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

ksdflsdjf
sfdsfd
sfsf

Data Science Project SoSe 2022

Autoreb	Max Mustermann, jklsdfksldkfslkdjf
Geburtsort	sdfjksafdls
Abgabetermin	20.07.2022

Contents

1 Introduction

1.1 Computational Tools

1.1.1 Dimension reduction

1.1.1.1 PCA

1.1.1.2 UMAP

1.1.2 Statistical analysis

1.1.2.1 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).

1.1.2.2 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).

1.1.2.3 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).

1.1.2.4 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).

1.1.3 Clustering

1.1.3.1 Kmeans

1.1.3.2 Hierarchial clustering

1.1.4 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)).

1.1.5 GSEA

1.1.6 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)).

2 Materials and Methods

2.1 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 (`DS2`). 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 `DS2` from 19624 to 19186 genes. In `DS1`, 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 `DS2`, resulting in gene expression data of 18 645 genes.

2.2 TCGA pan-cancer analysis

After the data cleaning, as described above, the pan-cancer analysis was performed on `DS1`. 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 p-value > 0.05 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. 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. For all KIRC patients of `DS1` 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).

2.3 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.

2.4 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.

2.5 Packages

3 Results

3.1 TCGA pan-cancer analysis

3.2 KIRC specific analysis

3.3 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	$1.70e - 09$

4 Discussion

5 Concluding remarks/Outlook

6 References

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

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

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

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.

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.

7 Appendix