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.

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.

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.

Preprocessing of expression data

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.

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.

Methods for descriptive analysis

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 **??**.

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.

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.

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.

Dimension reduction and pathway enrichment analysis

Principle Component Analysis (PCA)

For the PCA, the data obtained by the GSEA (tumor vs normal) or by GSVA (pan cancer) was used. After performing the PCA, the results were plotted to visualize the different clusters.

The PCA was performed for analyzing pathway and gene activity. PCA is performed by the built in r-function prcomp().

Uniform Manifold Approximation and Projection for Dimension Reduction (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 [**?**]). 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 "umap" [@umap].

Gene Set Enrichment Analysis (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.

GSEA is performed with the package "fgsea" [@fgsea]. 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.

Geneset Variation Analysis (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 @GSVA 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 [@GSVA].

GSVA is performed with the package "GSVA" [@gsva].

KANN RAUS m.M.n 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.

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

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.

Packages

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

Table 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

Package	Localisation	Usage	Link
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
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	descr_03	translating ensemble ids into gennames	https://bioconductor.org/packages/release/bioc/html/AnnotationDbi.html

Package	Localisation	Usage	Link
org.Hs.eg.db	chr_03	translating ensemble ids into gennames	https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html

Table 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/packages/release
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
dplyr	pre_04, pre_05	tidying and manipulating of dataframes	https://cran.r-project.org/web/packages/dplyr/index.htm
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.h
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/phea
vioplot	descr_02	creation of violinplots	https://cran.r-project.org/web/packages/vioplot/index.h
VennDiagram	descr_05	creation of VENN-diagrams	https://cran.r-project.org/web/packages/VennDiagram/V
dplyr	THCA_01, pan_01		
fgsea	THCA_01, pan_01	to do a GSEA	https://bioconductor.org/packages/release
GSVA	THCA_01, pan_03	to do a GSVA	https://bioconductor.org/packages/release
ComplexHeatmap	THCA_01, pan_03, pan_04	allows for the creation of heatmaps with more detailed options	https://bioconductor.org/packages/release
metaplot	THCA_02, pan_02, pan_04	data-driven plots	https://cran.r-project.org/web/packages/metaplot/index
gridExtra	THCA_02, pan_02, pan_04	"implementation of ""grid"" graphics "	https://cran.r-project.org/web/packages/gridExtra/index
umap	THCA_02, pan_02, pan_04	to do a UMAP	https://cran.r-project.org/web/packages/umap/index.htm
gage	pan_01	application of GSEA	https://bioconductor.org/packages/release
psych	pan_02	iterative factor analysis	https://cran.r-project.org/web/packages/psych/index.htm
cluster	pan_04	cluster analysis	https://cran.r-project.org/web/packages/cluster/cluster.p
MASS	neu_00	implementation of neural network	https://cran.r-project.org/web/packages/MASS/index.ht
neuralnet	neu_03	training of neural networks	https://cran.r-project.org/web/packages/neuralnet/neura
AnnotationDbi	descr_03	translating ensemble ids into gennames	https://bioconductor.org/packages/release
org.Hs.eg.db	descr_03	translating ensemble ids into gennames	https://bioconductor.org/packages/release

Table 3: Packages used in the analysis.

Package	Localisation	Usage
biomart	pre_02, pre_03, pre_05	renaming the genen
msigdbr	pre_03	downloading all of t
dplyr	pre_04, pre_05	tidying and manipu
ggplot2	pre_04, pre_05, descr_03, descr_04, THCA_01, THCA_02, pan_02, pan_04	allows for the creati
pheatmap	descr_01, pan_01, neu_02, neu_04	allows for the creati
vioplot	descr_02	creation of violinplo
VennDiagram	descr_05	creation of VENN-d
dplyr	THCA_01, pan_01	NA
fgsea	THCA_01, pan_01	to do a GSEA
GSVA	THCA_01, pan_03	to do a GSVA
ComplexHeatmap	THCA_01, pan_03, pan_04	allows for the creati
metaplot	THCA_02, pan_02, pan_04	data-driven plots
gridExtra	THCA_02, pan_02, pan_04	implementation of "
umap	THCA_02, pan_02, pan_04	to do a UMAP
gage	pan_01	application of GSE
psych	pan_02	iterative factor anal
cluster	pan_04	cluster analysis
MASS	neu_00	implementation of n
neuralnet	neu_03	training of neural ne
AnnotationDbi	descr_03	translating ensembl
org.Hs.eg.db	descr_03	translating ensembl