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 3
- 1.1 Computational Tools . . . 3
  - 1.1.1 Dimension reduction . . . 3
  - 1.1.2 Statistical analysis . . . 3
  - 1.1.3 Clustering . . . 4
  - 1.1.4 GSEA . . . 4
  - 1.1.5 Regression . . . 4

2 Materials and Methods 5
- 2.1 Data cleaning . . . 5
- 2.2 TCGA pan-cancer analysis . . . 5
- 2.3 Focused Analysis: Identifying subtypes in KIRC . . . 5
- 2.4 Packages . . . 5

3 Results 6
- 3.1 TCGA pan-cancer analysis . . . 6
- 3.2 KIRC specific analysis . . . 6

4 Discussion 7

5 Concluding remarks/Outlook 8

6 References 9

7 Appendix 10

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 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 H-test

1.1.2.4 Bonferroni correction

Multiple statistical testing results in an increased risk for 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 GSEA

1.1.5 Regression

2 Materials and Methods

2.1 Data cleaning

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. ((Anzahl gene und so eher in anderem teil oder ???)) 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 > 0.5 were considered as non-significant, hence their pathway activity 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.

2.4 Packages

3 Results

3.1 TCGA pan-cancer analysis

3.2 KIRC specific analysis

4 Discussion

5 Concluding remarks/Outlook

6 References

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

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