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

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).

1.2 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).

1.3 RNA-sequencing

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).

1.4 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).

1.4.0.1 Integrin

xxx

1.5 Computational tools

1.5.1 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 (Reimand *et al.*, 2019). 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).

1.5.2 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 (Hänzelmann *et al.*, 2013). The difference to the GSEA is that there no reference expression data is to perform the GSVA.

GSVA is performed with the package xxx.

1.5.3 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). 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.

1.5.4 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.

1.5.5 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 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.

1.7 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.

1.8 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, constructing 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

Implementing a neural network was performed with the package neuralnet xxx.

In general, a deep learning network consists of a 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 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, 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).

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 Material and Methods

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

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 (Durinck *et al.*, 2005).

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 **??** xxx.

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. UMAP is performed with the package (Konopka, 2022).

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.* (2013) 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 *et al.*, 2013).

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

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

2.5.7 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.

2.6 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
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

Package	Localisation	Usage	Link
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
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	disc_03	translating ensemble ids into gennames	https://bioconductor.org/packages/release/bioc/html/AnnotationDbi.html

Package	Localisation	Usage	Link
org.Hs.eg.db_3.0	org.Hs.egsdb	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

3 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.

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.

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.

Konopka, T (2022). Umap: Uniform manifold approximation and projection.

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.

4 Appendix

4.1 Plots

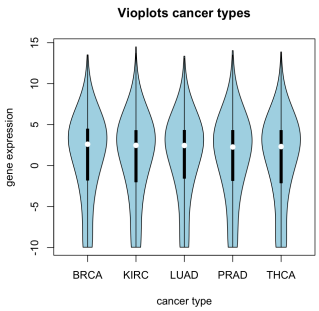

Figure 4.1: Mean-variance plot of cleaned TCGA expression data

4.2 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)
}
```