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 20.07.2022

1 Abstract

Contents

1 Abstract 2

2 List of abbreviations 5

3 Introduction 6
- 3.1 Computational Tools . . . 7
  - 3.1.1 Dimension reduction . . . 7
  - 3.1.2 Statistical analysis . . . 7
  - 3.1.3 Clustering . . . 9
  - 3.1.4 Immune deconvolution . . . 9
  - 3.1.5 GSEA . . . 9
  - 3.1.6 Binary logistic regression . . . 9

4 Materials and Methods 11
- 4.1 Data cleaning . . . 11
- 4.2 TCGA pan-cancer analysis . . . 11
- 4.3 Focused Analysis: Identifying subtypes in KIRC . . . 12
- 4.4 Predicting immune infiltration with logistic regression . . . 12
- 4.5 Packages . . . 13

5 Results 14
- 5.1 TCGA pan-cancer analysis . . . 14
  - 5.1.1 GSEA reveals the similarities and differences of tumor types in pan-cancer analysis (/in pathway activity) . . . 14
- 5.2 KIRC specific analysis . . . 15
- 5.3 Logistic regression . . . 15

6 Discussion 16

7 Concluding remarks/Outlook 17

8 References 18

9 Appendix 20

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 tumour 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 tumours 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 tumour types selected in the TCGA.

The activation of oncogenes and the loss of tumour suppressor gene function are the main cause for tumour development. This is often caused by genome instability of which there are two types: Chromosomal and microsatellite instability. Microsatellite instability is caused by point mutations, which cannot be repaired by mismatch-repair (MMR) in patients with MMR deficiency. This leads to reading frame shifts, therefore proteins with altered functions are expressed. Those frameshift neopeptides are highly immunogenic due to the formation of new immunogenic epitopes. Therefore, neopeptides may provide a new target for immune therapies (Woerner *et al.*, 2006) and in general MMR deficient or MSI tumours may need different therapeutical approaches (Baretti and Le, 2018).

Reprogramming cellular metabolism is also a driving force in tumorigenesis. This way cancer cells can cover their high needs for nutrients even in a nutrient-poor environment. For example, cancer cells deregulate glucose and amino acid uptake and use intermediates of Glycolysis and TCA Cycle to build more biomass needed for proliferation (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 tumour cells themselves, a tumour 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 tumour type.

3.1 Computational Tools

3.1.1 Dimension reduction

3.1.1.1 PCA

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 there is a difference in at least two populations **(oder: . . . that at least two populations show stochastic heterogeneity)** (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 Kmeans

3.1.3.2 Hierarchial clustering

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 but can't be optimized by least square method, as in linear regression, since the residuals strive towards infinity. Alternatively, the

data points 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," which is either computed by the product of the probability of all data points to have the observation 1 or the sum of the logarithmic probabilities. 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. Patients were compared 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)

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. Strikingly, immune response associated pathways, like Interferon and Interleukin singaling, were significantly upregulated in UCS, 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 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 metabolic gene sets that were analyzed in the PENG pathway GSEA clearly identified metabolic differences between tumor types. Noticeable, in the cluster with CHOL, PAAD, PCPG and SKCM lipid-dependent pathways were highly upregulated, while amino acid and carbohydrate pathways almost weren't enriched at all. For all tumor types, citric acid cycle pathways

weren't significantly enriched as well. Other peculiarities are the high downregulation of carbohydrate and nucleotide metabolism in DLBC, as well as high vitamin metabolism in CESC (Fig. XE). The MMR pathway enrichment that were examined tended to show a downregulation in almost all tumor types. Exceptions showed UCS, READ and THYM with positive enrichment of at least one pathway (Fig. XF). 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 downregulated 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).

5.2 KIRC specific analysis

5.3 Logistic regression

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	$1.70e-09$

6 Discussion

7 Concluding remarks/Outlook

8 References

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.

Baretti, M, and Le, DT (2018). DNA mismatch repair in cancer. Pharmacol Ther 189, 45–62.

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.

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.

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

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

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

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.

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.

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.

Woerner, SM, Kloor, M, Knebel Doeberitz, M von, and Gebert, JF (2006). Microsatellite instability in the development of DNA mismatch repair deficient tumors. Cancer Biomark 2, 69–86.

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

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.

9 Appendix