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 false positive rate (type I errors). Bonferroni correction is used to reduce this type I error rate. For this, the significance level is adjusted by dividing the critical p-value α 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 Regression

2 Materials and Methods

2.1 Data cleaning

The analysis was focused on two data sets. Pan-cancer analysis was performed on gene expression data of 9783 patients of 33 tumor types (Pan-cancer data set (DS1)). 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

2.3 Focused Analysis: Identifying subtypes in KIRC

For focused analysis, KIRC was examined in more detail. The analysis was based on gene expression data of normal and tumor tissue of 72 patients. 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 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. **Differences in pathway activity significantly defining these clusters were detected using Wilcoxon test for two clusters and H-test for three clusters, both with Bonferroni correction (?!)–> umformulieren**.

Pathways that were crucial for this clustering and were significantly different between those clusters were detected using Wilcoxon test for two clusters and H-test for three clusters, both with Bonferroni correction.

2.4 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	0.18303	0.01336	$< 2e-16$
PID_TCR_PATHWAY	0.09727	0.01844	$1.95e-07$
PID_IL1_PATHWAY	0.12211	0.01801	$3.23e-11$

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.

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.

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