Introduction

Biological background

In 2019 230,000 cancer deaths were documented in Germany[1]. 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 [@cell].

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 [@cancer_hallmarks].

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 [@cell].

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 [@cell]. The results from RNA-seq used for the analysis stem from data from the cancer genome atlas (TCGA).

Thyroid carcinoma (THCA) incidence increased dramatically over the past few years [@THCA]. The main tasks of the thyroid gland are synthesizing hormones and regulating body temperate and metabolism [@THgland1]. 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 [@THCA2]. Regarding the presented DTCs, PTCs have the best clinical prognosis [@PCA1], while TCV cancers have the worst clinical outcome [@PCA_aggressive]. 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 [@PCA3].

Integrin is a cellular adhesion molecule, that binds to laminin in the extracellular matrix [@integrin1]. 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 and alterations of integrin are very common in cancer cells [@Integrin2].

Computational tools

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. [@GSEA]

[1] https://www.krebsinformationsdienst.de/tumorarten/grundlagen/krebsstatistiken.php

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 [@GSVA] by following five steps. First, 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. Lastly, the enrichment score is computed based on the obtained ranked list by calculating the Kolmogorov-Smirnov-like rank statistic for each gene set. [@GSVA]

A Principal component analysis (PCA) xxx QUELLE 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.

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 [@tSNE]. 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. [@UMAP]

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.

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 [@lm]. 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.

A neural network was performed using the package "neuralnet" [@neuralnet]. In general, a deep learning network consists of an input layer, multiple hidden layers and an output layer [@neuronal]. 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.

The 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.

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.