Introduction

In 2019 230,000 cancer deaths were documented in Germany xxx

$[Krebsrate und Krebs-Sterberate in Deutschland (krebsinformationsdienst.de)](https://www.krebsinformationsdienst.d$

. 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].

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 [@cancer_hallmarks].

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 [@cell].

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 [@cell].

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

Thyroid carcinoma

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 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.

Computational tools

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 [@GSEA]. For this, two conditions with replicates are compared, requiring a reference of normal expression data. First, a gene list is defined. Then the statistically enriched pathways are identified and lastly, the results are visualized.

Gene Set Variation Analysis

The Gene Set Variation Analysis (GSVA) is performed with the same intention as the GSEA - 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 [@GSVA]. The difference to the GSEA is that there is no reference expression data required to perform the GSVA.

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 [@UMAP]. 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. Thereby, the identification of the clusters is a lot easier and the UMAP keeps the overall structure of the data set, which makes clustering easier. The disadvantage 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.

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 variance in the first PCs.

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 tissue and normal tissue.

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.

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.

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.

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 [@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.

Neural network

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.