Exploratory Data Analysis on The Cancer Genomics Atlas: A Case Study

Yuzhen Ke

Department of Statistics, Purdue University

1. Introduction

Over the past few decades, oncology scientists around the world have been conducting research and studies aimed at understanding the development of cancer and improving the effectiveness of treatments. Advances in technologies have helped make remarkable progress in cancer therapy: besides primary therapies like surgery, chemotherapy, and radiation therapy, novel approaches, such as targeted immunotherapies, have attracted attention from the public. However, cancer is still one of the leading causes of premature deaths and is a major public health problem worldwide. According to the latest estimates of global mortality data from the World Health Organization (WHO) in 2019, cancer is the first or second leading cause of death in 127 countries, and about 30% of people die from cancer among those cases with premature deaths associated with noncommunicable diseases.[1] Cancer is also a burden on public health in the United States. The American Cancer Society published their annual report and claimed that in 2022 there are projected 1,918,030 new cancer cases and 609,360 cancer deaths. [2]

What is worse, not all patients will fully recover from cancer; cancer may come back even after treatment or a period of time when cancer cannot be detected, which oncologists call recurrence or recurrent cancer. Recurrent cancer may occur in the same place as original cancer (local recurrence), in the lymph nodes or tissues near original cancer (regional recurrence), or in other organs and tissues far from original cancer (distant recurrence). [3] If the differences in gene expression between disease-free patients and recurrent patients can be discovered, oncologists will be able to apply more effective and patient-tailored treatments to original cancer, relieve patients of recurrent cancer and burdens physically, mentally and financially, and improve the life expectancy and survival rate.

This case study aims to identify the cancer recurrence-associated genes with datasets from The Cancer Genome Atlas (TCGA), which may help provide theoretical support for the advancement in diagnosis, treatment, and prevention of tumor recurrence. In this article, I will use pancreatic cancer (PAAD) and colorectal cancer (COAD) datasets to analyze and investigate whether there

are some specific significant genes that can serve as genomic biomarkers to classify potential recurrence or disease-free patients.

2. Demonstration of TCGA Data

The dataset used in this case study is obtained from The Cancer Genome Atlas (TCGA); TCGA Research Network has collected over 10,000 cases for more than 25 different primary cancer types and aims to identify oncogene-associated similarities among different types of tumors in order to encourage the development and implementation of advanced knowledge on cancer detection, diagnosis, therapy and prevention. Two data files are mainly used in this study, RNA (Final) file (EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2. geneExp.tsv) and TCGA-Clinical Data Resource (CDR) Outcome (TCGA-CDR-Supple mentalTableS1.xlsx), which are both publicly accessible online. [4][5]

The clinical dataset contains 11,160 specimens and 33 variables. This table includes important clinical information of each patient for further analysis, such as “type” (tumor type), “ajcc_pathologic_tumor_stage” (the stage of patient’s cancer), “PFI” (progression-free interval event), and “PFI.time” (progression-free interval time in days). More specifically, in “PFI”, 1 represents that the patient dies without a new cancer event or has a new cancer event, such as progression of the disease, local recurrence, and distant metastasis, while 0 represents that the patient is censored otherwise. For events, “PFI.time” is decided by either “new_tumor_event_dx_days_to” or “death_days_to”; for censored cases, “PFI.time” is either “last_contact_days_to” or “death_days_to”. The gene expression data consists of 11,070 observations and 20,531 genes (an 11,070 x 20,531 matrix), which will be used to analyze and investigate the significant oncology-related genes. [4]

The pancreatic cancer dataset is constituted by 165 observations over 20,564 covariates with 79 females and 95 males. There are 66 patients in the “censored” group (PFI = 0) and 99 patients in the “event” group (PFI = 1). On the other hand, the colorectal cancer dataset contains 338

observations with 164 females and 174 males. There are 239 patients in the “censored” group (PFI = 0) and 99 patients in the “event” group (PFI = 1) in the clinical colorectal cancer dataset.

The RNA (final) is the gene expression data used for discovering significant genes associated with patients’ groups (disease-free or recurrence). The original data contains 11,069 samples (patients) and 20,530 genes. It records the expression level of each oncology-related gene for each patient. This dataset contains patients who have different tumor types, but in this case study, only the subsets of pancreatic cancer and colorectal cancer-related gene expression data are used. [5]

Furthermore, genes of oncogenic signaling pathways and cholesterol are also used for the analysis of investigating whether there are effective biomarkers to identify patients to be in disease-free or recurrence groups. The report, Oncogenic Signaling Pathways in The Cancer Genome Atlas (April 2018), published the 10 canonical signaling pathways based on previous studies of The Cancer Genome Atlas (TCGA). [6] The researchers evaluated 10 canonical signaling pathways which are more likely to be tumor drivers or therapeutic targets. In these 10 signaling pathways, there are 265 genes in total. The list of 47 cholesterol genes is provided by Dr. Jingwu Xie at Indiana University School of Medicine, who is also one of the collaborators on cancer research at Purdue University.

3. Methods and Results

In this case study, significant genes are identified by analyzing the resulting p-values from the two-sample t-test and Cox proportional hazards regression. The two-sample t-test is conducted to decide whether each oncology-related gene is differentially expressed in two groups (recurrence and disease-free). The Cox proportional hazards regression model is performed to discover the relation between the survival time of patients and variables. In this case, the model is to evaluate whether specific factors (covariates and genes) significantly influence the rate of event (the recurrence in patients) happening at a specific time. However, because multiple hypothesis tests are conducted simultaneously, the percentage of type I errors increases and the p-values are not

valid. Therefore, the false discovery rate (FDR) is applied to improve the accuracy of hypothesis testing. By definition, the FDR is the ratio of the number of false discoveries to the total number of discoveries (rejections). Then, a subset of significant genes will be selected for further data exploration under specific p-values or adjusted p-values with a particular cutoff. The principal component analysis (PCA) is performed and the visualization of the PCA plot will help identify genomic biomarkers. The location of the individual patients will show whether some patients are clustering together, and the clustering in the PCA plot can indicate the grouping of observations sharing similar characteristics. Such characteristics are the goal biomarkers.

Before subsetting pancreatic and colorectal datasets from RNA (final), the data quality control is applied to this gene expression data file. In this cleaning data step, the R codes are provided by Dr. Zhongyuan Chen. There are 29 missing gene IDs and a duplicate gene ID, “SLC35E2”, to be removed. After the deletion of these 30 gene IDs, there are 11,069 samples and 20,530 genes left. Then, to check whether there are excessive missing values and obvious outliers in the data, the package “WGCNA” (Weighted Correlation Network Analysis Network Construction) in R is used in the data cleaning process. [6] The results from “WGCNA” show that there are some missing values in genes and samples, and also there are significant outliers. After removing offending genes and samples and obvious outliers, the final clean gene expression data contains 11,064 patients and 20,282 genes (an 11,064 x 20,282 matrix) ready for data analysis.

3.1. The PAAD Data

First of all, the “PFI” value is used to classify two different groups: those patients who die or have new events before censoring are in the recurrence group (PFI = 1), and others who are formally diagnosed as cancer-free before censoring are in the disease-free group (PFI = 0). These classifications will further help analyze and identify whether there are any significant recurrence-associated clinical covariates or oncogenes in pancreatic cancer for future development in curing recurrent tumors.

Prior to the preliminary analysis on the clinical dataset of PAAD, the clinical dataset is cleaned by only keeping observations with “TUMOR FREE” in the disease-free group and observations with “WITH TUMOR” in the recurrence group. For convenience, a group label called “recurrence” is created, in which 1 means recurrence and 0 means disease-free. The clean PAAD clinical dataset has a sample size of 68 in the disease-free group and 106 in the recurrence group. Then, investigate whether there are significant covariates among age, gender, race, stage, and histological grade. Age is treated as a continuous covariate and two-sample t-test is applied; others are treated as categorical covariates, and a contingency table and Chi-squared test are performed for analysis. The following preliminary analysis is done on the original PAAD clinical dataset with 174 observations and 5 selected covariates.

Age

The basic statistics for the covariate age are as follows (Table 1): the mean values for the disease-free group and recurrence group are 64.29412 and 64.70755 respectively, which are not much different from each other. The standard deviations of the two groups are 10.20544 and 11.59472 respectively; the recurrence group has a larger variance than the other one. The histogram plots (Figure 1) show that these two groups have a similar distribution, and the disease-free group has some outliers on the left tail. The dot plots with box-plot in Figure 2 support the previous claims that the means are nearly the same and the variance of the recurrence group is greater than the other one. The null hypothesis for the two-sample t-test is that there is no difference between the means of ages in the disease-free and recurrence groups, while the alternative hypothesis states the opposite. The Welch Two Sample t-test in Figure 3 shows that the p-value (0.8052) is greater than the significant level α (0.05). Therefore, it is not significant and the null hypothesis is accepted, which means that the difference in means between the two groups is not significant.

Recurrence <dbl>	mean <dbl>	sd <dbl>	samp_size <int>
0	64.29412	10.20544	68
1	64.70755	11.59472	106

Table 1. Basic Statistics (Age)

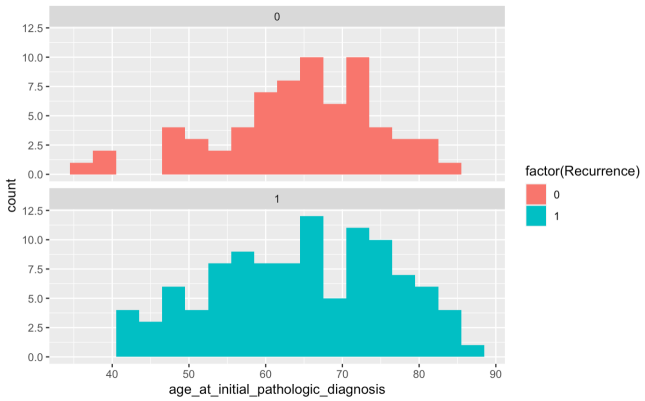


Figure 1. Histograms (Age)

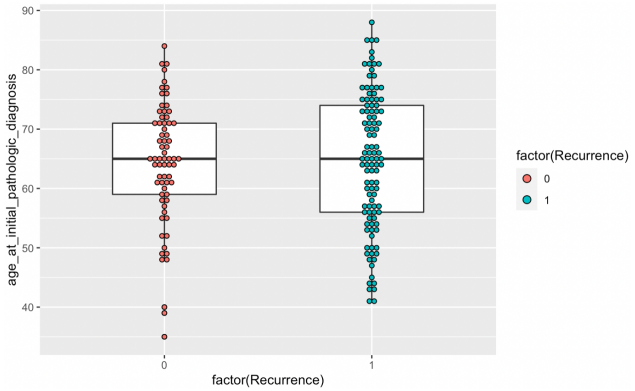


Figure 2. Dot Plots (Age)

```
        Welch Two Sample t-test

data:  age_at_initial_pathologic_diagnosis by
Recurrence
t = -0.24708, df = 155.75, p-value = 0.8052
alternative hypothesis: true difference in means
between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -3.718706  2.891847
sample estimates:
mean in group 0 mean in group 1
       64.29412        64.70755
```

Figure 3. Two Sample t-test (Age)

Gender

Based on the contingency table (Table 2) and mosaic plot (Figure 4), there is no obvious difference for covariate gender in the two groups. The Chi-squared test is applied, and the null hypothesis states whether the patient's group (disease-free or recurrence) is not associated with the covariate gender while the alternative hypothesis holds a differing opinion. From the result of the Chi-squared test (Figure 5), the p-value (0.9072) is much greater than the significance value α (0.05). Thus, it is not significant enough to reject the null hypothesis.

```
          0  1
FEMALE 30 49
MALE   38 57
```

Table 2. Contingency Table (Gender)

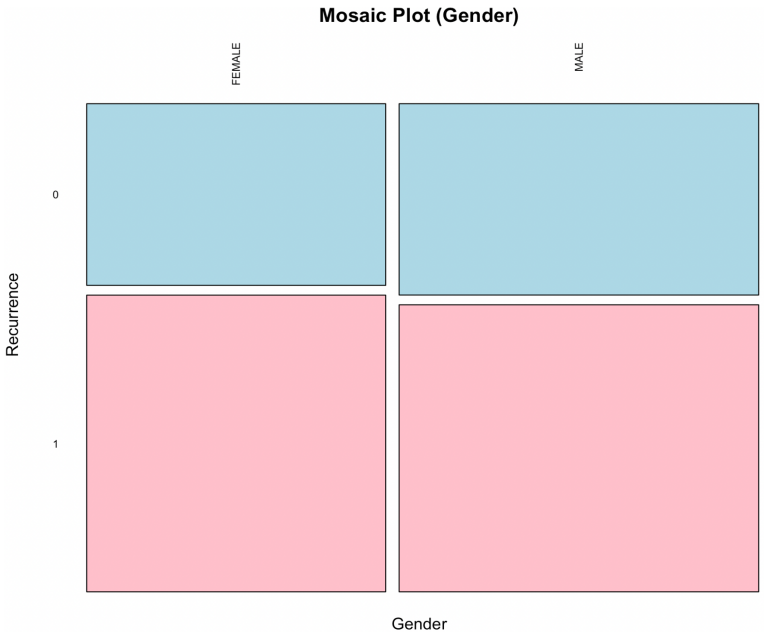


Figure 4. Contingency Table (Gender)

```
        Pearson's Chi-squared test with Yates'
continuity correction

data:  table.gender
X-squared = 0.01359, df = 1, p-value = 0.9072
```

Figure 5. Chi-squared Test (Gender)

Race

Table 3 shows the original contingency table, and Figure 6 represents the mosaic plot after removing 5 missing data; they do not indicate any significant pattern. However, because there are small counts in the cell of the contingency table (< 5) and the chi-squared approximation may be incorrect. After merging “ASIAN” and “BLACK OR AFRICAN AMERICAN” into one “Non-WHITE” group (Table 4), the result looks better. The p-value of Pearson’s Chi-squared test is 0.8959, which is much greater than the significance value α (0.05). It can be concluded that the patient’s group (disease-free or recurrence) is not associated with the covariate race.

	0	1
[Not Evaluated]	0	1
[Unknown]	0	4
ASIAN	5	6
BLACK OR AFRICAN AMERICAN	3	4
WHITE	60	91

Table 3. Contingency Table (Race)

race	recurrence 0	1
Non-WHITE	8	10
WHITE	60	91

Table 4. Merged Contingency Table (Race)

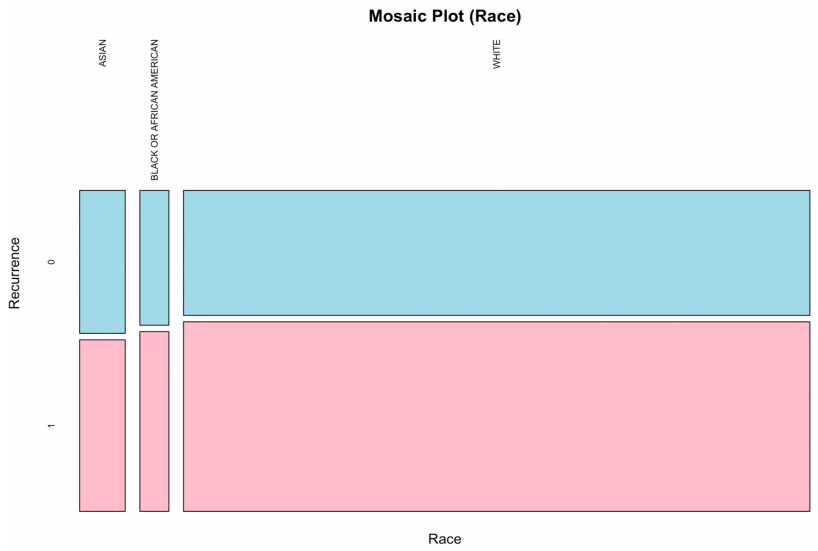


Figure 6. Mosaic Plot (Race)

```
        Pearson's Chi-squared test with Yates'
continuity correction

data:  new.table.race
X-squared = 0.017131, df = 1, p-value = 0.8959
```

Figure 7. Chi-squared Test (Race)

Stage

Table 5 and Figure 8 show the contingency table (before removing 3 missing data) and mosaic plot (after removing 3 missing data) respectively for the covariate stage. Because there are small values in Table 5, the Chi-squared approximation may be incorrect. Similarly, merge “Stage I”, “Stage IA” & “Stage IB” as “Stage I”, “Stage IIA” & “Stage IIB” as “Stage II”, and “Stage III” & “Stage IV” as “Stage III & IV”, and obtain a new contingency table in Table 6 and a new mosaic plot in Figure 9. Figure 9 presents that fewer patients have recurrent cancer in the early stage than in the late stage. So, there is a potentially significant relation between stage and recurrent cancer rates. Furthermore, even though there is a small value in the merged data contingency table (< 5), the p-value that the Chi-squared test (Figure 10) shows is 0.0229, which is smaller than the significance value α (0.05). The Fisher's exact test also presents the same conclusion: the p-value is 0.02374, which is also smaller than 0.05 (Figure 11). Therefore, the alternative hypothesis, which declares whether the patient is in disease-free group or recurrence group is associated with the covariate stage, is accepted.

	0	1
[Discrepancy]	2	0
[Not Available]	1	0
Stage I	1	0
Stage IA	3	2
Stage IB	9	5
Stage IIA	11	18
Stage IIB	39	74
Stage III	1	3
Stage IV	1	4

Table 5. Contingency Table (Stage)

stage	recurrence 0	1
Stage I	13	7
Stage II	50	92
Stage III & IV	2	7

Table 6. Merged Contingency Table (Stage)

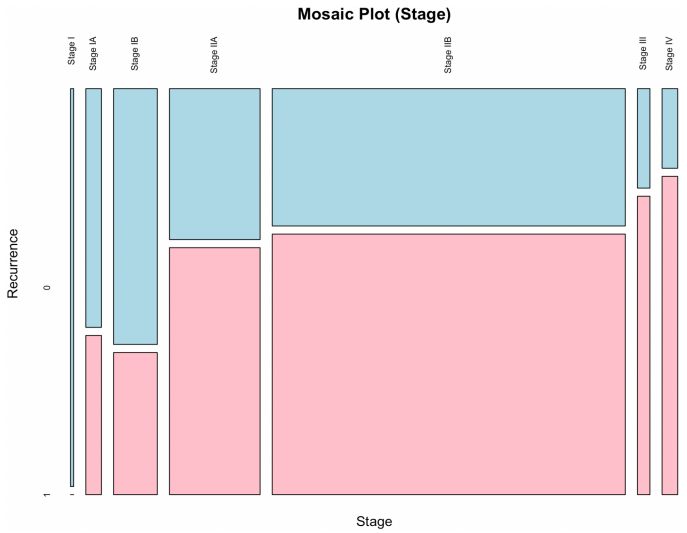


Figure 8. Mosaic Plot (Stage)

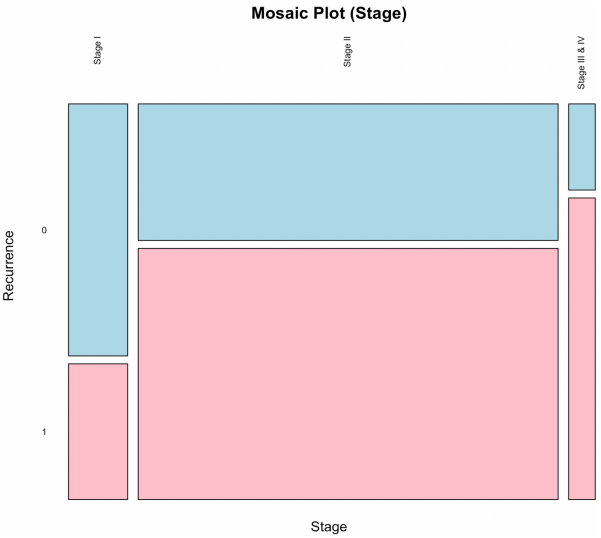


Figure 9. Merged Mosaic Plot (Stage)

```
Warning in chisq.test(new.table.stage) :
  Chi-squared approximation may be incorrect

        Pearson's Chi-squared test

data:  new.table.stage
X-squared = 7.6072, df = 2, p-value = 0.02229
```

Figure 10. Chi-squared Test (Stage)

```
        Fisher's Exact Test for Count Data

data:  new.table.stage
p-value = 0.02374
alternative hypothesis: two.sided
```

Figure 11. Fisher's Exact Test (Stage)

Histological grade

Histological grade data, shown in Table 7, also has small counts ("G4" and "GX") and thus is merged into "G3" (Table 8 & Figure 12). The result of Pearson's Chi-squared test in Figure 13 states that the p-value is 0.06298, which is greater than the significance level (0.05). Therefore, it is not significant enough to reject the null hypothesis, and it can be concluded that the patient's group (disease-free or recurrence) is not associated with the covariate grade.

	0	1
G1	17	14
G2	36	55
G3	13	34
G4	1	1
GX	1	2

Table 7. Contingency Table (Grade)

grade	recurrence 0	1
G1	17	14
G2	36	55
G3	15	37

Table 8. Merged Contingency Table (Grade)

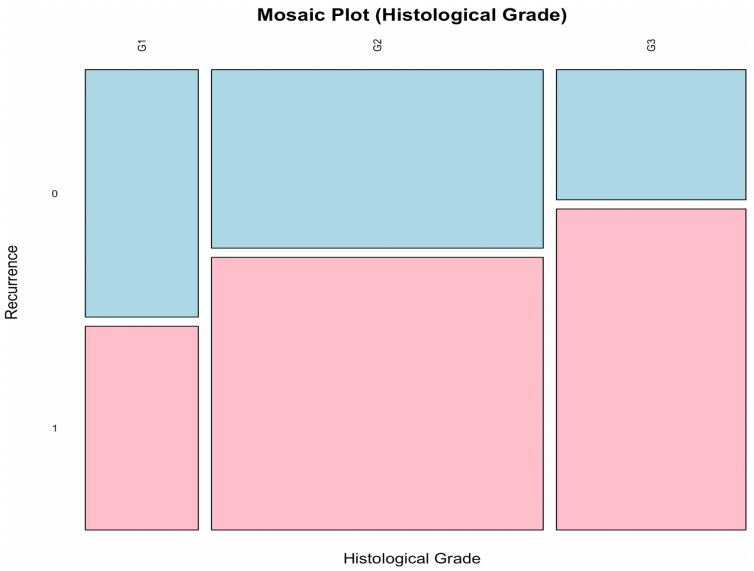


Figure 12. Merged Mosaic Plot (Grade)

```
        Pearson's Chi-squared test

data:  new.table.grade
X-squared = 5.5299, df = 2, p-value = 0.06298
```

Figure 13. Chi-squared Test (Grade)

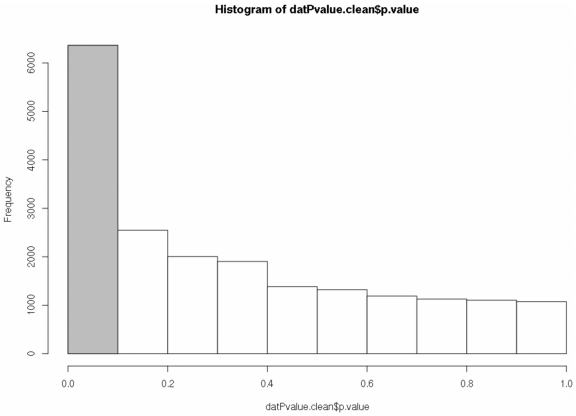
Therefore, among selected covariates (age, gender, race, stage, and histological grade), only stage is significant, but before merging different levels into one, the stage is insignificant. However, because this is a primary test, whether stage is considered an influencing factor is still to be tested in the following exploration.

Next, before further investigation, specimens are selected from the original gene expression data file, RNA (final), based on the pancreatic cancer type and the barcodes of patients in the clinical dataset. However, there are some duplicate observations: some patients have two sets of gene expression data in the RNA (final) file; one is collected from the patient's tumor tissues and the other one is collected from the patient's normal tissues. Because the gene expression data of tumor tissues is the main focus of this study, the data from normal tissues are considered uninformative and therefore deleted. Thus, a subset of data on PAAD cancer type is created for later gene analysis; this dataset contains 20,282 genes in columns and 165 observations in rows (a 165 x 20,282 matrix), including 66 patients in the disease-free group and 99 patients in the recurrence group.

To find the genes which are differentially expressed in disease-free and recurrence groups, the two-sample t-test is conducted for each gene; the null hypothesis of the two-sample t-test is that the gene is different between two groups, and the alternative hypothesis states the opposite. Because there are 20,282 genes to be tested simultaneously, a list of 20,282 p-values is created from the multiple comparisons. However, multiple comparisons cause an increased rate of false positivity and the p-values are not reliable. Thus, the false discovery rate (FDR) is conducted as an approach to controlling the number of false positives. Table 9 presents the number of significant genes from p-values or adjusted p-values under different cutoffs. Figure 14 and 15 shows histograms of p-values and adjusted p-values with cutoff of 0.10 respectively.

	p-value	adjusted p-value (FDR)
cutoff = 0.05	4,530	635
cutoff = 0.10	6,363	1,827

Table 9. Number of Significant Genes (PAAD)



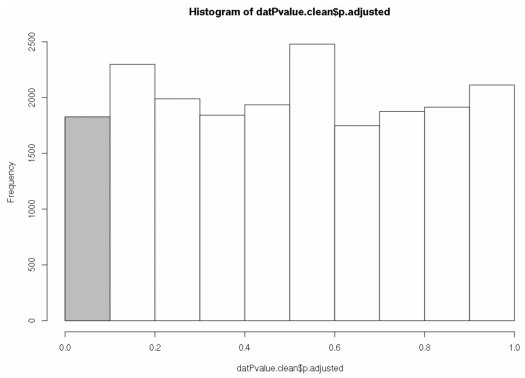


Figure 14 (left). Histogram of p-values with Cutoff 0.10 (PAAD) & Figure 15 (right). Histogram of adjusted p-values with Cutoff 0.10 (PAAD)

For further analysis, 1,827 significant genes are selected based on the standard of adjusted p-values smaller than 0.10. Then, among these 1,827 significant genes, there are 35 genes that are also pathway genes but there are only 2 genes that are also cholesterol genes. The numbers of these intersections of significant genes and pathway genes or cholesterol genes are too small and may not be good representations of patients. Next, Principal Component Analysis (PCA) is conducted to determine whether certain significant genes are a good representation of patients.

Figure 16 shows the PCA plot from the dataset which contains 165 observations and 1,827 significant genes (a 165 x 1,827 matrix), and the colored grouping labels are from “Recurrence” grouping labels: 0 (red dot) means the patient is in disease-free group and 1 (blue dot) means the patient is in recurrence group. From the PCA plot, there are 8 red dots that are obviously separated from the others, but it does not explain why other red dots in the disease-free group and blue dots are clustering together.

Next, instead of “Recurrence” grouping labels, a new type of grouping labels is applied and Figure 17 presents the new PCA plot. It is also from the dataset which contains 165 specimens and 1,827 significant genes, and Dr. Zhongyuan Chen at Purdue University contributes to these grouping labels. In Figure 17, 1 (blue dot) means the patient has significantly good survival, and 2 (red dot) means the patient does not have good survival. The separated 8 points on the right are all in the good survival group, and there are other 2 points in the good survival group clustering with the other group on the left. There are no other patient grouping structures in pancreatic cancer. However, the 10-patient size is too small and is not representative enough and there are 2 patients that cannot be separated. For further analysis of cancer recurrence, the colorectal cancer dataset is considered because this dataset has a larger sample size: 338 patients, including 289 patients in the disease-free group and 99 patients in the recurrence group.

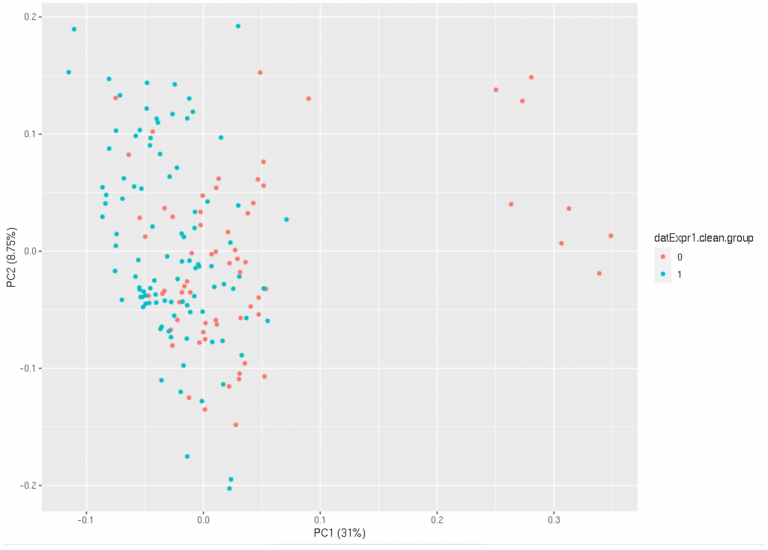


Figure 16. PCA Plot with “Recurrence” Grouping Labels (PAAD)

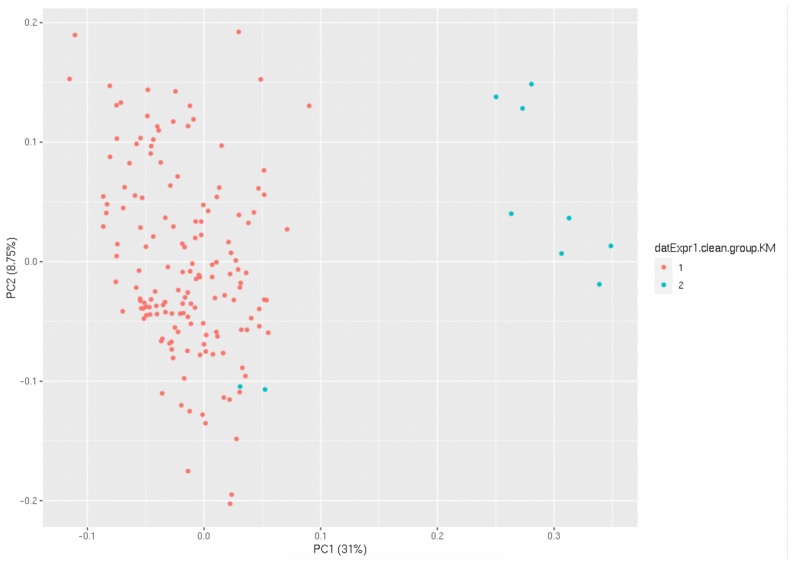


Figure 17. PCA Plot with “Good Survival” Grouping Labels (PAAD)

3.2. The COAD Data

There are 338 observations and 20,282 genes in the original subset of the colorectal gene expression dataset (a 338 x 20,282 matrix). Because there are lots of NA missing values, negative values, and 0, data quality control is required prior to any further data analysis on COAD data. The values for each gene expression are supposed to be non-negative, and negative values in this data file are uninformative and therefore treated as missing values. The first step of the data quality control process is to calculate the percentage of missing values (both NA and negative values) in each gene column and patient row, respectively. If the percentage is greater than 33%, the whole gene column or patient row with excessive missing values will be deleted; otherwise, the column or the row will be kept. The resulting percentages show that there are no columns or rows that have a percentage of missing values greater than 33%. So, all genes and patients are kept, the COAD data subset is a 388 x 20,282 matrix. Secondly, replace 0 with a small number, 1×10^{-5}, as it represents a low level of gene expression.

Next, the results, for each gene column, from the Shapiro Wilk test for normality present that the gene expression values in each column are not normally distributed, as the maximum p-value for the 20,282 gene list is 8.36×10^{-6}. Thus, the normal test shows that all gene columns have p-values smaller than the significance level α (0.05), which rejects the null hypothesis that the data

is normally distributed. Therefore, a log transformation is conducted in each cell in the dataset to conform to normality. After that, imputation is performed: missing values, including both negative values and NA values, are replaced by the log mean value of each gene column.

Then, by Dr. Zhongyuan Chen's code, a multivariate Cox proportional hazards regression is performed to select gene expression biomarkers associated with the patient's group (recurrence or disease-free). In this cox model, the survival object is created by "PFI.time " and "PFI", as "PFI.time" is used as time and "PFI" is used as the event (recurrence). Besides each gene, covariates (gender, age, stage, and subtype) are included in the model for investigation of how these factors jointly impact survival. The results from the cox model give a list of p-values for each gene. Again, since multiple hypotheses are tested simultaneously, the rate of false positivity is higher and the resulting p-values are not dependable. Thus, the false discovery rate (FDR) is conducted to obtain solid adjusted p-values. Table 10 presents the number of significant genes from p-values or adjusted p-values under different cutoffs. However, there are only 176 significant genes selected based on adjusted p-values with a cutoff of 0.10, which is a small sample size. Because additional covariates are included, "PFI.time" depends on both genes and covariates. For further analysis, marginal Cox models on each covariate, without genes, are performed, and Table 11 shows the p-values from each Cox proportional hazards regression of different covariates. Figures 18 - 21 present the resulting p-values from different Cox models respectively and only the covariate stage is significant enough. Therefore, the covariate stage has already helped separate patients, and the result of the previous multivariate Cox proportional hazards regression cannot fully explain and do clustering according to gene expressions.

	p-value	adjusted p-value (FDR)
cutoff = 0.05	1,892	99
cutoff = 0.10	2,898	176

Table 10. Number of Significant Genes (COAD)

```
Call:
coxph(formula = Surv(as.numeric(COAD_data$PFI.time), as.numeric(COAD_data$PFI)) ~
    COAD_data$gender)

  n= 338, number of events= 99

                      coef exp(coef) se(coef)     z Pr(>|z|)
COAD_data$genderMALE 0.3151    1.3704   0.2040 1.545    0.122
```

Figure 18. P-value from Cox Model (Gender)

```
Call:
coxph(formula = Surv(as.numeric(COAD_data$PFI.time), as.numeric(COAD_data$PFI)) ~
    as.numeric(COAD_data$age_at_initial_pathologic_diagnosis))

  n= 338, number of events= 99

                                                      coef exp(coef)  se(coef)      z Pr(>|z|)
as.numeric(COAD_data$age_at_initial_pathologic_diagnosis) -0.002246  0.997756  0.007767 -0.289    0.772
```

Figure 19. P-value from Cox Model (Age)

```
Call:
coxph(formula = Surv(as.numeric(COAD_data$PFI.time), as.numeric(COAD_data$PFI)) ~
    as.character(COAD_data$stage))

  n= 338, number of events= 99

                                  coef exp(coef) se(coef)     z Pr(>|z|)
as.character(COAD_data$stage)2  1.3484    3.8514   0.6043 2.231   0.0257 *
as.character(COAD_data$stage)3  1.8163    6.1489   0.6032 3.011   0.0026 **
as.character(COAD_data$stage)4  2.8790   17.7968   0.6062 4.749 2.04e-06 ***
```

Figure 20. P-values from Cox Model (Stage)

```
Call:
coxph(formula = Surv(as.numeric(COAD_data$PFI.time), as.numeric(COAD_data$PFI)) ~
    COAD_data$subtype)

  n= 338, number of events= 99

                          coef exp(coef) se(coef)     z Pr(>|z|)
COAD_data$subtypeGS   -0.3090    0.7342   0.3137 -0.985    0.325
COAD_data$subtypeMSI  -0.3540    0.7019   0.2843 -1.245    0.213
COAD_data$subtypePOLE -0.7354    0.4793   1.0081 -0.729    0.466
```

Figure 21. P-values from Cox Model (Subtype)

To investigate whether patients can be clustered according to gene expression, a univariate Cox proportional hazards regression is performed for each oncology-related gene from the COAD dataset. The survival object is created by “PFI.time ” and “PFI”, as “PFI.time” is used as time and “PFI” is used as the event (recurrence). The result creates a list of p-values but, again, because the multiple hypotheses are tested simultaneously, the rate of false positivity increases, and the p-values are not dependable. Thus, the false discovery rate (FDR) is performed. Table 11 presents the number of significant genes from p-values or adjusted p-values under different cutoffs. Figures 22 and 23 present histograms of p-values and adjusted p-values with a cutoff of 0.10 respectively.

	p-value	adjusted p-value (FDR)
cutoff = 0.05	1,329	1
cutoff = 0.10	2,539	35

Table 10. New Number of Significant Genes (COAD)

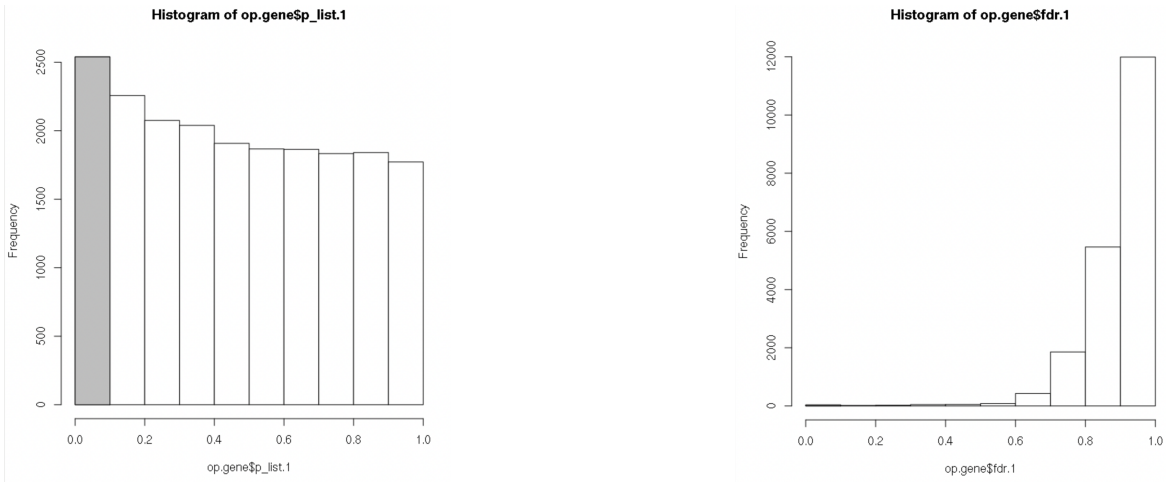


Figure 22 (left). Histogram of p-values with Cutoff 0.10 (COAD) & Figure 23 (right). Histogram of adjusted p-values with Cutoff 0.10 (COAD)

There are 35 significant genes selected based on adjusted p-values (FDR) with a cutoff of 0.10. Next, Principal Component Analysis (PCA) is performed to determine whether specific significant genes are a good representation of patients. Figure 24 presents the PCA plot from the dataset which contains 338 observations and 35 significant genes (a 338 x 35 matrix), and the colored grouping labels are from “Recurrence” grouping labels: 0 (red dot) means the patient is in disease-free group and 1 (blue dot) means the patient is in recurrence group. The PCA plot in Figure 24 shows that red dots and blue dots are clustering together on the left side and there are two outliers on the upper-right and lower-left corners. It does not show significant clustering or grouping structure among patients. Thus, these 35 significant genes are uninformative in identifying genetic biomarkers for colorectal cancer.

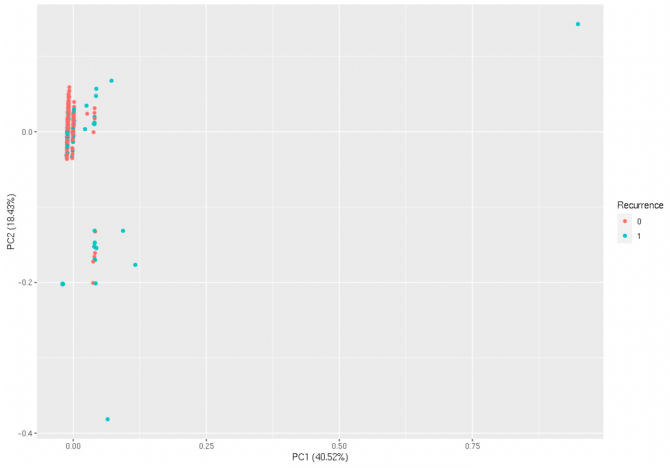


Figure 24. PCA Plot with 35 Significant Genes (COAD)

Then, among the 1,329 significant genes according to the p-values with a cutoff of 0.05, there are only 1 significant gene which are also in the cholesterol gene list. However, on the other hand, there are 21 significant genes that are also in the signaling pathway gene list. Figure 25 shows the PCA plot is based on the 21 overlapping genes and 338 patients (a 338 x 21 matrix). The plot indicates that the majority of the patients are clustering together and cannot be separated. However, in the upper-left part of the plot, there are some patients (red dots) in disease-free groups who are separated from the majority, which means that these 21 significant overlapping genes can identify some groups of patients from others.

Figure 25. PCA Plot with 21 Significant Genes (COAD)

4. Conclusion and Discussion

From the previous 4 PCA plots above, statements can be made: (1). The PCA plot of 35 significant genes (Figure 24) from the colorectal cancer dataset shows that most patients are clustering together, and the significant genes are uninformative; (2). The PCA plot of 21 overlapping genes, which are both pathway genes and significant genes (based on p-values with a cutoff of 0.05), presents that, even though the majority of patients cluster together, there are some particular patients in the disease-free group (red dots) separated in the upper-left part; (3). Figures 16 and 17 indicate that most of the patients cluster together, while some patients in the disease-free group, who have significantly good survival, are obviously separated. Therefore, the conclusions for

pancreatic cancer and colorectal cancer are the same, as the majority are grouped together while some patients with good responses are well separated. However, how to identify good patients (patients that have good responses or survival) is unknown for now. For further exploration of the identification and characteristics of these good patients, the following steps can be considered: (1). Classify the good patients (red dots) in the upper-left part of Figure 25 as one group; (2). Classify the bad patients in the recurrence group (blue dots) in the lower-left part of Figure 24 as the other group; (3). Compare the information of patients in these two groups and find the similarities and differences between the two groups.

5. References

[1] Global health estimates 2020: deaths by cause, age, sex, by country and by region, 2000 - 2019. Geneva, World Health Organization; 2020 (https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death, accessed 2 April 2022)

[2] Siegel RL, Miller KD, Fuchs HE, Jemal A (2022). Cancer statistics, 2022. *CA Cancer J Clin*. 2022; 72(1): 7 - 33. https://doi.org/10.3322/caac.21708

[3] “What Is Cancer Recurrence?” American Cancer Society, https://www.cancer.org/treatment/survivorship-during-and-after-treatment/long-term-health-concerns/recurrence/what-is-cancer-recurrence.html.

[4] Pan-Cancer Atlas. RNA (final) [Data set]. National Cancer Institute. https://gdc.cancer.gov/about-data/publications/pancanatlas

[5] Pan-Cancer Atlas. TCGA-Clinical Data Resource (CDR) Outcome [Data set]. National Cancer Institute. https://gdc.cancer.gov/about-data/publications/pancanatlas

[6] F. Sanchez-Vega, M. Mina, J. Armenia, W.K. Chatila, A. Luna, K.C. La, S. Dimitriadoy, D.L. Liu, H.S. Kantheti, S. Saghafinia, et al. Oncogenic signaling pathways in the cancer genome atlas. Cell, 173 (2018), pp. 321-337 e10

[7] Langfelder, P & Horvath, S. (2014). *Tutorial for the WGCNA package for R: I. Network analysis of liver expression data in female mice.*

https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/FemaleLiver-02-networkConstr-man.pdf