Material and Methods

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

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?

Metabolic pathway selection

From the Molecular Signature Database (MSigDB) [**?**] 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.

Preprocessing

Deleting Not Available Values (NA's)

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

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.

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 [@biomaRt].

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.

Descriptive analysis

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.

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. 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.

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.

Data Reduction and Pathway Activities

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.

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.

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.

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.

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

Neuronal Network

Packages Tabelle einfügen! –> sorry: habe leider die zeile gelöscht, hoffe du weißt noch, was du gemacht hast :(