PRECISION MEDICINE: GENE AND CLINICAL DATA ANALYSIS OF RENAL CANCER

by

SUMEYYE SU

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
May 2020

To my husband Taha and my son Tahir

And my parents, siblings

Without whom none of my success would be possible.

ACKNOWLEDGEMENTS

I would first like to express my deepest gratitude to my advisor, Dr. Leili Shahriyari, for her great support, guidance and caring. I always feel that being her student is one of the best things that happen in my life since she guides me to study my dream projects.

I also would like to thank my thesis committee Dr. Tuncay Aktosun, Dr.Hristo Kojouharov and Dr. Andrzej Korzeniowski for their time to review my dissertation and attend my defense. I also would like to thank the Department of Mathematics, University of Texas at Arlington to support me and provide a great environment for my studies.

In addition, I would like to emphasize my deepest appreciate to my husband, Taha Su, for his great encouragements and love to complete my PhD degree and my son Tahir, for being understanding and his endless love.

Last but not least, I would like to thank my parents, Fetiye and Yakup, and siblings Inci, Kenan, Humeyra, Rabia and Abdullah for their love, encouragement and prays from Turkey.

April 27, 2020

ABSTRACT

PRECISION MEDICINE: GENE AND CLINICAL DATA ANALYSIS OF RENAL CANCER

Sumeyye Su, Ph.D.

The University of Texas at Arlington, 2020

Supervising Professor: Leili Shahriyari

Recent advances in biotechnology led to generation of large complex biological and clinical data sets that can be used to infer the underlying mechanism of many diseases and arrive at personalized treatments. One of these data sets are the whole genome profiles, including a good collection of publicly available human gene expression data sets. In the first part of this study, we analyzed gene expression profiles of patients with renal cell carcinoma (RCC). We found that the regulator of G-protein signaling 5 (RGS5) might play a crucial role in initiation and progression of RCC, and it might be prognostic. We observed that a high expression level of RGS5 is associated with better survival months. Importantly, when the grade of tumor increases, the RGS5 expression level significantly decreases. Although there is no difference between expression level of RGS5 in male and female patients with primary tumors in the right kidney, among patients with primary tumors in the left kidney, females have a significantly higher RGS5 expression than male patients. Interestingly, we also observed a significant association between the high expression level of RGS5 and low serum calcium level and elevated white blood cells level.

Moreover, the outcome of cancer treatments especially immunotherapeutic interventions depends on tumor immune environments. In recent years, the development of various immunotherapies has improved overall survival months of some cancer patients, including renal cancer patients. However, for all immunotherapeutic interventions, only a small groups of patients respond to the treatments. Therefore, it is crucial to investigate and classify immune variations of tumors to identify the groups of patients who might benefit from each treatment option. In the second part of this study, we estimate the percentage of each immune cell type in 526 TCGA renal tumors using "digital mass cytometry". K-mean clustering of tumors based on their immune variations indicates the existence of four distinct classes of renal cell carcinoma: Cluster 1 ($CD4 < CD8 \approx M\Phi$), in which the numbers of macrophages and CD8+ T-cells are approximately the same, and the number of CD4+ T-cells is slightly less than the number of CD8+ T-cells; Cluster 2 ($CD8 < CD4 < M\Phi$), in which the number of macrophages is significantly higher than the number of CD4+ and CD8+ T-cells; Cluster 3 ($CD4 < M\Phi < CD8$), in which the number of CD8+ T-cells is significantly higher than the number of macrophages and CD4+ T-cells; and Cluster 4 ($CD8 < CD4 \approx M\Phi$) in which the numbers of macrophages and CD4+ T-cells are approximately the same, and the number of CD8+ T-cells is significantly less than CD4+ T-cells. Moreover, we observe a high positive correlation between the number of CD8+ T-cells and the expression levels of IFNG and PDCD1. Importantly, higher stage and grade of tumors have a significantly higher percentage of CD8+ T-cells in tumors. In addition, the primary tumors of patients, who were tumor free at the last time of follow up, have a higher percentage of NK cells and mast cells compared to the patients with tumors at the last time of follow up.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Renal Cell Carcinoma

Renal cell carcinoma (RCC) is the most commonly seen malignant tumor type in adult kidneys [1], and it is considered a morphologically and genetically heterogeneous tumor [2]. The incidence rate of RCC is around twice as common in males as in females [3, 4], and the most common risk factors are obesity, diabetes, and hypertension [5]. According to the US National Cancer Institute, there were approximately 533,204 people with kidney and renal pelvis cancer in 2016, and 73,820 people are estimated as a new cases in 2019 which is 4.2% of all new cancer cases. Although 74.8% of the patients survived five years or more in 2009-2015, 14,770 people are estimated to die in 2019 because of this disease [6].

To understand the process of initiation and progression of RCC and discover effective treatments, several data sets including, clinical and pathological information, genomic alterations, DNA methylation profiles, and RNA and proteomic profiles of RCC primary tumors have been collected and analyzed. For example, by analyzing these data sets, The Cancer Genome Atlas (TCGA) research network found a correlation between worsened prognosis in patients with the most common type of RCC, clear cell renal cell carcinoma (ccRCC), and a metabolic shift involving increased dependence on the pentose phosphate shunt, decreased AMPK, decreased Krebs cycle activity, increased glutamine transport and fatty acid production [7]. Additionally, a study of 103 RCC patients showed a significantly high FABP7 mRNA expression in

men and a high expression of BRN2 protein in women. They have also observed a poor prognosis in females with low FABP7 and high BRN2 expression [8].

1.2 Data Information

1.2.1 Gene Expression Data.

RNA-seq (RNA-sequencing) is a technique that uses next generation sequencing (NGS) to analyze the transcriptome of gene expression patterns encoded within a sample RNA [9]. In this study, we analyzed the RNA-seq data of the primary tumors of 534 patients with RCC to see if there is any signature of clinical or demographic information of the patients in the gene expression data of their primary tumors. We collected TCGA (The Cancer Genome Atlas) project data from cBioPortal. TCGA is a landmark cancer genomics program that describes over 20,000 primary cancer and matched normal samples spanning 33 cancer types. TCGA data sets are used for ability to diagnose, treat, and prevent cancer [10].

1.2.2 Clinical Data.

There is also clinical and demographic information of patients, namely clinical data, available in cBioportal. The available clinical data differ from study to study. Features we used in our study are:

Gender: Female, Male

Tumor grade: The grades of tumors, which is determined by the appearance of the cancer cells under the microscope, provide us with some insight about how the cancer might behave. Lower grades represent the cancer cells that look more like normal cells; G1 and G2 tumors tend to grow slowly and spread less. Higher grades represent the cancer cells that look more different from the normal cells. Cancer cells in G3 and G4 tend to grow quickly and spread fast [11].

Laterality: It designates the side on which the renal cancer originates, left or right kidney.

Tumor stage: The stage of tumors indicates the spread of tumors using the most common staging system, TNM, for kidney cancer (Table 1.1) [12].

TNM	Disease Status	Size
T1	Tumor limited to kidney	≤ 7 cm
T1a	Tumor limited to kidney	< 4 cm
T1b	Tumor limited to kidney	> 4 cm
T2	Tumor limited to kidney	> 7 cm
T2a	Tumor limited to kidney	< 10 cm
T2b	Tumor limited to kidney	≥ 10 cm
T3	Growing into the fat around the kidney	
T3a	Tumor invades to renal vein or fat but not beyond Gerota's fascia	
T3b	Tumor extends to vena cava below the diaphragm	
T4	Tumor invades beyond Gerota's fascia	

Table 1.1. TNM Staging for Kidney Cancer

Tumor status: It provides us with the state or condition of individuals' neoplasm at the time of last follow up, tumor free, with tumor.

Serum calcium level: Level of serum calcium in the blood, low or normal.

WBC: White Blood Cell Level in the blood, low or elevated.

Age: Age of patients at which the cancer was first diagnosed.

OS status: Overall patient survival status, living or deceased.

OS Months: Overall patients survival in months since initial diagonosis.

DFS Status: Disease free status since initial treatment, disease free, recurred/progressed.

CHAPTER 2

GENE EXPRESSION ANALYSIS OF PATIENTS WITH RENAL CELL CARCINOMA

To be able to analyze RCC patients' data sets, we combine clinical data of 534 RCC patients and RNA-Seq gene expression profiles of their primary tumors, which include 20,531 gene expression values for each patient. Tables in Figures 2-8 show an overview of demographic and clinical features of the patients. These tables provide the number of patients in each subcategory, including gender, type of tumor, survival status, etc. Differences in the numbers are due to missing information for some patients.

2.1 Method

2.1.1 Data Preparation.

2.1.1.1 Normalization.

The first crucial step in analyzing data sets is the normalization, which has a significant effect on the results. One of the most common normalization methods for gene expression data is the standardization of the values of each gene. However, it has been shown that this routine approach might not be a good one [13] for gene expression profiles of tumors. By analyzing the gene expression profiles of colon primary tumors, Shahriyari showed that the distribution of expression levels of genes across patients are very different from one gene to another one, while the distribution of gene expression levels of one patient is very similar to each other. We have also observed the same thing in RCC. The left subplot of Figure 2.1 shows the average

value of each gene and their standard deviations; as one can see the average and standard deviation of expression levels might be hugely different from one gene to another gene. However, the average and standard deviation of gene expression levels of patients are very similar (the right sub plot of the Figure 2.1). In other words, the distribution of expression levels of genes in one patient is very similar to another patient. Therefore, to avoid loosing statistical information of the data, we normalized the gene expression data sets from the primary tumor by scaling gene expression of each patient, separately [14].

Specifically, we have a data set $[p_1, \cdots, p_n]$, where n is the number of patients. Each p_i is a list of gene expression values, i.e. $p_i = [g_{i1}, \cdots, g_{im}]$, where each g_j is the expression value of gene j, and m is the total number of genes ($m = 20531$). That means the data set is an $n \times m$-dimensional matrix $D = [g_{ij}]$, where g_{ij} is the expression of gene j in patient i. To use simple feature scaling normalization method, we found maximum gene expression value of each patient, and then we divided the values of each gene by the maximum gene expression values for the patient; i.e $q_i = max([g_{i1}, \cdots, g_{im}])$, $\hat{p}_i = [g_{i1}/q_i, \cdots, g_{im}/q_i]$, where $\hat{p}_i$ is the normalized gene expression vector for the patient. After that, we calculated the standard deviation of each gene across samples, and RGS5 was one of the top 10 most variant genes across patients.

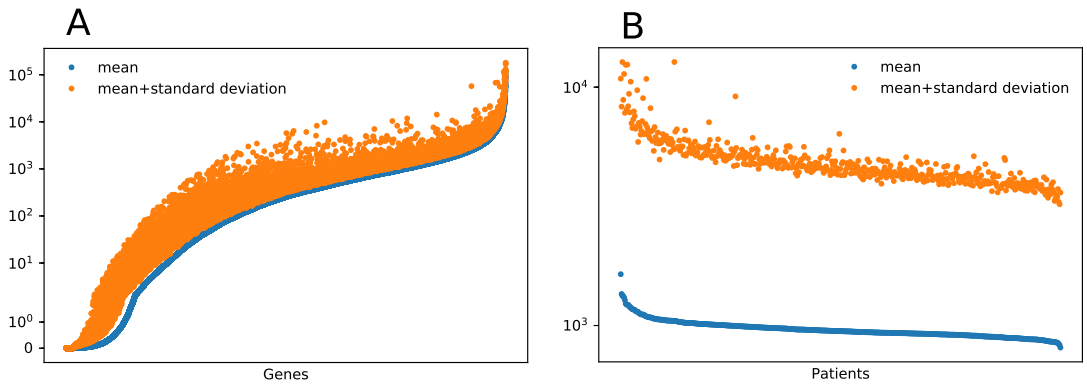


Figure 2.1. Sub-figure A shows the average expression levels of each gene across patients by blue dots and the corresponding average plus standard deviation by orange dots. Sub-figure B shows the average of gene expression profile of a primary tumor by blue dots, and the corresponding average plus standard deviation by orange dots..

2.1.1.2 Dimensionality Reduction.

Another challenging step in analyzing gene expression data sets is dimensionality reduction mainly because the number of genes (i.e. features) is much higher than the number of samples. The situation is worse for human data sets compared to animals or plants because of the lack of number of samples and existence of large number of genes. There are dimensionaltiy reduction methods that transform high-dimensional data into a significant representation of reduced dimensionality. Data with reduced dimensionality has the minimum number of parameters to observe properties of the data [15]. We apply two most common dimensionality reduction methods of Variance Threshold and Principal Component Analysis (PCA).

2.1.1.3 Variance Threshold.

Variance Threshold (Low Variance Filter) is a basic approach to select important features of a data set. First of all, normalization should be applied to make variance

values independent from the features domain range. Then, variance of each feature is calculated, and features with a variance value below a given threshold are removed from the data [16].

Here, we used variance threshold method to get top variant genes in the RNA-seq data. After normalization, we calculated the standard deviation of each gene across samples, plotted the sorted variance values and chose 0.04659080245779143 as threshold value Figure2.2. We found that the top 10 most variant genes across patients are B2M, CD74, EEF1A1L14, GAPDH, GPX3, IGFBP3, RGS5, SPARC, TGFB1, VIM. Among these genes, RGS5 looks very interesting (see the results).

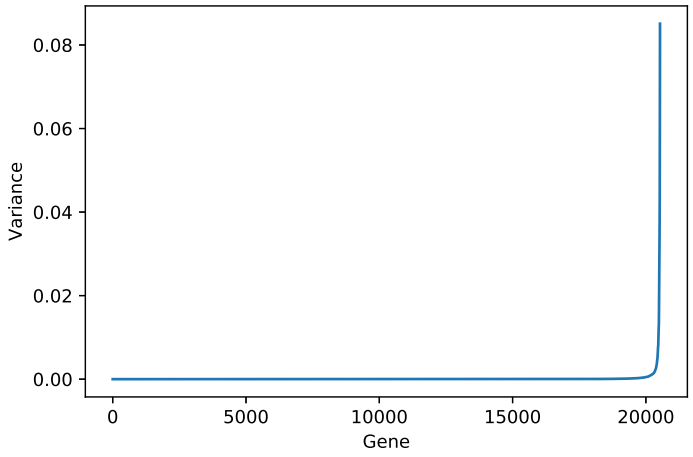


Figure 2.2. Sorted Variance of Genes.

2.1.1.4 Principal Component Analysis (PCA).

PCA is an orthogonal linear transformation that maps the data to a new subspace with low-dimensional representation of the data while retaining most of the variation of the data. In PCA, after standardize a d-dimensional data, eigenvectors and eigenvalues are obtained from the covariance or correlation matrix of the data,

or Singular Value Decomposition (SVD) is performed. Result of those three implementation of PCA are indeed the same but many of PCA implementations use SVD because of computational efficiency [17].

Eigenvectors (principal components) are chosen from covariance or correlation matrix, and eigenvectors with the lowest eigenvalues are dropped (choosing the top k eigenvectors where k is the number of dimensions of the new feature subspace, $k < d$), because the lowest eigenvalues do not give meaningful information about the data. In other words, principal components are the linear combinations of the original features. They indicate the directions of the new subspace, and the eigenvalues determine their magnitude. The first component in Figure 2.4 can be expressed in terms of the two genes, SPARC and EEF1A1L14 as $PC1 = -0.50SPARC - 0.86EEF1A1L14$.

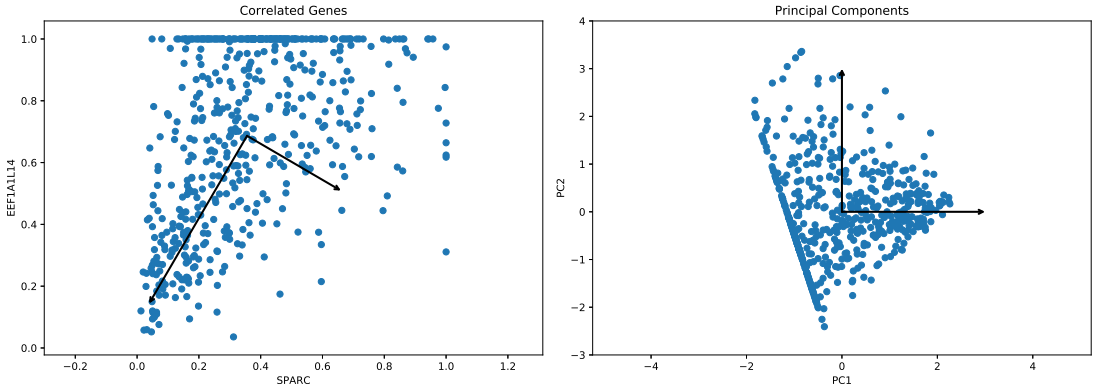


Figure 2.3. Two principal component for two dimensional gene expression profile.

Explained variance is an effective measure to decide the number of principal components, and it shows how much of the data can be explained by each of the principal components [18]. In the gene expression data set of RCC tumors, approximately 100 principal components contain almost all of the information so that 100 principal components can be used instead of 20,531 genes (Figure 2.4).

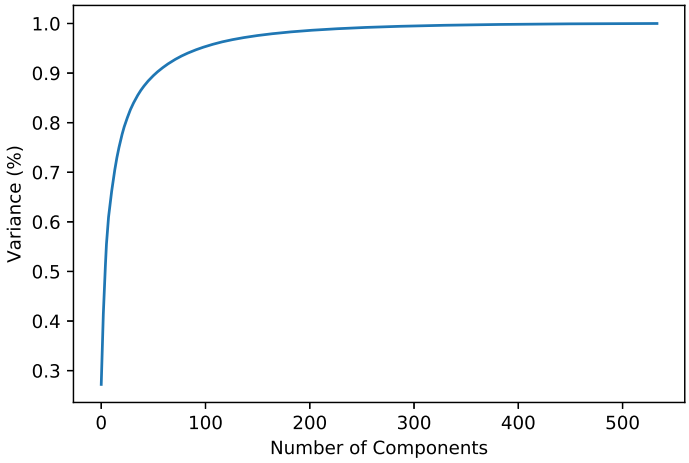


Figure 2.4. Cumulative Summation of the Explained Variance.

2.1.1.5 Comparison of PCA and Variance Threshold Method.

We separately applied PCA and Variance Threshold Method on RNA-Seq data of RCC tumors. When we compared PC1 and top ten gene expression values, we found that two genes from the top ten most variant genes are significantly correlated with PC1, those genes are ‘EEF1A1L14‘ and ‘SPARC‘ with Pearson correlation coefficient of 0.73 and 0.7, respectively. Moreover, RGS5 is also slightly positively correlated with PC1 with the correlation coefficient of 0.61.

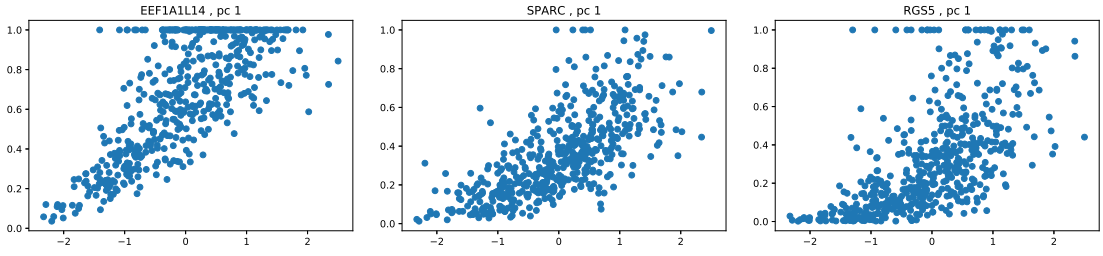


Figure 2.5. Comparison of PCA and Variance Threshold Methods.

2.1.2 Statistical Method.

We used several different graphical techniques to visualize quantitative data in our study. We used scatter plots to show the relationship between two continuous features (see Figure 2.5 as an example). In the first chapter, box-plots are mainly used to show relationship between categorical features and RGS5 gene expression level, where box plot is generated by calculating five relevant values: minimum, maximum, median, first quantile and third quantile of the data [19]. Furthermore, we used cluster heat map to simultaneously show row and column hierarchical cluster structure in data [20] (see Figure2.13).

For statistical analyses, we employed Mann-Whitney-Wilcoxon (MWW), also known as Mann-Whitney U-test. MWW is a non-parametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample be less than or greater than a randomly selected value from a second sample [21]. We chose MWW test since it works with non-normal distribution and ordered data as we have. For similar reason, we did not use t-test, because we have different number of patients for each subgroups and t-test works well for normally distributed data. We used star symbols to show the statistical significance between two categories in the box-plots, where ns: $0.05 < p \leq 1$, *: $0.01 < p \leq 0.05$, **: $0.001 < p \leq 0.01$, ***: $0.0001 < p \leq 0.001$, ****:$p \leq 0.0001$.

2.2 Results

2.2.1 RGS5 gene.

We found that RGS5 gene expression might be prognostic and might have a significant role in initiation and progression of RCC. The regulator of G-protein signaling (RGS) family regulates cellular signaling events downstream of G-protein

coupled receptors (GPCRs), which have been found to be associated with the initiation and progression of multiple cancers [22]. A high expression level of RGS5, a member of RGS family, has been detected in various human tissues such as heart, skeletal muscle, bladder, uterus, alimentary tract, and human umbilical vein endothelial cells (HUVECs) [23, 24, 25]. RGS5 is also involved in ion transport mechanisms in the kidney and has been associated with the regulation of blood pressure or hypertension [26].

In a study using Reverse Trascription-Polymerase Chain Reaction (RT-PCR), which is a strongly sensitive method for the detection and amplification of mRNA, a very weak or undetectable expression level of RGS5 has been observed in normal kidneys, while a high level of RGS5 expression has been detected in all RCCs [27]. Importantly, the same study indicates that the tumour endothelial cells are the main location of RGS5 in RCC [27]. Moreover, a study on DNA microarray data of 27 RCCs using Multivariate Cox analysis suggested RGS5, vascular cell adhesion molecule 1 (VCAM1), and endothelin receptor type B (EDNRB) as predictors of survival months [28].

On the other hand, a strong association has been observed between high RGS5 expression level and better survival in 51 non-small cell lung cancer (NSCLC) patients [29]. They detected a high RGS5 expression level in 47% of NSCLC patients and a link between low expression level of RGS5 and cancer vasculature invasion and lymph node metastasis [29]. In another study of 127 human paraffin-embedded epithelia ovarian cancer (EOC) tissue samples, RGS5 gene expression in EOC tissues was higher compared with normal ovaries, and EOC patients with high RGS5 expression had a better survival with progression-free [30]. In contrast to these two study, it has been found that over-expression of RGS5 gene in the human lung cancer cells is

associated with reductions in the survival rates and increases in the cytotoxic result of radiation [31].

2.2.2 Expression level of RGS5 significantly decreases when the grade of tumor increases.

We found that the grade of RCC tumors is a decreasing function of the normalized value of RGS5 (Figure 2.6). The grades of tumors, which is determined by the appearance of the cancer cells under the microscope, provide us with some insight about how the cancer might behave. Lower grades represent the cancer cells that look more like normal cells; G1 and G2 tumors tend to grow slowly and spread less. Higher grades represent the cancer cells that look more different from the normal cells. Cancer cells in G3 and G4 tend to grow quickly and spread fast [11]. We found that there is a significant reduction in the RGS5 expression from the G2 to the G4 grade. Note, we cannot make a confirm conclusion for grade one tumors, because there are not many G1 patients.

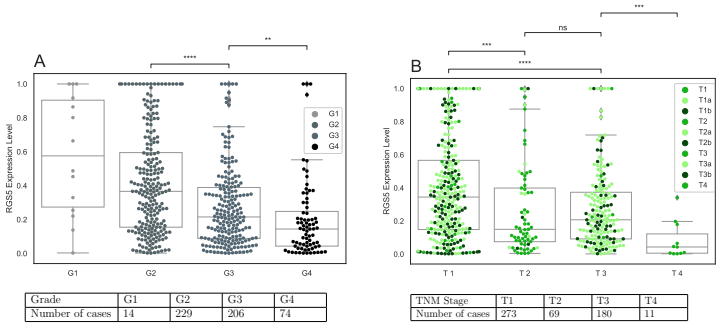


Figure 2.6. Expression level of RGS5 as a function of grade and TNM staging of RCC. The sub-figure A shows that the expression level of RGS5 in the patients' primary tumor decreases when the grade of tumor increases. The sub-figure B represents the expression level of RGS5 in primary tumors categorized based on TNM stage of RCC. For some patients with the stage Ti (green), $i = 1, 2, 3, 4$ cancer, we know if the patient had the stage Tia (light green) or Tib (dark green) cancer, for more information see Table 1.1. Tables in this figure indicate the number of patients in each category.

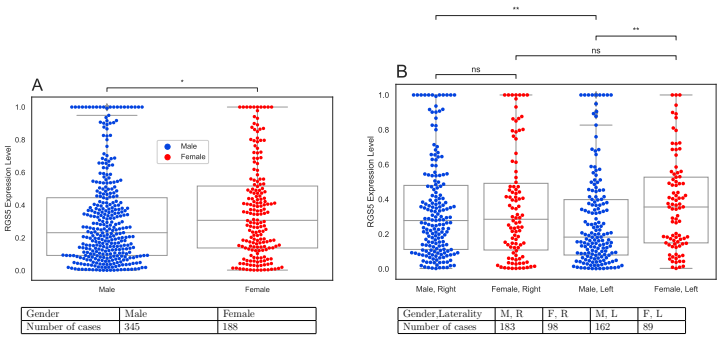
2.2.3 Stage1 tumors have the highest expression level of RGS5 compared to the other stages.

In Figure 2.6 , we grouped patients based on their tumor's stage using the most common staging system, TNM, for kidney cancer (Table 1.1) [12]. Since there are not many patients with stage 2 and 4 cancer, one may ignore the T2-T3 and T3-T4 differences and conclude a significant decrease in the RGS5 expression level from T1 to T3 (Figure 2.6).

2.2.4 Tumors originated in the left kidney have a significantly higher RGS5 expression in females than male patients.

We observed a significant high RGS5 expression in female patients versus male patients. However, further investigation revealed no differences in the expression level of RGS5 in males and females for tumors that originated in the right kidney. Inevitably, there is a significant difference in RGS5 expression between female and

male patients with tumors in the left kidney. Importantly, we found that male patients with a primary tumor in the right kidney have a higher level of the RGS5 expression compared to the left kidney. However, there is no significant difference between RGS5 expression of the female patients with a primary tumor in the left kidney compared to the right kidney (Figure 2.7).



Gender	Male	Female
Number of cases	345	188

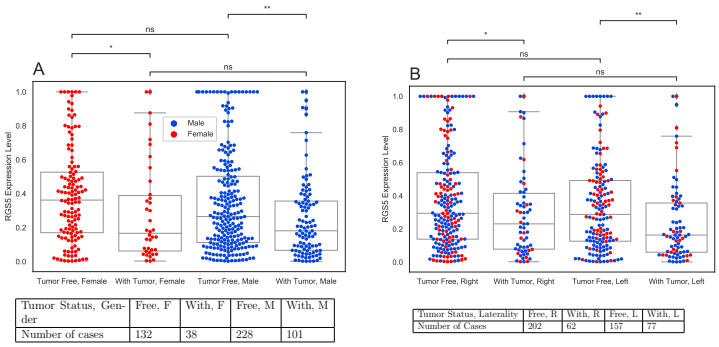
Gender,Laterality	M, R	F, R	M, L	F, L
Number of cases	183	98	162	89

Figure 2.7. Expression level of RGS5 as a function of gender and the location of the primary tumor. Sub-figure A shows that female patients have higher expression level of RGS5 than male patients. Grouping patients based on the location of their primary tumors (sub-figure B) shows that the difference between the expression level of RGS5 in female and male patients is due to the fact that male patients with primary tumors in the left kidney have lower expression level of RGS5 than other patients. Tables in this figure indicate the number of patients in each category.

2.2.5 Tumor free patients have a significantly higher level of RGS5 expression versus patients with tumors at the last time of follow up.

We further investigated the relationship between RGS5 expression in primary tumors and the patients' tumor status, which provide us with the state or condition of individuals' neoplasm at the last time of follow up. If patients did not have any tumor at the last time of follow up, they were categorized as "tumor free", and if they had any tumor then they were marked as "with tumor". We observed a higher level of RGS5 expression in tumor free patients compared to patients with tumor. When we divided kidney cancer patients into sub-groups based on their gender, location of

the primary tumor, and tumor status, the gene expression level reacts similarly to general results; aggressive tumors have a significantly lower expression level of RGS5 (Figure 2.8).



Tumor Status, Gender	Free, F	With, F	Free, M	With, M
Number of cases	132	38	228	101

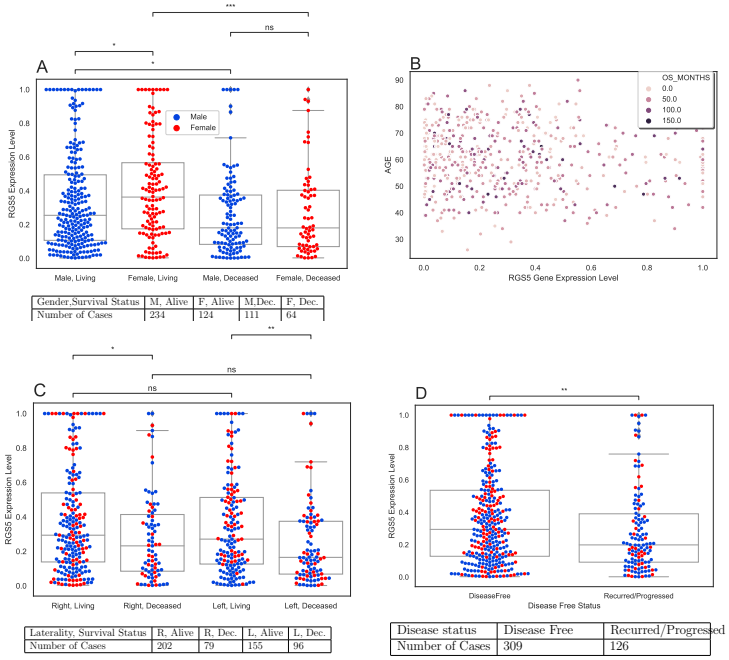
Tumor Status, Laterality	Free, R	With, R	Free, L	With, L
Number of Cases	202	62	157	77

Figure 2.8. Expression level of RGS5 gene in primary tumors as a function of tumor status at the time of last follow up. This figure indicates that the primary tumors of RCC patients with no tumors at the time of last follow up had a higher RGS5 expression level than patients with tumors at the time of last follow up regardless of patients' gender (sub-figure A) and the location of primary tumors (sub-figure B). Tables indicate the number of patients in each category.

2.2.6 RGS5 expression in primary tumors of living patients is higher than deceased patients.

When we considered overall patients' survival status, we see from Figure 2.9 that RGS5 expression level in primary tumors of patients who are live at the last time of follow up is higher than deceased patients. Also, when we divided the patients into subcategories based on gender, results are not affected; so living patients, regardless of their gender and even white blood cell levels, have a higher level of RGS5 expression compared to deceased patients. In addition, when we compared only living patients by their gender, we found that female living patients have a higher level of RGS5 expression compared to male living patients. Furthermore, the laterality does not affect the RGS5 expression level as much as the survival status (Figure 2.9). Moreover,

we see that there is a similar tendency in the RGS5 expression with disease-free status and overall survival status, as expected. However, there is no correlation between age at diagnosis and RGS5 expression level.



Gender,Survival Status	M, Alive	F, Alive	M,Dec.	F, Dec.
Number of Cases	234	124	111	64

Laterality, Survival Status	R, Alive	R, Dec.	L, Alive	L, Dec.
Number of Cases	202	79	155	96

Disease status	Disease Free	Recurred/Progressed
Number of Cases	309	126

Figure 2.9. Expression level of RGS5 in primary tumors as a function of survival status of patients at the time of last follow up. The left subplots (A and C) show that primary tumors of alive patients have a higher expression level of RGS5 compared to the deceased patients regardless of gender (sub-figure A) and the location of primary tumor (sub-figure C). Sub-figure B indicates that there is no linear correlation between age at diagnosis and RGS5 expression. Sub-figure D shows that primary tumors of patients, who were diseased free at the last time of follow up, have a higher expression level of RGS5 compared to the patients with reoccurred cancer. Tables indicate the number of patients in each category. In all sub-figures, female patients are represented by red color and males by blue.

2.2.7 Patients with higher level of RGS5 have higher survival months.

We observed no correlation between the level of RGS5 and overall survival months and also age at diagnosis (Figure 2.10). However, when we divided RGS5 expression level into two categories of low (patients with normalized RGS5$\leq$ 0.1) and

high (patients with normalized RGS5 > 0.1), we observed that the patients with a high RGS5 expression have better survival compared the patients with low RGS5 expression. This result aligns with our other observations regarding the grade of tumors (Figure 2.6), tumor status (Figure 2.8), and survival status (Figure 2.9), as we expected.

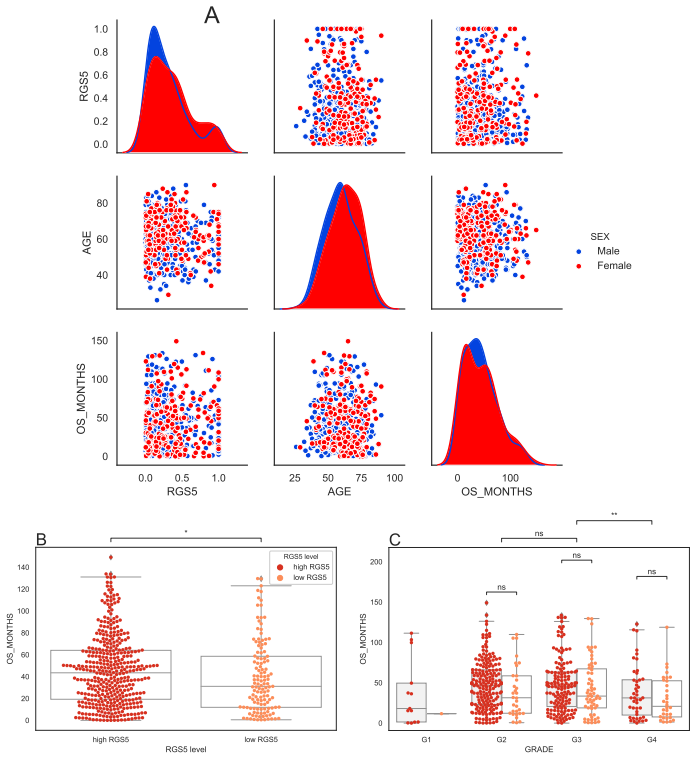


Figure 2.10. Overall survival months as a function of RGS5 expression level. Sub-figure A represents the values and distributions of RGS5 expression level, age at diagnosis, and survivals months for female and male patients. Sub-figures B and C show the difference between survival months of patients with low (normalized RGS5 expression value ≤ 0.1) and high (> 0.1) RGS5 expression levels. In sub-figure C, patients are grouped based on the grade of their tumors.

2.2.8 Expression level of RGS5 in patients with low serum calcium levels is significantly higher than those with normal calcium levels.

It has been observed that kidney disorders are highly correlated with low calcium levels in the blood (hypocalcemia) [32]. We found that patients with low calcium levels have a higher RGS5 expression level compared to patients with normal calcium levels. In addition, considering laterality, we observed that primary tumors in both right and left kidneys follow the same trend, although the difference is not significant in the left kidney. Similarly, when we grouped patients based on their gender, RGS5 expression of male patients who have low calcium levels is significantly higher than those with normal calcium levels, but it is not significant for female patients. Lastly, when we divided patients into subcategories based on overall survival status, we still observed a higher level of RGS5 in living patients compared to the diseased ones regardless of their calcium level. Importantly, deceased patients with normal calcium levels have the least level of RGS5 expression. Note, the percentage of living patients with low serum calcium levels ($138/(138+66) \approx 0.68$) is higher than those with normal levels ($89/(89+61) \approx 0.59$).

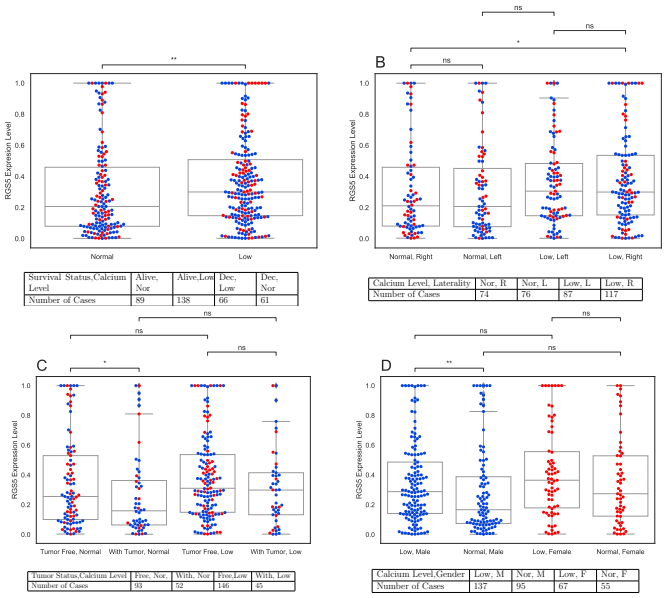
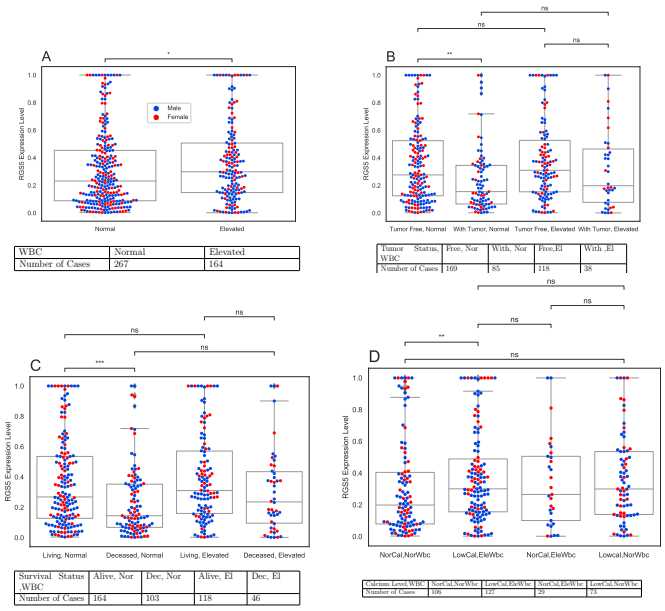


Figure 2.11. Expression level of RGS5 gene as a function of serum calcium level and living status (sub-figure A), laterality (sub-figure B), tumor status (sub-figure C), and gender (sub-figure D). This figure shows that primary tumor of patients with low calcium level have a higher expression level of RGS5 compared to the patients with normal calcium level. Tables indicate the number of patients in each category. In all sub-figures, female patients are represented by red and male patients by blue.

2.2.9 Patients with elevated white blood cells (WBC) have a high level of RGS5 expression.

Elevated white blood cells (WBC) is known as a common predictor of chronic kidney disease [33]. We observed that RGS5 expression in patients who have elevated WBC was greater than those in patients who have normal WBC. In addition, when we considered the tumor status of patients, again tumor-free patients have a higher RGS5 expression level than with tumor patients regardless of their WBC counts. Furthermore, the percentage of living patients with elevated WBC $(118/(118+46) \approx 0.72)$ is higher than those with normal levels $(164/(164+103) \approx 0.61)$. Lastly, when we divided the patients into sub-categories based on serum calcium level and WBC

level, we observed that calcium level is more important than WBC for predicting the expression level of RGS5.



WBC	Normal	Elevated
Number of Cases	267	164

Tumor Status, WBC	Free, Nor	With, Nor	Free,El	With ,El
Number of Cases	169	85	118	38

Survival Status, WBC	Alive, Nor	Dec, Nor	Alive, El	Dec, El
Number of Cases	164	103	118	46

Calcium Level,WBC	NorCal,NorWbc	LowCal,EleWbc	NorCal,EleWbc	LowCal,NorWbc
Number of Cases	106	127	59	73

Figure 2.12. Expression level of the RGS5 gene as a function of WBC (sub-figure A), WBC and tumor status (sub-figure B), WBC and living status (sub-figure C), and WBC and calcium level (sub-figure D). Tables indicate the number of patients in each category. In all sub-figures, female patients are represented by red color and male patients by blue.

2.2.10 Expression level of RGS5 might be prognostic.

To sum up, Figure 2.13 shows a hierarchically-clustered heat-map of the above-mentioned clinical and demographic features with the correlation metric. For the heat-map, we convert the categorical features to numerical values, and scale all them between 0 and 1. Stage and grade of tumor are valued from 1 to 4. Gender: Female: 0, Male: 1. Tumor Status: Tumor Free: 0, With Tumor:1. Laterality: Left: 0, Right: 1. Serum Calcium Level: Low: 0, Normal: 1. Disease Status: DiseaseFree: 0, Reoccurred: 1. Os Status: Living: 1, Deceased: 0. WBC: Normal: 0, Elevated:

1. We then use correlation metric to cluster. As a result, RGS5, survival months, survival status ended up in the same cluster, and then they grouped with WBC level and laterality. As one could expect grade and stage of the tumor are clustered with the tumor status and the disease status, and interestingly they all clustered with the serum calcium level.

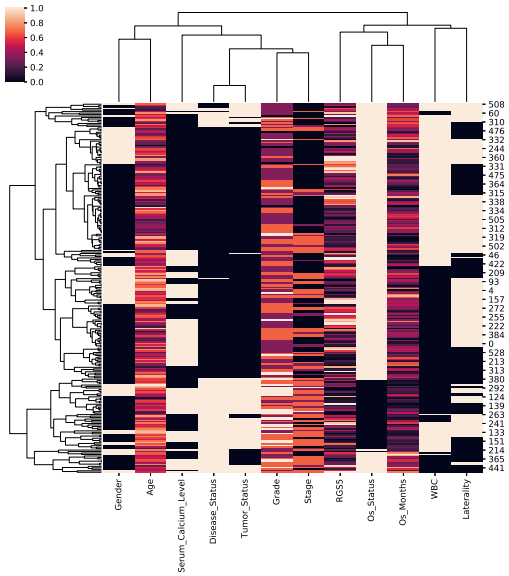


Figure 2.13. Hierarchically-clustered heat-map of features that have been investigated in this chapter and RGS5 gene.

2.3 Discussion

Recent advances in biotechnology have led to the development of new immunotherapeutic methods for treating various cancer types. One of the crucial requirements of these methods is explaining structure of tumor specific antigens (TSA) or tumor associated antigenes (TAA) that help to fight with malignant cells. Krüger et al. identified RGS5 as one of the RCC-associated antigens [34]. Boss et al. also found a significant up-regulation of RGS5 in a broad variety of malignant

cells, and by identifying two HLA-A2- and HLA-A3-binding peptides derived from the RGS5 protein, they suggested RGS5 peptides as candidates for designing cancer vaccines to target malignant cells and tumor vessels [35].

RGS family are known as signal transaction molecules that are associated with the arrangement of heterotrimetric G proteins by acting as GTPase activators. Moreover, RGS5 is a hypoxia-inducible factor-1 dependent involved in the induction of endothelial apoptosis [36]. Furthermore, it has been shown that the variants in RGS5, ATP1B1 and SELE genes were accounted for 2-5 mm Hg differences in mean systolic blood pressure levels, and the cumulative effect reaches 8-10 mm Hg [37]. Hence, RGS5 is an important contributor to high blood pressure [38], and hypertension is one of the most common risk factors of renal disease [39]. There is also an evidence that RGS5 can act as a physiological regulator of calcium sensing in the parathyroid gland as a inhibitor of calcium-sensing receptor (CaSR) signaling [40], and a study of patients with the stage 3-4 chronic kidney disease (CKD) determined that the lower serum calcium is associated with the higher risk of renal replacement therapy and rapid renal function progression [41]. Furthermore, it has been observed a higher incidence rate of early mortality in CKD patients with low blood calcium levels [42]. Here, we noticed a significantly higher expression level of RGS5 in primary tumors of patients with low serum calcium levels than those with normal calcium levels.

Additional to the various biological roles of RGS5, there are some contradictory observations about the prognostic behavior of RGS5. For instance, an association between high expression levels of RGS5 and poor outcomes have been observed in a study of human lung cancer [31], while studies on non-small cell lung cancer [29] and epithelia ovarian cancer [30] reported a link between a high expression levels of RGS5 and better survival months. Here, by analyzing the RNA seq data of primary tumors of 533 patients with RCC, we independently observed a strong association

between high expression levels of RGS5 and better outcomes. We found that the RGS5 expression level significantly decreases when the grade of tumor increases. Moreover, when we considered TNM staging system which gives us information about the size and location of the tumor (Table 1.1), we observed a significant decrease in the RGS5 expression level when the stage increases from T1 to T3.

Several studies show that men with RCC have higher stage and grade and ultimately worse overall survival than women with RCC [3, 43]. We also observed that the survived female patients have a significantly higher expression level of RGS5 compared to other categories (Figure 2.9). Moreover, some studies reported better clinical results such as low grade tumors and better survival in patients with right-sided RCC compared to the left-sided RCC [44, 45], but our result shows no significant difference between the expression level of RGS5 in female versus male patients with primary tumors located in the right kidney. However, among patients with primary tumors in the left kidney, female patients have a significantly higher RGS5 expression level (Figure 2.8).

There is a correlation between elevated WBC and kidney function deterioration [46], and a study of 362 advanced renal cell cancer patients showed that elevated WBC is associated with poor survival [47]. In our study, when we examined the relationship between RGS5 expression and WBC, we found that patients with elevated WBC have significantly higher RGS5 expression compared to those with normal WBC. Furthermore, we observed that the percentage of survived patients are higher in elevated WBC category than normal WBC group.

Although this study emphasizes that RGS5 plays a significant role in RCC, the different biological roles of RGS5 [48] and contradictory observations cause uncertainty about how RGS5 expression affects the outcome of treatments and how it could be used to develop better treatments. Therefore, future investigation is necessary to

explore the role of RGS5 in initiation, progression, and suppression of RCC to arrive at effective treatments.

CHAPTER 3

TUMOR DECONVOLUTION

The immune system consists of a complex network of cells, tissues, organs, and substances that they produce which helps the body fight infections and other diseases. Although the immune system is against cancer growth, cancer cells can block the immune system in several ways such as being invisible or having proteins that turn off immune cells. Activating or suppressing the immune system for the diseases, so called 'immunotherapy', is one of the popular type of cancer treatments. There are several type of immunotherapies such as immune checkpoint inhibitors and T-cell transfer therapy.

Several immunotherapeutic approaches have been recently used for treating patients with RCC [49, 50], which is considered a morphologically and genetically immunogenic tumor [2]. However, many patients do not respond to these treatments and develop adaptive or intrinsic resistance. We can increase the response rate to these treatments by identifying the characteristics of patients who will benefit from each of these therapies.

Several studies show that cancer cells and tumor-infiltrating immune cells (TIICs), which have important roles in both regulation of cancer progression and promotion of tumor development [51, 52], play an important role in the determination of malignant tumor types [53, 54]. Tumor-infiltrating lymphocytes (TILs), which include T-cells and B cells, are an important category of TICCs. CD4+ helper T-cells and cytotoxic CD8+ T-cells play a significant role in preventing tumor by targeting antigenic tumor cells [55], and CD8+ T-cells are linked with better clinical outcomes

and reaction to immunotherapy in many cancers [56, 57]. Furthermore, it has been recently observed that tumor associated B cells, which have significant roles in the immune system by producing antibodies and presenting antigens, could be predictors of survival and response to immune checkpoint blockade therapy [58]. Importantly, controversial roles of B cells have been observed in the tumor micro-environment such as causing tumor growth [59] and enhancing tumoricidal T-cell responses [60]. Additionally, a relationship between TIICs gene signatures and lower survival rates has been observed in RCC patients, and tumor-associated macrophages (TAM) and 22 T cell phenotypes are found to be correlated with clinical outcomes [61, 62]. These observations emphasize on importance of analyzing the cellular heterogeneity of tumors, including immune cells variations, to understand variations in immune responses, identify target tumors for each specific treatment, and design new effective cancer treatments [63].

3.1 Deconvolution Methods

There are some experimental approaches such as single cell analysis tools, including immunohistochemistry and flow cytometry to observe tumor immune infiltrates, however these methods are expensive and time consuming, and they are limited to analyzing a few immune cell types simultaneously [64]. For this reason, several computational methods have been recently developed to provide us with much less expensive and fast alternative ways to estimate the relative amount of each cell type from gene expression profiles of bulk tumors. We used most popular deconvolution methods and compared their results. The methods we studied are DeconRNASeq, CIBERSORT, ssGSEA, singscore and CIBERSORTx. Details of the methods are given in the subsections.

3.1.1 Deconvolution of mRNA-Seq (DeconRNASeq).

Deconvolution of mRNA-Seq (DECONRNASeq) is a method to estimate proportion of cell types from mRNA-seq data that is based on the following linear model:

$$y = WX, \tag{3.1}$$

where y is gene expression data of a sample (mixture data), w is the percentage of each cell in sample y, and X is the single cell gene expression of cells in w (signature matrix).

Assume y_{jk} is the expression level of gene j in a sample k, and the x_{ij} is the expression level of gene j in cell i, and w_{ki} is estimated proportion of cell i in the sample k, we can re-write the equation (3.1) in the following form:

$$y_{jk} = \sum_i w_{ki} x_{ij} \tag{3.2}$$

For each sample k, if the number of genes is less than the number of cell types, then the linear system (3.1) is undetermined. We need to have more genes than cell types to solve the system.

$$\begin{aligned}
y_{1k} &= w_{k1}x_{11} + w_{k2}x_{21} + \cdots + w_{kn}x_{n1} \\
y_{2k} &= w_{k1}x_{12} + w_{k2}x_{22} + \cdots + w_{kn}x_{n2} \\
&\vdots \\
y_{jk} &= w_{k1}x_{12} + w_{k2}x_{22} + \cdots + w_{kn}x_{nj}
\end{aligned}$$

Adding physical constraints, which are $w_{ki} \geq 0$ and $\sum_i w_{ki} = 1$, the above system of linear equation is over determined. In 2013, Gong et al introduced DECON-

RNASeq method [65] that use quadratic programming to solve the problem as linear least square sense and find optimal w_{ki} that minimizes the residuals for $WX - y = 0$:

$$\min_{A}(||WX - y||^2), s.t. \quad \begin{cases} \sum_{i=1}^{N} w_{ki} = 1, \quad w_{ki} \geq 0, \\ y_{jk} = \sum_{i=1}^{N} w_{ki} x_{ij}. \end{cases} \tag{3.3}$$

3.1.2 CIBERSORT Method.

Although DECONRNASeq works well for samples with well defined composition (e.g. blood), it performs weakly on samples with unknown content and noises (e.g. solid tumors) and cannot differentiate closely related cell types effectively [66]. In 2015, Newman et al. presented Cell-type Identification By Estimating Relative Subesets Of RNA Transcripts (CIBERSORT) that utilize a machine learning technique which is Nu-Support Vector Regression (ν-SVR) to estimate the cell frequency. CIBERSORT formulate problem(3.3) as an optimization problem such that:

$$\text{Linearfunction} \quad f(x, w) =< w, x > +b = \sum_{j=1}^{M} w_j x_j + b \quad \text{that}$$

$$\text{minimize}: \quad \frac{1}{2}||w||^2 + C(\nu\varepsilon + \frac{1}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^*))$$

$$\text{subject} \quad \text{to}: \quad \begin{cases} y_i - f(x_i) \leq \varepsilon + \xi_i \\ f(x_i) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

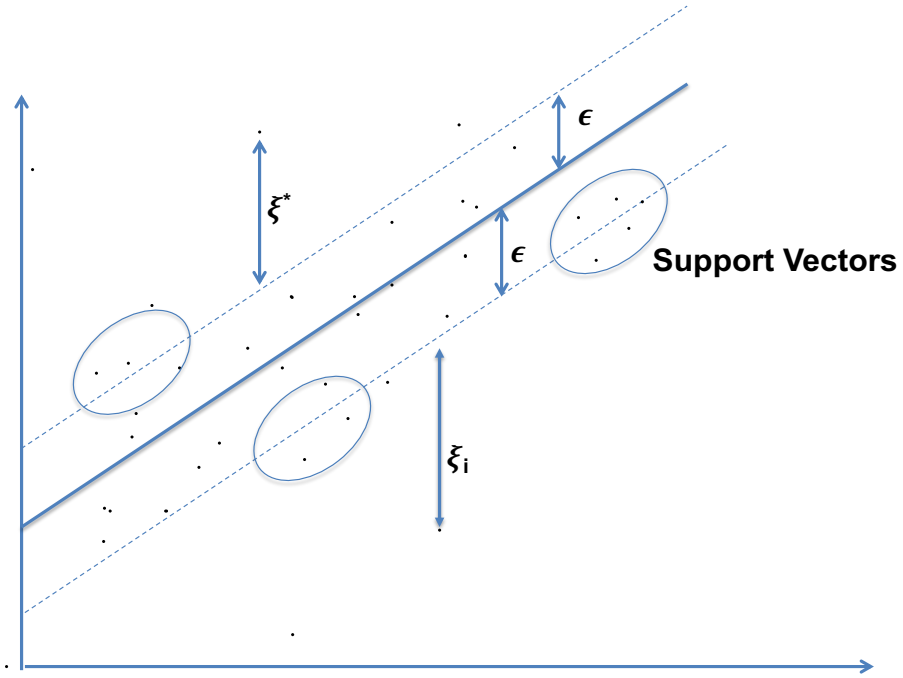


Figure 3.1. One dimensional linear SVR.

Matrix X in the equation (3.1) is determined by a hyperplane capturing the data points inside an ε-tube that is determined by support vectors (genes in the signature matrix). SVR penalizes the data points outside the ε-tube, and a small value is used for ν that determines the lower bound of support vectors and the upper bound of training errors (Figure 3.1). Regression coefficients of ν-SVR method gives the values (w_{ki}), however, the proportion of the cells require values that are non-negative and sum to one. Therefore, negative coefficients are set to 0, and they normalize the remaining coefficients so that their sum is 1 to estimate the percentage of each cell in the samples [66].

3.1.3 ssGSEA Method.

DECONRNASeq and CIBERSORT methods assume that the mixture data y consists of cells in the signature matrix W. While DeconRNA and CIBERSORT methods require a reference expression vectors for each cell types, ssGSEA is a rank-based method that estimate a score for a gene signature of interest relative to all other genes of samples [67]. In a study with multiple cancer types in 2013, immune and stromal infiltration levels of samples have been analyzed by implementation of single sample Gene Set Enrichment Analysis (ssGSEA) [68].

ssGSEA requires mixture data and cell signatures that each one contains a list of gene names that are highly expressed in interested cell type and provides us levels of cell populations rather than actual fraction of the cells. Algorithm of ssGSEA starts with ordering mixture data by absolute gene expression from highest to lowest and replaces gene expression values in mixture data by their ranks $L = \{r_1, r_2, ..., r_N\}$. Then, enrichment score is calculated for a given signature G with size N_G and single sample S, of the data of N genes and an enrichment score ES(G,S) is obtained by an integration of the difference between a weighted Empirical Cumulative Distribution Functions (ECDF) of the genes in the signature P_G^{ω} and the remaining genes P_{NG} [67]:

$$ES(G,S) = \sum_{i}^{N} \Big[P_G^{\omega}(G,S,i) - P_{NG}(G,S,i) \Big] \quad \text{where}$$

$$P_G^{\omega}(G,S,i) = \sum_{r_j \in G, j \leq i} \frac{|r_j|^{\alpha}}{\sum_{r_j \in G} |r_j|^{\alpha}} \quad \text{and} \quad P_{NG}(G,S,i) = \sum_{r_j \notin G, j \leq i} \frac{1}{N - N_G}.$$

3.1.4 SingScore Method.

The scores that are obtained by ssGSEA method do not exactly represent 'single-sample' and they are affected from differences between overall sample composition [69]. To avoid these weakness, a single-sample gene-set scoring method (SingScore)

has been developed by Foroutan et.all in 2018 that use a rank-based metric to evaluate the relative enrichment of a gene set within a sample gene expression, and normalized relative to the maximum and minimum theoretical scores for a gene set of a given size [70].

SingScore [70] produces a score for each cell type that represents a normalized mean percentile rank. The inputs of SingScore are the mixture data and bidirectional cell signatures including both Up-Regulated and Down-Regulated: highly expressed genes for the up set and lowly expressed genes for the down set. In the algorithm of SingScore, mixture data is ordered by gene expression levels from the highest to the lowest and top half of it is used as regulated genes of samples and bottom half of it is used as as down-regulated genes of samples. Then, these genes are ranked in ascending order for the up-set and in descending for the down-set. By using up-regulated gene set G_{up} and down regulated gene set G_{down} for a given cell type, the score (S) and normalized score ($\bar{S}$) are defined as:

$$S_{dir,i} = \left(\frac{\sum_{g \in G_{dir}} R^{g}_{dir,i}}{N_{dir,i}} \right), \quad \text{where}$$

dir: is the gene set direction (up or down regulated);

$S_{dir,i}$: is the score for sample i against the directed gene set;

$R^{g}_{dir,i}$: is the rank of gene g in the directed gene set;

$N_{dir,i}$ is the number of genes in G_{dir} that are observed within the data;

By using these, $\bar{S}$:

$$\bar{S}_{dir,i} = \frac{S_{dir,i} - S_{min,i}}{S_{max,i} - S_{min,i}} \quad \text{where}$$

$$S_{min,i} = \frac{N_{dir,i} + 1}{2} \quad \text{and} \quad S_{max,i} = \frac{2N_{total,i} - N_{dir,i} + 1}{2}$$

$\bar{S}_{dir,i}$: is the normalized score for sample i against genes in the signature;
$S_{min,i}$ and $S_{max,i}$: are the theoretical minimum and maximum mean ranks;
$N_{total,i}$: is the total number of genes in sample i

Calculated score for sample i is $S = \bar{S}_{up,i} + \bar{S}_{down,i}$

3.1.5 CIBERSORTx Method.

Newman et. all [71] have recently improved their method CIBERSORT by adding selection of batch correction modes so that the cross-platform variation between signature matrix and mixture data can be eliminated. They offer two mode: 1-) B-mode, which they first obtain deconvolution of cell fraction ($\hat{W}$) from the main method and they obtain reconstituted mixture data ($\hat{y}$) from $\hat{y} = \hat{W}^T X$. Then, they use ComBat, which is a batch correction method, to eliminate technical variation between $\hat{y}$ and y to build adjusted mixture data y_{adj}. Finally, they use y_{adj} in the main algorithm to estimate cell fraction W. 2-) S-mode, which signature matrix is directly adjusted $\hat{X}$ rather than mixture matrix using non-negative least square method with y_{adj} and $\hat{W}$. Thus $\hat{X}$ and original y are used to predict cell fraction of W [71].

3.1.6 TumorDecon Software.

TumorDecon Software is presented by a team of ShahriyariLab that includes four deconvolution methods of DeconRNASeq, CIBERSORT, ssGSEA, and Singscore and several signature matrices of various cell types, including LM22 which is generated by Newman et all [66]. It provides estimated cell proportion from chosen method and given gene expression profile of the tumors (mixture data). An example of mixture data and signature matrix is given in Figure 3.3.

A

Hugo_Symbol	TCGA-3L-AA1B-01	TCGA-4N-A93T-01	TCGA-4T-AA8H-01	TCGA-5M-AAT4-01	TCGA-5M-AAT5-01	TCGA-5M-AAT6-01	TCGA-5M-AATA-01	TCGA-5M-AATE-01
ABCB4	95.7447	7.2569	3.6556	6.8847	11.8707	13.7810	11.9284	13.0431
ABCB9	132.1710	127.8710	76.4760	174.2470	136.7060	118.5160	209.7770	96.7526
ACAP1	161.0250	73.0527	29.9762	19.3632	46.4936	273.6500	142.7440	15.7891
ACHE	80.2708	258.8290	184.9750	191.4800	118.2130	55.7800	133.5980	57.6644
ACP5	426.9830	72.5689	117.7120	167.3840	155.8030	1380.7200	910.5370	466.1200
ADAM28	235.9770	17.9003	15.3537	6.8847	16.8168	139.7780	44.9304	136.6100
ADAMDEC1	713.7330	149.0080	218.6070	561.5320	77.1596	449.5220	472.7630	407.0830
ADAMTS3	8.2205	2.4190	2.9245	2.5818	8.4084	46.5927	15.5070	2.0594
ADRB2	19.8259	32.4141	33.6319	33.1325	12.3653	84.6544	9.9404	3.4324
AIF1	191.4890	47.8955	40.9432	74.8709	66.2781	593.8930	181.7100	56.9779

B

Hugo_Symbol	B cells naive	B cells memory	Plasma cells	T cells CD8	T cells CD4 naive	T cells CD4 memory resting	T cells CD4 memory activated
ABCB4	555.713449	10.744235	7.225819	4.311280	4.605860	7.406442	8.043976
ABCB9	15.603544	22.094787	653.392328	24.223723	35.671507	30.048192	38.455423
ACAP1	215.305951	321.621021	38.616872	1055.613378	1790.097170	922.152747	340.883424
ACHE	15.117949	16.648847	22.123737	13.428288	27.187732	18.444927	13.441268
ACP5	605.897384	1935.201479	1120.104684	306.312519	744.656599	557.819820	248.546932
ADAM28	1943.742699	1148.120138	324.780800	22.689718	40.061713	21.233311	15.209620
ADAMDEC1	371.033593	318.478799	127.967448	44.616287	80.964500	47.433514	30.552976
ADAMTS3	146.195568	106.052311	74.339169	42.390416	79.241795	57.385906	53.413024
ADRB2	486.343816	510.081340	289.798450	899.648502	77.153955	595.210549	142.831911
AIF1	24.074298	20.321859	21.969573	742.815819	718.614787	35.172248	39.508760

Figure 3.2. The sub-figure A is an example of mixture data, where columns are patients id, rows are gene names. The sub-figure B is an example of signature matrix, where columns are cell types and rows are genes.

In this study, we used TumorDecon Software to determine which method has better accuracy among the deconvolution methods mentioned above, except CIBERSORTx. For the mixture and signature data, we used gene expression profiles of blood obtained from 12 healthy adults and LM22 signature matrix that are provided with CIBERSORTx software [71].

3.1.6.1 Visualization.

We added visualization function to TumorDecon software that provides 4 different plots: 1-) A bar chart plot that shows the cell frequencies of all samples, 2-) A box plot that shows cell frequencies in descending order. 3-) A hierarchical clustered heat map from cell frequencies in tumors. 4-) A pair plot that shows correlation between estimated cell frequencies in tumors.

Using visualization package, we compared results of deconvolution methods with ground truth cell proportion of the mixture data that is obtained by direct cytometry and fluorescence immunophenotyping [71]. We found that results obtained

by CIBERSORTx has the highest correlation with the ground truth data compared to the other deconvolution methods 3.3.

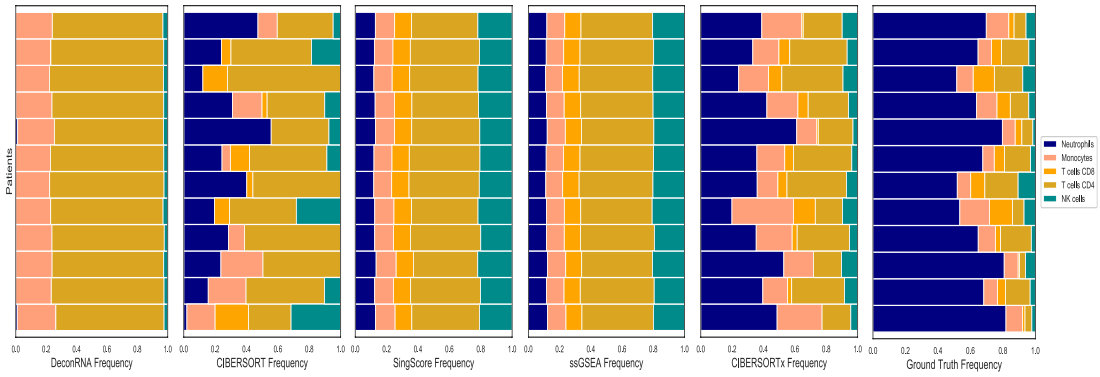


Figure 3.3. Fractions of each sample from each methods.

CHAPTER 4

IMMUNE CHARACTERISATION OF RENAL CELL CARCINOMA

In this chapter, we applied a powerful "digital mass cytometry" method called CIBERSORTx to estimate the fraction of each immune cell in RCC tumors to determine immune patterns of tumors (Figure 4.1) and investigate the association of these patterns with clinical features.

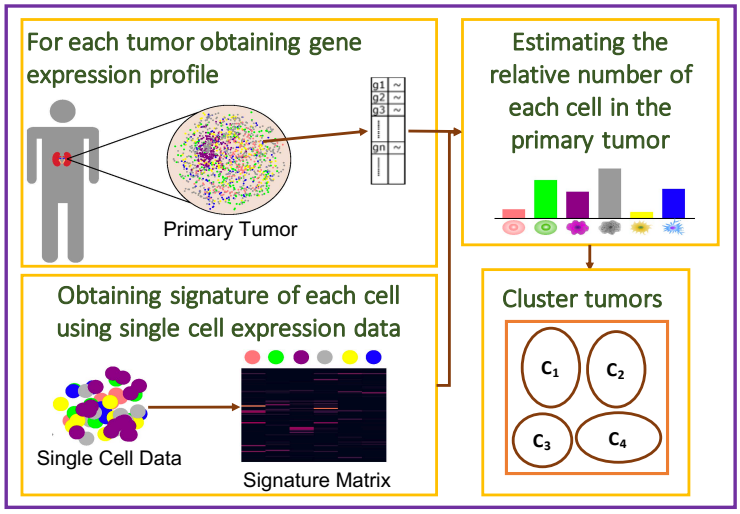


Figure 4.1. For each RCC tumor in TCGA data, we first estimate the percentage of each immune cell in the tumor using the gene expression profile of the tumor and the signature matrix of immune cells. We then cluster tumors based on the percentage of each immune cell in tumors.

4.1 Materials and Methods

Tumor infiltrating immune cells were estimated by using CIBERSORTx deconvolution method that has the most accurate results compared to other deconvolution methods we studied. To investigate the immune variations in renal cancer, we down-

loaded the TCGA data set of gene expression profiles of 607 RCC primary tumors from UCSC Xena to use as a mixture data y and used LM22 signature matrix, which is an immune cell signature matrix that include 547 genes differentiating 22 cell types that are B cells naive, B cells memory, Plasma cells, T cells CD8, T cells CD4 naive, T cells CD4 memory resting, T cells CD4 memory activated, T cells follicular helper, T cells regulatory (Tregs), T cells gamma delta, NK cells resting, NK cells activated, Monocytes, Macrophages M0, Macrophages M1, Macrophages M2, Dendritic cells resting, Dendritic cells activated, Mast cells resting, Mast cells activated, Eosinophils, Neutrophils. We then estimated cell fraction of RCC patients by using CIBERSORTx-B mode to remove technical differences between LM22 signature matrix (derived from microarray data) and RNA-seq mixture data.

4.1.1 K-Mean Clustering.

After we obtain estimated cell proportions, we included only cases with CIBERSORTx p-value < 0.05. We then applied unsupervised k-mean clustering algorithm to cluster patients based on their percentage of immune cells. The k-mean algorithm separates samples in k-group of equal variance by minimizing the inertia (distance between samples in the clusters and center of the clusters).

In the K-mean algorithm for a given integer k and a set of n data points in X, goal is determine k centers C so that:

$$\sum_{i}^{n} \min_{c \in C} ||x - c||^2 \quad \text{where} \quad c_i = \frac{1}{C_i} \sum_{x \in C_i} x$$

The k-means algorithm divides a set of samples into disjoint clusters, each described by the mean of the samples in the cluster. The means are commonly called the cluster "centroids"; note that they are not, in general, points from the data, although they live in the same space. To determine the optimal number of cluster

(k-value), we used elbow method that is plotting the number of clusters on the X-axis and the inertia on the Y-axis and then select the value of k for which we see a bend. We decide k=4 from Figure 4.2 [72].

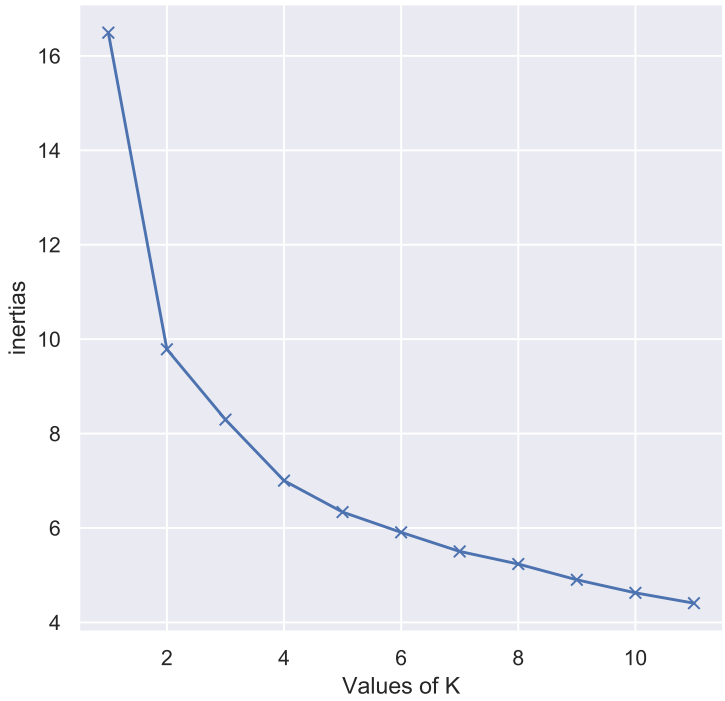


Figure 4.2. The Elbow Method using inertias.

We also use clinical information of patients from cBioPortal but dropped some patients in the gene expression data due to missing clinical information and continued our analysis with 526 patients. The patients' characteristics are given in Table 4.1.

Gender		Grade of Tumor				Stage of Tumor			
Female	Male	G1	G2	G3	G4	T1	T2	T3	T4
184	342	14	226	204	75	268	69	178	11

Tumor Status		Survival Status		Laterality	
Tumor Free	With Tumor	Alive	Deceased	Left	Right
356	137	355	171	247	278

Table 4.1. Patients' characteristics. Sub-tables indicate the number of patients in each category. Differences in the numbers are due to missing information for some patients

4.2 Results

To estimate the percentage of each cell type in RCC tumors, we apply "digital mass cytometry", CIBERSORTx, on TCGA gene expression profiles of RCC primary tumors. We compare the results of our "digital mass cytometry" analysis with the results of an experimental study of a large-scale mass cytometry-based immune cells analysis of 73 RCC patients [61]. Immune cells, which have been characterized in this experimental study done by Chevrier et al. [61] are macrophages, CD8+ T-cells, CD4+ T-cells, NK cells, B cells, plasma cells, dendritic cells (DC), CD45+ T-cells, double positive T-cells (DP_T-cells), double negative T-cells (DN_T-cells). To be able to compare our results, which includes 22 immune cell types given in LM22 signature matrix of CIBERSORTx, we combine cells that belong to the same family. For instance, since CD4+ naive T-cells, CD4+ memory resting T-cells, CD4+ memory activated T-cells, follicular helper T-cells, and regulatory T-cells are sub-types of CD4+ T-cells, we sum their numbers to estimate the total number of CD4+ T-cells. We do similar calculation for B cells, NK cells, DC cells, macrophages, and mast cells.

4.2.1 The most frequent immune cells in RCC tumors are macrophages, CD4+ T-cells, and CD8+ T-cells.

Results of experimental study done by Chevrier et al [61] show that macrophages are the most frequent immune cells in most RCC tumors with a mean of 31% followed by CD8+ T-cells and CD4+ T-cells, respectively (Figure 4.3D,B), which are in agreement with the results of CIBERSORTx applied on TCGA data set (Figure 4.3A,C).

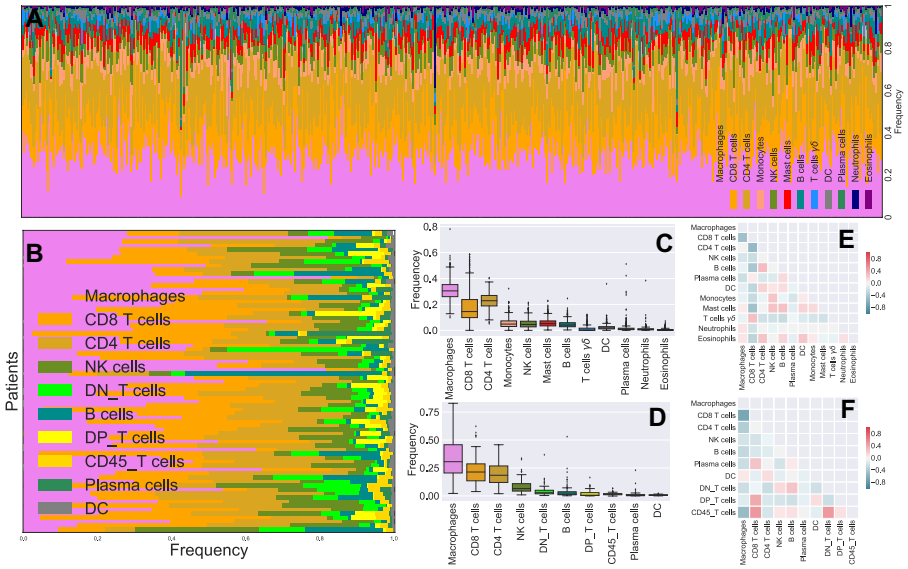


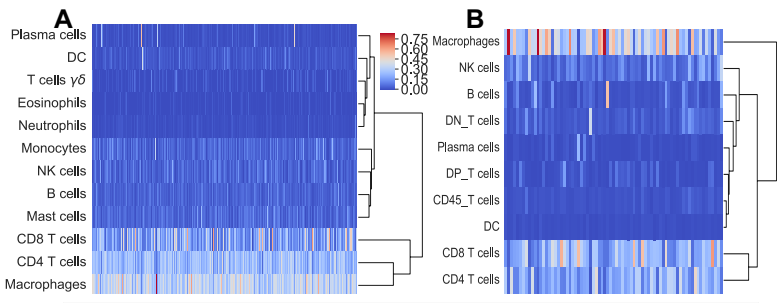
Figure 4.3. Sub-figures A and B respectively show the estimated percentage of each immune cell in RCC tumors obtained by applying CIBERSORTx on TCGA data (A) and mas cytometry analysis of 73 RCC patients done by Chevrier et al [61] (B). Sub-figures C and D respectively show the immune cell percentage in renal tumors in a box plot format for TCGA data (C) and mas cytometry analysis [61] (D). Sub-figures E and F indicate the correlation map of estimated immune cell frequencies in TCGA tumors (E) and the results of mas cytometry analysis [61] (F), respectively.

4.2.2 There is a negative correlation between the number of macrophages and CD8+ T-cells.

The results of mas cytometry analysis indicate a negative correlation between CD8+ T-cells and macrophages with Pearson correlation coefficients of -0.67. Importantly, the digital mass cytometry applied on TCGA data set confirms this negative correlation between the number of CD8+ T-cells and macrophages in RCC with a correlation coefficient of -0.46 (Figure 4.3E,F).

4.2.3 Variations of RCC tumors are mainly in the percentage of macrophages, CD8+ T-cells, and CD4+ T-cells compared to the other immune cell types.

Figure 4.4 shows a significant variations among the percentage of CD8+ and CD4+ T-cells and macrophages across RCC tumors, while there is a slight variation in the percentage of other immune cell types. Unsupervised hierarchical clustering of cell frequencies show that CD8+ T-cells and CD4+ T-cells are clustered together in the experimental results, and then they group with macrophages and other cells. The result of digital mass cytometry on TCGA data shows a kind of similar trend: CD4+ T-cells first clustered with macrophages, then they clustered with CD8+ T-cells and other cells (Figure 4.4B).



	CD4<CD8≈ $M\phi$	CD8<CD4<$M\phi$	CD4<$M\phi$<CD8	CD8<CD4≈ $M\phi$
Number of Patients	129	120	68	209

Figure 4.4. Immune cell frequency patterns of RCC tumors. Sub-figure A shows hierarchically-clustering of cell frequencies obtained by digital mass cytometry of TCGA data. Sub-figure B presents hierarchically-clustering of cell frequencies obtained by mas cytometry analysis [61]. In Sub-figure C, RCC tumors in TCGA data set are clustered using K-mean clustering method based on their immune cell frequencies. Table indicates the number of patients in each category.

4.2.4 There are four immune patterns of RCCs.

K-mean clustering of RCC tumors based on their immune cells' frequencies shows that there are four different immune classes: Cluster 1 ($CD4 < CD8 \approx M\Phi$), in which the numbers of macrophages and CD8+ T-cells are approximately the same, and the number of CD4+ T-cells is slightly less than the number of CD8+ T-cells; Cluster 2 called ($CD8 < CD4 < M\Phi$), in which the number of macrophages is significantly higher than the number of CD4+ and CD8+ T-cells; Cluster 3 ($CD4 < M\Phi < CD8$), in which the number of CD8+ T-cells is significantly higher than the number of macrophages and CD4+ T-cells; and Cluster 4 called ($CD8 < CD4 \approx M\Phi$) in which the numbers of macrophages and CD4+ T-cells are approximately the same, and

the number of CD8+ T-cells is significantly less than the number of CD4+ T-cells (Figure 4.4C).

4.2.5 Cluster ($CD8 < CD4 \approx M\Phi$) has the highest percentage of grade and stage 1 and 2 tumors.

Comparing clinical features of clusters show that Cluster ($CD8 < CD4 \approx M\Phi$) includes the highest percentage of grade 1 and grade 2 tumors and the lowest percentage of grade 4 tumors, and there is a similar trend for the stage of tumors (Figure 4.5A,B). Importantly, Cluster ($CD8 < CD4 \approx M\Phi$) has the highest proportion of patients who were tumor free and smallest percentage of the diseased patients at the last time of follow up compared to the other clusters (Figure 4.5D). Furthermore, this cluster has the highest frequency of mast cells, monocytes and B cells compared to the other clusters (Figure 4.4C). These results might imply that non-aggressive tumors include an approximately equal number of each immune cell type.

4.2.6 Cluster ($CD4 < M\Phi < CD8$) has the highest percentage of grade and stage 4 tumors compared to the other clusters.

The percentages of grade and stage 3 and 4 tumors are higher in Cluster ($CD4 < M\Phi < CD8$) compared to the other clusters (Figure 4.5A,B). Furthermore, this cluster includes the highest number of deceased patients and patients who had a tumor at the last time of follow up compared to the other clusters (Figure 4.5C,D). There is a significant difference among overall survival months of female and male patients in this cluster, female patients in the cluster ($CD4 < M\Phi < CD8$) have the highest overall survival months compared to the other clusters (Figure 4.5H). These results indicate that male patients' RCC tumors consisting of a significantly higher number of CD8+ T-cells than any other immune cell types might be aggressive.

4.2.7 There is no significant differences in overall survival months or age at diagnosis of patients in each cluster.

Figure 4.5 indicates no significant differences in the overall survival of patients in each cluster and some other interesting observations. For example, patients in Cluster $CD4 < CD8 \approx M\Phi$ with and without tumors at the last time of follow up have a similar overall survival months while in all other clusters patients with tumor have a significantly lower survival months than patients without tumors at the last time of follow up. Moreover, patients with tumor in this cluster have a significantly higher age at diagnosis compared to the patients with no tumors in this cluster. Furthermore female patients in this cluster have a significantly higher age at diagnosis but the same survival as male patients in this cluster. Additionally, female patients in Cluster ($CD4 < M\Phi < CD8$) have a significantly higher overall survival months than male patients in this cluster, while females have a slightly higher age at diagnosis than males in this cluster. Importantly, there is no significant differences in the age at diagnosis and survival months of patients in each cluster based on the location of their primary tumors, left and right kidneys.

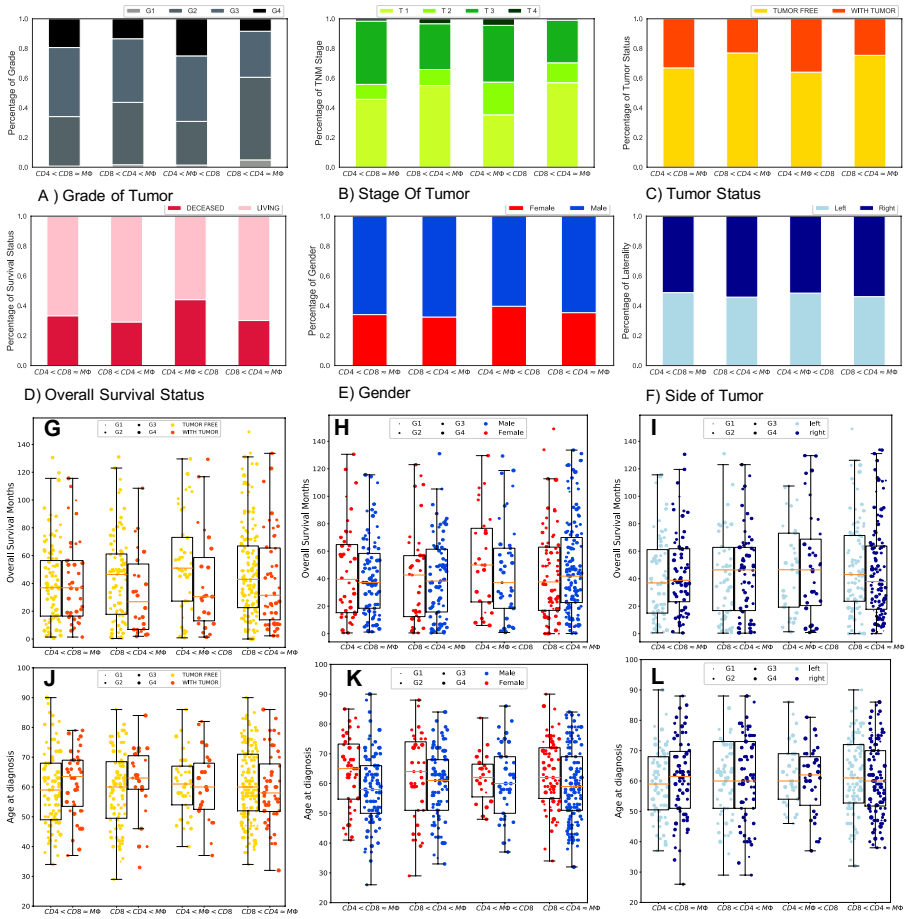


Figure 4.5. Sub-figures A-F show the clinical information of each group of RCC tumors. Sub-figures G, H and I show the overall survival months and sub-figures J, K and L represent the age of diagnosis of the patients in each cluster as a function of tumor status, gender and the location of the primary tumor, respectively; the size of markers indicates the grade of tumors.

4.2.8 Higher grade and stage of RCC tumors have higher percentage of CD8+ T-cells and lower percentages of mast cells and monocytes.

A study of 87 RCC patients indicates that the percentage of tumor infiltrating CD8+ T-cells co-expressing PD-1 and Tim-3 correlated with an aggressive phenotype and a larger tumor size at diagnosis [73]. Figure 4.6 also shows that the grade and stage 3 and 4 RCC tumors have a significantly higher percentage of CD8+ T-cells

compared to the stage and grade 1 and 2 tumors, which is consistent with the observations of Figure 4.5.

Figure 4.6 also indicates that the percentages of mast cells and monocytes in RCC tumors significantly decrease when the grade and stage of tumors increase. Note, Clusters ($CD8 < CD4 < M\Phi$) and ($CD8 < CD4 \approx M\Phi$) that have higher frequency of mast cells and monocytes and lower frequency of CD8+ T-cells have the least percentage of grade three and four tumors (Figures 4.4C and 4.5).

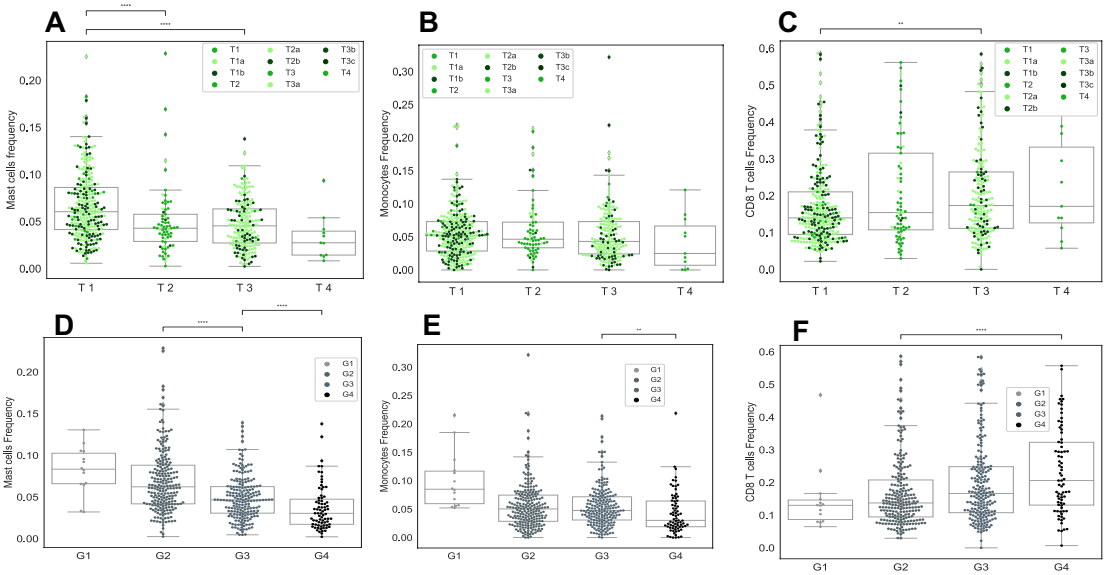


Figure 4.6. Percentage of mast cells, monocytes and CD8+ T-cells in RCC tumors as a function of grade and TNM staging. Sub-figures A, B, D , and E show that the percentages of mast cells and monocytes decrease when the grade and stage of tumor increase. Sub-figures C and F represent the percentage of CD8+ T-cells in primary tumors as functions of grade and stage of tumors. For some patients with the stage Ti (green), $i = 1, 2, 3, 4$ cancer, we have extra information about their stage: Tia (light green) or Tib (dark green), for more information see Table 1.1 [12] .

4.2.9 Tumor free patients have a significantly higher percentages of NK cells and mast cells.

Figure 4.7 shows that the percentage of NK cells and mast cells are significantly higher in primary tumors of tumor free patients versus patients with tumor at the last time of follow up in all clusters. Importantly, Cluster ($CD8 < CD4 \approx M\Phi$) has the highest percentage of mast cells and NK cells compared to the other clusters (Figures 4.4 C and 4.7). Note, this cluster has the highest percentage of grade and stage 1 and 2 tumors. Additionally, RCC tumors in Cluster ($CD4 < M\Phi < CD8$), which has the highest percentage of grade and stage 4 tumors, have the lowest amount of mast cells.

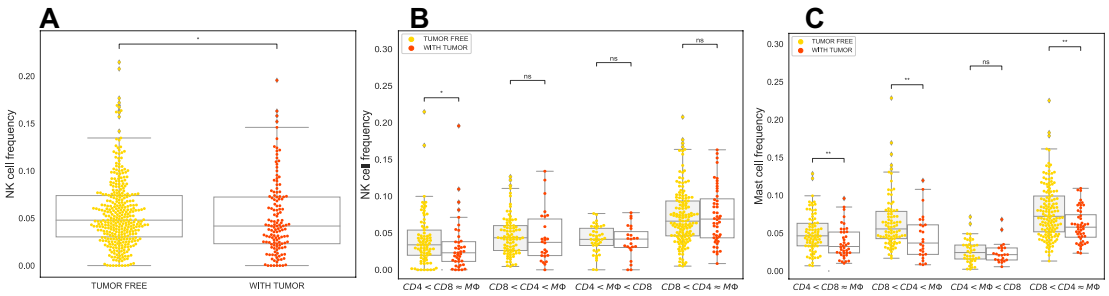


Figure 4.7. Sub-figure A shows that primary tumors of patients who were tumor free at the last time of follow up have higher percentage of NK cells than patients with tumor at the last time. Sub-figure B indicates the percentage of NK cell in each cluster. Sub-figure C shows that in all clusters of RCC tumors, the percentage of mast cells is significantly higher in tumor free patients than patients with tumor at the last time of follow up.

4.2.10 Genes expression levels of PDCD1 and INFG are significantly positively correlated with the percentage of CD8+ T-cells in RCC tumors.

Programmed cell death protein 1 (PD-1) is a type of protein that found on T-cells and it prevents T-cells from killing cancer cells when it binds to PD-1 ligand

(PD-L1) and PD-2 ligand (PD-L2) on cancer cells [74]. PDCD1 is a gene that encodes PD-1 proteins [75]. PDCD1 gene and CD8+ T-cells are highly positively correlated, with correlation coefficient of 0.85. Also, expression level of PDCD1 is the highest in the cluster ($CD4 < M\Phi < CD8$) and the lowest in the cluster ($CD8 < CD4 \approx M\Phi$) as a result of positive correlation with CD8+ T-cells (Figure 4.8C and F).

Interferon γ (INF_γ), encoded by INFG gene [76], is a cytokine that is essential for innate and adaptive immunity. It works as an activator of macrophages and stimulator of NK cells and neutrophils [77] and they are mostly produced by T-cells and NK cells as a reaction of a variety of inflammatory or immune stimuli [78]. Saliently, expression level of INFG is significantly positively correlated with the percentage of CD8+ T-cells and the expression level of PDCD1 in RCC tumors, with correlation coefficients of 0.79 and 0.87, respectively. In addition, cluster ($CD4 < M\Phi < CD8$) has the highest INFG expression level and cluster ($CD8 < CD4 \approx M\Phi$) has the lowest expression level of INFG as expected (Figure 4.8).

In contrast, there is a slightly positive correlation between the expression levels of CD274 and PDCD1LG2 genes, that encodes PD-L1 and PD-L2 respectively, with the expression levels of PDCD1 and INFG, and the percentage of CD8+ T-cells in RCC tumors (Figure 4.8F). In addition, cluster ($CD8 < CD4 \approx M\Phi$) has the lowest levels of CD274 and PDCD1LG2 compared to the other clusters (Figure 4.8B and D).

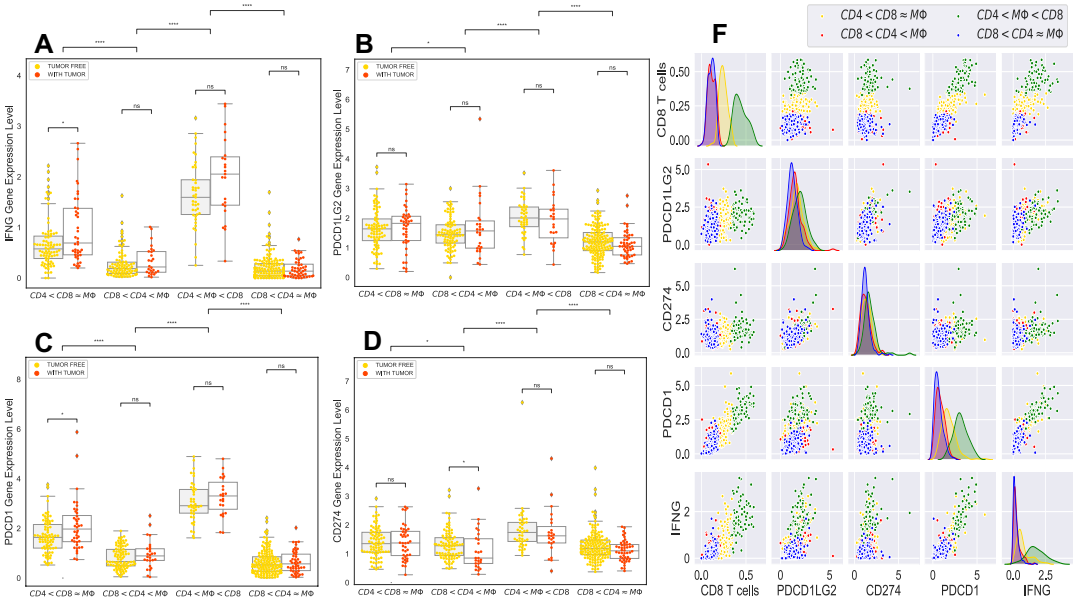


Figure 4.8. Sub-figures A, B, C and D respectively show INFG, PDCD1LG2, PDCD1 and CD274 genes expression values for the clusters based on tumor status. Sub-figure F represents the values and distribution of INFG, PDCD1LG2, PDCD1, CD274 genes expression and CD8+ T-cells in RCC primary tumors color coded by clusters that belong to.

4.2.11 Aggressive tumors are mostly in the clusters $(CD8 < CD4 < M\Phi)$ and $(CD4 < M\Phi < CD8)$.

When we use a sunburst chart to visualize hierarchical structures of clusters, where clusters are inner center, tumor status and grade of tumor are the outer rings in Figure 4.9A, survival status and stage of tumor are the outer rings in Figure 4.9B, we see that patients with stage 4 tumor and grade 4 tumor are mostly in the clusters $(CD8 < CD4 < M\Phi)$ and $(CD4 < M\Phi < CD8)$ with colors red and green in the figures respectively. Interestingly, these cluster have very distinct INGF and PDCD1 genes expression levels (Figure 4.8A,B).

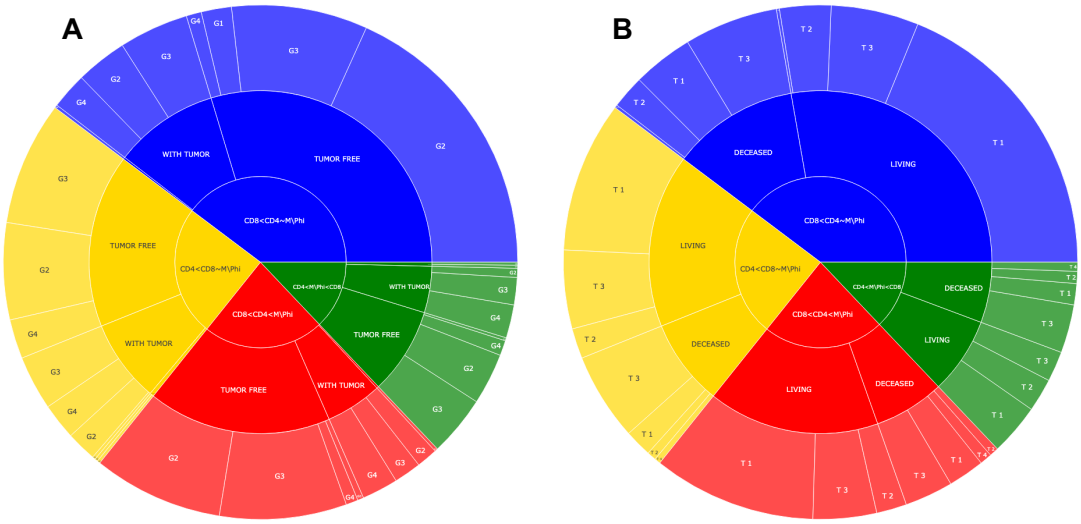


Figure 4.9. Hierarchical structure of clusters. This figure indicates percentage of patients with different clinical features in each cluster: tumor status and stage of tumor in Sub-figure A and overall survival status and stage of tumor in Sub-figure B.

4.2.12 Cluster ($CD8 < CD4 \approx M\Phi$) has the highest RGS5 gene expression level.

RGS5 is a member of the regulators of G protein signaling (RGS) family, and they are known as signal transaction molecules that are associated with the arrangement of heterotrimetric G proteins by acting as GTPase activators. Moreover, RGS5 is a hypoxia-inducible factor-1 dependent involved in the induction of endothelial apoptosis [36]. In Chapter 2, we found that a high RGS5 gene expression level is associated with better survival months for RCC patients and when the grade of RCC tumors increases, the RGS5 expression level significantly decreases [**?**]. Interestingly, cluster ($CD8 < CD4 \approx M\Phi$) has the highest RGS5 gene expression level and tumor free patients have higher level of RGS5 expression than patients with tumor (Figure 4.10A,B). Moreover, RCC tumors with a high expression level of RGS5 have higher percentage of mast cells and monocytes (Figure 4.10C and D).

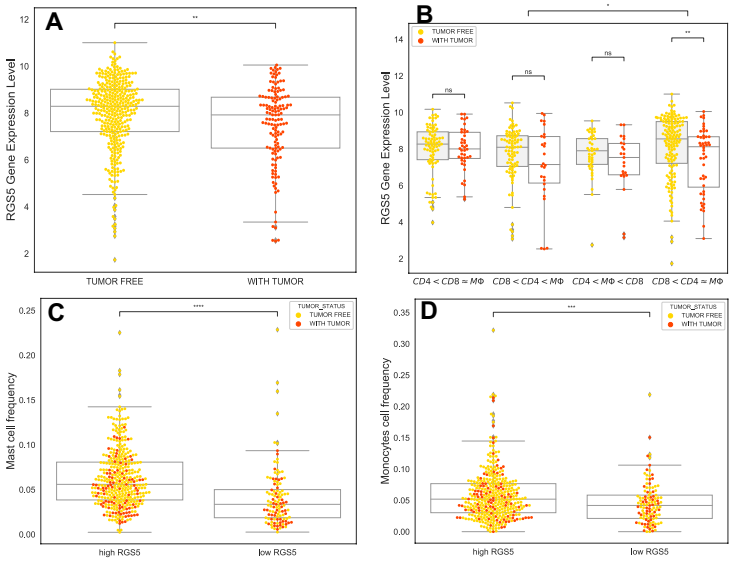


Figure 4.10. Sub-figure A shows expression level of RGS5 in RCC tumors as a function of tumor status at the time of the last follow up. Sub-figure B shows the relationship between RGS5 gene expression level and tumor status in each cluster. Sub-figures C and D indicate the relation between the level of RGS5 and the percentage of mast cells and monocytes in RCC tumors.

4.3 Discussion

Immune checkpoints are essential parts of immune system, and they are crucial to prevent autoimmune diseases. However, some tumors benefit from these checkpoints, because these checkpoints can prevent the immune system from killing cancer cells. One such immune checkpoint is programmed cell death 1 protein (PD-1), which binds to its ligand PD-L1 and inhibits immune cell activities, including T cell activities. One strategy for cancer immunotherapy is to block these checkpoints to promote anti-cancer T-cell activities [79, 80, 81, 82]. Immunotherapy such as targeting PD-1 pathway has improved overall survival months of several patients with metastatic cancers, including melanoma, head and neck cancer, renal cell carcinoma, non-small cell lung cancer (NSCLC), and colon cancer. However, there are many patients

who do not respond to these treatments, and some patients who initially respond to the treatments, they might develop resistance or experience severe adverse events [83, 84, 85]. For this reasons, further biomarkers of tumor cells such as PD-1 and PD-L1 and of tumor infiltrating immune cells such as T-cells and macrophages need to be established to identify the patients that can be treated by immunotherapy drugs [86].

In kidney cancer, common immunotherapy drugs such as nivolumab and avelumab target PD-1, PD-L1, and PD-L2 pathways [87]. Anti PD-1 drugs targets T-cells directly, while anti-PD-L1 drugs target tumor cells directly, and they may also target tumor associated macrophages that express PD-L1. Several studies indicate an increase of $INF\gamma$ production in the PD-1 inhibitors and other immune checkpoint blockade therapies that resulted in destruction of cancer cells [88, 89, 90], and a link between cancer immunotherapy improvement and an increase of $INF\gamma$ expression has been observed [78]. An increase in $INF\gamma$ gene expression has also been associated with better progression-free survival in NSCLC and urothelial cancer patients treated with a PD-L1 inhibitor [91].

Note, tumors in cluster ($CD4 < M\Phi < CD8$) have a high expression levels of INFG, the gene encoding $INF\gamma$, and PDCD1, the gene encoding PD-1, compared to the other clusters, and the expression levels of these genes are significantly correlated with the percentage of CD8+ T-cells in tumors. Importantly, it has been shown that $INF\gamma$ boosts the CD8+ T-cells expansion [92]. Thus, patients in the cluster ($CD4 < M\Phi < CD8$) might respond to the PD-1 inhibitors. In addition, since there is not a strong correlation between PDCD1LG2 and CD274 expression levels and levels of INFG and PDCD1 genes, PD-L1 and PD-L2 inhibitors might not be as effective treatments as the PD-1 inhibitors for the patients in this clusters. Although cluster ($CD8 < CD4 \approx M\Phi$) includes a high number of patients with lower grade

and patients who were tumor free in the last follow up time, tumors in this cluster have lower levels of INFG and PDCD1, therefore patients in this cluster may not be a good candidate for anti PD-1 therapies.

Anti-angiogenic agent (AA) is one of the main treatments in the aggressive RCC [93], because nutrients and oxygen are the main ingredients of the tumor growth which come from blood and angiogenes is the formation of blood vessels that carries blood to the tumor. Anti-angiogenics, also known angiogenesis inhibitors, are drugs that stop the growth of blood vessels (angiogenesis) that tumors need to grow [94]. A study of in vitro cell lines and in vivo mouse model of RCC shows that the recruitment of mast cells is related with increased RCC angiogenesis by modulating $PI3K \rightarrow AKT \rightarrow GSK3 \rightarrow AM$ signaling [95]. Since cluster ($CD8 < CD4 \approx M\Phi$) has the highest amount of mast cells compared to other clusters, angiogenesis inhibitors might be a good treatment option for the patients in this cluster. Moreover, mast cells are suggested as an independent prognostic factor in some studies of RCC patients [96, 97], and the amount of mast cells negatively correlated with 5-year survival [97]. However, our finding shows that the number of mast cells inversely correlated with the grade of tumors (Figure 4.6A,D), and the primary tumors of patients without tumors at the last time of follow up have higher percentage of mast cells than primary tumors of patients with tumor at the last time of follow up.

In addition to our first study results, we found that patients with higher RGS5 level have a higher percentage of mast cells and monocytes. Moreover, patients in cluster ($CD8 < CD4 \approx M\Phi$) have the highest amount of RGS5 expression. With the help of further investigation, RGS5 gene might also be a good target for patients in this cluster. For all above mentioned suggestions, further clinical and biological studies are required to test and validate them.

CHAPTER 5

CONCLUSION

Renal cell carcinoma (RCC) is the most frequently diagnosed malignant tumor type in the adult kidneys consisting of approximately 85% of kidney cancer cases [93]. Analysis of large complex biological and clinical data sets is one of the most popular research area to infer the underlying mechanism of many diseases, including renal cancer and arrive at personalized treatments.

In this study, we first analyzed gene expression profiles of 534 Renal Cell Carcinoma patients to find prognostic markers that can be used in treatments of RCC patients. We normalized gene expression of each patient separately and then used variance threshold method to reduce the dimensionality of the data. Among the top 10 most variant genes, RGS5 gene expression shows a strong association with some clinical features. For example, the grade of tumor is a decreasing function of RGS5 level and high RGS5 level is associated with better survival months. However, considering the different biological roles of RGS5 [48] and contradictory observations, this gene should be investigated more to be able to use it as targeted treatments.

Another popular research area about cancer treatments is investigating the tumor microenvironment that is essential for the immunotherapeutic interventions. Immunotherapies are costly treatments for certain cancer types, and they are not effective treatments for the considerable amount of patients and they have some important side effects [86]. For this reason, we analyzed immune cell proportion of RCC patients by using the same gene expression data we used in the first part. We first compared available tumor deconvolution methods to determine the best for our

analysis, and we found that CIBERSORTx has the highest correlation with ground truth data compared to other tumor deconvolution methods such as ssGSEA and SingScore. Using CIBERSORTx, we estimated the percentage of each immune cell type of 526 RCC patients. K-mean clustering of tumors based on their immune proportions showed the existence of four distinct classes of RCC tumors. Moreover, we observed a significant correlation between the number of CD8+ T cells and expression levels of IFNG and PDCD1. Importantly, higher stage and grade of tumors have a significantly higher percentage of CD8+ T cells in tumors. These results may be used to determine the group of patients who might benefit from anti PD-1 therapies or angiogenesis inhibitors.

REFERENCES

[1] Cohen HT, McGovern FJ. Renal-cell carcinoma. The New England journal of medicine. dec;(23):2477–90.

[2] Linehan WM, Zbar B. Focus on kidney cancer. Cancer Cell. 2004;6(3):223–228.

[3] Aron M, Nguyen MM, Stein RJ, Gill IS. Impact of Gender in Renal Cell Carcinoma: An Analysis of the SEER Database. European Urology. 2008;54(1):133–142.

[4] Stafford HS, Saltzstein SL, Shimasaki S, Sanders C, Downs TM, Sadler GR. Racial/ethnic and gender disparities in renal cell carcinoma incidence and survival. The Journal of urology. may;(5):1704–8.

[5] Moore LE, Wilson RT, Campleman SL. Lifestyle factors, exposures, genetic susceptibility, and renal cell cancer risk: A review. Cancer Investigation. 2005;23(3):240–255.

[6] The Surveillance, Epidemiology, and End Results (SEER) Program;. https://seer.cancer.gov/statfacts/html/kidrp.html.

[7] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. jul;(7456):43–9.

[8] Tan C, Takayama T, Takaoka N, Fujita H, Miyazaki M, Sugiyama T, et al. Impact of gender in renal cell carcinoma: The relationship of FABP7 and BRN2 expression with overall survival. Clinical Medicine Insights: Oncology. 2014;8:21–27.

[9] Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics: Abstract: Nature reviews genetics. Nature Reviews Genetics;(1):57–63.

[10] TCGA. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga;.

[11] Novara G, Martignoni G, Artibani W, Ficarra V. Grading Systems in Renal Cell Carcinoma. Journal of Urology. feb;(2):430–436.

[12] Swami U, Nussenzveig RH, Haaland B, Agarwal N. Revisiting AJCC TNM staging for renal cell carcinoma: quest for improvement. Annals of Translational Medicine. mar;(S1):S18–S18.

[13] Shahriyari L. Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. Briefings in Bioinformatics. 2017;(September):1–10.

[14] Shahriyari L, Abdel-Rahman M, Cebulla C. BAP1 expression is prognostic in breast and uveal melanoma but not colon cancer and is highly positively correlated with RBM15B and USP19. PLoS ONE. 2019;14(2):1–14.

[15] Young IT. Introduction To Statistical Pattern Recognition. Proc Natl Electron Conf. 1974;29:349–352.

[16] KNIME. Seven Techniques for Data Dimensionality Reduction — KNIME;p. 1–21.

[17] Bartholomew DJ. Principal components analysis. International Encyclopedia of Education. 2010;p. 374–377.

[18] Ghodsi A. Dimensionality Reduction A Short Tutorial. Science;p. 25.

[19] Frigge M, Hoaglin DC, Iglewicz B. Some implementations of the boxplot. American Statistician. 1989;43(1):50–54.

[20] Wilkinson L, Friendly M. History corner the history of the cluster heat map. American Statistician. 2009;63(2):179–184.

[21] Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics. mar;(1):50–60.

[22] Hurst JH, Hooks SB. Regulator of G-protein signaling (RGS) proteins in cancer biology. Biochemical Pharmacology. 2009;78(10):1289–1297.

[23] Seki N, Sugano S, Suzuki Y, Nakagawara A, Ohira M, Muramatsu MA, et al. Isolation, tissue expression, and chromosomal assignment of human RGS5, a novel G-protein signaling regulator gene. Journal of Human Genetics. 1998;43(3):202–205.

[24] Bell SE, Mavila A, Salazar R, Bayless KJ, Kanagala S, Maxwell SA, et al. Differential gene expression during capillary morphogenesis in 3D collagen matrices: regulated expression of genes involved in basement membrane matrix assembly, cell cycle progression, cellular differentiation and G-protein signaling. Journal of cell science;(Pt 15):2755–73.

[25] Hakak Y, Shrestha D, Goegel MC, Behan DP, Chalmers DT. Global analysis of G-protein-coupled receptor signaling in human tissues. FEBS Letters. 2003;550(1-3):11–17.

[26] Deckers IAG, Van Den Brandt PA, Van Engeland M, Van Schooten FJ, Godschalk RWL, Keszei AP, et al. Potential role of gene-environment interactions in ion transport mechanisms in the etiology of renal cell cancer. Scientific Reports;(January):1–10.

[27] Furuya M, Nishiyama M, Kimura S, Suyama T, Naya Y, Ito H, et al. Expression of regulator of G protein signalling protein 5 (RGS5) in the tumour vasculature of human renal cell carcinoma. Journal of Pathology. 2004;203(1):551–558.

[28] Yao M, Huang Y, Shioi K, Hattori K, Murakami T, Sano F, et al. A three-gene expression signature model to predict clinical outcome of clear cell renal carcinoma. International Journal of Cancer. 2008;123(5):1126–1132.

[29] Huang G, Song H, Wang R, Han X, Chen L. The relationship between RGS5 expression and cancer differentiation and metastasis in non-small cell lung cancer. Journal of Surgical Oncology. 2012;105(4):420–424.

[30] Wang D, Xu Y, Feng L, Yin P, Song SS, Wu F, et al. RGS5 decreases the proliferation of human ovarian carcinoma-derived primary endothelial cells through the MAPK/ERK signaling pathway in hypoxia. Oncology Reports. 2019;41(1):165–177.

[31] Xu Z, Zuo Y, Wang J, Yu Z, Peng F, Chen Y, et al. Overexpression of the regulator of G-protein signaling 5 reduces the survival rate and enhances the radiation response of human lung cancer cells. Oncology Reports. 2015;33(6):2899–2907.

[32] O'Neill WC. Targeting serum calcium in chronic kidney disease and end-stage renal disease: Is normal too high? Kidney International;(1):40–45.

[33] Erlinger TP, Tarver-Carr ME, Powe NR, Appel LJ, Coresh J, Eberhardt MS, et al. Leukocytosis, hypoalbuminemia, and the risk for chronic kidney disease in US adults. American Journal of Kidney Diseases;(2):256–263.

[34] Krüger T, Schoor O, Lemmel C, Kraemer B, Reichle C, Dengjel J, et al. Lessons to be learned from primary renal cell carcinomas: Novel tumor antigens and HLA ligands for immunotherapy. Cancer Immunology, Immunotherapy. 2005;54(9):826–836.

[35] Boss CN, Grünebach F, Brauer K, Häntschel M, Mirakaj V, Weinschenk T, et al. Identification and characterization of T-cell epitopes deduced from RGS5,

a novel broadly expressed tumor antigen. Clinical cancer research : an official journal of the American Association for Cancer Research. jun;(11):3347–55.

[36] RGS5 regulator of G protein signaling 5 [Homo sapiens (human)];. https://www.ncbi.nlm.nih.gov/gene/8490.

[37] Chang YPC, Liu X, Kim JDO, Ikeda MA, Layton MR, Weder AB, et al. Multiple Genes for Essential-Hypertension Susceptibility on Chromosome 1q. The American Journal of Human Genetics. 2007;80(2):253–264.

[38] Gu S, Cifelli C, Wang S, Heximer S. RGS proteins: identifying new GAPs in the understanding of blood pressure regulation and cardiovascular function. Clinical Science. 2009;116(5):391–399.

[39] Rubagotti A, Martorana G, Boccardo FM. Epidemiology of Kidney Cancer. European Urology, Supplements. 2006;5(8):558–565.

[40] Koh J, Dar M, Untch BR, Dixit D, Shi Y, Yang Z, et al. Regulator of G Protein Signaling 5 Is Highly Expressed in Parathyroid Tumors and Inhibits Signaling by the Calcium-Sensing Receptor. Molecular Endocrinology. 2011;25(5):867–876.

[41] Lim LM, Kuo HT, Kuo MC, Chiu YW, Lee JJ, Hwang SJ, et al. Low serum calcium is associated with poor renal outcomes in chronic kidney disease stages 3-4 patients. BMC Nephrology. 2014;15(1):1–9.

[42] Kovesdy CP, Kuchmak O, Lu JL, Kalantar-Zadeh K. Outcomes associated with serum calcium level in men with non-dialysis-dependent chronic kidney disease. Clinical Journal of the American Society of Nephrology. 2010;5(3):468–476.

[43] Woldrich JM, Mallin K, Ritchey J, Carroll PR, Kane CJ. Sex Differences in Renal Cell Cancer Presentation and Survival: An Analysis of the National Cancer Database, 1993?2004. Journal of Urology. may;(5):1709–1713.

[44] Strauss A, Uhlig J, Lotz J, Trojan L, Uhlig A. Tumor laterality in renal cancer as a predictor of survival in large patient cohorts: A STROBE compliant study. Medicine. 2019;98(17):e15346.

[45] Guo S, Yao K, He X, Wu S, Ye Y, Chen J, et al. Prognostic significance of laterality in renal cell carcinoma: A population-based study from the surveillance, epidemiology, and end results (SEER) database. Cancer Medicine. 2019;8(12):5629–5637.

[46] Fan F, Jia J, Li J, Huo Y, Zhang Y. White blood cell count predicts the odds of kidney function decline in a Chinese community-based population. BMC Nephrology. 2017;18(1):1–9.

[47] Fox P, Hudson M, Brown C, Lord S, Gebski V, De Souza P, et al. Markers of systemic inflammation predict survival in patients with advanced renal cell cancer. British Journal of Cancer;(1):147–153.

[48] Altman MK, Nguyen DT, Patel SB, Fambrough JM, Beedle AM, Hardman WJ, et al. Regulator of G-Protein Signaling 5 Reduces HeyA8 Ovarian Cancer Cell Proliferation and Extends Survival in a Murine Tumor Model. Biochemistry Research International;p. 1–9.

[49] Escudier B. Emerging immunotherapies for renal cell carcinoma. Annals of Oncology. sep;p. viii35–viii40.

[50] Considine B, Hurwitz ME. Current Status and Future Directions of Immunotherapy in Renal Cell Carcinoma. Current Oncology Reports. apr;(4):34.

[51] Grivennikov SI, Greten FR, Karin M. Immunity, Inflammation, and Cancer. Cell. mar;(6):883–899.

[52] Kitamura T, Qian BZ, Pollard JW. Immune cell promotion of metastasis. Nature Reviews Immunology. feb;(2):73–86.

[53] Candido J, Hagemann T. Cancer-Related Inflammation. Journal of Clinical Immunology. jan;(S1):79–84.

[54] Swann JB, Smyth MJ. Immune surveillance of tumors. Journal of Clinical Investigation. may;(5):1137–1146.

[55] Vesely MD, Kershaw MH, Schreiber RD, Smyth MJ. Natural Innate and Adaptive Immunity to Cancer. Annual Review of Immunology. apr;(1):235–271.

[56] Yao J, Xi W, Zhu Y, Wang H, Hu X, Guo J. Checkpoint molecule PD-1-assisted CD8+ T lymphocyte count in tumor microenvironment predicts overall survival of patients with metastatic renal cell carcinoma treated with tyrosine kinase inhibitors. Cancer Management and Research. sep;p. 3419–3431.

[57] Pagès F, Kirilovsky A, Mlecnik B, Asslaber M, Tosolini M, Bindea G, et al. In Situ Cytotoxic and Memory T Cells Predict Outcome in Patients With Early-Stage Colorectal Cancer. Journal of Clinical Oncology. dec;(35):5944–5951.

[58] Griss J, Bauer W, Wagner C, Simon M, Chen M, Grabmeier-Pfistershammer K, et al. B cells sustain inflammation and predict response to immune checkpoint blockade in human melanoma. Nature Communications. 2019;10(1).

[59] Bodogai M, Chang CL, Wejksza K, Lai J, Merino M, Wersto RP, et al. Anti-CD20 antibody promotes cancer escape via enrichment of tumor-evoked regulatory B cells expressing low levels of CD20 and CD137L. Cancer Research. 2013;73(7):2127–2138.

[60] DiLillo DJ, Yanaba K, Tedder TF. B Cells Are Required for Optimal CD4 + and CD8 + T Cell Tumor Immunity: Therapeutic B Cell Depletion Enhances B16 Melanoma Growth in Mice. The Journal of Immunology. apr;(7):4006–4016.

[61] Chevrier S, Levine JH, Zanotelli VRT, Silina K, Schulz D, Bacac M, et al. An Immune Atlas of Clear Cell Renal Cell Carcinoma. Cell. may;(4):736–749.e18.

[62] Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, et al. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. Cell Reports. apr;(1):313–326.e5.

[63] Becht E, Giraldo NA, Dieu-Nosjean MC, Sautès-Fridman C, Fridman WH. Cancer immune contexture and immunotherapy. Current Opinion in Immunology. 2016;39:7–13.

[64] Heath JR, Ribas A, Mischel PS. Single-cell analysis tools for drug discovery and development. Nature Reviews Drug Discovery. 2016;15(3):204–216.

[65] Gong T, Szustakowski JD. DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013;29(8):1083–1085.

[66] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods. 2015;12(5):453.

[67] Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. nov;(7269):108–112.

[68] Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications. dec;(1):2612.

[69] Bhuva DD, Foroutan M, Xie Y, Lyu R, Cursons J, Davis MJ. Using singscore to predict mutation status in acute myeloid leukemia from transcriptomic signatures. F1000Research. oct;p. 776.

[70] Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. BMC Bioinformatics. dec;(1):404.

[71] Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology. jul;(7):773–782.

[72] Arthur D, Vassilvitskii S. K-means++: the advantages of careful seeding. In: In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms; 2007. .

[73] Granier C, Dariane C, Combe P, Verkarre V, Urien S, Badoual C, et al. Tim-3 Expression on Tumor-Infiltrating PD-1 + CD8 + T Cells Correlates with Poor Clinical Outcome in Renal Cell Carcinoma. Cancer Research. mar;(5):1075–1082.

[74] PD-1. Programmed Death-1 https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pd-1;.

[75] PDCD1 programmed cell death 1 https://www.ncbi.nlm.nih.gov/gene/5133;.

[76] IFNG. IFNG interferon gamma https://www.ncbi.nlm.nih.gov/gene/3458;.

[77] Tau G, Rothman P. Biologic functions of the IFN-γ receptors. Allergy: European Journal of Allergy and Clinical Immunology. 1999;54(12):1233–1251.

[78] Ni L, Lu J. Interferon gamma in cancer immunotherapy. Cancer Medicine. sep;(9):4509–4516.

[79] Chen DS, Mellman I. Oncology meets immunology: The cancer-immunity cycle. Immunity. 2013;39(1):1–10.

[80] Motzer RJ, Escudier B, McDermott DF, George S, Hammers HJ, Srinivas S, et al. Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. New England Journal of Medicine. nov;(19):1803–1813.

[81] Dine J, Gordon R, Shames Y, Kasler M, Barton-Burke M. Immune checkpoint inhibitors: An innovation in immunotherapy for the treatment and management of patients with cancer. Asia-Pacific Journal of Oncology Nursing;(2):127.

[82] Motzer RJ, Tannir NM, McDermott DF, Arén Frontera O, Melichar B, Choueiri TK, et al. Nivolumab plus Ipilimumab versus Sunitinib in Advanced Renal-Cell Carcinoma. New England Journal of Medicine. apr;(14):1277–1290.

[83] Gettinger SN, Horn L, Gandhi L, Spigel DR, Antonia SJ, Rizvi NA, et al. Overall survival and long-term safety of nivolumab (anti-programmed death 1 antibody, BMS-936558, ONO-4538) in patients with previously treated advanced non-small-cell lung cancer. Journal of Clinical Oncology. 2015;33(18):2004–2012.

[84] Koyama S, Akbay EA, Li YY, Herter-Sprie GS, Buczkowski KA, Richards WG, et al. Adaptive resistance to therapeutic PD-1 blockade is associated with upregulation of alternative immune checkpoints. Nature Communications. 2016;7:1–9.

[85] Hofmann L, Forschner A, Loquai C, Goldinger SM, Zimmer L, Ugurel S, et al. Cutaneous, gastrointestinal, hepatic, endocrine, and renal side-effects of anti-PD-1 therapy. European Journal of Cancer. 2016;60:190–209.

[86] Maleki Vareki S, Garrigós C, Duran I. Biomarkers of response to PD-1/PD-L1 inhibition. Critical Reviews in Oncology/Hematology. aug;p. 116–124.

[87] Aggen DH, Drake CG, Rini BI. Targeting PD-1 or PD-L1 in Metastatic Kidney Cancer: Combination Therapy in the First-Line Setting. Clinical Cancer Research. jan;.

[88] Chen H, Liakou CI, Kamat A, Pettaway C, Ward JF, Tang DN, et al. Anti-CTLA-4 therapy results in higher CD4 + ICOS hi T cell frequency and IFN-γ levels in both nonmalignant and malignant prostate tissues. Proceedings of the National Academy of Sciences. feb;(8):2729–2734.

[89] Dulos J, Carven GJ, van Boxtel SJ, Evers S, Driessen-Engels LJA, Hobo W, et al. PD-1 Blockade Augments Th1 and Th17 and Suppresses Th2 Responses in

Peripheral Blood From Patients With Prostate and Advanced Melanoma Cancer. Journal of Immunotherapy;(2):169–178.

[90] Peng W, Liu C, Xu C, Lou Y, Chen J, Yang Y, et al. PD-1 Blockade Enhances T-cell Migration to Tumors by Elevating IFN- Inducible Chemokines. Cancer Research. oct;(20):5209–5218.

[91] Higgs BW, Morehouse CA, Streicher K, Brohawn PZ, Pilataxi F, Gupta A, et al. Interferon Gamma Messenger RNA Signature in Tumor Biopsies Predicts Outcomes in Patients with Non–Small Cell Lung Carcinoma or Urothelial Cancer Treated with Durvalumab. Clinical Cancer Research. aug;(16):3857–3866.

[92] Sercan Ö, Hämmerling GJ, Arnold B, Schüler T. Cutting Edge: Innate Immune Cells Contribute to the IFN-γ-Dependent Regulation of Antigen-Specific CD8 + T Cell Homeostasis. The Journal of Immunology. jan;(2):735–739.

[93] Choueiri TK, Motzer RJ. Systemic Therapy for Metastatic Renal-Cell Carcinoma. New England Journal of Medicine. jan;(4):354–366.

[94] Al-Husein B, Abdalla M, Trepte M, DeRemer DL, Somanath PR. Antiangiogenic Therapy for Cancer: An Update. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy. dec;(12):1095–1111.

[95] Chen Y, Li C, Xie H, Fan Y, Yang Z, Ma J, et al. Infiltrating mast cells promote renal cell carcinoma angiogenesis by modulating PI3K$\rightarrow$AKT$\rightarrow$GSK3$\beta\rightarrow$AM signaling. Oncogene. may;(20):2879–2888.

[96] Fu H, Zhu Y, Wang Y, Liu Z, Zhang J, Wang Z, et al. Tumor Infiltrating Mast Cells (TIMs) Confers a Marked Survival Advantage in Nonmetastatic Clear-Cell Renal Cell Carcinoma. Annals of Surgical Oncology. may;(5):1435–1442.

[97] Cherdantseva TM, Bobrov IP, Avdalyan AM, Klimachev VV, Kazartsev AV, Kryuchkova NG, et al. Mast Cells in Renal Cancer: Clinical Morphological

Correlations and Prognosis. Bulletin of Experimental Biology and Medicine. oct;(6):801–804.

BIOGRAPHICAL STATEMENT

Sumeyye Su was born in Schaffhausen, Switzerland, in 1989. She received her Bachelor of Science degrees in Mathematics and Economics from Yildiz Technical University in Istanbul Turkey in 2012 and 2014, respectively. She joined the Ph.D. program in Mathematics at the University of Texas at Arlington in the Fall of 2017.