Materials and Methods

In the course of this project two separate analysis are performed: a pan-cancer analysis focusing on differences between cancer types and a focused analysis investigating THCA.

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 data frame with 37 clinical annotations regarding tumor type, tumor stage, gender, age, etc. for all patients.

The third piece of data 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, 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 data frame complements 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.

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

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 below a threshold were discarded to reduce dimensionality.

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 in which over 99% of their respective genes were present in the filtered expression data were selected as final pathways.

Methods for descriptive analysis

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 **??**.

Jaccard index is a method to describe the similarity between two quantities. To compute it, the intersection of all gene EnsembleIDs 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.

A volcano plot is used to identify genes displaying significantly different expression in cancerous 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 the 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

The GSEA was used to identify enriched pathways in THCA tumor tissue. Here, GSEA was performed with the package "fgsea" [@fgsea]. First, the expression values were ranked in decreasing order by Log2FC for every patient. Log2FC was chosen as the ranking metric as it is easy to compute and shows a high sensitivity. **xxx Quelle: Ranking metrics in gene set enrichment analysis: do they matter?** Secondly, using the ranked Log2FC vectors, the enrichment score of each pathway was calculated for each patient with the `fgseamultilevel()` function.

As no normal tissue reference data was provided for the TCGA expression data, pathway activities were computed via GSVA. The analysis was performed with the `gsva()` function from the "GSVA" package [@gsva]. To give a general overview over the differences in expression of THCA and homeostatic thyroid tissue GSVA, the THCA expression data were also analysed by GSVA. To do so, tumor and normal expression data were combined into a singular data frame of which enrichment scores were computed with `gsva()`. Then, the GSVA data was split again and the log2FC between the two matrices was computed and taken as pathway activity.

PCA was performed to provide an uncorrelated data set for the subsequent UMAP. For the TCGA GSVA pathway activity data the `prcomp()` function was used. To verify the results, PCA was performed on TCGA expression data, as well. In this case `Seurat::RunPCA()` from the Seurat package was used to minimize computation time [@seurat].

UMAP analysis was used to identify and visualize clusters in TCGA GSVA and expression data. This was achieved with the `umap()` function from the package "umap" [@umap] running on all PCs from TCGA GSVA ans expression data.

Regression analysis

For THCA pathway activity regression analysis a highly variant and significantly altered pathway was selected. To prepare the data appropriately the THCA GSEA data set was divided into a training and test data set containing 44 and 15 samples respectively. A linear regression analysis was performed on the training data with the `glm()` function. To do so, the correlation of all pathways was computed and pathways with high correlations are omitted. Subsequently, the 10% of most variant pathways are selected as variables for the regression model. A second model was introduced by computing the p-values of all coefficients and selecting only those pathways contributing significantly to the model were kept.

A neural network was implemented to predict the pathway activity using the `neuralnet()` function from the "neuralnet" package [@neuralnet]. For identification of the best initial conditions, 25 different networks are generated, each with 2 hidden layers and different combinations of neurons per layer. For each combination the networked was trained on the min-max-scaled training data and the best network was determined by the lowest mean squared error (MSE) in the test data.

Environment

The R version 4.0.1 was used, the table of used packages is attached in the appendix (see table @ref(tab:packagesused)).