. (Theoretisch kann man hier ja noch schreiben, dass nicht nur das Erreichen des Minimums, sondern auch die Geschwindigkeit für das Erreichen des Minimums relevant sind) xxx. For the next samples those steps are repeated to reach the minimum of the cost function.

$$Costfunction = \frac{1}{2m} \sum_{i=1}^{m} (x - y)^2$$

m is the number of samples, y the output and x the expected output.

2 Materials and Methods

For the means of this project two separate analysis are performed: a pan-cancer analysis focusing on differences between cancer types and a focused analysis investigating THCA.

2.1 Description of the underlying data

For the analysis four data sets were provided. For pan-cancer analysis a gene expression data frame with normalized and log2 transformed bulk RNA-seq expression data for 60,489 genes in 9741 patients with 33 different forms of cancer was used. The data was derived from The Cancer Genome Atlas (TCGA). Complementing the TCGA expression data is an annotation dataframe with 37 clinical annotations regarding tumor type, tumor stage, gender, age, etc. for all patients.

The third object is a list containing five lists for the focused analysis, one list for each tumor type (BRCA, KIRC, LUAD, PRAD, THCA). For our focused analysis, the only the THCA data were used. The THCA list consists of three data frames: The first two contain normalized and log2 transformed bulk RNA-seq expression data for 19,624 genes in 59 THCA patients for carcinogenic and homeostatic tissue. The third dataframe complementes the data with the respective clinical annotations.

The last object contains 46 pathways associated with the hallmarks of cancer in form of a list of string vectors.

2.2 Metabolic pathway selection

To perform enrichment analysis later on, 6366 canonical pathways were selected from the Molecular Signatures Database (MSigDB)

$$@msigdb$$

with the `msigdbr::msigdbr()` function. As not to introduce a bias during enrichment analysis, the similarity of MSigDB pathways among themselves as well as with the hallmark pathways was computed with the Jaccard index. Pathways with a Jaccard index greater than the 1σ range were discarded.

2.3 Preprocessing of expression data

2.3.1 Data cleaning

All expression data were checked for missing values with the `na.omit()` function. Subsequently, low variance filtering was performed for TCGA and THCA tumor expression data. The variances of expression were computed for every gene across all samples and then, genes with variances blow a threshold were discarded to reduce dimensionality.

2.3.2 Biotype filtering

Next, biotype filtering was performed for pan-cancer and THCA expression data to reduce dimensionality further. Only genes sharing biotypes with the hallmark pathways were kept for the the following analysis. The biotypes of the genes were retrieved using the `biomart::getBM()` function from the biomaRt package (**biomart?**). To allow for an appropriate comparison within all pathways, only MSigDB pathways where over 99% of their respective genes were present in the filtered expression data were selected as final pathways.

2.4 Methods for descriptive analysis

2.4.1 Mean-variance plot

In a mean-variance plot the variance is plotted over the mean of expression values of single genes across all patients. Thus, the variance and mean were calculated for each gene in the THCA expression data. The final plot was created with the package ggplot2 **??**.

2.4.2 KANN RAUS m.M.n 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. To compute it, the intersection of all gene ENSEMBL-IDs from two compared pathways was divided by their union. We used this method to determine the degree in which pathways are similar to each other.

2.4.4 Volcano plot

A volcano plot is used to identify genes displaying significantly different expression in carcinogenic versus homeostatic tissues. First, the log2 fold change (Log2FC) is calculated for each gene across all samples in the THCA expression data in the following way:

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

Next, a two-sided t-test was performed with the `t.test()` function to determine the significance of a difference in expression. To avoid the accumulation of type one 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 lower p-value than the corrected significance level α are significantly differently expressed. If the Log2FC is additionally positive, the genes are significantly overexpressed in tumor tissue, if the Log2FC is negative, the genes are significantly underexpressed in tumor tissue.

2.5 Dimension reduction and pathway enrichment analysis

The GSEA was used to identify enriched pathways in THCA tumor tissue. Here, GSEA was performed with the package "fgsea" (**fgsea?**). First, the expression values were ranked decreasingly by log2FC for every patient. Log2FC was chosen as the ranking metric as it easy to compute and shows a high sensitivity. **xxx Quelle: Ranking metrics in gene set enrichment analysis: do they matter?** Secondly, using the ranked log2FC vectors, the enrichment score of each pathway was calculated for each patient with the `fgseamultilevel()` function.

As no normal tissue reference data was provided for the TCGA expression data, pathway activities were computed via GSVA. The analysis was performed with the `gsva()` function from the "GSVA" package (**gsva?**). To give a general overview over the differences in expression of THCA and homeostatic thyroid tissue GSVA, the THCA expression data were also analysed by GSVA. To do so, tumor and normal expression data were combined into a singular dataframe of which enrichment scores were computed with `gsva()`. Than, the GSVA data was split again and the log2FC between the two matrices was computed and taken as pathway activity.

PCA was performed to provide an uncorrelated dataset for the subsequent UMAP. For the TCGA GSVA pathway activity data the `prcomp()` function was used. To verify the results, PCA was performed on TCGA expression data, as well. In this case `Seurat::RunPCA()` from the Seurat package was used to minimized computation times. **xxx Quelle: seurat package**

UMAP analysis was used to identify and visualize clusters in TCGA GSVA and expression data. This was achieved with the `umap()` function from the package "umap"

$$@umap$$

running on all PCs from TCGA GSVA ans expression data.

2.6 Regression analysis

2.6.1 Linear Regression

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

2.6.2 Neuronal Network

A neural network was used to predict the activity of REACTOME_INTERLEUKIN_36_PATHWAY based on the activity of other pathways. Therefore, the network was trained with the pathway activity of 45 xxx patients from the THCA data for focused analysis. The other 15 patients were used to validate the network, obtaining a mean squared error (MSE) value, to evaluate the precision of the network.

For identification of the best initial conditions, 25 different networks are generated, each one with 2 hidden layers and different combinations of neurons per layer. For each combination the MSE is calculated and the 3 combinations with the lowest MSE are selected for selection of the best initial conditions regarding the weights and biases. For each of the 3 networks 100 random initial conditions are tested, resulting in one network with the lowest MSE.

2.7 Packages

```
## Warning: Paket 'readxl' wurde unter R Version 4.1.3 erstellt
```

Table 2.2: Packages used in the analysis.

Package	Localisation	Usage	Link
biomaRt	pre_02, pre_03, pre_05	renaming the genenames from the hallmarkpathways-dataframe into ensembleIDs	https://bioconductor.org/packages/release/bioc/html/biomaRt.html

Package	Localisation	Usage	Link
msigdbr	pre_03	downloading all of the canonical pathways and the genes which they include in homo sapiens from the msigbdr data base	https://bioconductor.org/packages/release/data/experiment/html/msigdb.html
dplyr	pre_04, pre_05	tidying and manipulating of dataframes	https://cran.r-project.org/web/packages/dplyr/index.html
ggplot2	pre_04, pre_05, descr_03, descr_04, THCA_01, THCA_02, pan_02, pan_04	allows for the creation of plots with more detailed options	https://cran.r-project.org/web/packages/ggplot2/index.html
pheatmap	descr_01, pan_01, neu_02, neu_04	allows for the creation of heatmaps with more detailed options	https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf
vioplot	descr_02	creation of violinplots	https://cran.r-project.org/web/packages/vioplot/index.html
VennDiagram	descr_05	creation of VENN-diagrams	https://cran.r-project.org/web/packages/VennDiagram/VennDiagram.pdf
dplyr	THCA_01, pan_01	NA	NA
fgsea	THCA_01, pan_01	to do a GSEA	https://bioconductor.org/packages/release/bioc/html/fgsea.html
GSVA	THCA_01, pan_03	to do a GSVA	https://bioconductor.org/packages/release/bioc/html/GSVA.html
ComplexHeatmap	THCA_01, pan_03, pan_04	allows for the creation of heatmaps with more detailed options	https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html

Package	Localisation	Usage	Link
metaplot	THCA_02, pan_02, pan_04	data-driven plots	https://cran.r-project.org/web/packages/metaplot/index.html
gridExtra	THCA_02, pan_02, pan_04	implementation of “grid” graphics	https://cran.r-project.org/web/packages/gridExtra/index.html
umap	THCA_02, pan_02, pan_04	to do a UMAP	https://cran.r-project.org/web/packages/umap/index.html
gage	pan_01	application of GSEA	https://bioconductor.org/packages/release/bioc/html/gage.html
psych	pan_02	iterative factor analysis	https://cran.r-project.org/web/packages/psych/index.html
cluster	pan_04	cluster analysis	https://cran.r-project.org/web/packages/cluster/cluster.pdf
MASS	neu_00	implementation of neural network	https://cran.r-project.org/web/packages/MASS/index.html
neuralnet	neu_03	training of neural networks	https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf
AnnotationDbi	diff_03	translating ensemble ids into gennames	https://bioconductor.org/packages/release/bioc/html/AnnotationDbi.html
org.Hs.eg.db	diff_03	translating ensemble ids into gennames	https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html

Table 2.1: Packages used in the analysis.

Package	Localisation	Usage	Link
biomart	pre_02, pre_03, pre_05	renaming the genenames from the hallmarkpathways-dataframe into ensembleIDs	https://bioconductor.org/pac
msigdbr	pre_03	downloading all of the canonical pathways and the genes which they include in homo sapiens from the msigbdr data base	https://bioconductor.org/pac
dplyr	pre_04, pre_05	tidying and manipulating of dataframes	https://cran.r-project.org/web/packages/dp
ggplot2	pre_04, pre_05, descr_03, descr_04, THCA_01, THCA_02, pan_02, pan_04	allows for the creation of plots with more detailed options	https://cran.r-project.org/web/packages/gg
pheatmap	descr_01, pan_01, neu_02, neu_04	allows for the creation of heatmaps with more detailed options	https://cran.r-project.org/web/packages/ph
vioplot	descr_02	creation of violinplots	https://cran.r-project.org/web/packages/vio
VennDiagram	descr_05	creation of VENN-diagrams	https://cran.r-project.org/web/packages/Ve
dplyr	THCA_01, pan_01		
fgsea	THCA_01, pan_01	to do a GSEA	https://bioconductor.org/pac
GSVA	THCA_01, pan_03	to do a GSVA	https://bioconductor.org/pac
ComplexHeatmap	THCA_01, pan_03, pan_04	allows for the creation of heatmaps with more detailed options	https://bioconductor.org/pac
metaplot	THCA_02, pan_02, pan_04	data-driven plots	https://cran.r-project.org/web/packages/me
gridExtra	THCA_02, pan_02, pan_04	"implementation of ""grid"" graphics "	https://cran.r-project.org/web/packages/gri
umap	THCA_02, pan_02, pan_04	to do a UMAP	https://cran.r-project.org/web/packages/um
gage	pan_01	application of GSEA	https://bioconductor.org/pac
psych	pan_02	iterative factor analysis	https://cran.r-project.org/web/packages/psy
cluster	pan_04	cluster analysis	https://cran.r-project.org/web/packages/clu
MASS	neu_00	implementation of neural network	https://cran.r-project.org/web/packages/MA
neuralnet	neu_03	training of neural networks	https://cran.r-project.org/web/packages/neu
AnnotationDbi	descr_03	translating ensemble ids into gennames	https://bioconductor.org/pac
org.Hs.eg.db	descr_03	translating ensemble ids into gennames	https://bioconductor.org/pac

Table 2.3: Packages used in the analysis.

Package	Localisation
biomart	pre_02, pre_03, pre_05
msigdbr	pre_03
dplyr	pre_04, pre_05
ggplot2	pre_04, pre_05, descr_03, descr_04, THCA_01, THCA_02, pan_02, pan_04
pheatmap	descr_01, pan_01, neu_02, neu_04
vioplot	descr_02
VennDiagram	descr_05
dplyr	THCA_01, pan_01
fgsea	THCA_01, pan_01
GSVA	THCA_01, pan_03
ComplexHeatmap	THCA_01, pan_03, pan_04
metaplot	THCA_02, pan_02, pan_04
gridExtra	THCA_02, pan_02, pan_04
umap	THCA_02, pan_02, pan_04
gage	pan_01
psych	pan_02
cluster	pan_04
MASS	neu_00
neuralnet	neu_03
AnnotationDbi	descr_03
org.Hs.eg.db	descr_03

3 Discussion

3.0.1 Pan-cancer analysis

Our findings from pan-cancer expression data show promising results. Via GSVA analysis we identified four clusters in the cancer types correlating strongly to the associated histological type. Glioblastoma seem to take a special role as they are predominantly characterized by the high activity of neural crest differentiation pathways and receptor tyrosine kinases. This is in line with previous studies showing that glioblastomas derive from neural crest cells.

xxx cite(Bednarczyk, J.L., McIntyre, B.W. Expression and ligand-binding function of the integrin a4b1 (VLA-4) on neural-crest-derived tumor cell lines. Clin Exp Metast 10, 281–290 (1992). https://doi.org/10.1007/BF00133564).

This was also found for some melanoma like UVM, which explains the observed clustering of UVM with other glioblastoma. Also, the high receptor tyrosine kinase activity has been linked to the formation of UVM and glioblastoma and suggested as a possible target for therapy.

xxx cite Jo, D.H., Kim, J.H. & Kim, J.H. Targeting tyrosine kinases for treatment of ocular tumors. Arch. Pharm. Res. 42, 305–318 (2019). https://doi.org/10.1007/s12272-018-1094-3 https://doi.org/10.1111/febs.12109

Further, especially liver and kidney adenocarcinoma seemed to form a strong subcluster within the other adenocarcinoma. They are characterized by exceptionally high activity of metabolic pathways such as carbohydrate metabolism, lipid, and amino acid synthesis. Again, this change in metabolism was previously found in hepatocellular carcinoma.

xxx cite Sangineto M, Villani R, Cavallone F, Romano A, Loizzi D, Serviddio G. Lipid Metabolism in Development and Progression of Hepatocellular Carcinoma. Cancers. 2020; 12(6):1419. https://doi.org/10.3390/cancers12061419

The most significant classification we found was the clustering of tumor types by their differentiation stage. Poorly differentiated tumors like leukemia and squamous cell carcinoma show an upregulation of pathways associated with embryonic stem cell-like expression signatures. In contrast highly differentiated tumors like most adenocarcinoma as well as most glioblastoma underexpress these gene sets. Such a clustering by differentiation stage was previously described by Ben-Porath et al.. However, these findings cannot be verified directly as provided annotation data did not contain information regarding the differentiation stage.

xxx cite Ben-Porath, I., Thomson, M., Carey, V. et al. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet 40, 499–507 (2008). https://doi.org/10.1038/ng.12

Taken together our results are in line with current research and allow for the following hypothesis: The expression profile of a given cancer type depends highly on its differentiation stage and its histological type but little on the actual tumor type itself. Understanding how these changes in expression link to mutational signatures might help in developing druggable targets for therapy.

3.0.2 THCA focused analysis

From our GSEA and pan-cancer GSVA results, we identify two separate ways of carcinogenesis in THCA. The follicular subtype upregulates proliferative signaling through mTOR/PI3K and MAPK signaling pathways. This was previously shown by Furuya et al.

Quelle: (Fumihiko Furuya, Changxue Lu, Mark C. Willingham, Sheue-yann Cheng, Inhibition of phosphatidylinositol 3-kinase delays tumor progression and blocks metastatic spread in a mouse model of thyroid cancer, *Carcinogenesis*, Volume 28, Issue 12, December 2007, Pages 2451–2458, https://doi.org/10.1093/carcin/bgm174)

A second way of carcinogenesis by signaling through alpha6beta4, RAS, JAK/STAT, and EWSR1/FLI1-fusion mediated pathways was observed in the data. This way of carcinogenesis was linked to non-follicular types of THCA.

Quelle: Effect of beta 4 Integrin Knockdown by RNA Interference in Anaplastic Thyroid Carcinoma

Quelle: https://doi.org/10.1002/jcp.27199

Quelle: https://doi.org/10.1007/s00428-017-2095-1

Pan-cancer GSVA shows three distinct clusters in the expression data, upregulating either one or both ways of proliferative signaling. While the follicular subtype seemed to strongly correlate with one cluster, a similar process was not observed in tall-cell and classical phenotypes. With more detailed annotation data it might be possible to link anaplastic and papillary histological subtypes of THCA to the two yet unassigned clusters.

Despite differences in proliferative signaling, all clusters share an upregulated hedgehog signaling pathway which is consistent with the literature.

Quelle (https://doi.org/10.1007/s12020-013-0015-y)

Also, metabolic changes in line with the Warburg effect were observed in all clusters.

3.0.3 Regression of IL-36 pathway activity

Our data suggest that our neuronal network is well suited to predict pathway activities from GSEA data. The model shows an excellent fit to data and produces only minor errors. However, both linear models struggle in predicting the data accurately. This might be since GSEA pathway activity data usually clusters into an up- and downregulated group with no values in between. Since the REACTOME_INTERLEUKIN_36_PATHWAY also shows this problem, the two clusters might produce larger correlation values that might impact the accuracy of the regression coefficients and the intercept. Secondly, the correlation of the residuals with the test data values did not approach zero, thus, our linearity assumption is not met. Therefore, it can be concluded that a linear regression model is not well suited to predict the REACTOME_INTERLEUKIN_36_PATHWAY activity accurately.

4 Outlook

Futher ways of analysis could be the prediction of the histological type of THCA as well as the way of carcinogenesis with a neuronal network. This might be possible with a larger training data set as well as more detailed and specified annotations. Furthermore, it might be possible to link whole genome sequencing and methylation data to pathway activity. In that way, one could suggest a suitable targeted therapy option for a THCA patient based only on sequencing data from a small biopsy sample.

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.

Fritsch, S, Guenther, F, and Wright, MN (2019). Neuralnet: Training of neural networks.

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

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

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

Liberzon, A, Birger, C, Thorvaldsdóttir, H, Ghandi, M, Mesirov, JP, and Tamayo, P (2015). The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst 1, 417–425.

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

Lunt, M (2013). Introduction to statistical modelling: Linear regression. Rheumatology 54, 1137–1140.

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.

Reimand, J et al. (2019). Pathway enrichment analysis and visualization of omics data using g:profiler, GSEA, cytoscape and EnrichmentMap. Nature Protocols 14, 482–517.

Riedmiller, MA Rprop - description and implementation details.

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

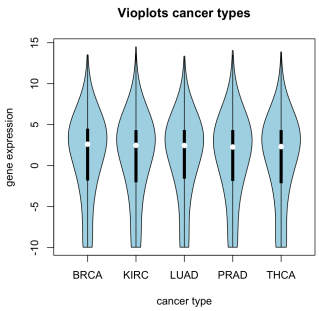

Figure 6.1: Mean-variance plot of cleaned TCGA expression data

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)
}
```