Federal State Autonomous Educational Institution for Higher Education
National Research University Higher School of Economics

Faculty of Computer Science
Educational Program
Applied Mathematics and Information Science

# BACHELOR'S THESIS

## Research project

## "Detecting Key Genes Associated With Clear Cell Renal Cell Carcinoma Immunotherapy Resistance"

Prepared by the student of group 191, 4th year of study,
Kolchina Anastasia Konstantinovna

Supervisor:
Doctor of Philosophy, Research Fellow, Grigory Andreevich Puzanov

Moscow 2023

# Contents

# Abstract

Clear cell renal cell carcinoma (ccRCC) is a highly malignant type of kidney cancer associated with poor prognosis. This disease is extremely difficult to treat due to its chemo- and radiotherapy resistance. Despite several novel methods of immunotherapy, a high percentage of patients does not respond to it. In order to discover mechanisms of treatment resistance and therefore to choose the appropriate therapy for a particular patient, it is highly important to investigate all possible reasons. One of the perspective methods is detecting key genes - a group of genes with the highest impact on the immunotherapy failure. The present work is aimed at searching and studying such genes by bioinformatics analysis, which consists of differential expression and single-cell RNA-seq analysis, clustering, building linear prognostic model and investigating biological processes associated with the found genes.

# Аннотация

Светлоклеточный рак почки (скПКР) - высокозлокачественный тип почечного рака, ассоциированный с неблагоприятным прогнозом. Болезнь тяжело лечится из-за ее устойчивости к химио- и радиотерапии. Несмотря на новый метод лечения - иммунотерапию, довольно высокий процент пациентов на нее не отвечает. Чтобы открыть механизмы устойчивости к терапии и впоследствии выбрать подходящее лечение для пациента, важно изучить все возможные причины. Один из перспективных методов - это выявление ключевых генов, то есть группы генов, несущих наибольший вклад в развитие резистентности к иммунотерапии. Данная работа направлена на поиск и изучение таких генов с помощью биоинформатического анализа, состоящего из анализа дифференциальной экспрессии и одноклеточного секвенирования РНК, кластеризации, построения линейной прогностической модели и выявления биологических процессов, задействующих обнаруженные гены.

# Keywords

bioinformatics, cancer, immunotherapy, treatment resistance, differentially expressed genes, single-cell analysis

# Abbreviations

- ccRCC - clear cell Renal Cell Carcinoma
- DEG - Differentially Expressed Genes
- DNA - Deoxyribonucleic Acid
- ES - Enrichment Score
- GSEA - Gene Set Enrichment Analysis
- ICI - Immune Checkpoint Inhibitors
- KICH - Kidney Chromophobe
- KIRC - Kidney Renal clear cell Carcinoma
- ncRNA - non-coding RNA
- RNA - Ribonucleic Acid
- RNA-seq - RNA-sequencing
- TKI - Tyrosine Kinase Inhibitors

# Introduction

Clear cell renal cell carcinoma (ccRCC) is the most common subtype or renal cell carcinoma (RCC), whereas RCC itself accounts up to 90% of all kidney cancers (Hsieh et al. 2017). Overall, kidney cancer is one of the most common cancers in the world, affecting mostly people aged 60 and over ( 2023). Up to one half of patients develop metastases (Janzen et al. 2003), which seriously limits the possibility of local surgical treatment. Chemotherapy cannot be applied in case of ccRCC (Moreira et al. 2020). With the development of cancer immunotherapy, more options of treatment became available. The first generation of medications: cytokines interleukin-2 (IL-2) and interferon not only had a poor response rate (about 10%) , but also toxic adverse effects. Over the last two decades tyrosine kinase receptor inhibitors (TKI) were in common usage due to their higher response rate (up to 27%) (Wang et al. 2022a). Nevertheless, most patients develop drug resistance to it on the first year of treatment. Recently emerged immune checkpoint inhibitors (ICI) are considered to be the best choice, either in monotherapy or in a combination with TKI. ICI medications encourage the patient's own immune system to destroy cancer by blocking immune checkpoint proteins from binding with their partner proteins ( 2022), which allows T-cells to kill cancer cells (Fig. 0.1).

Most popular ICI drugs act against a checkpoint protein called PD-1. There are several names of them, for example pembrolizumab and nivolumab, which can all be assembled into one group of anti-PD-1 immunotherapy medications. Despite this sensational discovery in cancer treatment, the problem of immunotherapy resistance remains unsolved. In addition, many patients experience serious adverse events, which occurence is highly individual. These are the main reasons why ICI treatment should be thoroughly studied so that clinicians could decide whether the patient would benefit from a particular medicine. Though drug resistance depends on many factors like sex, gut microbiota and tumor microenvironment (Moreira et al. 2020), the genetic aspect remains to be highly informative. Thus,

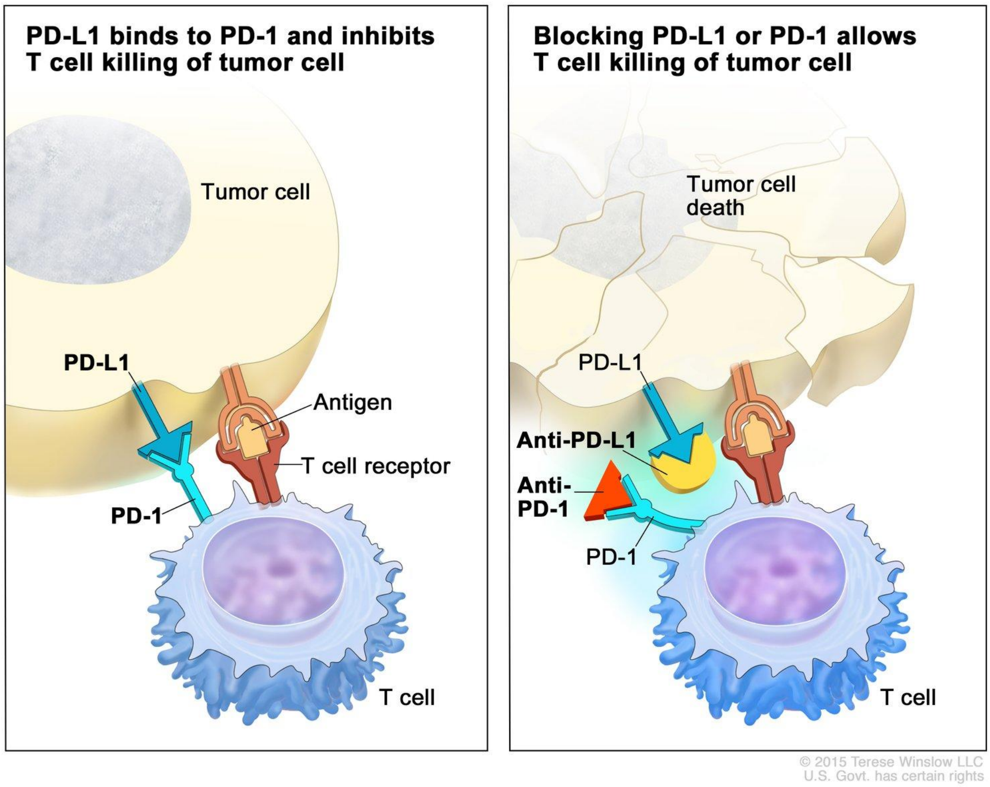
PD-L1 binds to PD-1 and inhibits T cell killing of tumor cell
Tumor cell
PD-L1
Antigen
T cell receptor
PD-1
T cell
Blocking PD-L1 or PD-1 allows T cell killing of tumor cell
Tumor cell death
PD-L1
Anti-PD-L1
Anti-PD-1
PD-1
T cell
© 2015 Terese Winslow LLC
U.S. Govt. has certain rights

Figure 0.1: Checkpoint proteins, such as PD-L1 on tumor cells and PD-1 on T cells, help keep immune responses in check. The binding of PD-L1 to PD-1 keeps T cells from killing tumor cells in the body (left panel). Blocking the binding of PD-L1 to PD-1 with an immune checkpoint inhibitor (anti-PD-L1 or anti-PD-1) allows the T cells to kill tumor cells (right panel).
*Credit: © Terese Winslow*

in the current work we aim at searching prognostic biomarkers for the anti-PD1 therapy, which has never been done for this type of ICI treatment specifically for ccRCC.

The main goal of this research is to find a group of key genes, which have the greatest impact on immunotherapy resistance, that is, a negative response to the treatment. The initial task list expanded during the work and now can be stated as following:

1 Data preprocessing

2 Finding DEG - Differentially Expressed Genes

3 Filtering DEG using Logistic Regression

4 Clustering

5 Single-cell analysis

6 Biological research using bioinformatics tools and databases

7 GSEA - Gene Set Enrichment Analysis

The main result of the work is discovering 44 genes, justifiably responsible for the resistance to anti-PD1 therapy against ccRCC. Though among these genes there are known for their impact on overall survival, some are unknown and long ncRNAs - noncoding RNAs. Moreover, during the process another result emerged: it appeared that the genes can be perfectly divided into 2 clusters, which may indicate the presence of 2 subtypes of ccRCC in the dataset or something different, but still meaningful.

Due to the novelty of ICI immunotherapy (the first medication of this class was registered in 2011 - see Cameron, Whiteside, and Perry 2011), this scientific field is insufficiently explored. Another problem is data collection, as currently there is a lack of big datasets of cancer patients undergoing a particular therapy. Nevertheless, we managed to shed light on the genetic aspect of anti-PD1 therapy for ccRCC, so we hope these results will be used in future researches on similar topics or in drug design against kidney cancer.

It was decided to divide this paper into two parts (main chapters). In chapter "Bioinformatics Workflow" the reader can follow a step-by-step data analysis by reading the corresponding subsections. We consider this section to be more technical, containing mostly data analysis itself. The second chapter "Biological Analysis" was made to state all the used biological databases and instruments to understand the results obtained from chapter one, and a short resume concerning the found key genes' properties and function.

# Literature Review

## Immunotherapy Resistance

In general, response to the therapy means benefitting from it. Talking about immunotherapy, it is considered that a patient must develop his defensive immune reaction to fight cancer. Resistance, at the same time, means having no response to the medication. Resistance itself can be of 2 types: primary and acquired. Primary resistance is when a patient does not respond to the drug at all, while acquired means that a patient was having positive response for some time, after which he relapsed as the medication stopped working (Nowicki, Hu-Lieskovan, and Ribas 2018).

There are several reasons for developing resistance for both types. Interferon resistance, loss of function of T-cells and immunosuppressive tumor microenvironment are only some of them. However, knowing the reasons is not enough to overcome immunotherapy resistance. Choosing a proper treatment without wasting time and experiencing serious adverse effects can be life-saving for oncology patients, so it would be wise to make the selection process more personalised, not only based on the protocols and clinical recommendations, but also the genomic pattern of the patient. Luckily with the development of new technologies it is possible to discover genes and genetic processes serving as prognostic biomarkers for the therapy. As written in Lavacchi et al. 2020, some of them are already discovered, for example PBRM1, B7-H1 and PD-L1 expression. It is also interesting that not only genes affect the prognosis, but also ncRNAs (Cheng et al. 2022), which is important, as among the 44 genes we obtained a high percentage is ncRNAs. We hope that the results of our work can be useful for estimating new prognostic biomarkers (in our case - key genes) for the PD1 therapy prognosis, particularly for ccRCC.

## Key Genes Detection

The task of detecting key genes is quite common in bioinformatics. As for cancer, it is usually done with the goal of predicting therapeutic effect or building a prognosis model using risk scores. Huang et al. 2021 revealed genes associated with ccRCC by conducting bioinformatics analysis, including several steps like selecting genes by differential expression, multivariate Cox regression, single-cell analysis and more. As a result, three hub genes out of 2492 were selected. This information may serve as a base for the development of new medications against ccRCC.

The work made by Puzanov 2022 was aimed at identifying key genes associated with ccRCC subtypes with the worst survival rates. To begin with, all the studied samples were clustered by survival rate. After that, hub genes were selected for each cluster. At the same time, it was revealed that expression of several genes from the cluster with the worst prognosis affects the response answer to certain immunotherapy medications. Research methods included k-means clustering, ROC-analysis and construction of a predictive model by Multivariate Cox regression. Networks of protein–protein interactions were also built to find out the functionality and connection of the filtered genes.

Another research concerning ccRCC biomarkers was also conducted by Wang et al. 2022b, but this time associated with ICI immunotherapy prognosis in general from over 600 cancer samples from several databases. Here, like in other similar studies, differential expression analysis and multivariate Cox regression were done, as well as some statistical analysis using R. Finally, five key genes were selected, with whose help the prognostic risk model was constructed and the patients were divided in two groups of low- and high-risk. The given groups were thoroughly studied, which led to some notable conclusions regarding abundance in cells of different types, as the cell composition varied between the groups.

# Conclusions

As for our work, we explore the mechanisms of immunotherapy resistance for particularly anti-PD1 therapy in case of ccRCC, whilst similar works, like the ones we referred to before, usually do not focus on a concrete type of treatment or a disease. In other words, not only did we choose a particular group of medicines (ICI - immune checkpoint inhibitors), but also anti-PD1 therapy from this group and, moreover, a particular type of cancer. We have chosen anti-PD1 because it is one of the most popular drug group of the new generation immunotherapy. Here we have to mention that we do not take into the account which particular medication was used (the most popular anti-PD1 drugs are Pembrolizumab, Nivolumab and Cemiplimab), as they act the same way as described in Introduction, the only difference is in the age of creation, manufacturer, drug tolerance and the type of cancer they are used against.

Some methods used in this work are the same as in the articles mentioned above, like finding DEG and clustering, nevertheless, our other methods were not used there, but our decision to use them emerged depending on the situation while the working process and intermediate results.

# 1 Bioinformatics Workflow

## 1.1 Data Preprocessing

The bulk transcriptome RNA-seq data was obtained from TIGER (Tumor Immunotherapy Gene Expression Resource) database, ID RCC-Braun_2020, originally from Braun et al. 2020. There are 311 patients, who received mono anti-PD1 therapy or in combination with Everolimus (immunosuppressive drug), of which 44 responded to the therapy and 237 did not. The dataset itself consists of expression and clinical data.

Expression data is a classical table, where we have genes as rows and samples as columns. Each cell therefore is a number, indicating the level of genetic expression. The table is already log2-normalised.

Clinical data provides some information on each sample (patient): age, gender, type of therapy, overall survival and most important - response.

First of all, we had to filter the samples (311 patients = 311 samples at the start), deleting those, who were combinated with Everolimus (as we wanted to examine mono anti-PD1 therapy) or had an unknown response. After this we had 172 samples.

Secondly, gene list required some filtering, too. We started with 56269 genes and deleted all NaNs, which resulted in 43765 genes left in total. See 1.1 for visual demonstration.

All the work on this step was done using Pandas (v.1.5.3) on Python.

Table 1.1: Genes & samples count on different steps

| | Genes | Samples |
|---|---|---|
| Original | 56269 | 311 |
| Preprocessed | 43765 | 172 |
| After DEG analysis | 366 | 172 |
| After ML filter | 44 | 172 |

## 1.2 Searching for DEG

DEG, in our case, are genes statistically significantly expressing in a different way in samples with negative response to the therapy compared to positive-responded ones. Several tools exist to find such genes, most of them containing a statistical model inside. Limma package from Bioconductor on R (Ritchie et al. 2015) was used for such purpose, as it can handle already normalised gene expression, whereas other similar tools require original read count. After using Limma (v.3.56.1) with p-value 0.01, 366 genes were obtained.

## 1.3 Machine Learning in gene filtering

Our initial plan was the following. Firstly, we scaled each gene's expressions by the standart scaling via substracting the mean and dividing by the standart deviation. For each gene out of the 366, an own Logistic Regression model was made. On this step, scikit-learn (v.1.2.2) was used. In terms of Machine Learning our objects are samples, our only feature is the expression of this gene and target is the answers to therapy in all samples. This data matrix was splitted into train and test data (test size 0.3) and, finally, AUC score was calculated on the test data. It was expected to save only those genes, whose AUC was higher than a certain threshold, for example 0.6. Unfortunately, it was impossible to do due to the overfitting of the models. We do not have many objects (only 172, of which train accounts for 120 samples) and we could not expand the dataset due to the lack of them, as we mentioned before. Traditional methods to overcome overfitting are not applicable in our case, as we cannot throw out any samples for no reason or add any synthetic data. As a result, top AUCs were higher than 0.9 and the remaining numbers were around 0.8, which is too much to use for filtering genes - simply not common in bioinformatics.

Logistic Regression was, however, used to reduce the number of genes in another way. We constructed a traditional Logistic Regression with L1 regularisation on a whole dataset, so we had samples as objects and genes as features, responses as target. Thanks to L1 many weights became zero, so we saved only genes with nonzero coefficients in the model, as it means that these features are of the main importance. As a result, 44 genes remained.

## 1.4 Clustering

It was decided to cluster the genes by their expression. To do this, K-means from scikit-learn was used, as this algorithm is one of the most popular and usually provides decent results. We iterated over the number of clusters from 2 to 10 and studied the illustrations in reduced space. It was clearly seen that we have exactly 2 clusters, not more. PCA from scikit-learn and UMAP from package umap-learn v.0.5.3 (McInnes, Healy, and Melville 2020) were used to reduce space and then make illustrations using matplotlib (v.3.7.1). UMAP gave better results than PCA, so in the final verison we used it. It is interesting that the cluster division is correct for both 366 and 44 genes, even the proportion visually remains the same (see 3.1 (Appendix) and 1.1 for clusters visualisation), while the initial 43765 genes after the preprocessing cannot be clustered.

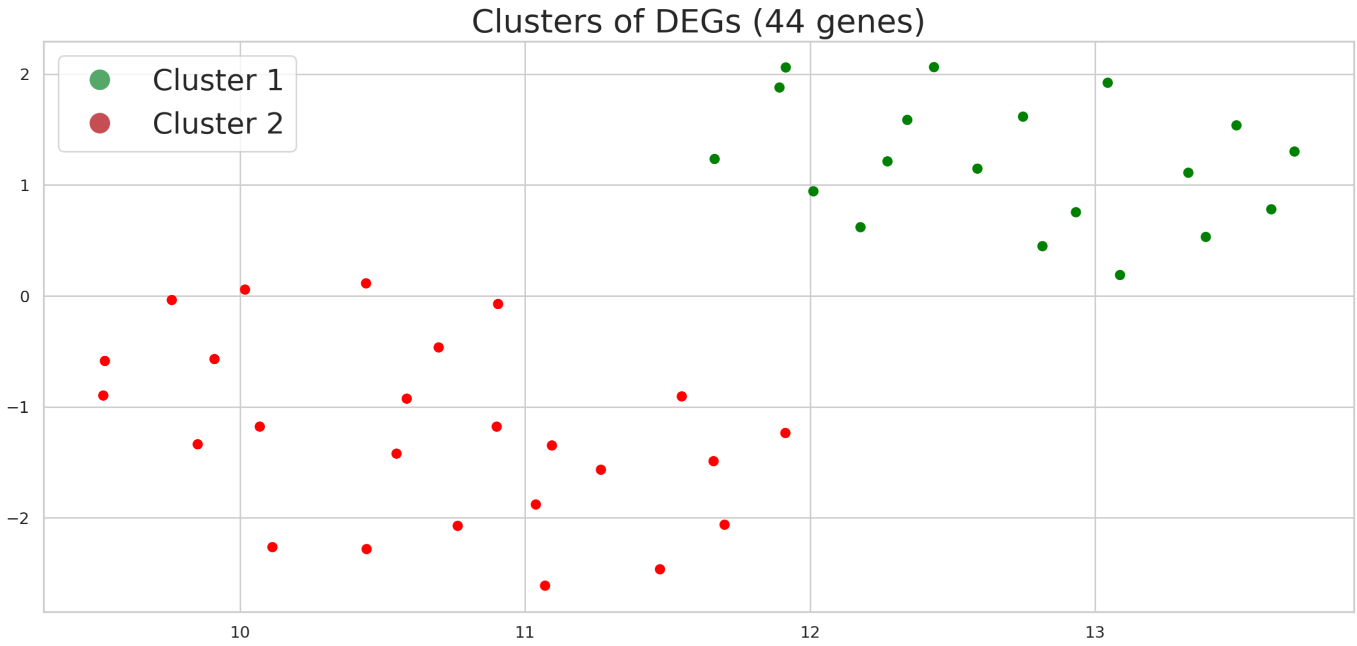

Figure 1.1: Clusters of the 44 genes, based on their expression.

Another interesting observation is that AUCs in the second cluster are mostly higher, than in the first one. It was proved by making bubble plots on 3.2 (Appendix), 1.2 and can also bee seen from the table 3.1. Moreover, the second cluster contains more coding genes.

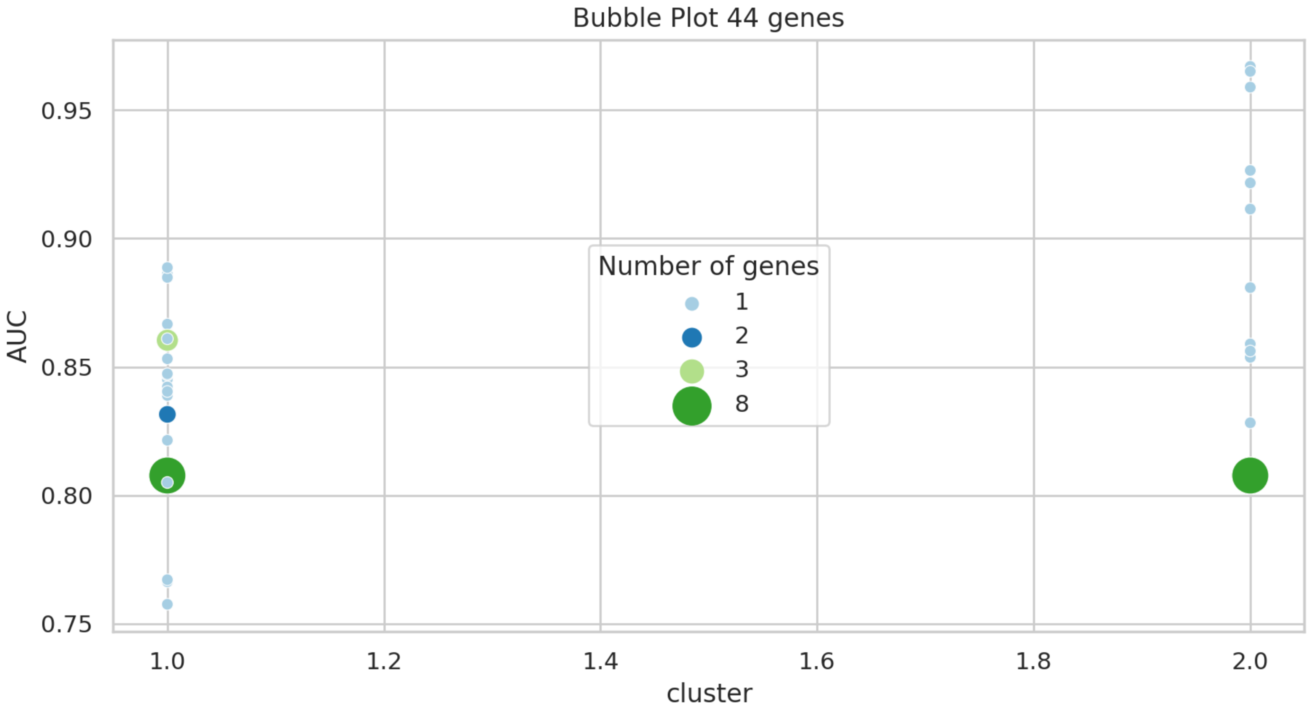

Figure 1.2: Bubble plot, illustrating 44 genes by their belonging to clusters. The size of each bubble indicates the number of genes with the same AUC value. It can be seen that AUCs from the second cluster are significantly higher.

## 1.5 Single-Cell RNA-seq Analysis

We found it interesting to conduct a single-cell RNAseq analysis. Single-cell is a sequencing method aimed at determining the nucleic acid sequence, allowing to obtain gene expression from separate cells. It is possible to assign cell types and make visualisations of clusters for each type to see in which cell clusters the gene expresses the most. Therefore we wanted to check if the found genes are linked to any particular cell type. To do this, Seurat (v.4.3.0) package on R was used (Hao et al. 2021). We used data from NCBI GEO (Edgar, Domrachev, and Lash 2002) (access code GSM4630029), which represents a sample of tissue from ccRCC, sequenced by the single-cell technology. All the necessary steps, including data preprocessing, clustering, dimension reduction (UMAP) and cell type assignment were done following the tutorial from Andrews et al. 2021. We plotted all the coding genes from the list 3.1, but did not manage to find any connection to particular cell types. It can be seen on 1.3 - 1.4 for the gene EXTL3 that it expresses mostly in Endothelial cells, which is probably the only visible linkage to a cell type (and what we desired to see for the other genes), but still we can see that it expresses almost in all cell types. The situation with the other genes is not better and they express in any cell types without clearly visible gatherings, which did not give us any knowledge about our genes' connection to cell types.

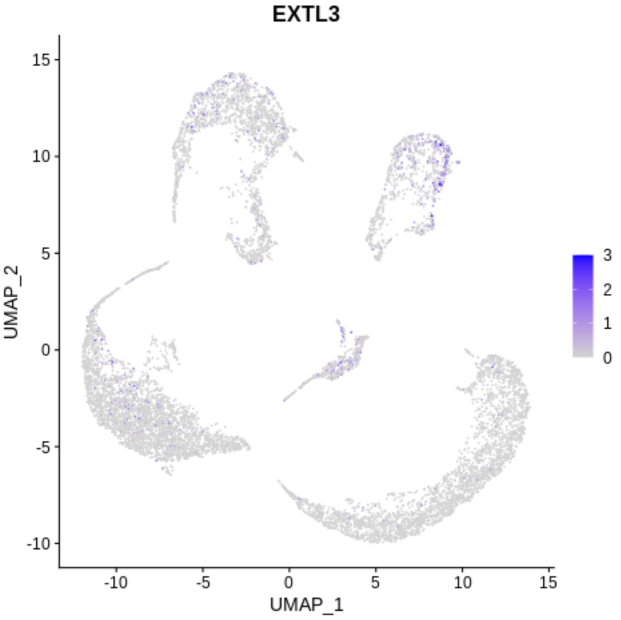

Figure 1.3: Cells in a reduced dimension. Violet colour represents expression of the gene EXTL3.

## 1.6 Results and Conclusions

Following the data analysis pipeline, including preprocessing, DEG searching and constructing linear models, it became possible to reduce the number of genes we are interested at from 56269 to 44. Now we can confidently say that the genes from this group are considered most significant in terms of their regulation of response to anti-PD1 therapy. All the 44 genes, including their AUCs, belonging to a cluster and coding function can be seen from the table 3.1. Here we list again our main observations:

- the genes perfectly divide into 2 clusters of similar size

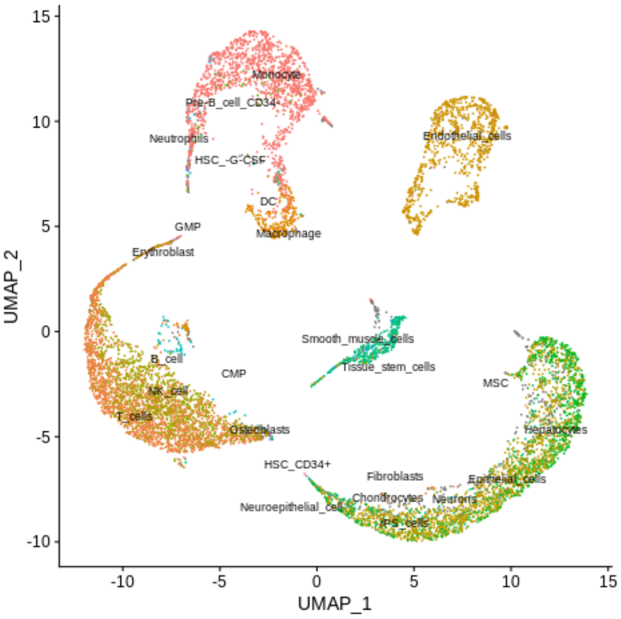

Figure 1.4: Cells in a reduced dimension with cell type annotation from Seurat (parameter hpca.fine).

- the cluster division is correct both for 366 and 44 genes
- AUCs for one cluster are higher and this cluster contains mostly coding genes
- no connection to any particular cell type was found for these genes

We hope that this group of genes will be studied closer by researchers. However, we suppose that such clustering may indicate that either we have 2 subtypes of ccRCC here or the coding genes have more impact of the therapy response. Both assumptions are notable and should be verified. In addition, one of the clusters has higher AUCs and content of coding genes, which makes us think that this cluster's genes are more important for the immunotherapy resistance.

# 2 Biological Analysis

This chapter describes all the biological databases and tools which were used for the purpose of investigating the function of the found 44 genes.

## 2.1 STRING

STRING (Szklarczyk et al. 2023) database (v.11.5) is a helpful tool to show physical and functional interactions between proteins in a visual form of graphs. It is often used to check if there is any connection between coding genes in a group. We used STRING to illustrate protein-protein interactions for our 44 key genes. Only 17 of them encode a protein, so we have 17 nodes in a graph. We do not have any edges here, which means that no existing of predicted interactions were found.

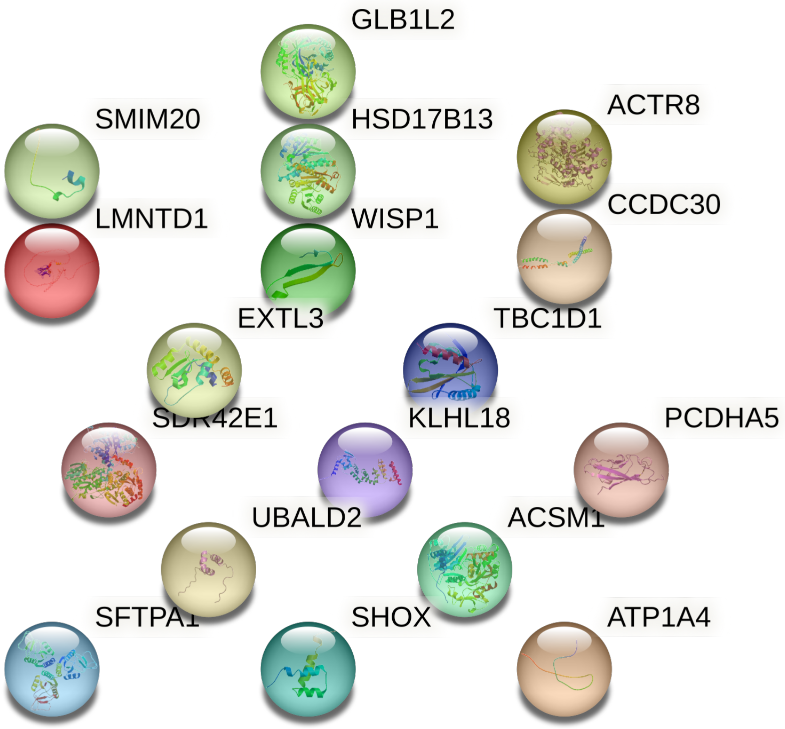

Figure 2.1: Graph of protein-protein interactions made in STRING. No edges means no connections between the proteins encoded by the genes.

## 2.2 NDEx & Cytoscape

We used NDEx (v.2.5.4) search tool (Pratt et al. 2015) integrated with Cytoscape (Shannon et al. 2003, v.3.10.0) to see in which pathways and processes do the genes from our set take part. However, the result did not meet our expectations as there were no processes in which more than 2 genes from the set were present. Moreover, the pathways shown in Cytoscape were quite common and not specific for ccRCC, so it did not make us interested.

## 2.3 NPInter

As among the found key genes some are non-coding RNAs, we wanted to check whether they interact with the proteins encoded by the rest, coding ones. In such case NPInter (Teng et al. 2020) database (v.4.0) can be used, as it shows various types of physical connections between RNAs and proteins, RNAs and DNAs. We were only interested in RNA-protein interactions, so browsed there our key genes. Unfortunately, no connections between them was found.

## 2.4 GEPIA 2

GEPIA 2 (Tang et al. 2019) is a resource which uses cancer-related databases to study gene expression in tumors. In our case it was used to make survival analysis. Survival plot is a plot illustrating percents of survived patients through months. We choosed KIRC - kidney renal clear cell carcinoma as a dataset for the plot and ran the analysis on key coding genes. Successfully, several genes have impact on the overall survival of patients with KIRC. Three of them: SMIM20, TBC1D1 and EXTL3 are illustrated below (2.2 - 2.4).

GEPIA 2 also provides bodymaps of gene expression in tumor and normal samples. We also checked our genes here and discovered that TBC1D1' expression is the highest in kidney tumors, to be precise, in KICH (kidney chromophobe). We hope it may serve as another approval for the importance of this gene in our task (see illustrations 2.5 - 2.6).

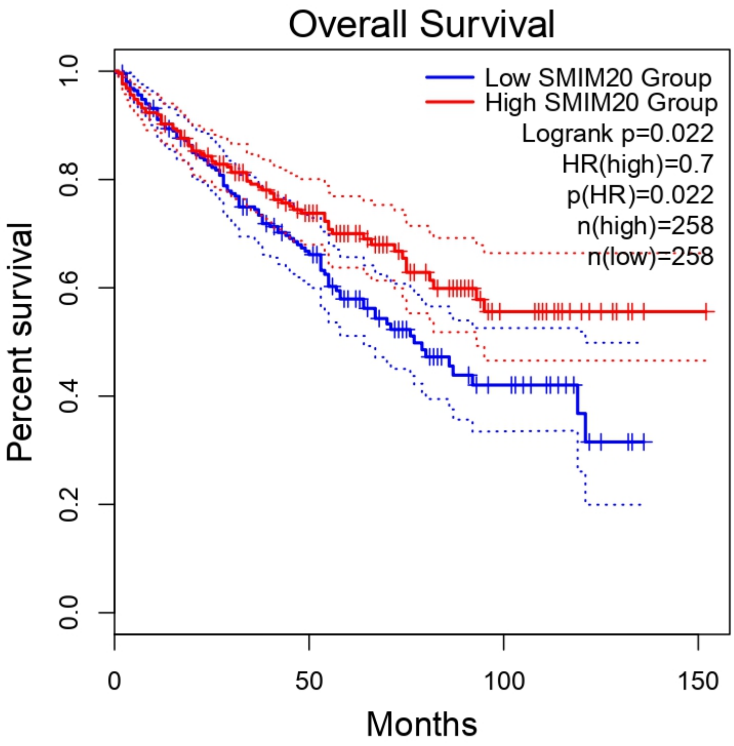

Figure 2.2: Survival plot for patients with KIRC for gene SMIM20. Red line means high expression of SMIM20 and blue is low. We can see here that patients with high expression of SMIM20 live longer.

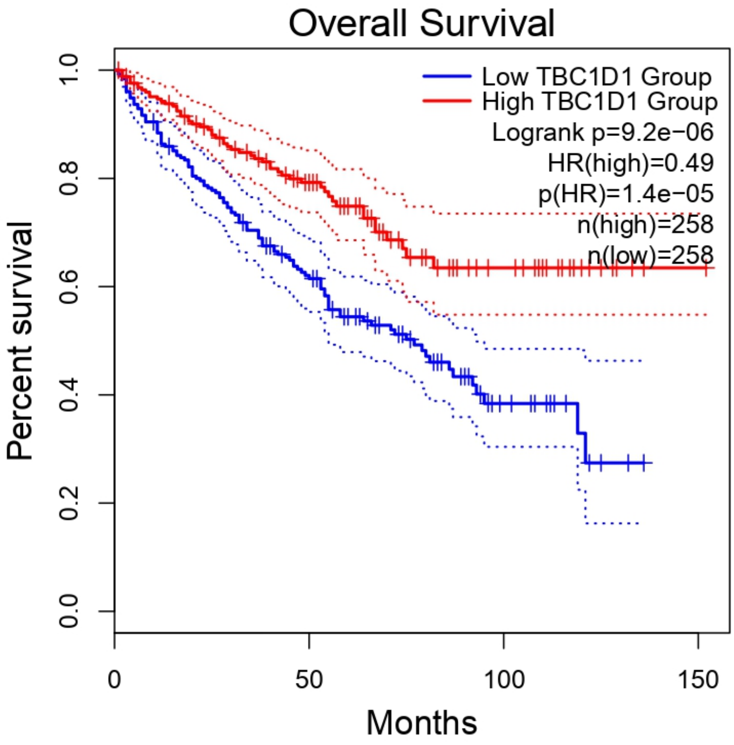

Figure 2.3: Same as 2.2, but for gene TBC1D1

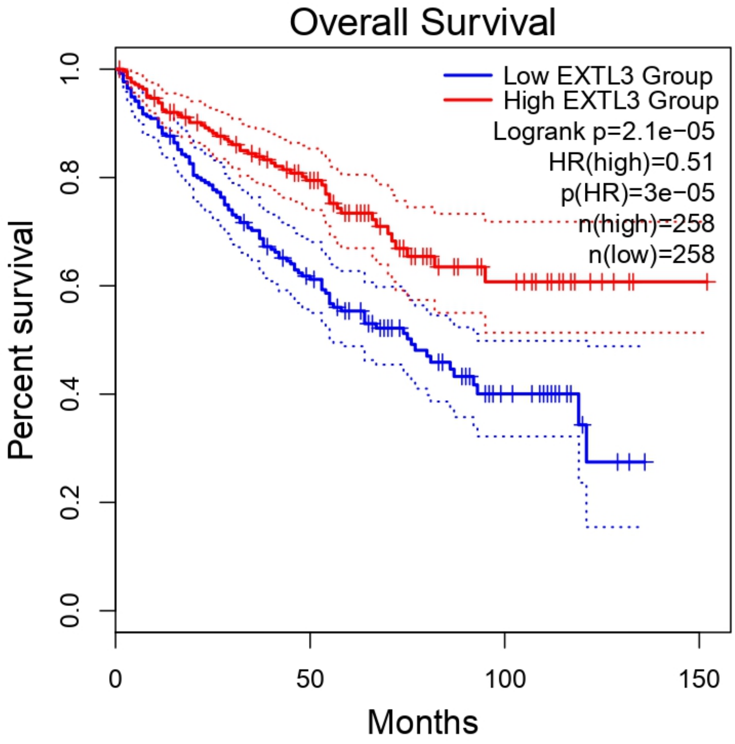

Figure 2.4: Same as 2.2, but for gene EXTL3

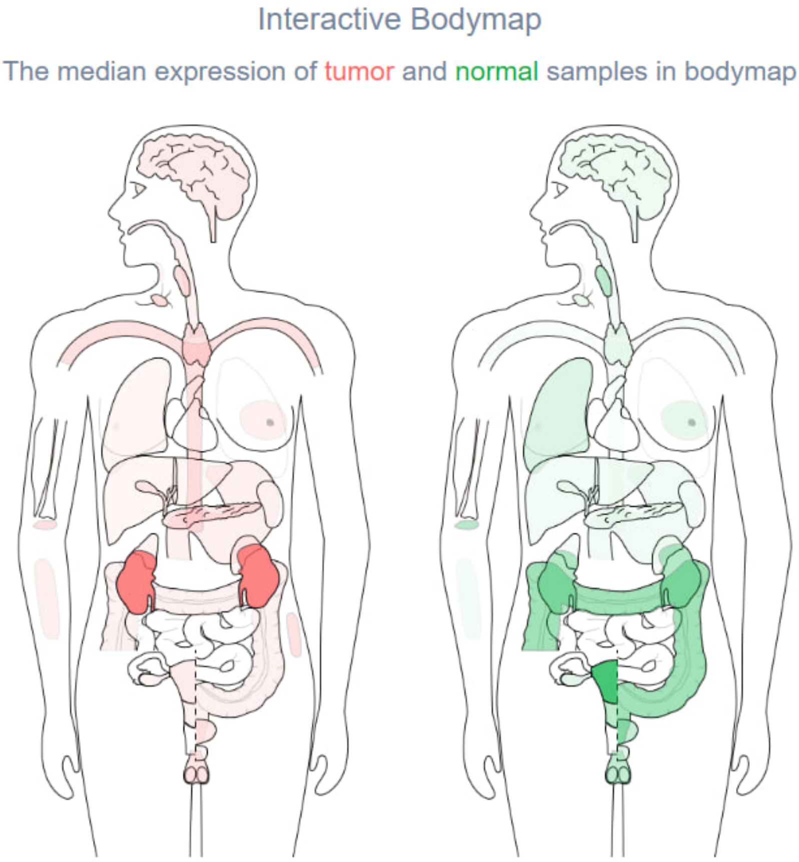
Interactive Bodymap
The median expression of tumor and normal samples in bodymap

Figure 2.5: Bodymap, illustrating expression of TBC1D1 in tumor and healthy samples. The brighter the colour, the higher the expression. It can be seen that in tumor samples the highest expression is in kidneys.

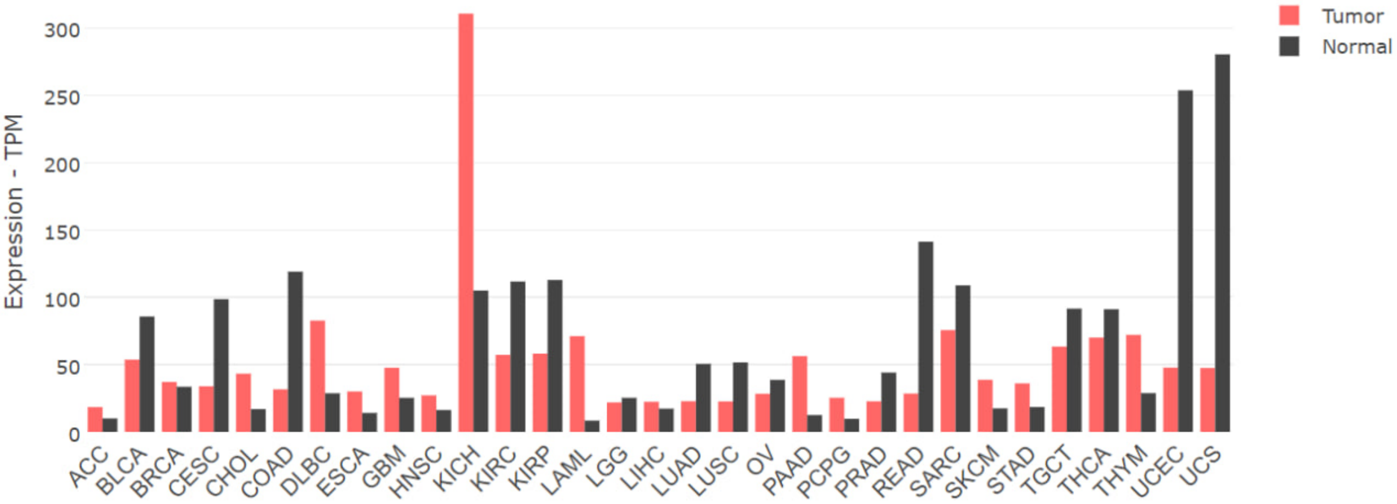

Figure 2.6: Bar plot illustrating expression of TDC1D1 in tumor samples and their healthy pairs.

## 2.5 GSEA

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether a defined set of genes shows statistically significant differences between two biological states - phenotypes. (Subramanian et al. 2005). It is possible to use different biological databases to find pathways involving the gene set. It was decided to use a Python implementation of the GSEA analysis called GSEApy (Z. Fang, Liu, and Peltz 2023, v.1.0.4).

Firstly, the samples need to be splitted into 2 phenotypes by some condition. Our approach is: mark samples with positive response to the therapy as the first phenotype and negative responded as the second one. As a result, 172 samples were splitted into 2 groups of 39 responders and 133 non-responders. Next we constructed a table for the 44 genes with their expression in samples and renamed the columns (samples) according to their response. Finally, we ran GSEA analysis in GSEApy, changing the parameter of the minimal count of genes in a set to 2 from the default 15, as we have only 44 genes and the default parameter is aimed at thousands of them. As a final part, we had to try several gene sets from the library to see in which pathways some of our genes are represented.

The result of the GSEA analysis is shown here 2.1.

As a gene set we used "Jensen_DISEASES" (Pletscher-Frankild et al. 2015),

| Name | Term | ES | NES | NOM p-val | FDR q-val | FWER p-val | Tag % | Gene % | Lead_genes |
|---|---|---|---|---|---|---|---|---|---|
| gsea | Carcinoma | 0.649 | 1.568 | 0.051 | 0.082 | 0.033 | 09/10 | 40.91% | TBC1D1;<br>EXTL3;<br>ACTR8;<br>ACSM1;<br>ATP1A4;<br>GLB1L2;<br>PCDHA5;<br>KLHL18;<br>CCDC30 |
| gsea | Kidney cancer | 0.600 | 1.239 | 0.220 | 0.216 | 0.15 | 4/5 | 29.55% | TBC1D1;<br>EXTL3;<br>ATP1A4;<br>PCDHA5 |

Table 2.1

it shows gene association to diseases. Though the table showed 2 diseases: carcinoma and kidney cancer, we cannot consider kidney cancer to be significant due to its p-value $> 0.05$ (column NOM p-val). Even for carcinoma we can see p-value slightly bigger than accepted, but still we decided to show it as a result. Looking at the columns again, we would like to clarify that ES and NES are enrichment score and normalised enrichment score respectively. The genes in the set are ranked by their expression, and enrichment score shows how much the genes in a gene set are overrepresented at the top/bottom of the ranked list.

If we then look at the GSEA plot 2.7, we may notice thin blue sticks in the center of the picture - these are our genes ranked by their expression in a gene set, a total of 10. The green mountain is plotted by going through the ranked list and calculating expression change. It is better when most genes are concentrated before the peak and this is exactly what we see here. This information is also illustrated in a column "Tag %", which means that 9 of the 10 genes are in the right place, they are also shown in the column "Lead_genes". "FDR q-val" column is simply false discovery rate. It is not that high and, moreover, the plot 2.7 is not that bad in GSEA (if it looked like normal disctribution, we would not include it and mark as bad result), so that is why we illustrate our GSEA results and hope that they may serve as slight, yet not that confident, but still approval that at least the whole work we have done makes sense.

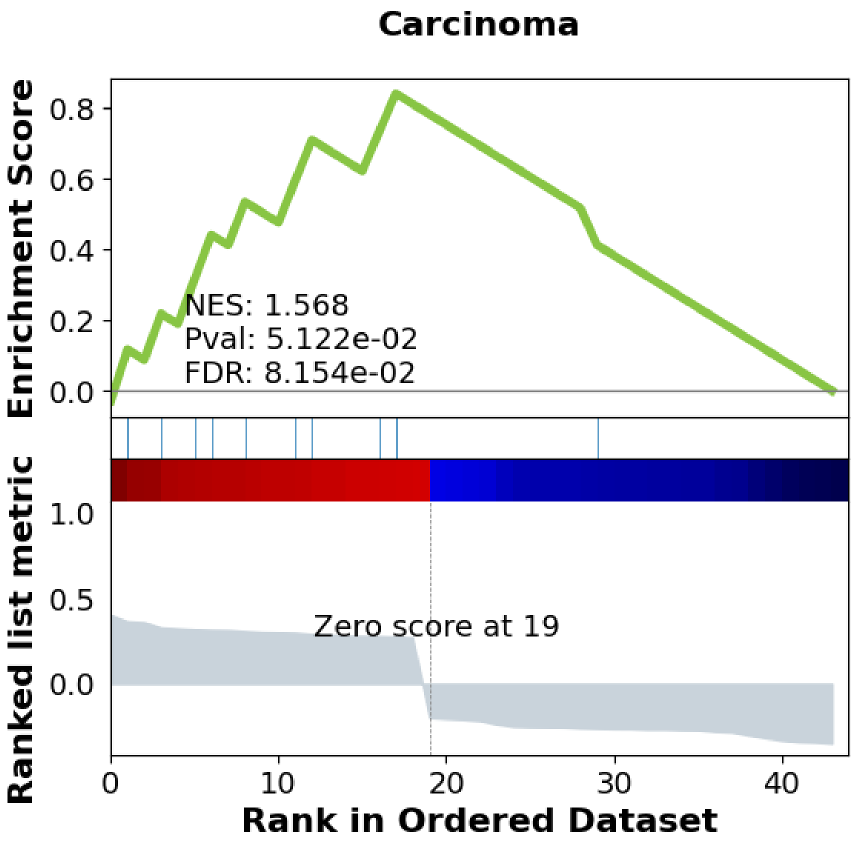

Figure 2.7: Bar plot illustrating expression of TDC1D1 in tumor samples and their healthy pairs.

## 2.6 Results and Conclusions

In this chapter we described the biological analysis, conducted to learn more about the found genes and somehow prove their importance. Unfortunately, some tools did not provide us with novel information as in the gene list contains many ncRNAs. However, we consider such content significant, cause as we mentioned before, ncRNAs can serve as predictive biomarkers for the anti PD-1 therapy prognosis. As for the coding genes obtained, they impact the overall survival of patients with kidney cancer. GSEA analysis results have shown connection of the genes to carcinoma, even though the p-value is arguable.

# Conclusion

The whole research was conducted to find genetic prognostic biomarkers - key genes, which may then be used in personalised medicine or drug design. Though the results require verification and future investigation, we hope they will be useful due to the narrowness of the task: as we have browsed, this is the only bioinformatics analysis for searching key genes specially for the prognosis of anti-PD1 therapy against ccRCC. It was not an easy task due to the lack of proper databases, but, nevertheless, we did our best to overcome the troubles. We proved that among the 44 found genes some are of obvious importance, while others are insufficiently explored, which makes it even more challenging. In addition, the discovery of 2 clusters based on the gene expression seems interesting, which should definitely studied further. In general, the real prospects of the future work is simply studying closer the gene list we provide, which requires good biological knowledge. In the end, however, there is a real prospect of adding new prognostic biomarkers in ccRCC immunotherapy treatment.

# References

1. Apr. 2022. URL: https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/checkpoint-inhibitors.

2. Mar. 2023. URL: https://www.cancer.net/cancer-types/kidney-cancer/statistics#:~:text=In%5C%202023%5C%2C%5C%20an%5C%20estimated%5C%2081%5C%2C800,most%5C%20common%5C%20cancer%5C%20for%5C%20men..

3. Tallulah S Andrews et al. "Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data". In: *Nature protocols* 16.1 (2021), pp. 1–9.

4. David A Braun et al. "Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma". In: *Nature medicine* 26.6 (2020), pp. 909–918.

5. Fiona Cameron, Glenn Whiteside, and Caroline Perry. "Ipilimumab: first global approval". In: *Drugs* 71.8 (2011), pp. 1093–1104.

6. Chao Cheng et al. "Overcoming resistance to PD-1/PD-L1 inhibitors in esophageal cancer". In: *Frontiers in Oncology* 12 (2022).

7. Ron Edgar, Michael Domrachev, and Alex E Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic acids research* 30.1 (2002), pp. 207–210.

8. Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. "GSEApy: a comprehensive package for performing gene set enrichment analysis in Python". In: *Bioinformatics* 39.1 (2023), btac757.

9. Yuhan Hao et al. "Integrated analysis of multimodal single-cell data". In: *Cell* (2021). DOI: 10.1016/j.cell.2021.04.048. URL: https://doi.org/10.1016/j.cell.2021.04.048.

10. James Hsieh et al. "Renal cell carcinoma". In: *Nature Reviews Disease Primers* 3 (Mar. 2017), p. 17009. DOI: 10.1038/nrdp.2017.9.

11. Hao Huang et al. “Identification of hub genes associated with clear cell renal cell carcinoma by integrated bioinformatics analysis”. In: *Frontiers in Oncology* (2021), p. 3857.

12. Nicolette K Janzen et al. “Surveillance after radical or partial nephrectomy for localized renal cell carcinoma and management of recurrent disease”. In: *Urologic Clinics* 30.4 (2003), pp. 843–852.

13. Daniele Lavacchi et al. “Immune checkpoint inhibitors in the treatment of renal cancer: current state and future perspective”. In: *International Journal of Molecular Sciences* 21.13 (2020), p. 4691.

14. Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. arXiv: 1802.03426 [stat.ML].

15. Marco Moreira et al. “Resistance to cancer immunotherapy in metastatic renal cell carcinoma”. In: *Cancer Drug Resistance* 3.3 (2020), p. 454.

16. Theodore S Nowicki, Siwen Hu-Lieskovan, and Antoni Ribas. “Mechanisms of resistance to PD-1 and PD-L1 blockade”. In: *Cancer journal (Sudbury, Mass.)* 24.1 (2018), p. 47.

17. Sune Pletscher-Frankild et al. “DISEASES: Text mining and data integration of disease–gene associations”. In: *Methods* 74 (2015), pp. 83–89.

18. Dexter Pratt et al. “NDEx, the Network Data Exchange”. In: *Cell Systems* 1 (Oct. 2015), pp. 302–305. DOI: 10.1016/j.cels.2015.10.001.

19. Grigory Andreevich Puzanov. “Identification of key genes of the ccRCC subtype with poor prognosis”. In: *Scientific Reports* 12.1 (2022), p. 14588.

20. Matthew E Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47–e47.

21. Paul Shannon et al. “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. In: *Genome research* 13.11 (2003), pp. 2498–2504.

22. Aravind Subramanian et al. “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles”. In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

23. Damian Szklarczyk et al. “The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D638–D646.

24. Zefang Tang et al. “GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis”. In: *Nucleic acids research* 47.W1 (2019), W556–W560.

25. Xueyi Teng et al. “NPInter v4. 0: an integrated database of ncRNA interactions”. In: *Nucleic acids research* 48.D1 (2020), pp. D160–D165.

26. Qi Wang et al. “Immune-associated gene signatures serve as a promising biomarker of immunotherapeutic prognosis for renal clear cell carcinoma”. In: *Frontiers in Immunology* 13 (2022).

27. Qi Wang et al. “Immune-associated gene signatures serve as a promising biomarker of immunotherapeutic prognosis for renal clear cell carcinoma”. In: *Frontiers in Immunology* 13 (2022).

# 3 Appendix

## 3.1 Packages & Databases versions

- Pandas v.1.5.3
- Limma v.3.56.1
- scikit-learn v.1.2.2
- umap-learn v.0.5.3
- matplotlib v.3.7.1
- Seurat v.4.3.0
- STRING v.11.5
- NDEx v.2.5.4
- Cytoscape v.3.10.0
- NPInter v.4.0
- GSEApy v.1.0.4

## 3.2 Illustrations

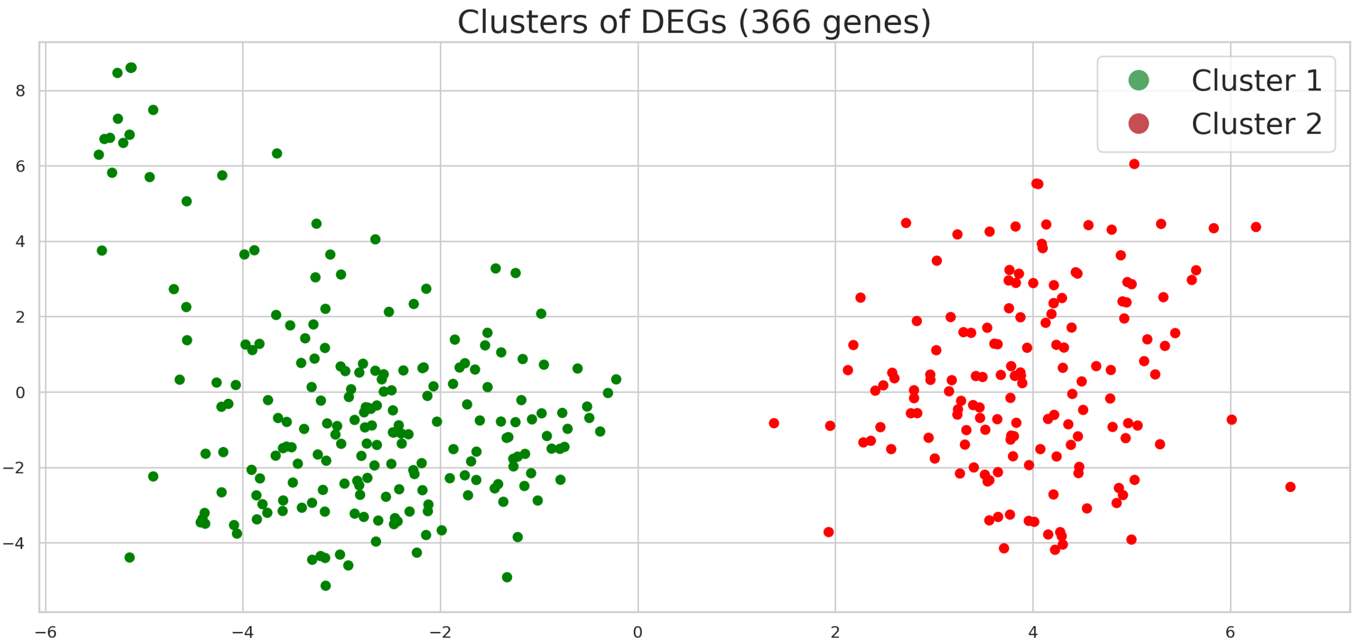

Figure 3.1: Clusters of the 366 genes, based on their expression.

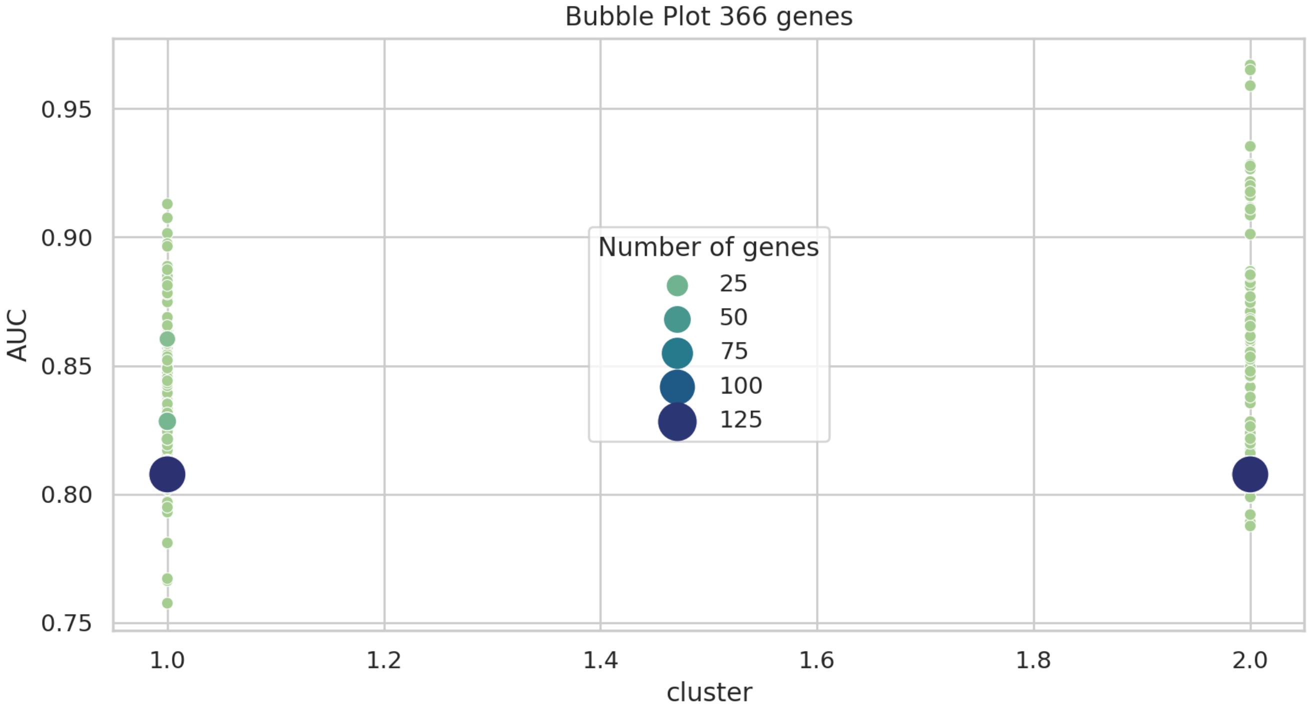

Figure 3.2: Bubble plot, illustrating 366 genes by their belonging to clusters. The size of each bubble indicates the number of genes with the same AUC value.

## 3.3 Tables

Table 3.1: 44 selected genes, their AUCs, cluster membership and whether it is a coding gene (codes a protein or not)

| Gene | AUC | Cluster | Protein |
|---|---|---|---|
| AC133965.1 | 0.954148674419627 | 2 | No |
| GLB1L2 | 0.941081780765721 | 2 | Yes |
| SMIM20 | 0.939105541904592 | 2 | Yes |
| EXTL3 | 0.934914336991141 | 2 | Yes |
| ACSM1 | 0.933937556631466 | 2 | Yes |
| RP11-152N13.16 | 0.927909816688834 | 2 | No |
| PPIAP26 | 0.927482562656981 | 2 | No |
| PHBP13 | 0.921443759221532 | 1 | No |
| PLBD1-AS1 | 0.919134786240593 | 2 | No |
| ATP1B3-AS1 | 0.917372881677255 | 1 | No |
| TBC1D1 | 0.916081963281663 | 2 | Yes |
| UBE2CP3 | 0.905062359738298 | 2 | No |
| ATP1A4 | 0.896214674246012 | 2 | Yes |
| LINC01293 | 0.893605282246404 | 1 | No |
| UBALD2 | 0.892724967027187 | 2 | Yes |
| SDR42E1 | 0.891494767790371 | 2 | Yes |
| CCDC30 | 0.882959999111795 | 2 | Yes |
| TUBBP2 | 0.878022009876825 | 1 | No |
| LMNTD1 | 0.868785193292271 | 1 | Yes |
| PCDHA5 | 0.865141440681886 | 2 | Yes |
| HSD17B13 | 0.858139575229036 | 2 | Yes |
| RP5-1142A6.5 | 0.847438179376452 | 1 | No |
| KLHL18 | 0.845979971430293 | 2 | Yes |
| RNA5SP216 | 0.841393587033122 | 1 | No |
| RP11-769N22.1 | 0.841393587033122 | 1 | No |

Table 3.1: 44 selected genes, their AUCs, cluster membership and whether it is a coding gene (codes a protein or not)

| Gene | AUC | Cluster | Protein |
|---|---|---|---|
| SHOX | 0.841390539801833 | 1 | Yes |
| RP11-463O9.9 | 0.838966475081386 | 1 | No |
| RP11-118H4.1 | 0.833105010089338 | 1 | No |
| RP11-367O10.1 | 0.83184381529556 | 1 | No |
| HSPE1P12 | 0.826923076923077 | 1 | No |
| RP11-383J24.2 | 0.826923076923077 | 1 | No |
| RP11-959F10.4 | 0.826923076923077 | 1 | No |
| RP11-5N11.3 | 0.826923076923077 | 1 | No |
| AC005077.14 | 0.826923076923077 | 1 | No |
| RP11-140H17.1 | 0.826923076923077 | 1 | No |
| RP11-382A20.1 | 0.826923076923077 | 2 | No |
| ACTR8 | 0.826923076923077 | 2 | Yes |
| SFTPA1 | 0.826923076923077 | 1 | Yes |
| RP11-2N1.2 | 0.813706982109992 | 1 | No |
| WISP1-OT1 | 0.807420023790944 | 1 | Yes |
| RP11-664I21.5 | 0.805260243632337 | 1 | No |
| HRAT17 | 0.805260243632337 | 1 | No |
| SLC25A36P1 | 0.800484496124031 | 1 | No |
| CTC-218B8.3 | 0.795527728085868 | 1 | No |