Ruprecht-Karls-Universität Heidelberg
Fakultät für Biowissenschaften
Bachelorstudiengang Molekulare Biotechnologie

ksdflsdjf
sfdsfd
sfsf

Data Science Project SoSe 2022

Autoren Anna von Bachmann, Linda Blaier, Maja Glotz, Tim Wenzel
Heidelberg 18.07.2022

1 Abstract

Contents

2 List of abbreviations

3 Introduction

Cancer is one of the most prevalent diseases in the world and responsible for almost 10 million deaths each year (Ferlay *et al.*, 2021). Therefore, it is a big interest to further understand the characteristics of cancer in more detail hoping to find new treatment options. Pan cancer analysis aims to find similarities and differences across different tumor types. These types of analysis are often based on The Cancer Genome Atlas (TCGA) which is a comprehensive genome sequencing database consisting of bulk samples of over 11 000 tumors from the 33 most prevalent forms of cancer(Cooper *et al.*, 2018).

It is known that cancer cells have certain characteristics that distinguish them from normal cells, the so-called hallmarks of cancer. These enable them to evade human protection mechanisms and sustain uncontrolled cell proliferation. Cancer hallmarks often are a consequence of altered gene expression patterns (Hanahan and Robert, 2011). Among other things in this project, we want to identify and compare pathways with altered gene expression who are primarily responsible for establishing the cancer hallmarks across the 33 tumor types selected in the TCGA.

The activation of oncogenes and the loss of tumor suppressor gene function are the main cause for tumor development. Both can lead to reprogramming of cellular metabolism. This way cancer cells can cover their high needs for nutrients even in a nutrient-poor environment. For example, cancer cells change their energy metabolism to deregulate glucose and amino acid uptake. They also use anaerob glycolysis as preferred way of generating energy which among other things like the use of intermediates of Glycolysis and TCA Cycle leads to increased biomass needed for proliferation (Warburg *et al.*, 1927; Natalya and Craig, 2016). Therefore, we also did a pan cancer analysis on metabolic pathways for which we used pathways curated by the Kyoto Encyclopedia of Genes and Genomes (KEGG).

Beyond the tumor cells themselves, a tumor type is also characterized by its microenvironment. Dependent on the composition of immune cells in the microenvironment, it can either promote or suppress tumorigenesis. For instance, Cytotoxic T cells (CD8+) are usually associated with a positive prognosis while macrophages are associated with a poor prognosis

for the cancer patient (Anderson and Simon, 2020). To investigate immune infiltration, we used a gene set from the Pathway interaction database (PID). The respective pathways mainly concern signalling especially regarding cell cycle and immune response.

Kidney cancer is one of the most common kinds of cancer. Most kidney cancer patients are diagnosed with Renal Cell Carcinoma (RCC) which can further be classified depending on the cell type the cancer originates from into renal papillary cell carcinoma (KIRP), chromophobe carcinoma (KICH) and renal clear cell carcinoma (KIRC). The treatment of choice for early stage RCC patients is surgical removal. While this is relatively successful with an overall survival of 60-70%, the prognosis for late stage RCC is less than 10%. This can be attributed to RCC being an intratumorally heterogenous form of cancer. Additionally, KIRC is resistant to traditional treatment methods like chemotherapy and radiotherapy (Dimitrieva *et al.*, 2016). Therefore, novel individualized treatment options need to be developed (Zhang *et al.*, 2019). Since KIRC is the most prevalent form of RCC and has the worst prognosis with a 5-year survival rate of 55-60% (Tabibu *et al.*, 2019) we set the focus of our analysis on this tumor type.

3.1 Computational Tools

3.1.1 Dimension reduction

3.1.1.1 PCA

Principal component analysis is used to reduce a multidimensional set of data to a lower number of informative dimensions called principal components (PCs). The principle components are linear combinations of the original variables and are identified using matrix diagonalisation therefore, PCs have no correlation to each other. Each PC explains a portion of the total variance of the dataset, while they are ordered by decreasing proportion of explained variance. This way PCA aids to analyse and visualize structures in datasets (Abdi and Williams, 2010).

3.1.1.2 UMAP

3.1.2 Statistical analysis

3.1.2.1 Shapiro-Wilks test

Shapiro-Wilks (SW) test is a normality test based on regression and correlation. It tests the null hypothesis that the data follows a normal distribution. Small values of SW test statistic indicate no normality of the data thus the null hypothesis is rejected. SW values of one suggest normality (Yap and Sim, 2011).

3.1.2.2 Wilcoxon rank-sum and signed-rank test

Wilcoxon rank-sum test and Wilcoxon signed-rank test both are non-parametric statistical hypothesis tests that can be used when the data does not follow a normal distribution. Wilcoxon signed-rank test is used to analyze matched-pair or one-sample data. It tests the null hypothesis that there is no difference in probability distribution of first and second sample, hence the distribution of pairwise differences is centered at zero. The test is based on ranked absolute values of differences (Woolson, 2007). Wilcoxon rank-sum test is performed when analyzing unpaired-data and is likewise based on ranked values. The null hypothesis states that there is no association between the two samples (Rey and Neuhäuser, 2011).

3.1.2.3 Kruksal-Wallis

Kruksal-Wallis test is a rank-based non-parametric hypothesis test. It is an extension of Wilcoxon rank-sum test and can allows comparing more than two independent data sets. Kruksal-Wallis test tests the null hypothesis that there is no difference in distributions of all k data sets. Alternative hypothesis states that at least two populations show stochastic heterogeneity (Vargha and Delaney, 1998; Ostertagová *et al.*, 2014).

3.1.2.4 Bonferroni correction

Multiple statistical testing results in an increased risk for type I errors (false positives). Bonferroni correction is used to reduce this type I error rate. For this, the significance level is adjusted by dividing the `False decovery rate (FDR)` by the number of tests (Armstrong, 2014).

3.1.3 Clustering

3.1.3.1 K-means

Clustering is used to find similarities in a data set and to group the data points accordingly. K-means is a numerical, unsupervised, non-deterministic iterative clustering method (Na *et al.*) which determines the group affiliation based on a minimization of the within-cluster sum of squares.

3.1.3.2 Hierarchial clustering

Hierarchical clustering is a bottom-up clustering method. Therefore, each object starts in its own cluster and similar clusters are merged iteratively to identify subcategories in the data set. There are different linkage methods that determine which clusters are most similar one of which is the Ward method. It groups clusters with regard to minimizing the variance within clusters (Susan Holmes, 2019).

3.1.4 Immune deconvolution

The immune deconvolution package (Merotto and Sturm (2022)) is used to obtain immune cell fractions from bulk RNA-sequencing data. The input is a matrix with genes as rows and samples as columns, containing gene expression values as transcripts per million (TPM). The deconvolution algorithm models the expression of a single gene as a linear combination of the expression of that gene across the different cell types. An equation for each gene is set up that contains terms of the matrix multiplication of a signature matrix S and an immune cell fraction vector F. The signature matrix S contains all average gene expression profiles for each gene in the immune cell types, respectively. The output using the "quanTIseq" method is a matrix containing the immune cell fractions for each sample that have been re-normalized to sum up to one (Finotello and Trajanoski (2018)).

3.1.5 GSEA

3.1.6 Binary logistic regression

Binary logistic regression is a statistical method used to predict a dichotomous dependent variable from one or several independent variables. In contrast to linear regression, logistic

regression does not require a linear relationship between the independent and the dependent variable.
A logistic regression model is created by training on one dataset and testing it on another independent one. First, the observations of the dependent variable of the training dataset are converted into 0 an 1 for each outcome, respectively. These data points can be plotted into a coordinate system and subsequently be moved to positive and negative infinity through a $log(odds)$-transformation. The odds ratio is calculated using the logit function:

$$log(odds) = log(\frac{p}{1-p})$$

A regression line is now fitted to these data points and these are projected onto the fitted line and thus get a new $log(odds)$-value. In order to plot the sigmoidal logistic graph, the new $log(odds)$ value of each data point is inserted into the logistic function, which can be derived by rearranging the $log(odds)$-equation:

$$p = \frac{e^{log(odds)}}{1 + e^{log(odds)}}$$

The result of the function is the probability between 0 and 1 for each data point to have a 1 as the outcome of the dependent variable. The best fitting linear regression line can be found using the "maximum likelihood" that is computed with the probabilities of each point. The calculation of the likelihood is made for all possible regression lines, by rotating the line minimally for every new computation. The best fitting line is the one with the maximum value of the likelihood.
For the evaluation of the regression model R^2-values and statistical tests can be considered. Since the residuals cannot be used to compute R^2 in logistic regression, several pseudo-R^2, such as the "McFadden's" were introduced, which compare the maximum likelihood of the model to a null model. A Wald-test and a $chi2$-test can be performed to test for the significance of each coefficient and of the overall model (Sperandei (2014), Peng *et al.* (2002)).

4 Materials and Methods

4.1 Data cleaning

The analysis was focused on two data sets containing bulk-cell sequencing data. Pan-cancer analysis was performed on gene expression data of 9783 patients of 33 tumor types (Pan-cancer data frame `DF1`). Focused analysis was based on gene expression data of normal and tumor tissue of 72 KIRC patients (`DF2`). The data obtained was already normalized by $log2$(Transcripts per million).
Both data sets were filtered for protein-coding genes using biomaRt package. This reduced gene expression data in `DS1` from 60 498 to 19 624 genes and in `DF2` from 19624 to 19186 genes. In `DF1`, variance was computed for each gene over all samples and the lower p50-quantile was subsequently removed. Furthermore, constantly expressed genes only were removed in `DF2`, resulting in gene expression data of 18 645 genes.

4.2 TCGA pan-cancer analysis

After the data cleaning, as described above, the pan-cancer analysis was performed on `DF1`. With dimensional reduction methods like PCA and UMAP, patients from all tumor types could be visualized in a two-dimensional coordinate system.
In order to determine the activity of hallmark, KEGG, PID, PENG and MMR pathways in each patient, GSEA was carried out for each tumor type, respectively. Genes were ranked according to the z-score after z-normalization of each gene across all samples within every tumor type. Pathways with an adjusted $FDR > 0.05$ using `Benjamini & Hochberg` correction method were considered as non-significant and their NES was set to zero. Pathway activity matrices that contain these NES were generated for all pathway sets and could be depicted in heatmaps. Starting from these matrices, hierarchical clustering was performed to find groups of tumor types with similar regulation. The variance of the NES for each pathway across all tumor types was computed in order to identify the top 30 pathways with the highest variance and hence most differences in regulation between tumors. Results

of the GSEA with PID pathways were analyzed in more detail by hierarchical clustering of patients based on the NES of PID pathways within each tumor type. Significantly different enriched pathways between the resulting clusters were identified using Kruksal-Wallis test (FDR: 5%) and Bonferroni correction.
For all KIRC patients of `DF1` the immune cell fractions of the bulk samples were estimated using the `immunedeconv` package. The `quanTIseq` method was used and therefore normalized gene expression values were transformed back into transcripts per million (TPM).

4.3 Focused Analysis: Identifying subtypes in KIRC

For focused analysis, KIRC was examined in more detail. The analysis was based on `DS2`. Data cleaning was performed as described above.
First, differential gene expression in tumor tissue compared to normal tissue was analyzed by calculation of Foldchange $FC = (meancondition1)/(meancondition2)$ for each gene, where condition one represents gene expression of tumor tissue and condition two gene expression of normal tissue. Statistical significance was determined using Wilcoxon signed-rank test with Bonferroni correction, as Shapiro-Wilks test indicated no normality of the data.
The second part of this analysis was focused on differential pathway activity. Pathway activity matrices were determined using GSEA, where genes were ranked based on their foldchange. Pathways with an adjusted p-value > 0.05 were considered as non-significant, hence their NES was set to zero. Hallmark pathways, KEGG pathways and PID pathways were analyzed. For PID and KEGG pathways, patients were clustered based on their pathway activity. Subclusters of patients were visualized for each geneset by UMAP calculated on PCA-results and subsequently identified by k-means clustering. Optimal number of clusters was determined using elbow method and silhouette method. Pathways that were crucial for this clustering and were significantly different between those clusters were detected by Wilcoxon rank-sum test for two clusters and Kruksal-Wallis test for three clusters, both with Bonferroni correction.

4.4 Predicting immune infiltration with logistic regression

A binary logistic regression model was created based on the pathway activity of KIRC samples from `DS1` and tested on `DS2`. The dependent variable to be predicted was the

"Immune infiltration" of samples. The independent variables were pathways chosen from the most differentially significantly expressed PID pathways between the three clusters that have emerged in the PID clustering of KIRC patients in `DS1`. All patients of the cluster with highest pathway activity were marked as immune infiltrated with a "1" and all other patients with a "0." The logistic regression was performed to predict the immune infiltration in the samples of `DS2`.

4.5 Packages

5 Results

5.1 TCGA pan-cancer analysis

5.1.1 GSEA reveals the similarities and differences of tumor types in pan-cancer analysis (/in pathway activity)

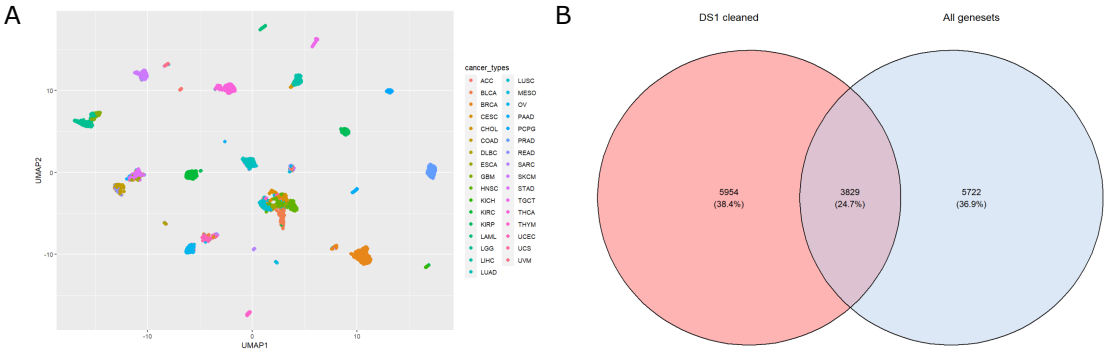

Figure 5.1: This is a figure.

The results of the GSEAs that were performed on `DS1` showed diverse enrichment of pathways across tumor types. To demonstrate the output, exemplary results of KIRC patient `TCGA-B8-5163-01` are depicted in Figure XA and XB. Curves of GSEA enrichment scores for upregulated, neutral and downregulated pathways show different running sums of the weighted enrichment score (Fig. XA). The GSEA barplot shows the most enriched PID pathways, where the black lines indicate the position of the gene in the ranked gene list of the example patient.

In order to visualize the enrichment results, heatmaps of the pathway activity matrices were plotted. Four main clusters of the 33 tumor types could be identified in the hallmark GSEA. In general, many pathways did not show a remarkable enrichment across all tumor types. Strikingly, immune response associated pathways, like Interferon and Interleukin singaling, were significantly upregulated in PAAD, OV and STAD and downregulated

in for example LGG, UVM and ACC. Moreover, pathways promoting cell proliferation like E2F and G2M-checkpoint were exceptionally upregulated in TGCT and THYM and downregulated in KIRP, PRAD and LGG (Fig. XC).
The heatmap from KEGG GSEA results showed four clusters as well. Noticeably, O-glycan biosynthesis was highly upregulated in every single tumor type, while pathways for one-carbon metabolism, homologous recombination and cell cycle were downregulated. The cluster with UVM, SARC, SKCM, ACC and PCPG showed high positive enrichment in metabolic pathways, like carbohydrate and drug metabolism. Pathways associated with antigen presentation and autoimmune responses were significantly downregulated in the cluster with DLBC, LAML and KICH, as well as CHOL and upregulated in almost all other tumor types (Fig. XD).
The top 30 variance pathways across all tumor types identified several clusters. The pathways included immune-, metabolic- and hallmark-associated pathways. The highest variance overall was computed for CD8 TCR, TCR and CD8 TCR downstream pathways. Interferon, interleukin and complement signaling pathways were often either up- or down-regulated together in tumor types. An exceptional regulation signature could be seen in THYM, where TCR- and cell-cycle-related pathways were highly upregulated. CHOL showed a unique pattern as well with high NES in interferon- and inflammatory- response gene sets (Fig. XG).

5.1.2 Identifying immune signaling (expression pattern) subtypes in 24 tumor types

In addition to expression patterns that differ between the 33 tumor types, we next focused on assessing heterogeneity within each tumor type. For this purpose, we used PID signaling pathways to detect whether subtypes format based on our GSEA results (FIG). Hence, we identified three subtypes, “upregulated,” “downregulated” and “neutral” in 24 out of 33 tumor types using hierarchical clustering. Significant pathways were filtered using Kruskal-Wallis-Test with an $FDR < 0.05$ and Bonferroni correction. Those subtypes are only relative within one cancer type and are not representative compared to normal tissue. The “neutral” subtype is the most common in 22 of 24 tumor types, the only exception being LUAD and PAAD, which are dominated by the “downregulated” subtype. Generally, the “neutral” subtype is present in more than 50% of the cancer patients upon the 24 tumor types characterized by successful subtyping. The “downregulated” and “upregulated” subtype however are approximately equally distributed. The subtypes were particularity distinguished by CD8+ T cell receptor signaling in all 24 tumor types.

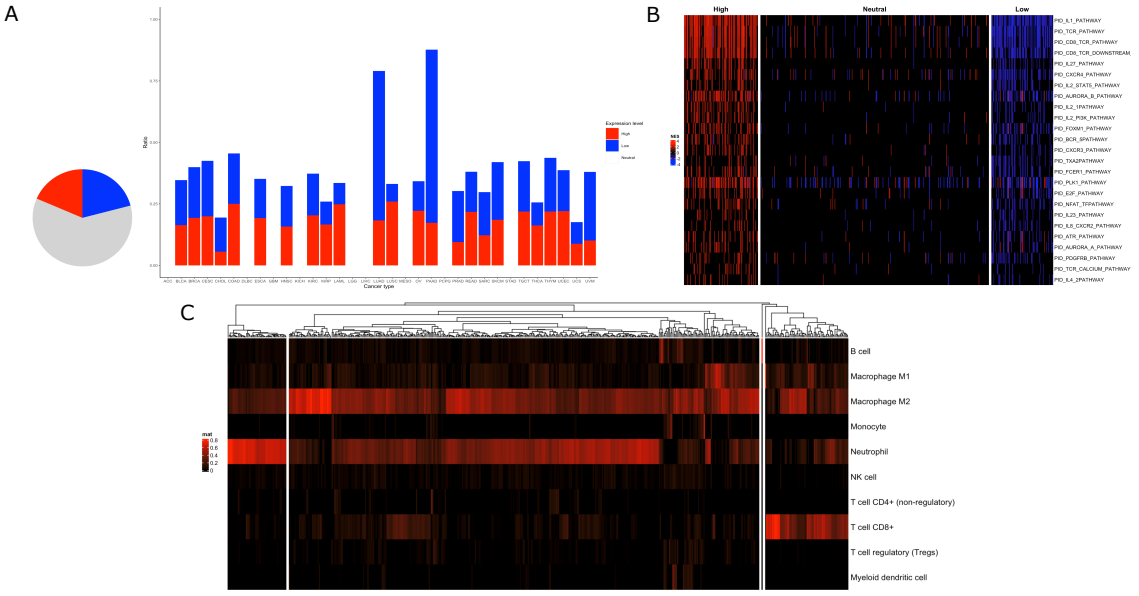

Figure 5.2: This is a figure.

In (FIG) an exemplary result of the subtype identification is shown for the cancer type KIRC, where the "neutral" subtype is predominant. The top 20 pathways ranked according to their $p - value$ are displayed. The differently expressed pathways with the highest significance are Interleukin 1 and 27 pathways, as well as pathways related to T cell specific receptor signaling. Beyond immune signaling pathways, cell cycle regulatory pathways including E2F, a transcription factor associated with cell proliferation, and Aurora B, which plays a role in mitosis, appear to differ between the 3 subtypes. (SOURCE)Therefore, the subtype with the most immune-cell related expression patterns show the highest upregulation in cell cycle promoting pathways, characterizing this subtype as high proliferating. An analogous correlation appears in the "downregulated" and "neutral" subtype.

Based on PID pathway activities, we conducted a more in-depth analysis of the tumor microenvironment in KIRC. Therefore, we performed immune deconvolution to assess the total amount of immune cells present in tumor tissue (FIG). Patients in the "upregulated" subtype could be mapped to high CD8+ T-cell infiltration, whereas macrophages M2 and neutrophiles showed lower infiltration compared to the other two subtypes. Nonetheless, it is visible that other types of immune cells are outstandingly present in the tumor microenvironments of the "neutral" and "downregulated" subtype. The latter subtype showed highest infiltration rates in neutrophils, whereas the "neutral" subtype is characterized by high macrophage, as well as medium neutrophile infiltration. Other immune cells such as NK cells and B cells were not detectable upon all subtypes. Consequently, the subtypes

differ not in immune infiltration in general, as the pathways expression patterns suggest, but are characterized by the presence of different immune cell types.

5.2 Focused analysis on KIRC

5.2.1 Characterization of metabolic and immune signaling landscape in KIRC

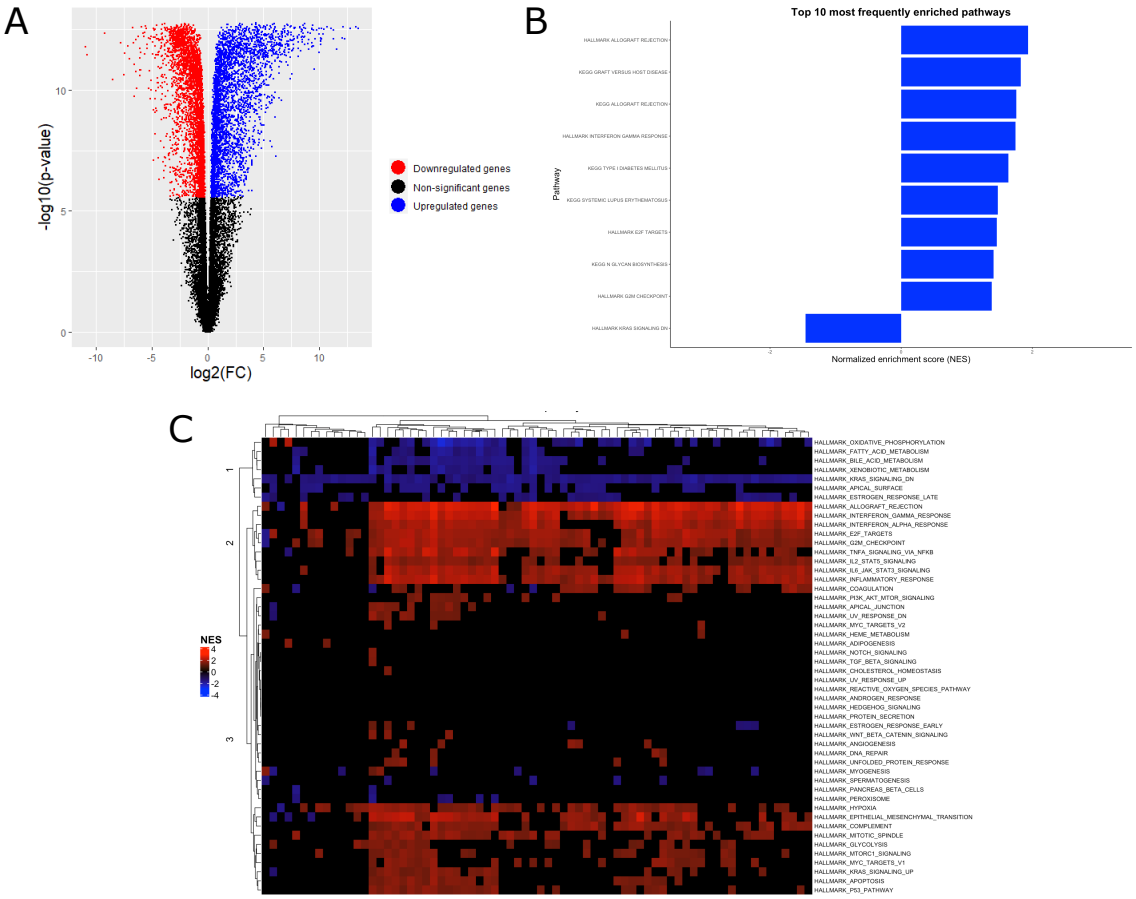

Figure 5.3: This is a figure.

After characterizing differences within Kidney renal clear cell cancer patients, we next wanted to identity differentially expressed genes and pathways between tumorous and healthy tissue. Using the binary logarithm of the mean Fold change between both tissue types, we encountered that the tumorous tissue features both up and down regulated genes (FIG). To further analyze which parameters are crucial in distinguishing tumorous from healthy tissue, we used GSEA on 3 gene sets: PID, KEGG, and Hallmark.

The most frequently enriched pathways, which showed differential expression in at least 50 of 72 patients, were mapped to allograft rejection, interferon gamma response and cell cycle regulators, such as E2F, G2M and KRAS. Consequently, both immune response and cell cycle regulators appear to play an important role in tumor development and hence occur in the majority of KIRC patients. Noticeably, all oncogenes besides KRAS are upregulated. (FIG)

5.2.1.1 Cancer hallmark pathways form subclusters in KIRC

To characterize subtypes within the 72 KIRC patients, we next looked at each gene set individually. We categorized the hallmark pathways into three subclusters, where certain pathways are either upregulated, downregulated or not significantly differentially expressed at all. The downregulated subtype was mainly mapped to fatty acid metabolism, whereas the upregulated subtype consists of cell cycle regulatory pathways as well as immune signaling pathways. Nonetheless, the expression patterns within these 3 subtypes of hallmark gene sets are not homogenously distributed upon all KIRC patients. Using hierarchical clustering, two clusters could be identified, one smaller cluster, which shows expression levels close two zero, and one bigger cluster with highly fluctuating expression patterns.

5.2.1.2 Identification of two KIRC subtypes that differ in pathways related to cell cycle and immune infiltration

To determine signaling pathways that distinguish KIRC tumor tissue from normal tissue, PID pathways were analyzed in more detail. GSEA revealed that significantly enriched pathways are mainly upregulated and indicated sample subclusters (supp? Figure 5A). K-means clustering based on UMAP allowed identification of two distinct clusters (Figure 5B).
Next, Wilcoxon ranked-sum test with Bonferroni correction determined 14 pathways that significantly differ between the two clusters, hence are crucial for formation of these two clusters. Figure 5C and supplementary? Figure 5D show distinct differences in NES scores of these pathways in both clusters and indicate an “upregulated” and “neutral” subtype. These differences include pathways that play an important role in cell cycle regulation, such as E2F and P53 pathway (Polager and Ginsberg, 2009). Noticeably, a majority of significantly differently pathways are related to immune response as interleukin pathways (e.g. IL1, IL2) or T cell specific receptor signaling. More precisely, top significant pathways refer to CD8+ T cells, thus potentially differential numbers of these cells between both

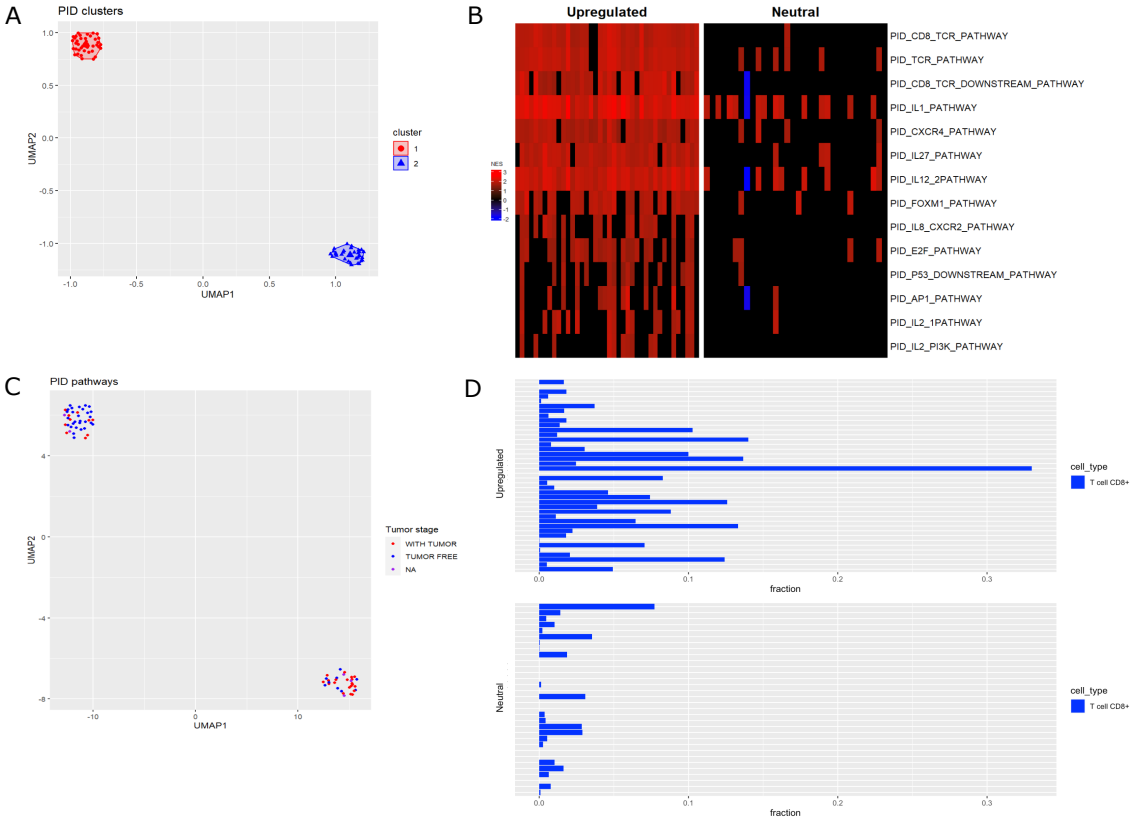

Figure 5.4: This is a figure.

clusters. This lead was confirmed by immune deconvolution, where the "upregulated" subtype is associated with higher CD8+ T cell fractions (Figure 5E). Consequently, the "upregulated" cluster shows higher CD8+ T cell infiltration compared to the "neutral" subtype characterizing two different tumor microenvironments.
Additional patient data was provided, such as gender and tumor stage and was examined with regard to PID clusters. Only a correlation between PID subtypes and tumor stage could be observed, as 70.0% of patients in the "upregulated" cluster are tumor free, while in the "neutral" cluster, only 28.1% of patients characterize as tumor free (Figure 5F).

5.2.1.3 Immune infiltration subtypes can be projected to KEGG signaling and metabolic pathways

Next, KEGG pathways were investigated more precisely. GSEA shows that enriched KEGG pathways in tumor compared to normal tissue were both, up- and downregulated (Figure

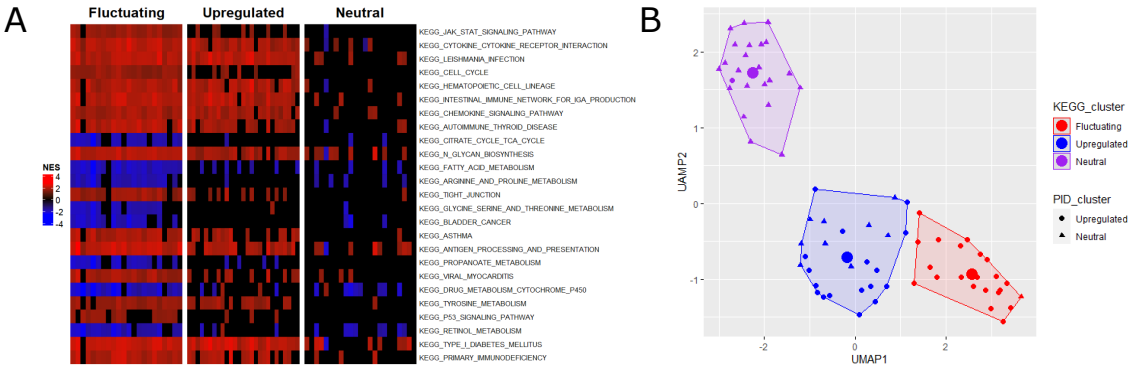

Figure 5.5: This is a figure.

6A). Based on UMAP, patients could be divided into three subtypes by k-means clustering. Pathways, that are significantly different between clusters, were determined using Kruksal-Wallis test with Bonferroni correction. This allowed identification of a "fluctuating" cluster showing both up- and downregulation, one "upregulated" cluster with high or neutral NES and one "neutral" cluster mainly without pathway enrichment (Figure 6B). KEGG pathways, that are upregulated in the "fluctuating" and "upregulated" subtypes, but not the in "neutral" subtype, include pathways that are mainly associated with immunological processes but also cell cycle. In this context, immunological processes refer to disease related pathways, such as diabetes mellitus, as well as cytokine and chemokine signaling and antigen processing and presentation.
Moreover, pathways that are downregulated in the "fluctuating" subtype but neutral in both other subtypes relate mostly to metabolic pathways, such as fatty acid, amino acid, and drug metabolism, as well as TCA cycle. Moreover, KEGG clusters show similar subdivision of samples as PID clusters. In Figure 6C it can be seen that the "fluctuating" KEGG cluster mainly matches patients of the "upregulated" PID cluster, while the "neutral" KEGG subtype includes mostly patients of the "neutral" PID subtype. Additionally, the "upregulated" KEGG cluster corresponds to both PID clusters.

5.3 Logistic regression

The binary logistic regression model predicts immune infiltration in patients based on the pathway activity (NES) of `PID_TCR_PATHWAY` and `PID_IL1_PATHWAY`. Both independent variables significantly contribute to the model as can be seen in Table XX, which depicts the model properties.

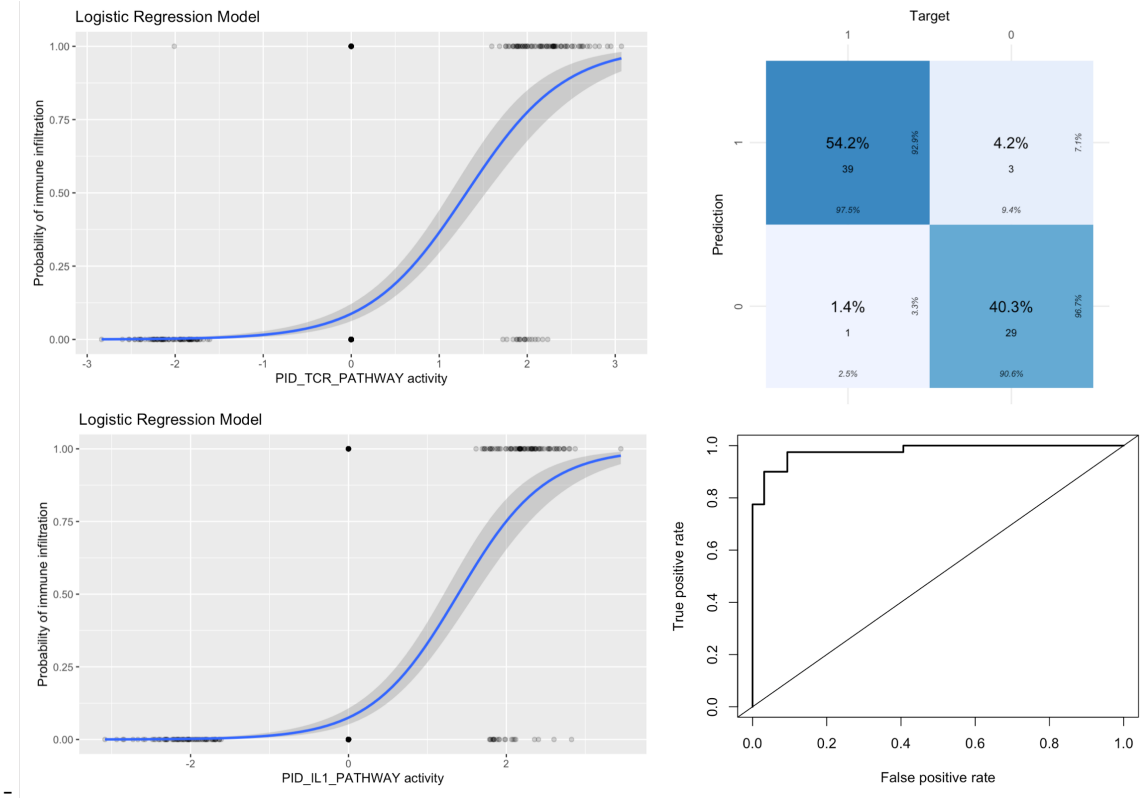

Figure 5.6: This is a figure.

Logistic regression model results			
	Estimate	**Standard error**	**p-value**
Intercept	-2.5856	0.2059	$< 2e-16$
PID_TCR_PATHWAY	0.9727	0.2027	$1.59e-06$
PID_IL1_PATHWAY	1.1651	0.1934	$1.70e-09$

By taking the exponent of the coefficients in Table XX, the returned value implies that for every increase of 1 in NES of the TCR and the IL-1 pathway, the odds-ratio for a patient being immune infiltrated increases by the factor 2.64517(164.5%) and 3.206149(220.6%), respectively. The logistic function plots of each independent variable show that a higher enrichment score indicates a higher probability of a patient being classified as "Immune infiltrated" (FIG. XA and XB). The contribution of each coefficient to the model was evaluated and confirmed with a Wald's test, where both TCR ($p = 1.590e-06$) and IL-1 pathway ($p = 1.701e-09$) were significant. The significance of the whole model in contrast to a null-model was confirmed by McFadden's pseudo-R^2 ($R^2_{Mc} = 0.5074508$) and the p-value of the overall chi-square statistic ($p = 7.860549e-60$).

The model was tested on KIRC patients in `DS2` and classified 42 out of 72 patients as "Immune infiltrated." The accuracy of the model was verified by the clustering of the patients according to PID pathway activity, which showed two distinct clusters. One cluster showed upregulation in immune pathways and hence all these patients were assigned to the immune infiltrated group. In 94.4% of the patients the prediction and the cluster assignment corresponded correctly. Moreover, the performance of the model was evaluated with a confusion table that compares predictions and immune infiltration according to the cluster as well. It revealed a false-negative rate of 2.5%, a false-positive rate of 9.4% and a true-positive rate (recall) of 97.5% (FIG. XC). The ROC-curve that was plotted for the model (Fig. XD) showed an AUC of 0.9789062.

6 Discussion

Cancer hallmarks define abnormalities in metabolism, immune infiltration, and cell proliferation regardless of the cancer type(Hanahan and Robert, 2011). Since all these aspects are rather complex and (de-)regulation takes place on transcription as well as (post-) translational levels, we tried to assess whether bulk sequencing data allow us to observe these alterations. Therefore, we utilized 5 different gene sets, more precisely hallmark, PID, KEGG, MMR and PENG. Notably, even though tumor pathology differs strongly among the 33 tumor types of interest, tumors always tend to cluster based on the aspects of metabolism and immune signaling. Contrary to expectations, many essential hallmark gene sets, such as angiogenesis were not enriched throughout the tumor types at all (Li *et al.*, 2017). Notably high E2F upregulation was observed in TGCT and THYM, which is associated with cell transformation and beneficial for both tumor cell proliferation, as well as immune cell proliferation in the TME (Yang and Sladek, 1995). However, downregulation of E2F in KIRP could not be confirmed with literature (Wang *et al.*, 2020).

To gain deeper insights into tumor metabolism, KEGG pathways reveled upregulation of O-glycan biosynthesis in all 33 tumor types. This pathway is crucial in cancer cell transformation and therefore is representative for metabolic alterations (Brockhausen, 1999). Beyond that higher drug metabolism in some cancer types could imply higher drug resistance and selective advantages for tumor cells (Michael and Doherty, 2005), increasing the necessity for alternative treatment options. Downregulation of antigen presentation in some cancer types is consistent with immune evasion according to literature (Cornel *et al.*, 2020) Following the lead, we used PID signaling pathways to examine the influence of immune cells in the TME. Hence, subclusters based on PID signaling pathways were identified within each tumor type and mapped to immune cell related expression patterns in 24 out of 33 tumor types. These tumor types were divided in “upregulated,” “downregulated” and “neutral” subtypes.

To determine the foundation of the subtypes, we looked at KIRC in more detail. Utilizing immune deconvolution on KIRC patients, we were able to assign the “upregulated” subtype to high CD8+ T cell infiltration, the “downregulated” subtype to high neutrophile infiltration and the “neutral” type to only medium immune infiltration. These characteristics can

be mapped to different compositions of the TME during tumor progression. Early stage KIRC is often associated with macrophage M2 infiltration, also known as tumor-associated macrophages, which create an immunosuppressive microenvironment. This tumor stage is equivalent to the "neutral" type we discovered in KIRC. KIRC progression, however, is characterized by increased CD8+ T cell infiltration, which we observed in the "upregulated cluster" (Zheng *et al.*, 2021a). Considering the coexistence of macrophages in this cluster, which are known for expression of immune checkpoints like PDL-1, these T cells are most likely to be exhausted, losing their anti-tumorous capacity (Gordon *et al.*, 2017). Contrary to most tumor types, macrophage and T cell infiltration correlate with poor prognosis in KIRC(Fridman *et al.*, 2017). The "downregulated" subtype we discovered was mainly characterized by high neutrophil expression. Neutrophil expression in KIRC was associated with lung metastasis and therefore poor prognosis as well(Nishida *et al.*, 2020).

Since immune infiltration and the composition of the TME fluctuates among tumor types, no general conclusion regarding prognosis can be drawn from the 24 immune signaling pathways. Further analysis on the immune cells present in each tumor types using immune deconvolution remains to be performed. Additionally, Z-scoring across one gene for all samples as a substituting method for the fold change might have synthetically produced 3 subtypes. Nonetheless, immune deconvolution, performed on the TPM values was used to confirm parts of our results. The strict cutoff for the ´FDR´ used for filtering the GSEA results is another factor, which led to rather sparse pathway activity matrices might have created the "neutral" subtype as well. Consequently, mean pathway expression values could have been distorted, either creating expression pattern clusters between the tumor types or cause them to vanish.

Beyond the relative differences within KIRC, tumorous tissue was compared with healthy tissue. Noticeably, allograft rejection and interferon gamma response as well as cell cycle regulator upregulation were the most commonly enriched pathways, distinguishing tumorous from healthy tissue. These pathways characterize an aggressive and often metastatic KIRC subtype(Lin *et al.*, 2022). Since these hallmark pathways are upregulated in the majority of KIRC patients, the dominance of this subtypes is standing to reason and is supported by further analysis. On the first look KRAS appeared to be downregulated, which is rather untypical for an oncogene, especially since KRAS is effected in the majority of tumors. The ´Hallmark KRAS signaling DN´ pathway, however, consist of genes, which are downregulated in presence of KRAS activation, which is concurrent to literature(Kumar *et al.*, 2018). Even though no mutational analysis was performed, this pathway hints at a loss of function mutation in most KIRC patients.

Further analysis, based on PID signaling pathways, revealed an “upregulated” and “neutral” subtype. The “upregulated” subtype was mapped to high CD8+ T cell infiltration. These subtypes were commonly found in analysis of the KIRC microenvironment, classifying the infiltrated subtype as immunologically as “hot” and the other one as “cold.” Once again, the immunologically “cold” subtype was associated with better prognosis(Xu *et al.*, 2021). Nonetheless, the T cell infiltrated subtype showed an increased number of “Tumor free” patients. Since therapeutic procedures are not defined for most patients, additional patient data are needed to assess, whether these observations are truly contradictory or not. Besides immune signaling pathways, the “upregulated” subtype showed high expression in cell cycle regulators, such as E2F. Considering an association between the IL PI3K pathways and activation of E2F in T cells, proliferation might not only be upregulated in the tumor itself, but in T cells especially, leading to high infiltration rates(Brennan *et al.*, 1997). Furthermore, immune signaling subtypes were mapped onto metabolic and signaling KEGG clusters. Noticeably, TCA cycle was downregulated in most patients. This observation is associated with the Warburg effect, where glycolysis is used for the main energy generation. This leads to increased lactate production and release, creating a low pH-value, which is known to be immunosuppressive. Additionally, the risk for apoptosis due to lesser mitochondria presence is beneficial for tumor growthLiberti and Locasale (2016).

Lastly, a tumor regression model was built to determine, whether a random KIRC patient shows expression patterns of the infiltrated or not infiltrated subtypes based on the expression of ´PID_TCR_PATHWAY´ and ´PID_IL1_PATHWAY ´. This led to a very significant model confirmed by McFadden’s Rˆ2 (0.51), which indicates an “excellent model” with values between 0.2-0.4 (McFadden, 1977). In the training DF(DF1) medium and low expression patterns were merged and thus no differentiation between “neutral” and “downregulated” considered. Accuracy of prediction was compared to cluster number. Since clustering was based on signaling PID pathways, some patients could be clustered to “immune infiltrated” due to differential expression of cell cycle regulators etc. However, both factors did not appear to lower the significance of the model. Training on bigger sample size (here: 531) could significantly improve the model either, lowering both FPR and FNR.

Challenges, encountered we encountered in both pan-cancer and KIRC-specific analysis, were sparse pathway activity matrices as described above. Further, bulk sequencing data made it hard to map certain pathway activities to the cells present in the TME. Next steps in the analysis would be to define the coherence between immune system and metabolic pathways in KIRC more precisely and project these results on our pan-cancer analysis,

ideally allowing us to characterize the 24 immune signaling subtypes in depth. Besides sequencing data, mutational analysis would be needed gain a deeper insight into the genetic alterations associated with tumor progression. Correlations between survival could enhance both reliability and significance of our analysis as well.

7 Concluding remarks/Outlook

Immune signaling pathway panel could be broadly introduced in diagnostic and are already used for reliable subtyping. Based on these results, the predicted immune infiltration could be used for therapeutic application, since the immune infiltrated subtype is known to be more perceivable to Immunotherapies such as Immune-Checkpoint-inhibitors(Krishna *et al.*, 2021). This could lead to more successful treatments in early stages of the cancer. Moreover, using mutational analysis and proteomics, neo peptide antigens could be identified. In case of broadly similar antigens, mRNA vaccines are standing to reason, which would be especially beneficial for the immunologically cold phenotype, offering possible therapeutic procedures, individual for each subtype and adding prophylactic treatments as the goal of cancer treatment(Xu *et al.*, 2021).

8 References

Abdi, H, and Williams, LJ (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2, 433–459.

Anderson, NM, and Simon, MC (2020). The tumor microenvironment. Current Biology 30, R921–R925.

Armstrong, RA (2014). When to use the bonferroni correction. Ophthalmic Physiol Opt 34, 502–508.

Brennan, P, Babbage, JW, Burgering, BM, Groner, B, Reif, K, and Cantrell, DA (1997). Phosphatidylinositol 3-kinase couples the interleukin-2 receptor to the cell cycle regulator E2F. Immunity 7, 679–689.

Brockhausen, I (1999). Pathways of o-glycan biosynthesis in cancer cells. Biochim Biophys Acta 1473, 67–95.

Cooper, LA, Demicco, EG, Saltz, JH, Powell, RT, Rao, A, and Lazar, AJ (2018). PanCancer insights from the cancer genome atlas: The pathologist's perspective. The Journal of Pathology 244, 512–524.

Cornel, AM, Mimpen, IL, and Nierkens, S (2020). MHC class i downregulation in cancer: Underlying mechanisms and potential targets for cancer immunotherapy. Cancers (Basel) 12.

Dimitrieva, S, Schlapbach, R, and Rehrauer, H (2016). Prognostic value of cross-omics screening for kidney clear cell renal cancer survival. Biol Direct 11, 68.

Ferlay, J, Colombet, M, Soerjomataram, I, Parkin, DM, Piñeros, M, Znaor, A, and Bray, F (2021). Cancer statistics for the year 2020: An overview. International Journal of Cancer 149, 778–789.

Finotello, F, and Trajanoski, Z (2018). Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunology, Immunotherapy 67, 1031–1040.

Fridman, WH, Zitvogel, L, Sautès–Fridman, C, and Kroemer, G (2017). The immune contexture in cancer prognosis and treatment. Nature Reviews Clinical Oncology 14, 717–734.

Gordon, SR et al. (2017). PD-1 expression by tumour-associated macrophages inhibits phagocytosis and tumour immunity. Nature 545, 495–499.

Hanahan, D, and Robert (2011). Hallmarks of cancer: The next generation. Cell 144, 646–674.

Krishna, C et al. (2021). Single-cell sequencing links multiregional immune landscapes and tissue-resident t cells in ccRCC to tumor topology and therapy efficacy. Cancer Cell 39, 662–677.e6.

Kumar, A, Kumari, N, Gupta, V, and Prasad, R (2018). Renal cell carcinoma: Molecular aspects. Indian J Clin Biochem 33, 246–254.

Li, M, Sun, Q, and Wang, X (2017). Transcriptional landscape of human cancers. Oncotarget 8, 34534–34551.

Liberti, MV, and Locasale, JW (2016). The warburg effect: How does it benefit cancer cells? Trends Biochem Sci 41, 211–218.

Lin, E et al. (2022). Integrative analysis of the genomic and immune microenvironment characteristics associated with clear cell renal cell carcinoma progression: Implications for prognosis and immunotherapy. Front Immunol 13, 830220.

McFadden, D (1977). Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments, Cowles Foundation for Research in Economics, Yale University.

Merotto, L, and Sturm, G (2022). Immunedeconv: Methods for immune cell deconvolution.

Michael, M, and Doherty, MM (2005). Tumoral drug metabolism: Overview and its implications for cancer therapy. Journal of Clinical Oncology 23, 205–229.

Na, S, Xumin, L, and Yong, G Research on k-means clustering algorithm: An improved k-means clustering algorithm. In: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 63–67.

Natalya, and Craig (2016). The emerging hallmarks of cancer metabolism. Cell Metabolism 23, 27–47.

Nishida, J, Momoi, Y, Miyakuni, K, Tamura, Y, Takahashi, K, Koinuma, D, Miyazono, K, and Ehata, S (2020). Epigenetic remodelling shapes inflammatory renal cancer and neutrophil-dependent metastasis. Nat Cell Biol 22, 465–475.

Ostertagová, E, Ostertag, O, and Kováč, J (2014). Methodology and application of the kruskal-wallis test. Applied Mechanics and Materials 611, 115–120.

Peng, J, Lee, K, and Ingersoll, G (2002). An introduction to logistic regression analysis and reporting. Journal of Educational Research - J EDUC RES 96, 3–14.

Polager, S, and Ginsberg, D (2009). p53 and E2f: Partners in life and death. Nature Reviews Cancer 9, 738–748.

Rey, D, and Neuhäuser, M (2011). Wilcoxon-signed-rank test. In: International Encyclopedia of Statistical Science, ed. M Lovric, Berlin, Heidelberg: Springer Berlin Heidelberg, 1658–1659.

Sperandei, S (2014). Understanding logistic regression analysis. Biochem Med (Zagreb) 24, 12–18.

Susan Holmes, WH (2019). Modern statistics for modern biology.

Tabibu, S, Vinod, PK, and Jawahar, CV (2019). Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. Scientific Reports 9.

Vargha, A, and Delaney, HD (1998). The kruskal-wallis test and stochastic homogeneity. Journal of Educational and Behavioral Statistics 23, 170–192.

Wang, H, Wang, X, Xu, L, Zhang, J, and Cao, H (2020). Integrated analysis of the E2F transcription factors across cancer types. Oncol Rep 43, 1133–1146.

Warburg, O, Wind, F, and Negelein, E (1927). THE METABOLISM OF TUMORS IN THE BODY. J Gen Physiol 8, 519–530.

Woolson, RF (2007). Wilcoxon signed-rank test. Wiley Encyclopedia of Clinical Trials, 1–3.

Xu, H, Zheng, X, Zhang, S, Yi, X, Zhang, T, Wei, Q, Li, H, and Ai, J (2021). Tumor antigens and immune subtypes guided mRNA vaccine development for kidney renal clear cell carcinoma. Mol Cancer 20, 159.

Yang, XH, and Sladek, TL (1995). Overexpression of the E2F-1 transcription factor gene mediates cell transformation. Gene Expr 4, 195–204.

Yap, BW, and Sim, CH (2011). Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation 81, 2141–2155.

Zhang, S, Zhang, E, Long, J, Hu, Z, Peng, J, Liu, L, Tang, F, Li, L, Ouyang, Y, and Zeng, Z (2019). Immune infiltration in renal cell carcinoma. Cancer Science 110, 1564–1572.

Zheng, Y, Wen, Y, Cao, H, Gu, Y, Yan, L, Wang, Y, Wang, L, Zhang, L, and Shao, F (2021b). Global characterization of immune infiltration in clear cell renal cell carcinoma. Onco Targets Ther 14, 2085–2100.

Zheng, Y, Wen, Y, Cao, H, Gu, Y, Yan, L, Wang, Y, Wang, L, Zhang, L, and Shao, F (2021a). Global characterization of immune infiltration in clear cell renal cell carcinoma. Onco Targets Ther 14, 2085–2100.

9 Appendix