UNIVERSITY OF CALIFORNIA,
IRVINE

Support Vector Machine for Kidney Cancer Classification

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Biomedical Engineering

by

Liyu Cao

Thesis Committee:
Associate Professor James Brody, Chair
Associate Professor Michelle Digman
Professor Frithjof Kruggel

2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Professor James Brody, who has the attitude and the substance of a genius: he continually and convincingly conveyed a spirit of adventure in regard to research and scholarship, and an excitement in regard to teaching. Without his guidance and persistent help this thesis would not have been possible.

I would like to thank my committee members, Professor Michelle Digman and Professor Frithjof Kruggel , whose knowledge demonstrated to me that target on biology and statistics should always be appreciated and is really helpful in my master thesis.

ABSTRACT OF THE THESIS

Machine Learning and Cancer Classification

By

Liyu Cao

Master of Science in Biomedical Engineering

University of California, Irvine, 2018

Professor James Brody, Chair

Renal cancer is the 12th leading cause of cancer death, accounting for 2.4 percent of all cancers in the United States. Two of most common types of kidney tumors are clear cell and papillary carcinoma. In this study, data of Kidney Renal Clear Cell Carcinoma (KIRC) and Kidney renal papillary cell carcinoma (KIRP) projects from *the Cancer Genome Altas* database were analyzed. These two data were combined and most part of it was used as a training set for candidate biomarker identification and 50 of the dataset was used to test the accuracy of classification model based on the identified biomarker. As a result, the accuracy of classification was 96.0% when the model based on top 5 biomarkers was trained by gene expression data. For DNA methylation level, the performance of model based on top biomarkers was the test with an overall accuracy of 100.0%. In conclusion, some informative biomarkers that could accurately discriminate KIRC and KIRP were identified in this study.

INTRODUCTION

In the United States, renal cancer is the 12th leading cause of cancer death, accounting for 2.4 percent of all cancers (Siegel, Miller, & Jemal, 2016). It is estimated that 14240 Americans died from renal cancer in 2016, corresponding to about 39 deaths per day (Siegel et al., 2016). In the United States, 62700 new cases of renal cancer are expected to be diagnosed in 2016 (Siegel et al., 2016). Renal cancer is the 7th common cancer expected to occur in men, accounting for 5 percent among all cancers, 10th common cancer for women, accounting for 3 percent (Siegel et al., 2016). Renal cell carcinomas (RCCs) arise from the renal epithelium and accounting for about 85 percent of renal cancers (Cohen & McGovern, 2005). RCC consists of several histological cell types. Both clear cell and papillary RCC are thought to arise from the epithelium of proximal tubule (Cairns, 2011). Chromophore RCC, oncocytoma, and collecting duct RCC are believed to arise from the distal nephron, probably from the epithelium of the collecting tubules (Cairns, 2011). The most common histological type is clear cell carcinoma, also called conventional RCC, which represents 75– 80% of RCC. Papillary (10–15%), chromophobe (5%) and other more rare forms such as collecting duct carcinoma (< 1%) comprise the remainder (Cairns, 2011). In this study, I focused on clear cell and papillary cell carcinoma.

Recent years, microarray technology becomes very popular and allows researchers to measure the quantity of mRNA of thousands of genes simultaneously (Brazma et al., 2001; Schena, Shalon, Davis, & Brown, 1995; Smyth, 2004). For example, like the layout of Fig.1, a collection of gene expression data can be viewed abstractly as a table with rows representing genes, columns representing various samples and each position in the table

describing the measurement for a particular gene in a particular sample (Brazma et al., 2001). The microarray is a powerful device, many studies on analyzing DNA microarray data reveal unknown biological processes (Bigger, Brasky, & Lanford, 2001; Hihara, Kamei, Kanehisa, Kaplan, & Ikeuchi, 2001; McCormick et al., 2001). In most cases, a microarray data includes thousands of genes expression level, and a limited number of samples were analyzed in each study. Among all the genes, many are irrelevant, insignificant or redundant to the discriminant problem under investigation (Zhou & Tuck, 2007). Therefore, a method to identify the informative genes from other redundant genes is of fundamental and practical importance to the researches of classification problems, such as classification of cancer subtypes.

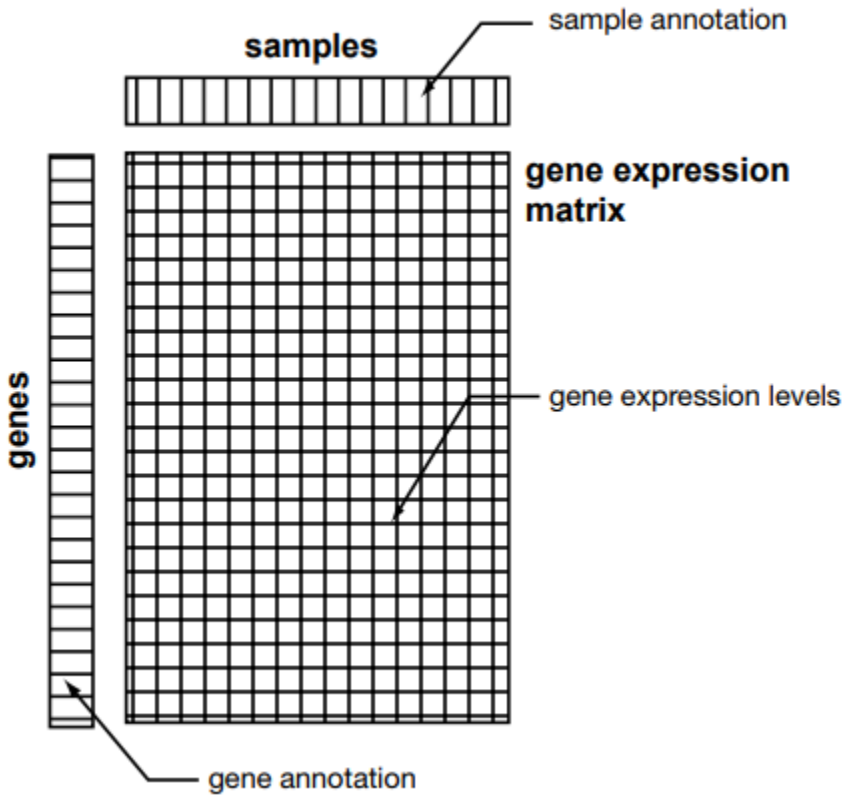

Fig.1

The well-known support vector machines (SVMs) have been widely used for solving classification problems (Eskin, Weston, Noble, & Leslie, 2003; Leslie, Eskin, & Noble, 2002). In research of Guyon (Guyon, Weston, Barnhill, & Vapnik, 2002), linear SVMs are used in a recursive feature elimination procedure, and the algorithm is therefore referred as support vector machine-recursive feature elimination (SVM-RFE). The SVM-RFE (Guyon et al., 2002) was proposed in to conduct gene selection for cancer classification.

Sometimes, it is difficult for doctors to diagnose the origin of the tumor in patients when cancers metastasized or spread to other organs or throughout the body. Therefore, the most effective treatment for the specific types of cancers cannot be applied to patients. For improving the effectiveness of treatment of cancer patients, my aim of study was to narrow down the biomarkers to classify different types of cancers for doctors to diagnose. In this study, microarray data of Kidney Renal Clear Cell Carcinoma (KIRC) and Kidney Renal Papillary Cell Carcinoma (KIRP) projects from The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) were analyzed. This study was comprised of two sections. The first section was to identify novel gene biomarkers for the classification of KIRC and KIRP's left and right subtype which stands for the tumor in left and right kidney. In the second section, the gene biomarkers for the classification of KIRC and KIRP were identified. For solving the discriminant problems, I performed SVM-RFE to find informative genes. Then the selected genes after SVM-RFE were validated and tested with samples in test data set from TCGA.

Materials and Methods

Data Collection

The sequence-based gene expression and DNA methylation data of KIRC and KIRP were generated on the Illumina HiSeq 2000 RNA sequencing platform (https://support.illumina.com/sequencing/sequencing_instruments/hiseq_2000.html) and Illumina HumanMethylation450 platform (https://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit.html) respectively. And the data sets were retrieved from cBioPortal (http://www.cbioportal.org/) for Cancer Genomics.

The pathology reports of confirmed KIRC and KIRP cases were downloaded from TCGA. Each pathology report was looked through to find the tumor site of each case. Each case of KIRC and KIRP was then labeled as left and right subtype.

Z-score Calculation of Two Populations

Assume there are two normally distributed but independent populations, the difference level between the population means can be estimated by calculating z- score:

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) - \left(\frac{\sigma_2^2}{n_2}\right)}}$$

Support Vector Machine-Recursive Feature Elimination

In this paper, I limit myself to two-class SVM classifiers which means it can be applied to classify sample into two classes. For a given training set of a number of features {x_1, x_2, x_3, … x_k, … x_l} with known labels { y_1, y_2, y_3, … y_k, … y_l}, $y_k \in \{-1, +1\}$. Here x is a vector of n components, y is the label of it. The data of training set are used to establish a decision function $D(x)$ (Guyon, Weston, Barnhill, & Vapnik, 2002):

$$D(x) > 0 \Rightarrow x \in class(+)$$

$$D(x) < 0 \Rightarrow x \in class(-)$$

$$D(x) = 0\,, decison\ boundary.$$

The decision functions are weighted sums of all features plus bias:

$$D(x) = w^T x + b$$

But data overfitting arises when the number n of the features is large and the number of cases of the training set is small. To overcome the overfitting problem, the number of features should be eliminated to reduce the dimensional complexity. Support Vector Machine-Recursive Feature Elimination (SVM-RFE) is an elegant method to abandon redundant features and find significant features. With the SVM-RFE, a rank list of features is generated and the one feature with the lowest rank is eliminated in each step of the recursive feature elimination process.

Leave-one-out, Cross-Validation

In k-fold cross-validation, sometimes called rotation estimation, the dataset D is randomly split into k mutually exclusive subsets (the folds) $D_1; D_2; \ldots; D_k$ of approximately equal size (Kohavi, 1995). The model is trained and tested k times; each time $t \in \{1,2,\cdots,k\}$, it is trained on all subsets except D_t, and then tested on D_t. The cross-validation estimate of

accuracy is the overall number of correct classifications, divided by the number of instances in the dataset (Kohavi, 1995).

Gene Signature Identification and Performance Assessment

Gene expression and DNA methylation data were analyzed using R software and e1071 package from The Comprehensive R Archive Network (CRAN). To identify gene signature, I applied support vector machine-recursive feature elimination (SVM-RFE) algorithm for feature selection and using the selected features to establish SVM classifier with the training set. The stability of the model was validated by 10-fold cross-validation. The predictive error of the SVM classifier was then generated with the test set.

Results

The z-score distribution of difference between left and right subtype of KIRC and KIRP in gene expression and DNA methylation

For looking through the difference between KIRC and KIRP cases in gene expression and DNA methylation, the gene expression data and DNA methylation data of KIRC and KIRP from TCGA was queried.

Totally 60483 gene expression and 397665 gene methylation data were extracted and calculated to z-score for a quick look for the difference between left and right subtype.

For the KIRC project, 503 cases (241 left, 262 right) have been studied in my research.

Figure 2 shows the z-score distribution of differences between left and right subtypes in gene expression level. The z-score of differences in gene expression is small. There is no such a gene whose expression level has the difference z-score larger than 4. Since larger the

absolute value of z-score of difference, larger the difference of gene expression or DNA methylation between left and right subtypes, it indicates the difference between left and right subtypes of KIRC project is not statistically significant. In the same way, from the Figure 3, the differences between left and right subtypes in DNA methylation level is not significant either.

For the KIRP project, there are 287 cases (159 left, 128right). The distribution of z-score in Figure 4 and Figure 5 shows the differences between left and right subtypes of KIRP project. The distributions are similar to the distribution of KIRC project which I discussed before. In other words, the z-sore is not big enough to help us find the informative gene to identify left and right subtype.

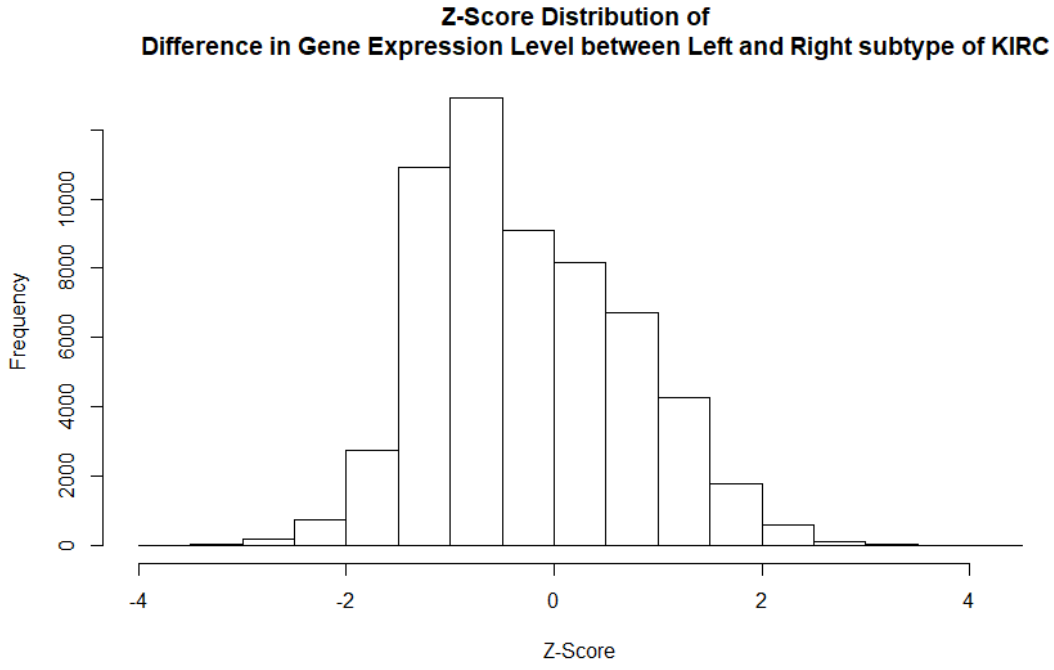

Fig.2

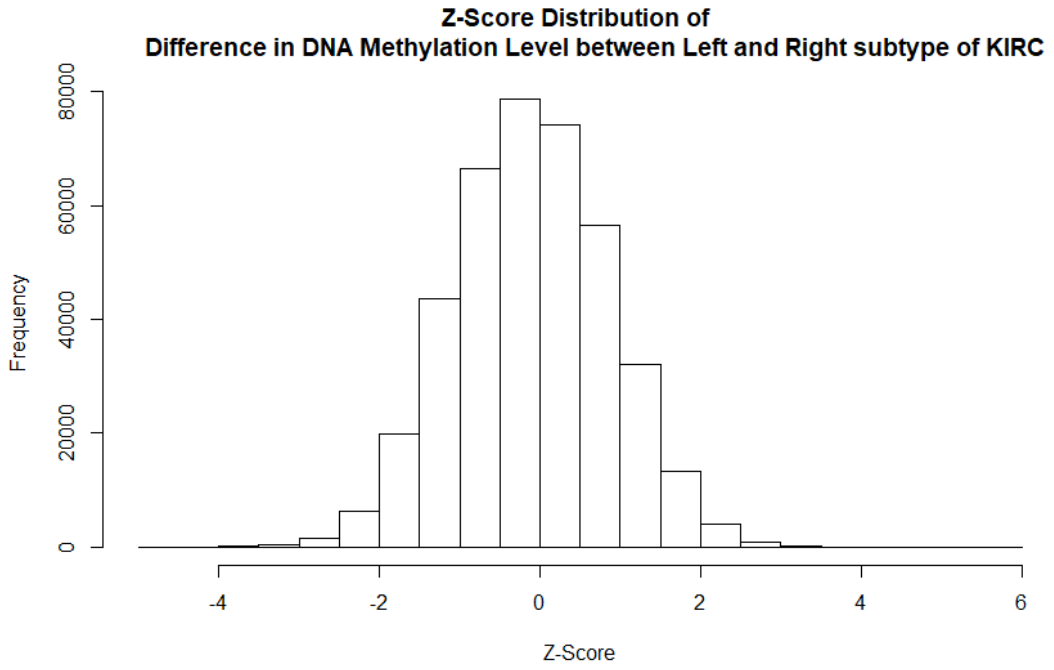
Z-Score Distribution of
Difference in DNA Methylation Level between Left and Right subtype of KIRC
Frequency
0
20000
40000
60000
80000
-4
-2
0
2
4
6
Z-Score

Fig.3


Z-Score Distribution of
Difference in Gene Expression Level between Left and Right subtype of KIRP
Frequency
0
2000
4000
6000
8000
10000
-4
-2
0
2
4
Z-Score

Fig.4

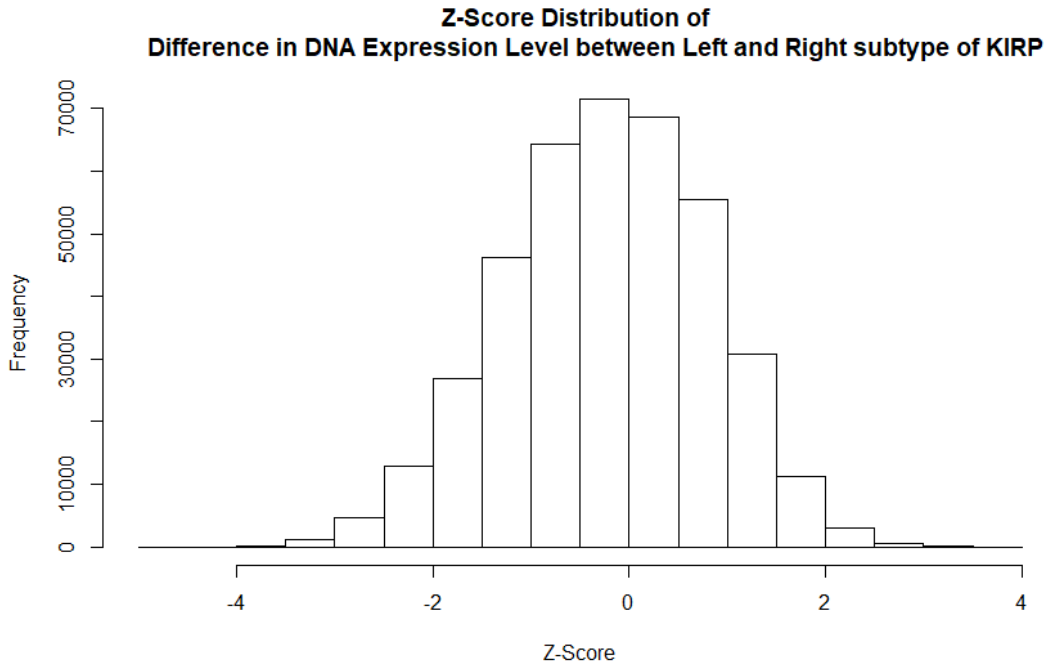

Fig.5

Validation and test of the SVM classifier of left and right subtype of KIRC in gene expression and DNA methylation

Since there is no way to find the informative biomarkers to classify kidney cancer from calculating the z-score of differences, I decided to utilize a new method to tackle this discrimination problem. I performed SVM-RFE algorithm to find the most informative biomarkers. As the output of SVM-RFE algorithm, a ranked list of all features is generated. The top 1 means the most informative feature.

The most informative biomarkers were selected to set up a SVM classifier trying to discriminate left and right subtypes of kidney cancer. The stability of the model was then validated by 10-fold cross-validation. The predictive error of the SVM classifier was then generated with the test set.

The Figure 6 shows the 10-fold validation error of the SVM classifier in gene expression level by choosing a different number of top features. The validation error of this classifier is near 0.5 no matter how many top features are used in the classifier. Then I performed test procedure, the test set of 50 cases which have not been used in modeling and validation is used to generate the predictive error. The results are described in Figure 7. The test error is approximate 0.5 which means the classifier is random and useless.

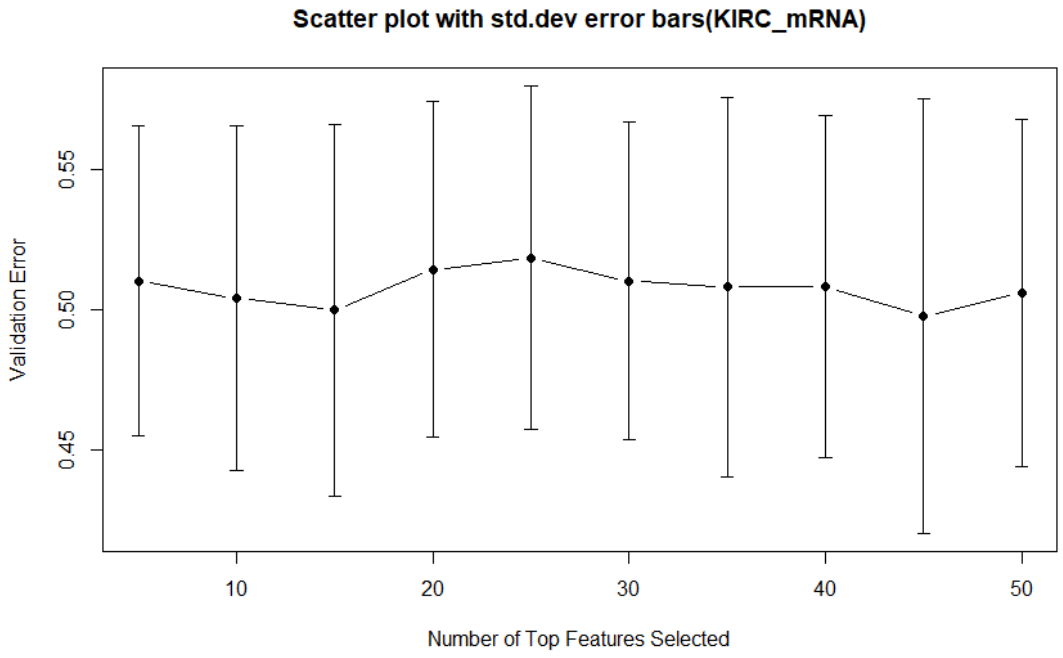

Fig.6

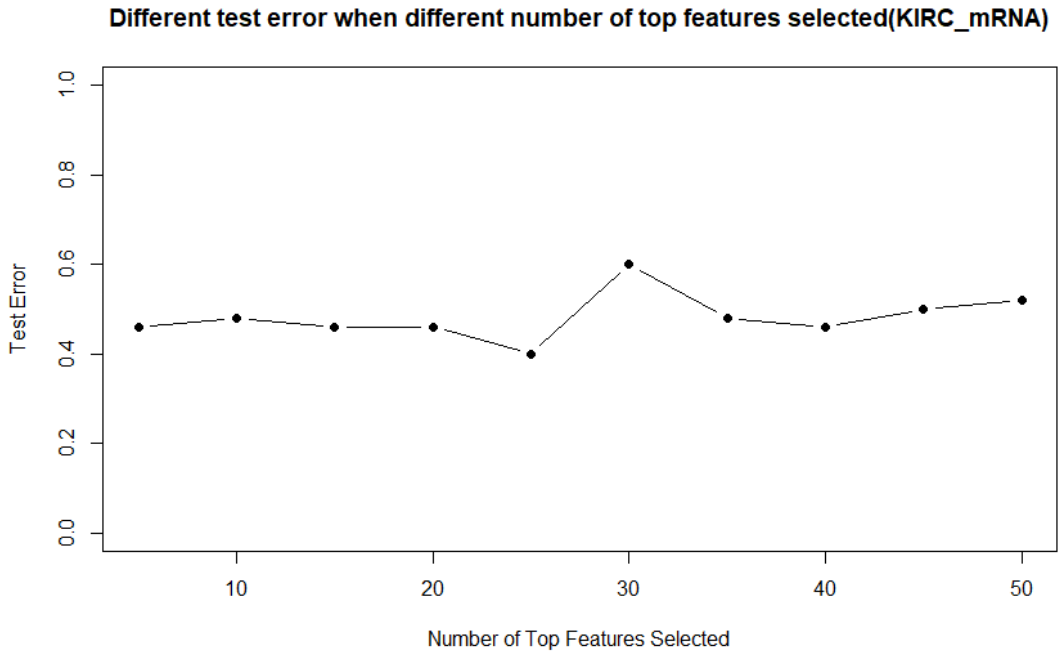

Fig.7

Next, I trying to find the biomarker by using DNA methylation data. In the same procedure, I performed SVM-RFE algorithm and set up SVM classifier with DNA methylation data. Figure 8 and Figure 9 are the results of validation and test of SVM classifier. The error is also approximate 0.5.

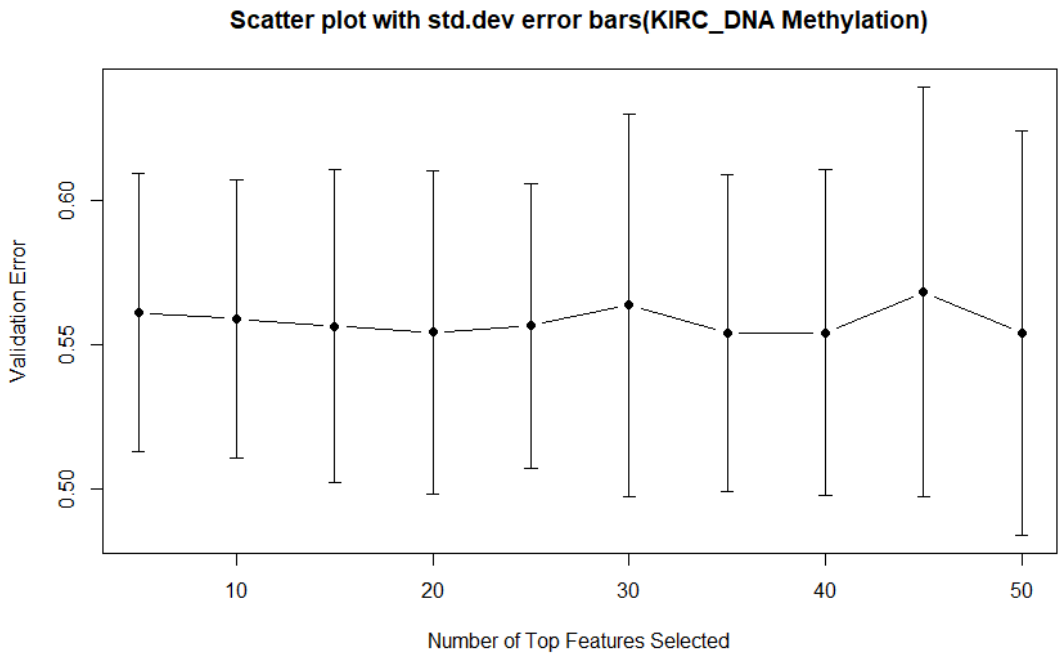
Scatter plot with std.dev error bars(KIRC_DNA Methylation)
Validation Error
0.50
0.55
0.60
10
20
30
40
50
Number of Top Features Selected

Fig.8

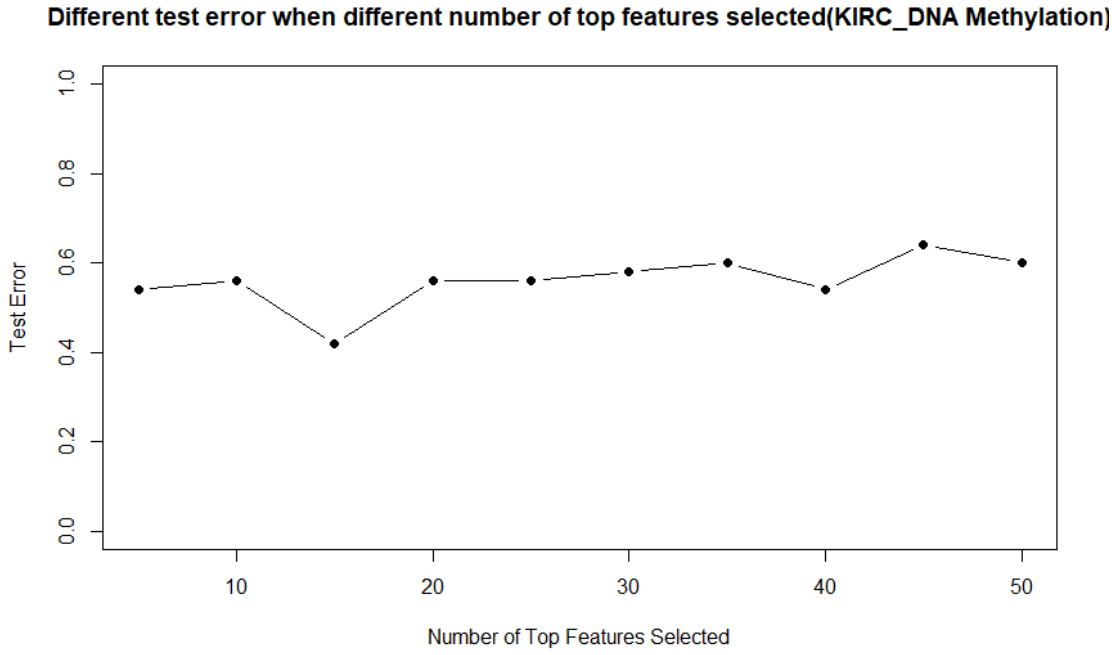
Different test error when different number of top features selected(KIRC_DNA Methylation)
Test Error
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
Number of Top Features Selected

Fig.9

Validation and test of the SVM classifier of left and right subtype of KIRP in gene expression and DNA methylation

After KIRC project, I turned to study the KIRP project. The gene expression and DNA methylation data were used in my research. The classifiers are unsatisfactory since the validation error and test error is large (Fig 10,11,12,13).

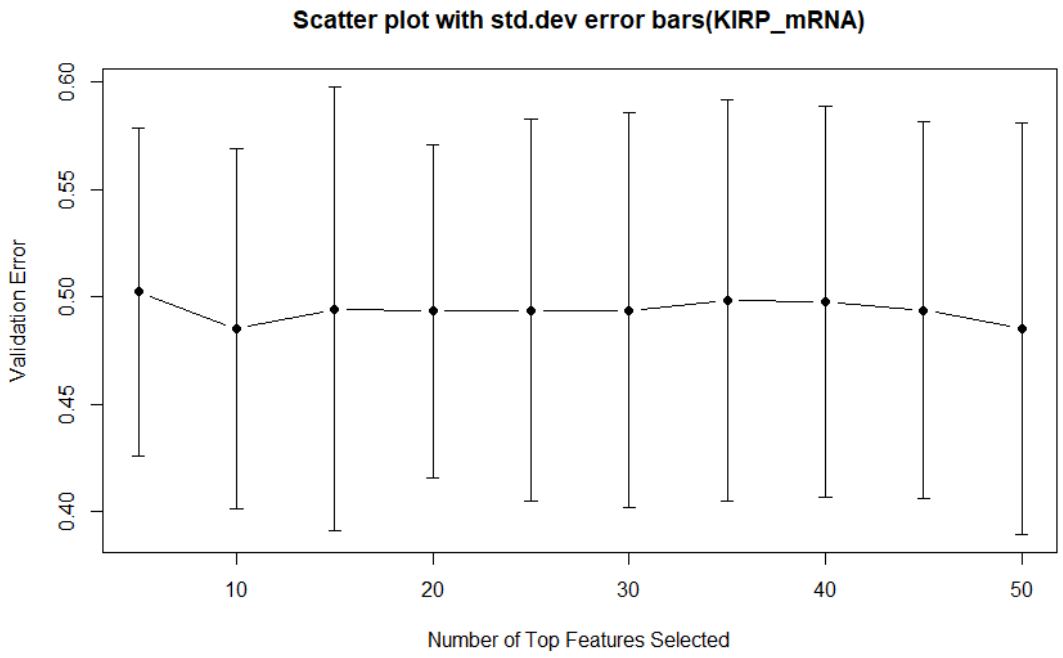

Fig.10

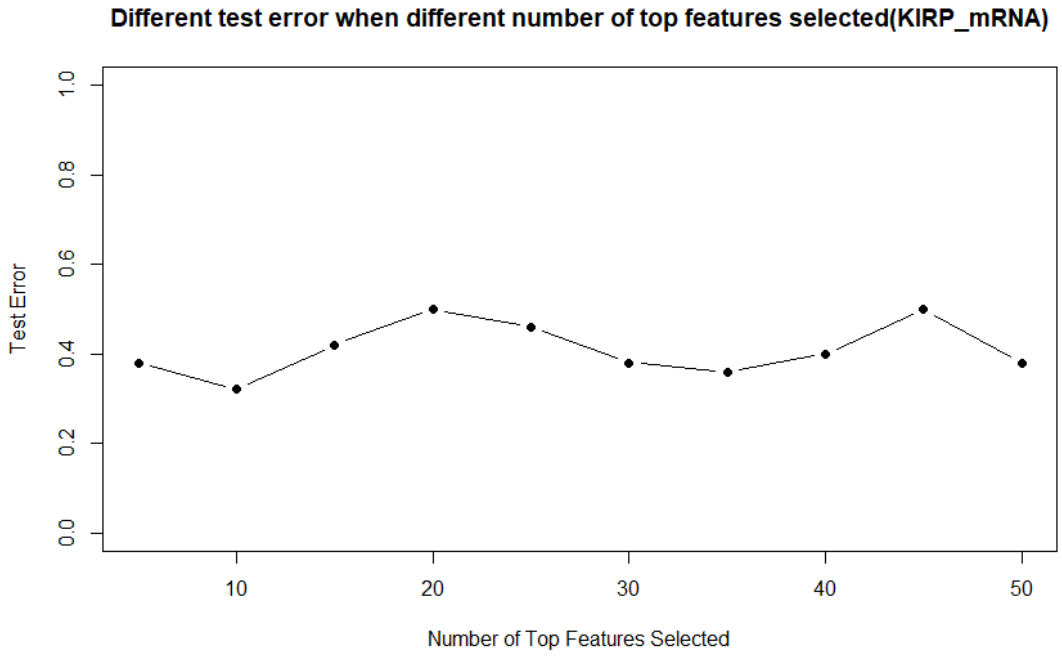
Different test error when different number of top features selected(KIRP_mRNA)
Test Error
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
Number of Top Features Selected

Fig.11

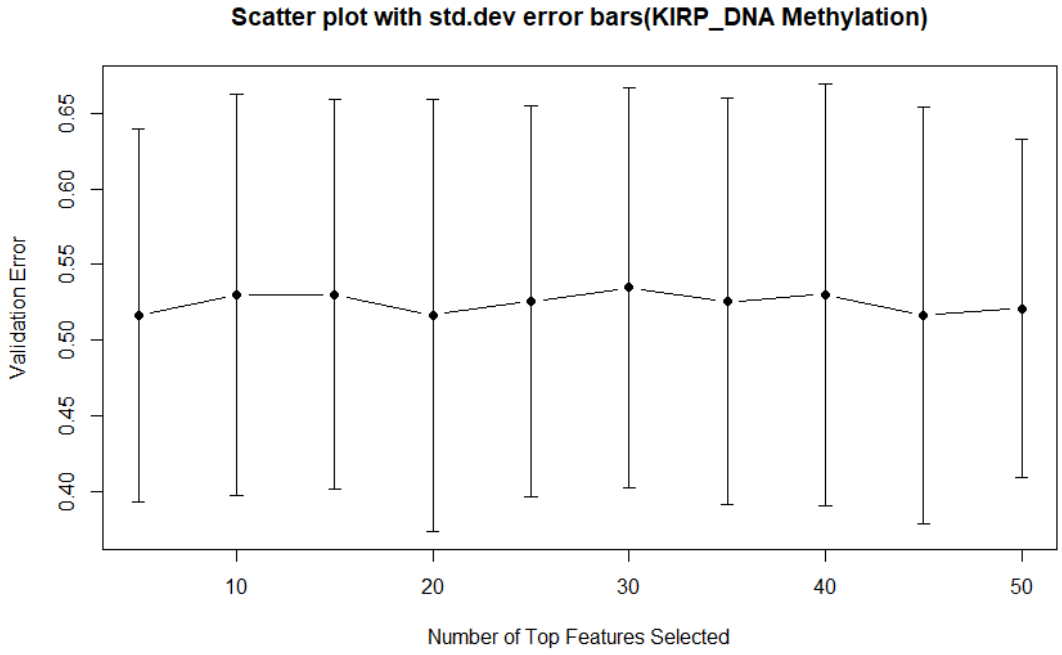
Scatter plot with std.dev error bars(KIRP_DNA Methylation)
Validation Error
0.40
0.45
0.50
0.55
0.60
0.65
10
20
30
40
50
Number of Top Features Selected

Fig.12

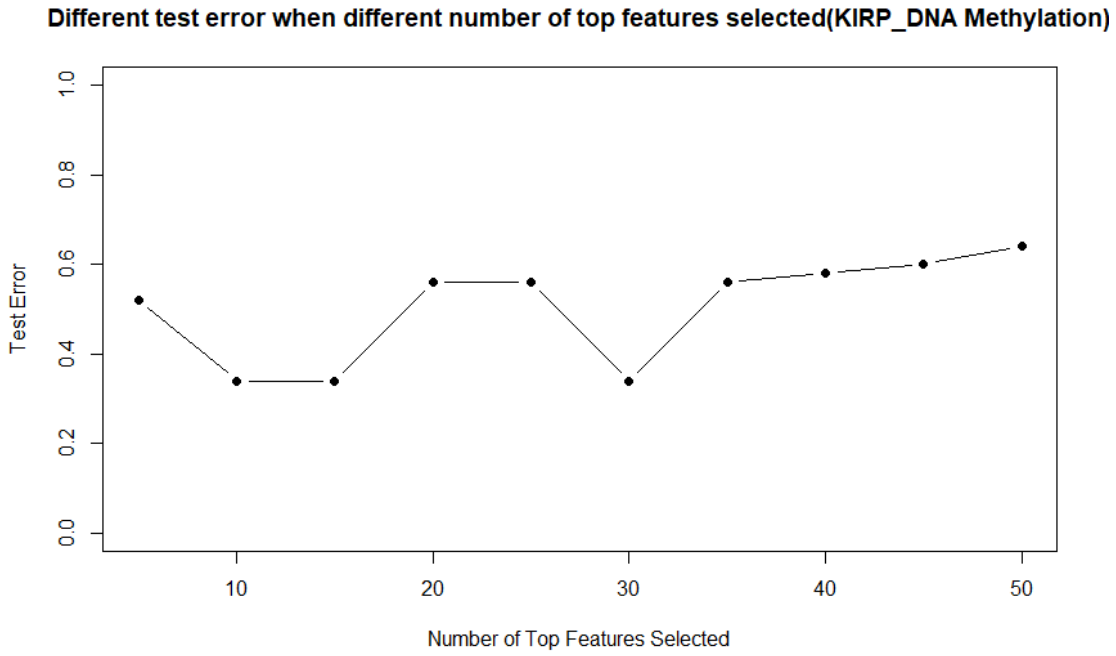

Fig.13

Validation and test of the SVM classifier of KIRC and KIRP in gene expression and DNA methylation

In this section, I implemented SVM-RFE to select informative features to identify Kidney Renal Clear Cell Carcinoma and Kidney Renal Papillary Cell Carcinoma. The rank lists of top 10 features in gene expression level are included in Table 1. Higher the rank, more important the expression data of this gene to identify Kidney Renal Clear Cell Carcinoma and Kidney Renal Papillary Cell Carcinoma. To make the rank reliable, leave-one-out 10-fold validation was applied to obtain 10 rank values for each gene. Based on the results, the average rank was calculated and listed as AvgRank in Table 1. The Top 10 features of methylation level of genes are listed in Table 2.

Table 1. Top 10 features selected by SVM-RFE algorithm for gene expression

Rank	FeatureName	AvgRank
1	NDUFA4L2	12.9
2	TMEM104	66.3
3	REEP3	88.7
4	APLN	89.8
5	HRC	90.3
6	MPI	91.7
7	MLLT1	95.7
8	MYH14	114.7
9	TNF	125.3
10	PCNXL3	136.4

Table 2. Top 10 features selected by SVM-RFE algorithm for DNA methylation

Rank	FeatureName	AvgRank
1	ZNF202	1.5
2	TMEM198	3.1
3	NDUFS7	13.5
4	C1QBP	15
5	MRPL45	15.5
6	PSMA2	16.8
7	C12orf76	17.3
8	LRWD1	19.5

9	C12orf62	19.6
10	ZNF778	21.2

Based on the results of SVM-RFE, the SVM classifiers with a different number of top features were built by a training set of 740 cases. And 10-fold cross-validation was performed with a validation set to obtain the validation error. At last, the SVM classifier was tested by a test set of 50 cases which is the data never used in modeling and validation steps.

For the expression level of genes, as shown in Figure 13 and Figure 14, the validation error is about 3.5% and the test error ranges from 4% to 6% by choosing a different number of top features from 5 to 50. In the same way, when the methylation level of genes was used as factors to classification, the error is even smaller. For the validation step, the error is zero always. And for the test step, the error ranges from 0 to 2%.

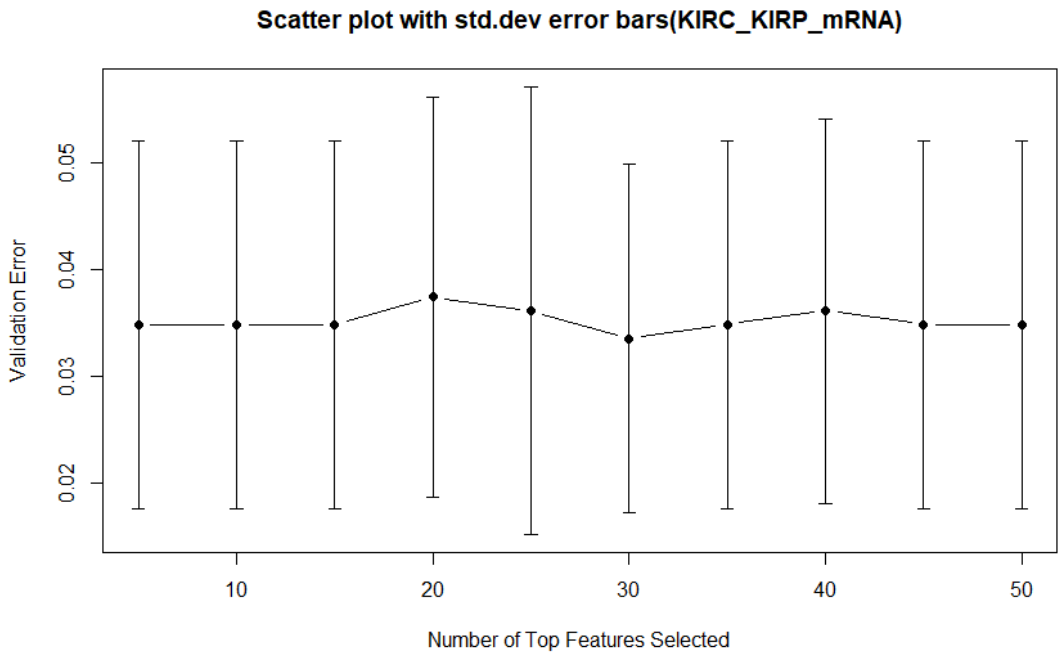

Fig.14

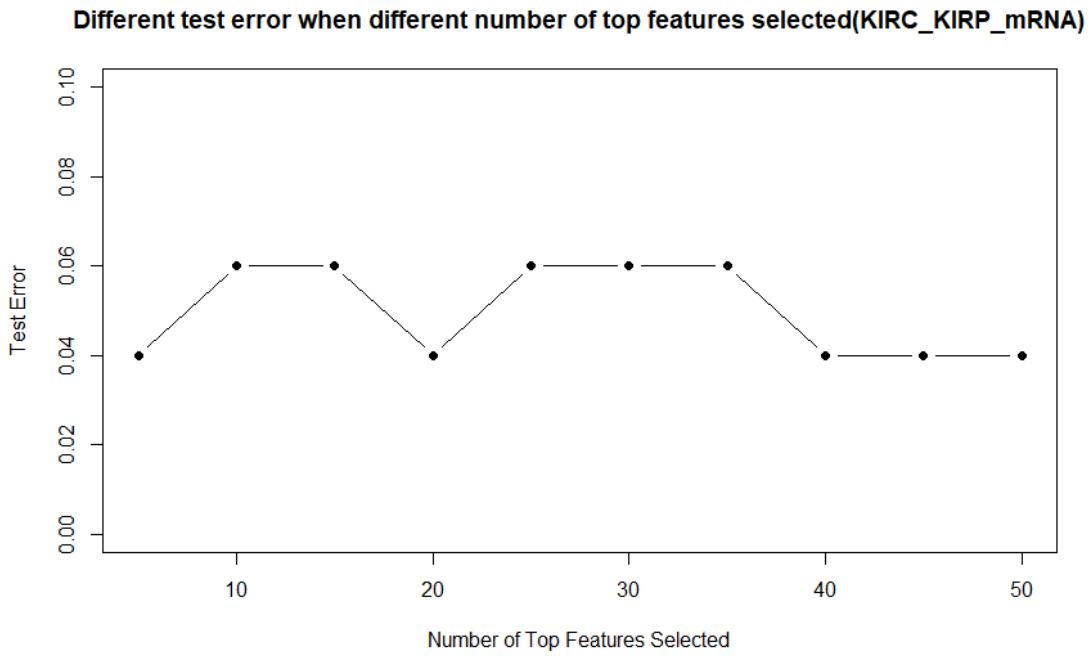

Fig.15

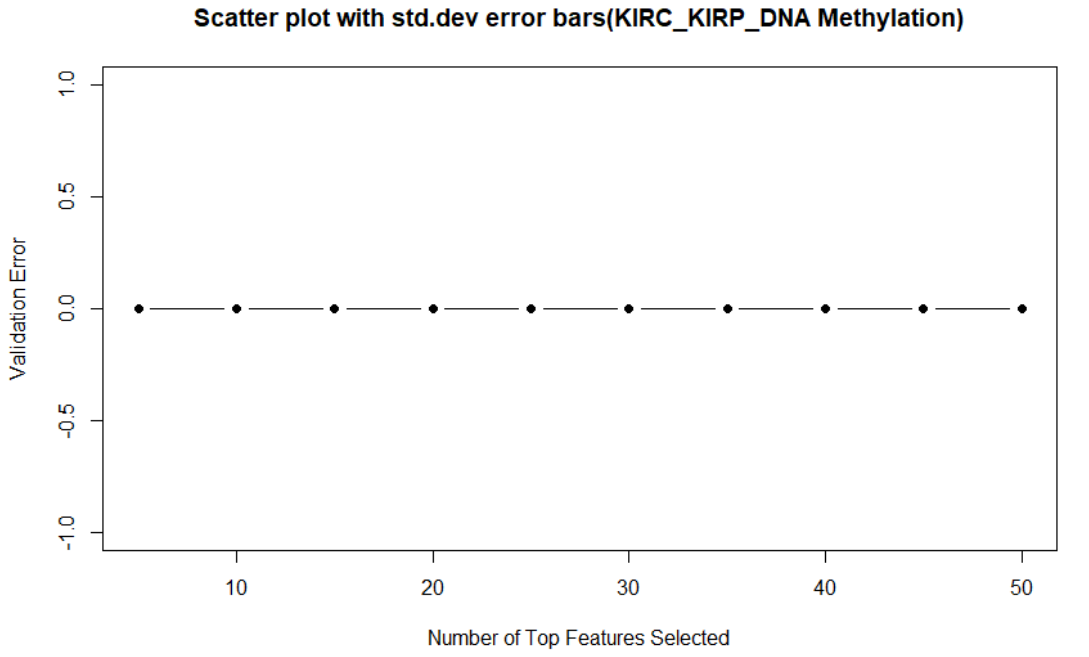

Fig.16

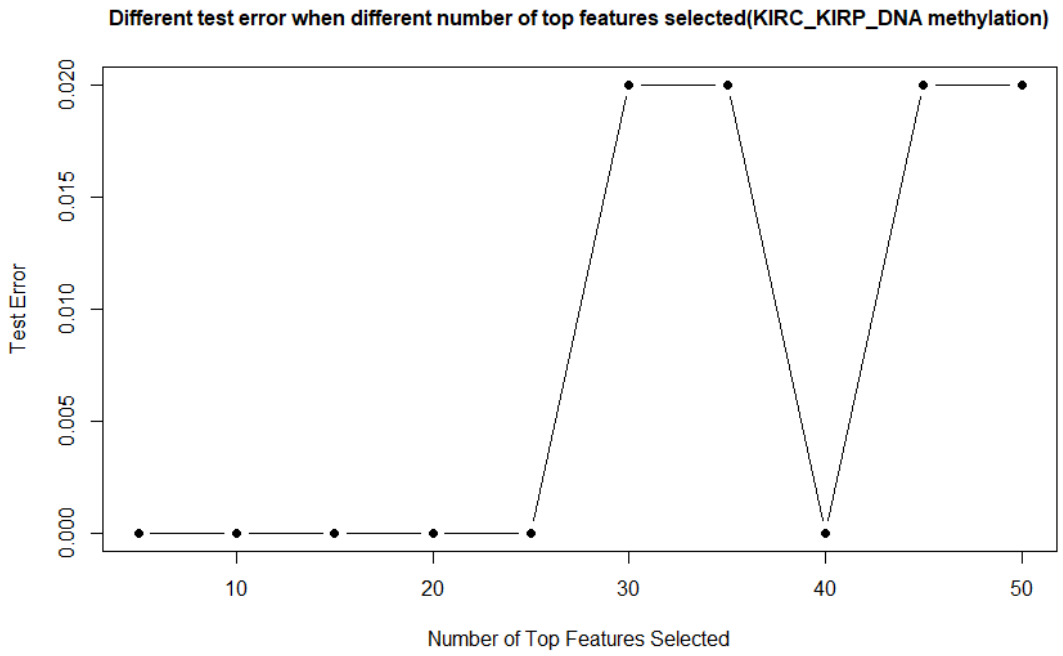

Fig.17

Discussion

Due to the successful development of high-throughput microarray and next-generation sequencing technologies, cancer genomic data like TCGA were utilized in my study. I firstly tried to find the biomarker to the classification of renal carcinoma of left and right subtypes. But the results showed there is no big difference between left and right renal carcinoma.

And in this study, the informative biomarkers were identified for the accurate and robust classification of KIRC and KIRP. For the gene expression level, the validation error is about 3.5% and the test error ranges from 4% to 6% by choosing a different number of top features from 5 to 50. And for the gene methylation level, the validation error is zero

always. For the test step, the error ranges from 0 to 2%. Furthermore, the reason why the test error goes up when the number of feature increases at 30, 35, 45 and 50 needs to be studied later.

Additionally, the biological features of top 5 genes were collected. Recently, a study(L. Liu et al., 2016) has shown NDUFA4L2 expression, which is top 1 genes in my SVM classifier based on gene expression level, was an independent prognostic factor for ccRCC patients. The results of the article showed that NDUFA4L2 protein expression was found to be higher in clear cell renal cell carcinoma (ccRCC) tissues 81.4% (70/86) than in normal tissues 26.7% (23/86) ($p = 0.021$)(L. Liu et al., 2016). The average level of NDUFA4L2 mRNA expression was found to be 122.23 ± 6.018 and 21.34 ± 1.036 in ccRCC tissue and adjacent normal tissue ($p < 0.001$) (L. Liu et al., 2016).

The official name of TMEM104 is transmembrane protein 104. The function of this gene is unclear now.

REEP3 ensure endoplasmic reticulum clearance from metaphase chromatin and making sure correct progression through mitosis and proper nuclear envelope architecture (Schlaitz et al., 2013).

According to RefSeq (https://www.ncbi.nlm.nih.gov/refseq/), APLN encodes a peptide that functions as an endogenous ligand for the G-protein coupled apelin receptor. The encoded preproprotein is proteolytically processed into biologically active C-terminal peptide

fragments. These peptide fragments activate different tissue-specific signaling pathways that regulate diverse biological functions including fluid homeostasis, cardiovascular function, and insulin secretion. This protein also functions as a coreceptor for the human immunodeficiency virus 1.

HRC encodes a luminal sarcoplasmic reticulum protein identified by its ability to bind low-density lipoprotein with high affinity. The protein interacts with the cytoplasmic domain of triadin, the main transmembrane protein of the junctional sarcoplasmic reticulum (SR) of skeletal muscle. The protein functions in the regulation of releasable calcium into the SR. And HRC promotes growth of hepatocellular carcinoma in vitro and in vivo (J. Liu et al., 2015).

Furthermore, the biological features of top 5 genes of methylation level were searched. The ZNF202 gene product is a transcriptional repressor that binds to elements found predominantly in genes that participate in lipid metabolism (Wagner et al., 2000).

In mammalian cells, TMEM198 is required for Wnt signaling and casein kinase 1-induced LRP6 phosphorylation (Liang et al., 2011). Wnt signaling pathway is associated with cancer, and defects in the Wnt pathway can lead to cancer (Polakis, 2000). And this protein has also been identified as the p32 subunit of pre-mRNA splicing factor SF2, as well as a hyaluronic acid-binding protein.

According to RefSeq, the NDUFS7 gene encodes a protein that is a subunit of one of the complexes that forms the mitochondrial respiratory chain. This protein is one of over 40 subunits found in complex I, the nicotinamide adenine dinucleotide (NADH): ubiquinone oxidoreductase. This complex functions in the transfer of electrons from NADH to the respiratory chain, and ubiquinone is believed to be the immediate electron acceptor for the enzyme. Mutations in this gene cause Leigh syndrome due to mitochondrial complex I deficiency, a severe neurological disorder that results in bilaterally symmetrical necrotic lesions in subcortical brain regions.

Interactome analysis reveals that C1QBP (complement component 1, q subcomponent binding protein) is associated with cancer cell chemotaxis and metastasis (Zhang et al., 2013). Early studies suggested that the androgen receptor (AR) might play important roles to promote the renal cell carcinoma (RCC) progression (Yue et al., 2017). And C1QBP Regulates YBX1 to Suppress the Androgen Receptor (AR)-Enhanced RCC Cell Invasion (Yue et al., 2017).

The official name of MRPL45 is mitochondrial ribosomal protein L45 which helps in protein synthesis within the mitochondrion.

In summary, renal carcinoma subtypes are characterized by different genetic mutations and methylation variations, therefore presenting different gene expression and methylation profiles. In the list of informative genes, NDUFA4L2 is an independent prognostic factor for ccRCC patients (L. Liu et al., 2016). REEP3 is included in cell mitosis.

TMEM198 is required for Wnt signaling pathway which is associated with cancer (Liang et al., 2011; Polakis, 2000). And a study suggested that C1QBP Regulates YBX1 to Suppress the Androgen Receptor (AR)-Enhanced RCC Cell Invasion (Yue et al., 2017). NDUFS7 and MRPL45 are important in function of mitochondrion. Recently, mitochondria are now new promising target for cancer therapy (Fulda, Galluzzi, & Kroemer, 2010).

In conclusion, I developed gene-based biomarkers for the classification of KIRC and KIRP. Hopefully, my study can help doctors to diagnose the subtype of renal carcinoma of patients. And the further study will be collecting more data of patients to validate and test my model and apply other machine learning algorithm to build better models.

References

Afshari, C. A., Nuwaysir, E. F., & Barrett, J. C. (1999). Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. *Cancer Research*, *59*(19), 4759–4760.

Bigger, C. B., Brasky, K. M., & Lanford, R. E. (2001). DNA microarray analysis of chimpanzee liver during acute resolving hepatitis C virus infection. *Journal of Virology*, *75*(15),

7059–7066.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., … Causton, H. C. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, *29*(4), 365–371.

Cairns, P. (2011). Renal cell carcinoma. *Cancer Biomarkers*, *9*(1–6), 461–473.

Cohen, H. T., & McGovern, F. J. (2005). Renal-cell carcinoma. *New England Journal of Medicine*, *353*(23), 2477–2490.

Eskin, E., Weston, J., Noble, W. S., & Leslie, C. S. (2003). Mismatch string kernels for SVM protein classification. In *Advances in neural information processing systems* (pp. 1441–1448).

Fulda, S., Galluzzi, L., & Kroemer, G. (2010). Targeting mitochondria for cancer therapy. *Nature Reviews Drug Discovery*, *9*(6), 447–464.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389–422.

Hihara, Y., Kamei, A., Kanehisa, M., Kaplan, A., & Ikeuchi, M. (2001). DNA microarray analysis of cyanobacterial gene expression during acclimation to high light. *The Plant Cell*, *13*(4), 793–806.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145). Stanford, CA.

Leslie, C. S., Eskin, E., & Noble, W. S. (2002). The spectrum kernel: A string kernel for SVM protein classification. In *Pacific Symposium on Biocomputing* (Vol. 7, pp. 566–575). Hawaii, USA.

Liang, J., Fu, Y., Cruciat, C.-M., Jia, S., Wang, Y., Tong, Z., … Wu, W. (2011). Transmembrane

Protein 198 Promotes LRP6 Phosphorylation and Wnt Signaling Activation. *Molecular and Cellular Biology*, *31*(13), 2577–2590. https://doi.org/10.1128/MCB.05103-11

Liu, J., Han, P., Li, M., Yan, W., Liu, J., He, J., … Tian, D. (2015). Histidine-rich calcium binding protein promotes growth of hepatocellular carcinoma in vitro and in vivo. *Cancer Science*, *106*(10), 1288–1295. https://doi.org/10.1111/cas.12743

Liu, L., Lan, G., Peng, L., Xie, X., Peng, F., Yu, S., … Tang, X. (2016). NDUFA4L2 expression predicts poor prognosis in clear cell renal cell carcinoma patients. *Renal Failure*, *38*(8), 1199–1205. https://doi.org/10.1080/0886022X.2016.1208517

McCormick, S. M., Eskin, S. G., McIntire, L. V., Teng, C. L., Lu, C.-M., Russell, C. G., & Chittur, K. K. (2001). DNA microarray reveals changes in gene expression of shear stressed human umbilical vein endothelial cells. *Proceedings of the National Academy of Sciences*, *98*(16), 8955–8960.

Polakis, P. (2000). Wnt signaling and cancer. *Genes & Development*, *14*(15), 1837–1851. https://doi.org/10.1101/gad.14.15.1837

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270,* 467–470

Schlaitz, A.-L., Thompson, J., Wong, C. C. L., Yates, J. R., & Heald, R. (2013). REEP3/4 ensure endoplasmic reticulum clearance from metaphase chromatin and proper nuclear envelope architecture. *Developmental Cell*, *26*(3), 315–323. https://doi.org/10.1016/j.devcel.2013.06.016

Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*, *66*(1), 7–30.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1–25.

Wagner, S., Hess, M. A., Ormonde-Hanson, P., Malandro, J., Hu, H., Chen, M., … Bowen, B. R. (2000). A Broad Role for the Zinc Finger Protein ZNF202 in Human Lipid Metabolism. *Journal of Biological Chemistry*, *275*(21), 15685–15690. https://doi.org/10.1074/jbc.M910152199

Yue, D., Wang, Y., Sun, Y., Niu, Y., & Chang, C. (2017). C1QBP Regulates YBX1 to Suppress the Androgen Receptor (AR)-Enhanced RCC Cell Invasion. *Neoplasia (New York, N.Y.)*, *19*(2), 135–144. https://doi.org/10.1016/j.neo.2016.12.003

Zhang, X., Zhang, F., Guo, L., Wang, Y., Zhang, P., Wang, R., … Chen, R. (2013). Interactome Analysis Reveals that C1QBP (complement component 1, q subcomponent binding protein) Is Associated with Cancer Cell Chemotaxis and Metastasis. *Molecular & Cellular Proteomics*, *12*(11), 3199–3209. https://doi.org/10.1074/mcp.M113.029413

Zhou, X., & Tuck, D. P. (2007). MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics*, *23*(9), 1106–1114.