Prediction of Overall Survival in Gastric Cancer by a Six-Gene Cox Proportional Hazards Model

Zihan Fang

International School of Beijing, Beijing 101300, China.

Abstract: This study aims to investigate prognostic genes that are correlated with overall survival in gastric cancer. A list of 43 critical genes is first selected from literature review and is narrowed down using marginal analysis, false discovery rate (FDR). This study used the 16 differential genes selected by FDR to create a cox proportional hazards regression model. Using a stepwise approach, the cox model is refined. From the 16 genes, 6 genes significantly correlated with overall survival are chosen and kept in the cox model. A principal component regression model was also constructed based on the principal component analysis results. The concordance of the principal component regression model is then compared to the concordance of the cox proportional hazards regression model. The final 6 identified genes are NRP1, STK11, MCM2, MARCKS, CTS6, C5. Of the 6 genes, MARCKS, NRP1, STK11, MCM2 are in line with previous research. CTS6 and C5, although studied in other cancers, are comparatively novel in the field of gastric cancer.

Keywords: Prognostic Genes; Gastric Cancer; Six-Gene Cox Proportional Hazards Model

1. Introduction

Gastric cancer (GC) is the fifth most common cancers worldwide, accounting for 35% of all cancer-related deaths (bray et al, Arnold et al). The American cancer society estimates that there are about 26,380 new cases of GC in United States during 2022. More than 90% of GC have been reported to be adenocarcinomas which develop from mucosa, the inner most layer of the stomach (Ilic and Ilic).

As a multifactorial disease, there are well-established non-omics risks factors that contribute to the progression of GC. One of the best-known risk factors is H. pylori infection, bacteria in the digestive tract that attacks the stomach lining. Tobacco, obesity, radiation and dietary factors, such as consuming a high in-take of salt-preserved food and low in-take of vegetables and fruits, are also potential risk factors (Ilic and Ilic). Overall, people are likely to be diagnosed with the cancer around 60-80 years of age (Julita et al). It uncommon for patients to be diagnosed with GC under the age of 45 (Ferlay et al, Howlader et al). In addition, the frequency of being diagnosed with GC in men is double the frequency of being diagnosed in women (Ilic and Ilic).

Over the past decade, research have been carried out to study critical genes and biomarkers to understand the disease's progression and survival rates. Historically, genes such as CLDN1, THBS2 and SPOCK 1 are shown to be upregulated in GC and to be associated with decreased survival (Marimuthu et al, Jung et al, Pan et al, Eftang et al). More recently, genes such as MARCKS, NRP1, COL10A1 and CD109 have been previously identified to be correlated with overall survival, and many prognostic gene models have been made accordingly (Sun et al, Quan et al, Wang et al, Huang et al, Dai et al). However, the identification and confirmation of prognostic genes are still incomplete due to the complexity of the genetic interactions. Some new biomarkers such as CST6 have been identified to be correlated with the survival in GC but was previously known to be a critical gene in breast cancers (Li et al). Recently, genes such as P3H2 and C5 genes, previously not reported, have been identified in Zhou et al study. This emphasizes the importance to continue investigating into genetic factors that could influence the overall survival.

To investigate the genetic factors behind cancer, researchers use public data bases to conduct bio-informational analysis. A large and new data set with high quality is essential to reduce errors in research and come to more applicable, accurate and reliable conclusions. The TCGA database is widely used in the study of cancer as it includes a bigger sample size, data with higher quality and newer data overall. Using gene and clinical data from TCGA, this study identified four genes that are significantly correlated with overall survival in GC and established a cox proportional-hazards regression model.

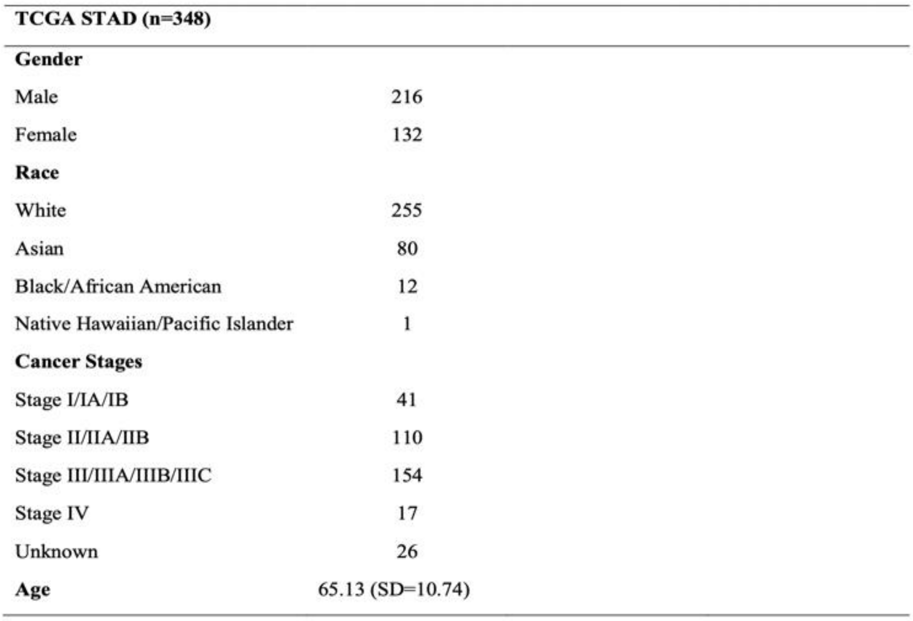
2. Methods

2.1 Data acquisition

This study is based on the public database provided by The Cancer Genome Atlas (TCGA). Over a 12-year period, TCGA analyzed and collected cancer samples from over 11,000 patients, generating data points including clinical information, molecular analyte metadata and molecular characterization data such as gene expression values.

Gene expression data and the corresponding clinical data were downloaded from the TCGA STAD database. All cases containing unreported, unspecified and unknown data points are excluded from this study. This study considered data from both genders and vital statuses. In addition, this study took into consideration of all reported races in GC including white, Asian, black/African American and native Hawaiian/pacific islander. Because the sample size of Asian, Black/African American and Native/Hawaiian islander cancer patients small compared to the sample size of white patients, this study separated race in two main categories: white and non-white. In total, 348 cases are analyzed. Details regarding the sample size can be seen in table 1.1 below and the appendix A.

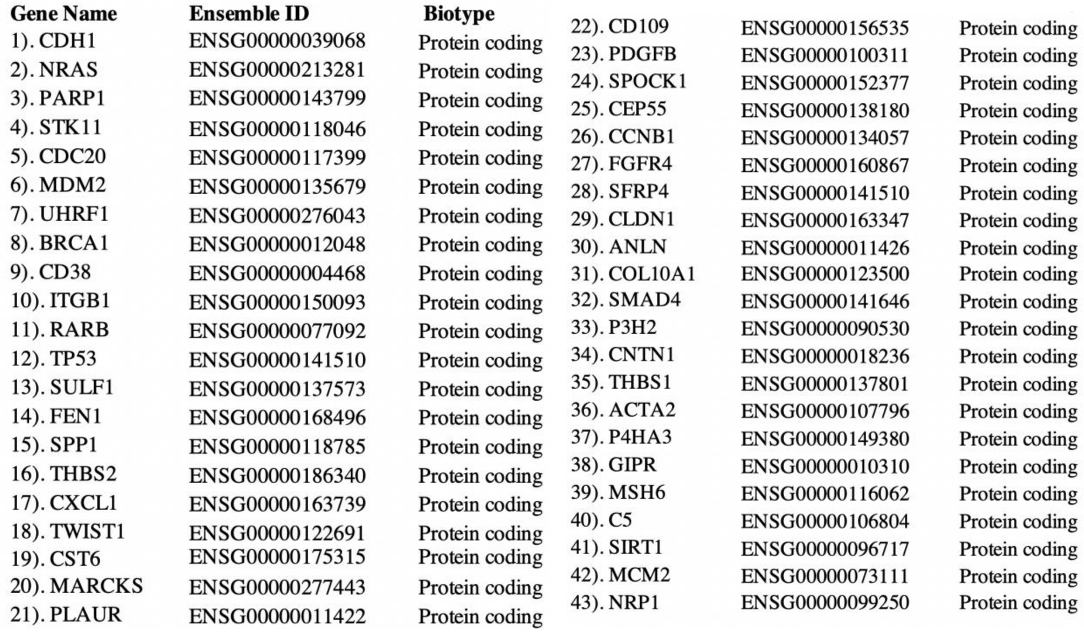
Table 1 TCGA STAD



TCGA STAD (n=348)	
Gender	
Male	216
Female	132
Race	
White	255
Asian	80
Black/African American	12
Native Hawaiian/Pacific Islander	1
Cancer Stages	
Stage I/IA/IB	41
Stage II/IIA/IIB	110
Stage III/IIIA/IIIB/IIIC	154
Stage IV	17
Unknown	26
Age	65.13 (SD=10.74)

A list of 43 critical genes regarding GC is selected from conducting a literature review of published paper on gene expression and overall survival of GC. 39 articles are selected from journals including the National Library of Medicine, Nature, American Association for Cancer Research, Science Direct, Karger, Wiley Online Library, Springer Link, Frontiers, Hindawi, PeerJ and MDPI. Across the articles, there are some overlaps, though not strong, between the genes identified. The list of 43 genes then was narrowed down to identify the genes that most significantly correlate with survival in GC.

Table 2 List of 43 Genes



Gene Name	Ensemble ID	Biotype
1). CDH1	ENSG00000039068	Protein coding
2). NRAS	ENSG00000213281	Protein coding
3). PARP1	ENSG00000143799	Protein coding
4). STK11	ENSG00000118046	Protein coding
5). CDC20	ENSG00000117399	Protein coding
6). MDM2	ENSG00000135679	Protein coding
7). UHRF1	ENSG00000276043	Protein coding
8). BRCA1	ENSG00000012048	Protein coding
9). CD38	ENSG00000004468	Protein coding
10). ITGB1	ENSG00000150093	Protein coding
11). RARB	ENSG00000077092	Protein coding
12). TP53	ENSG00000141510	Protein coding
13). SULF1	ENSG00000137573	Protein coding
14). FEN1	ENSG00000168496	Protein coding
15). SPP1	ENSG00000118785	Protein coding
16). THBS2	ENSG00000186340	Protein coding
17). CXCL1	ENSG00000163739	Protein coding
18). TWIST1	ENSG00000122691	Protein coding
19). CST6	ENSG00000175315	Protein coding
20). MARCKS	ENSG00000277443	Protein coding
21). PLAUR	ENSG00000011422	Protein coding
22). CD109	ENSG00000156535	Protein coding
23). PDGFB	ENSG00000100311	Protein coding
24). SPOCK1	ENSG00000152377	Protein coding
25). CEP55	ENSG00000138180	Protein coding
26). CCNB1	ENSG00000134057	Protein coding
27). FGFR4	ENSG00000160867	Protein coding
28). SFRP4	ENSG00000141510	Protein coding
29). CLDN1	ENSG00000163347	Protein coding
30). ANLN	ENSG00000011426	Protein coding
31). COL10A1	ENSG00000123500	Protein coding
32). SMAD4	ENSG00000141646	Protein coding
33). P3H2	ENSG00000090530	Protein coding
34). CNTN1	ENSG00000018236	Protein coding
35). THBS1	ENSG00000137801	Protein coding
36). ACTA2	ENSG00000107796	Protein coding
37). P4HA3	ENSG00000149380	Protein coding
38). GIPR	ENSG00000010310	Protein coding
39). MSH6	ENSG00000116062	Protein coding
40). C5	ENSG00000106804	Protein coding
41). SIRT1	ENSG00000096717	Protein coding
42). MCM2	ENSG00000073111	Protein coding
43). NRP1	ENSG00000099250	Protein coding

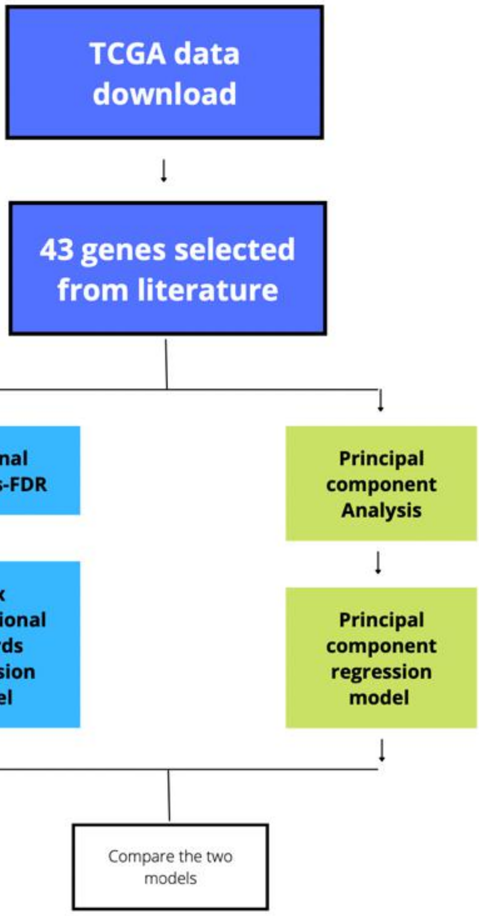
2.2 Statistical Testing

All statistical testing was conducted using the R software. The overall survival distribution was shown through plotting the relationship between the percent surviving and the survival time in days. To create the cox proportional hazards regression model, I implemented marginal analysis of false discovery rate (FDR) to narrow down the list of 43 genes by identifying the genes that are truly significant. The FDR uses adjusted p-value to eliminate false positive results and is defined as FDR=E(Q), where Q=V/R if R> 0. If R= 0, Q=0. This study set FDR adjusted values at the threshold of 0.05 and looked at the genes whose adjusted p-value is lower or equal to 0.05.

To then investigate into the correlation between non-omics risk factors as well as the significant genes identified, this study constructed multivariate cox proportional hazard regression model to determine the hazard ratios and the 95% confidence intervals for the association between overall survival and genetic and non-omics risk factors. For all variables, P<0.05 is established as the cut off value. The model has the overall function of h(t)=h0(t)×exp(b1x1+b2x2+...+bpx). To improve the accuracy of the model, this study implemented the stepwise approach to choose which variables to keep using the step function in R.

To compare the accuracy of the cox proportional hazards regression model based on the genes identified by FDR to the principal component regression model, this study then conducted principal component analysis on the original list of 43 genes using the princomp function in R. The number of principal components selected for regression was based on the percent of the original data they contain. Majority of the original data must be present in the number of principal components selected for the regression to be accurate. A principal component regression model was then created with principal components selected as the independent variables; the dependent variable of survival time and status remained the same as the original cox proportional hazards regression model. Lastly, the concordance of the two models is compared to determine which model more accurately predicts the influence of genetic and non-omics risk factors on overall survival. The overall procedure is summarized in the flow chart below.

Figure. 1 Flowchart of Overall Procedure

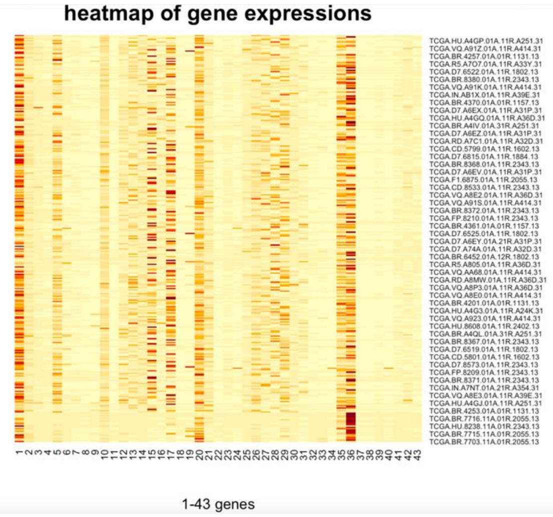

TCGA data download
43 genes selected from literature
Marginal Anlaysis-FDR
Principal component Analysis
Cox proportional hazards regression model
Principal component regression model
Compare the two models

3. Results

3.1 EDA

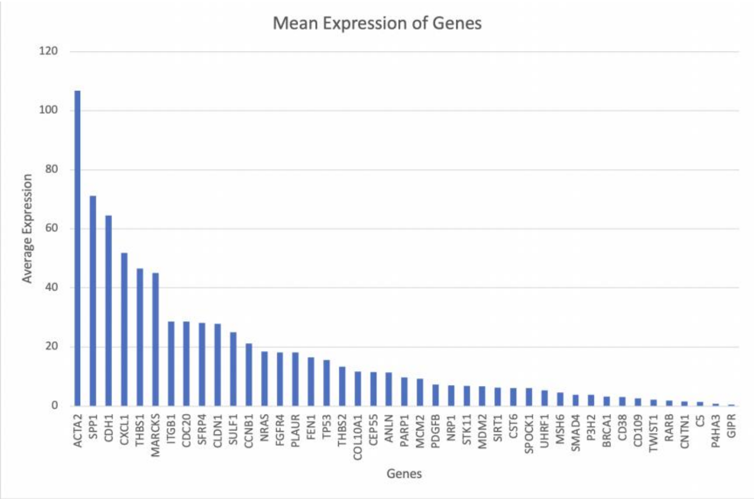
The heatmap shown below shows the overall gene expression for the list of 43 genes across 348 cases. The numbers on the x axis correspond to the 43 genes in the order of table 2.1.2. The Y-axis shows 348 cases in which each gene’s expression data is collected.

Figure. 2 Heatmap of Gene Expressions



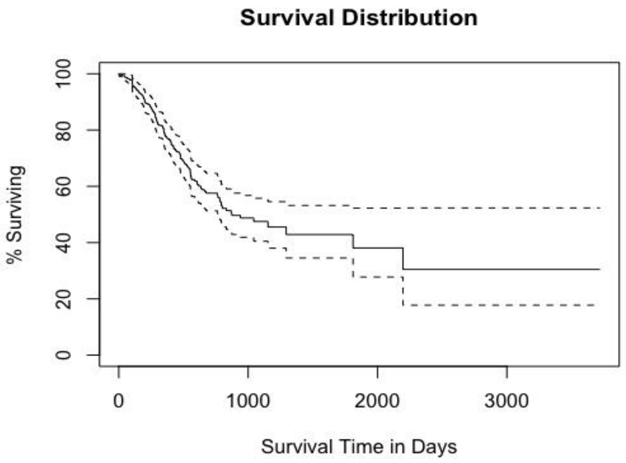
In the bar graph summarizing the average expression of each gene, there are no gene that is not expressed at all. The gene expression varies greatly, ranging from 106.8 to 0.5.

Figure. 3 Mean Expression of Genes



The survival distribution across the sample clearly shows that as time progresses the percent of people surviving decreases quite significantly. By the 3000's day, the percent of surviving people is less than 40. Overall, the survival distribution suggests that the mortality rate of GC is relatively high. Based on this assumption, this study then built models to determine the factors that influence survival.

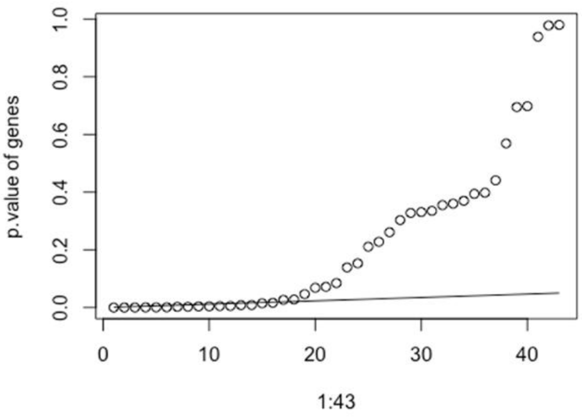
Figure. 4 Survival Distribution



3.2 Differential Genes

FDR results conducted on the 43 genes indicate that only 16 genes are truly significant according to their adjusted p-value at the cut off p<0.05. Three non-omics factors, age, gender and race, are also included the FDR analysis to be considered as covariates alongside with critical genes. In the plot of adjusted p-value versus the 1-43 gene shown below, the datapoints under the linear line, which indicates the cut off value, are interpreted to be truly significant. The 16 significant genes identified in the order of their p-value rank are CD109, NRP1, SPOCK1, P4HA3, STK11, MARCKS, UHRF1, CTS6, CNTN1, MCM2, THBS2, FEN1, P3H2, C5, ACTA2 and SDRP4 (see appendix B).

Figure. 4 P.Value of Genes

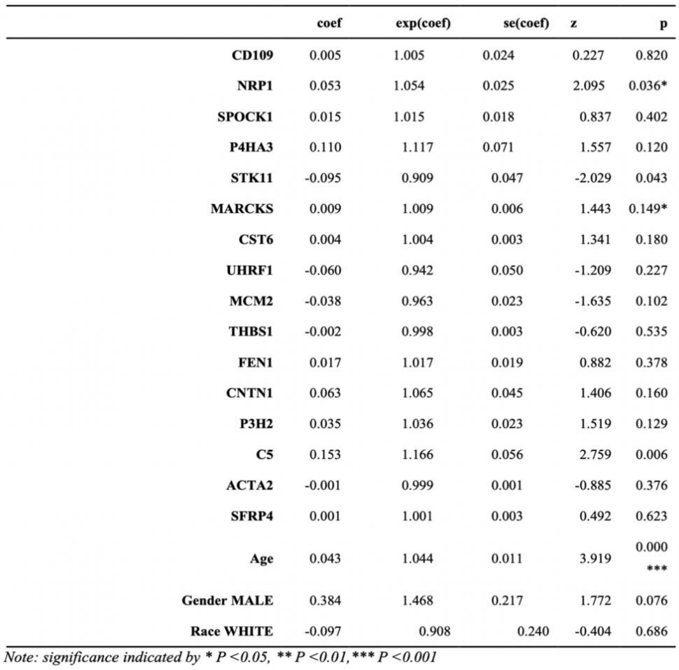


3.3 Cox Proportional Hazards Regression Models

In the original cox model that includes all 16 genes identified by FDR as well as non-omics risk factors including gender, race and age, the result of the model indicates that not all variables are significant. Only gene NRP1, STK11, and P3H2 are significant as they have p values less than 0.05. NRP1 and P3H2 are shown to be risky candidate genes as there is a positive correlation between their increasing expression and increasing risk of death. NRP1 has the hazard ratio of 1.059 which indicates that as its gene expression increases, the risk of death or estimated hazard increases 1.059 times or 5.9%. P3H2 has the hazard ratio of 1.048 which indicates that as its gene expression increases, the estimated hazard increases by 4.8%. On the other hand, STK11 is shown to be a protective candidate gene as there is a correlation between its increasing expression and decreasing risk of death. STK11 has the hazard ratio of 0.90 which suggests that as its gene expression increases, the risk of death decreases by 0.10 times or 10%.

The only non-omics risk factor that is shown to be significant age. In fact, age has a p value that is 0.001 which makes it the most significant variable in this model. In line with previous research and understanding of cancer, the hazard ratio of age is 1.043 which means that as age increases by one, the risk of estimated hazard increases by 4.3%.

Table 3 Cox Model for 16 Genes



	coef	exp(coef)	se(coef)	z	p
CD109	0.005	1.005	0.024	0.227	0.820
NRP1	0.053	1.054	0.025	2.095	0.036*
SPOCK1	0.015	1.015	0.018	0.837	0.402
P4HA3	0.110	1.117	0.071	1.557	0.120
STK11	-0.095	0.909	0.047	-2.029	0.043
MARCKS	0.009	1.009	0.006	1.443	0.149*
CST6	0.004	1.004	0.003	1.341	0.180
UHRF1	-0.060	0.942	0.050	-1.209	0.227
MCM2	-0.038	0.963	0.023	-1.635	0.102
THBS1	-0.002	0.998	0.003	-0.620	0.535
FEN1	0.017	1.017	0.019	0.882	0.378
CNTN1	0.063	1.065	0.045	1.406	0.160
P3H2	0.035	1.036	0.023	1.519	0.129
C5	0.153	1.166	0.056	2.759	0.006
ACTA2	-0.001	0.999	0.001	-0.885	0.376
SFRP4	0.001	1.001	0.003	0.492	0.623
Age	0.043	1.044	0.011	3.919	0.000 ***
Gender MALE	0.384	1.468	0.217	1.772	0.076
Race WHITE	-0.097	0.908	0.240	-0.404	0.686

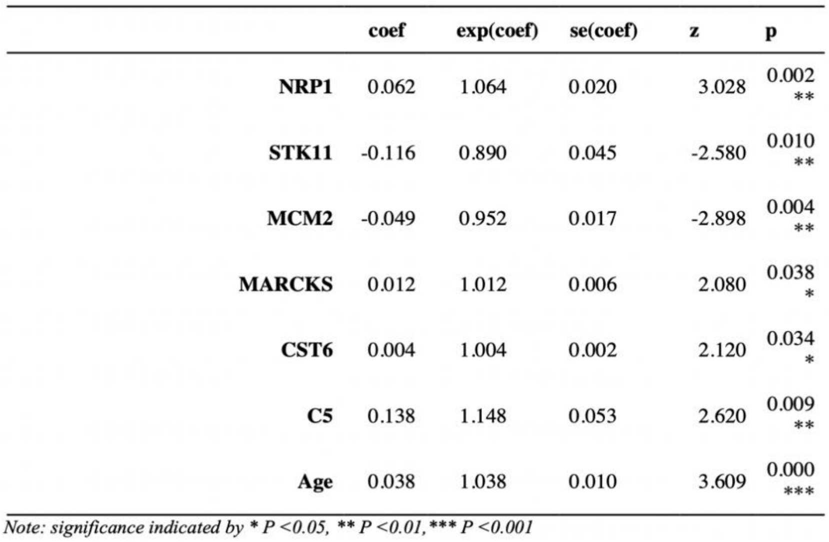
*Note: significance indicated by * P <0.05, ** P <0.01,*** P <0.001*

After implementing the stepwise approach using the step function in R to create a more accurate model by selecting more significant variables, the 16 genes in the original model were narrowed down to 6 including NRP1, STK11, MCM2, MARCKS, CST6 and C5. In this updated model, non-omics risk factors are narrowed down to only age. In terms of significance, age remains to be the most significant with a p value less than 0.001. Genes such as NRP1, STK11, MCM2 and C5 are the second most significant with p values less than 0.01. MARCKS and CST6 have p values less than 0.05. All 6 genes in this model are significant in predicting the overall survival in GC. Consistent with the original version of the model, NRP1 is a risky candidate gene. MARCKS, CST6 and C5 are all risky candidate genes. As the level of their gene expressions increase, the risk of death increases.

The hazard ratio for NRP1 is 1.064 which means that as the gene expression increases, the estimated hazard increases by 6.4%. MARCKS has a hazard ratio of 1.012; as its expression increases, the risk of death increases by 1.2%. The hazard ratio of CST6 is 1.004, suggesting that as its gene expression increases, the risk of death increases by 0.4%. Finally, C5 is shown to have a hazard ratio of 1.148, which means that as its gene expression increases, the risk of death increases by 14.8%. From the hazard ratios of the risky candidate genes, we can see that the increasing gene expression of C5 increases the risk of death by the highest percentage, 14.8%, when compared to the other genes.

On the other hand, STK11 and MCM2 are protective candidate genes as its increasing gene expression is correlated with a decrease in estimated hazard. The hazard ratio for STK11 is 0.890 which demonstrates that as its gene expressing increases, the estimated hazard decreases by 11% or 0.890 times. Finally, the hazard ratio for MCM2 is 0.952 which suggests that as its gene expression increases, the estimated hazard decreases by 4.8%. From the hazard ratios of the protective candidate genes, we can see that the increasing gene expression of STK11 decreases the risk of death by the highest percentage, 11%, when compared to MCM2. The detailed results of this 6-gene prognostic model are shown in the table below.

Table 4 Cox Model For 6 Genes



	coef	exp(coef)	se(coef)	z	p
NRP1	0.062	1.064	0.020	3.028	0.002 **
STK11	-0.116	0.890	0.045	-2.580	0.010 **
MCM2	-0.049	0.952	0.017	-2.898	0.004 **
MARCKS	0.012	1.012	0.006	2.080	0.038 *
CST6	0.004	1.004	0.002	2.120	0.034 *
C5	0.138	1.148	0.053	2.620	0.009 **
Age	0.038	1.038	0.010	3.609	0.000 ***

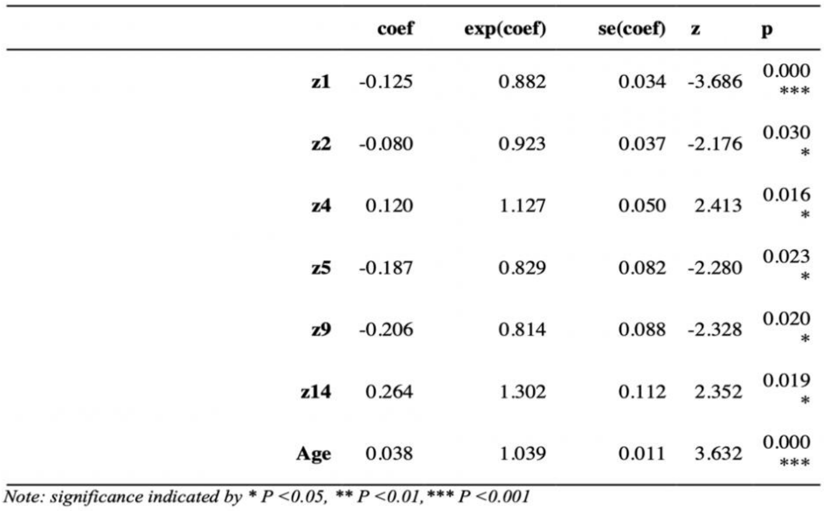
*Note: significance indicated by * P <0.05, ** P <0.01,*** P <0.001*

3.4 Principal Component Regression

In the principal component analysis conducted on the original list of 43 genes, the screeplot of the principal components versus variance shows that only the first three principal components need to be taken into consideration (see appendix C). However, when viewing the summary of the importance of components using the princomp function, the first three principal components only account for 36.8% of all the gene data which does not cover enough of my data to produce accurate results in the principal component regression model. Therefore, this study chose to take the first 15 principal components into consideration when creating the regression model as the first 15 principal components covers the majority, 75%, of my data.

The results of the regression model demonstrating the correlation between first 15 principal components, age, and survival show that not all components are significant. Principal component one is the most significant with a P values less than 0.001. Principal component four is the second most significant with a p values less than 0.01. Principal component two and five are significant with a p values less than 0.05. After implementing the stepwise approach to select the most significant variables to include, 15 principal components are narrowed down principal component one, two, four, five, nine and fourteen. Age as a variable also remains significant. In this model, principal component one and age remain to be the most significant with p values less than 0.001. The rest of the principal components are significant with p values less than 0.05. The details results are shown in the table below.

Table 5 Principal Regression Cox Model



	coef	exp(coef)	se(coef)	z	p
z1	-0.125	0.882	0.034	-3.686	0.000 ***
z2	-0.080	0.923	0.037	-2.176	0.030 *
z4	0.120	1.127	0.050	2.413	0.016 *
z5	-0.187	0.829	0.082	-2.280	0.023 *
z9	-0.206	0.814	0.088	-2.328	0.020 *
z14	0.264	1.302	0.112	2.352	0.019 *
Age	0.038	1.039	0.011	3.632	0.000 ***

*Note: significance indicated by * P <0.05, ** P <0.01,*** P <0.001*

Overall, the cox proportional hazards regression model has a concordance of 0.0692 and a standard error of 0.024. On the other hand, the principal component regression model has a concordance of 0.663 and a standard error of 0.027. The cox proportional hazards regression has a concordance closer to 0.07 and a smaller standard error compared to the principal component regression model. Therefore, this study will focus on the 6-gene cox proportional hazards regression model to predict overall survival instead of the principal component regression's model.

5. Discussion

Overall, the significant correlations between MCM2, MARKS and NRP1 and overall survival in GC identified in this study is in line with previous research. The overexpression of MARCKS and NRP1 is shown to be correlated with poor prognosis in GC (sun et al, Zhang et al, Quan et al, Dai et al, Huang et al). MCM2 have been consistently identified as an important gene across all cancer types (Yuan et al). C5 and CTS6 identified in this study, however, have not been identified repeatedly and researched on extensively in the field of GC before (Zhou et al, Wang et al).

MARCKS was identified to be a potential prognostic biomarker and a therapeutic target for GC patients. As a protein that is membrane-associated, MARCKS plays an important part in different cellular functions including cell motility, cytoskeletal control, motility and inflammatory pathways and can further increase metastasis, leading to high risk of death (Quan et al). Furthermore, MARCKS exacerbates GC and progression depended on the EMT pathway whose activation triggers metastasis (Quan et al). NRP1 is involved in tumorigenesis, development, invasion and metastasis of GC cells (Sun et al). In addition, research demonstrates that the overexpression of NRP1 is significantly correlated malignant phenotype of GC (Wang et al). According to Zhang et al study, the novel tumor-homing peptide iRGD can improve 5-FU (standard chemotherapy drug from metastatic GC) effect on GC through NRP1. Thus, NRP1 can be also potential therapeutic target. The findings of this study are consistent with these studies and research.

STK11 is a tumor suppressor gene. Thus, its high expression decreases the risk of death and is protective against metastasis, which is supported by the results of this study (Liang et al). However, it is important to note that STK11 is complicated by its potential indirect correlation with the risk of GC. The susceptibility to GC is influenced by germline genetic syndromes including juvenile polyposis syndrome, li-Fraumeni syndrome, Lynch syndrome, familial adenomatous polyposis and Peutz-Jeghers syndrome (Slavin et al). It is shown that a disrupted STK11 gene causes Peutz-Jeghers syndrome (Slavin et al). According to the NCCN clinical practice guidelines in oncology, Peutz-Jeghers syndrome increases the risk of GC up to 29%. Furthermore, STK11 disruption is also shown to be related to other types of cancers such as breast and pancreatitis cancer (NCCN). Thus, disrupted STK11 gene becomes risk as it is indirectly correlated with increasing risk of GC..

MCM2, minichromosomal maintenance 2, is a part of the monochromical group of proteins which focuses regulating DNA replication and cell cycle (Tsaniras et al, Fraggos et al, Li and Xu). Extensive amount of previous research show that MCM2 is important in cancer cell replication and the development of cancer across different types of cancers (Yuan et al). Furthermore, MCM2's gene expression is deeply associated with immune-related molecule expression and immune cell infiltration in various cancers, suggesting that MCM2 is a potential biomarker for immune therapy (Yuan et al). According to the findings of this study, MCM2 is identified to be significantly correlated with the overall survival in GC. This is consistent with previous research that demonstrates MCM2 as important gene across all types of cancer.

Although having been identified as associated with GC before, CTS6 and C5 identified in this study has not been extensive researched on in the field of GC (Zhou et al, Wang et al). CST6 was previously identified to be a critical gene for breast cancer. According to emerging research, high gene expression of CST6 is associated with poor prognosis in Triple-Negative breast cancer and is also correlated with lymph-node metastasis (Li et al). As a subtype of cystatin, CTS6 is shown to be significantly more up regulated in the TNBC tissues compared to healthy breast tissues (Li et al). This also appears to be the case for GC. According to the results of this study, as the gene expression of CST6 increases, the risk of death also increases which suggests that CST6 is a risky candidate gene promoting GC. Future research needs to be conducted to confirm the correlation between CTS6 and overall survival. In addition, the biological processes behind CTS6 that are associated with GC need to be further investigated.

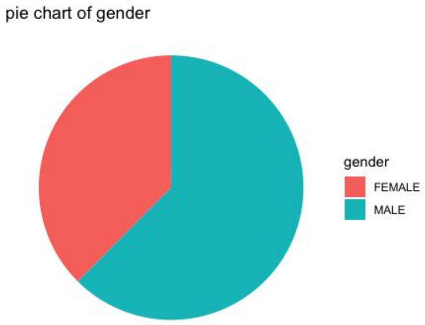
C5 gene have not been extensively identified in GC before, but however is shown to be associated with the tumor genesis and cancer progression in other types of malignant tumors (Zhou et al). C5 is shown to play a critical part in the coding complement system's components (DeMartino et al). The gene is also shown to be related to pathways such as immune response, GPCR signaling pathway and lectin-induced complement pathway (Zhou et al). It was suggested that the cell C5a release from C5 contributes to cancer progression due to its new mechanism of self-activation of C5aR-expressing cancer cells enhancing invasion and microenvironment favorable for the progression of cancer (Nitta et al). Because there are not enough reports on the role of C5 in GC, future research is needed to investigate and confirm its relationship with the overall survival and development of GC.

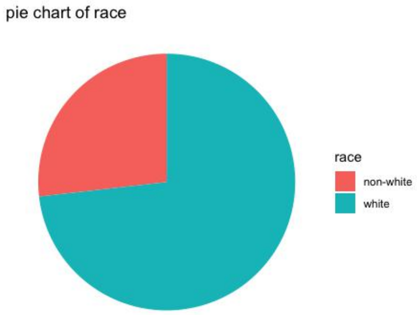
To evaluate the findings of this study, it is important to acknowledge that although TCGA is a relatively large database with new and accurate data, it is not representative of all GC cases across the world. For example, TCGA data tends to have a significant amount of data on white patients but less information on patients of other races. In addition, this study did not consider many of the non-omics risk factors that previous studies have shown to be associated with over survival such as dietary factors, radiation, and H. pylori infection because TCGA's information on other risk factors are limited. Therefore, the 6-gene prognostic model could have been more accurate if all potential risk factors are included. However, it is very difficult to have a complete set of data on every single factor that potentially influences a type of cancer. Therefore, despite the dataset does not include all non-omics risk factors, the results of the model are still important due to the new and big sample of genetic information provided in the TCGA database. The findings build onto our understanding of critical genes in GC.

Conclusion

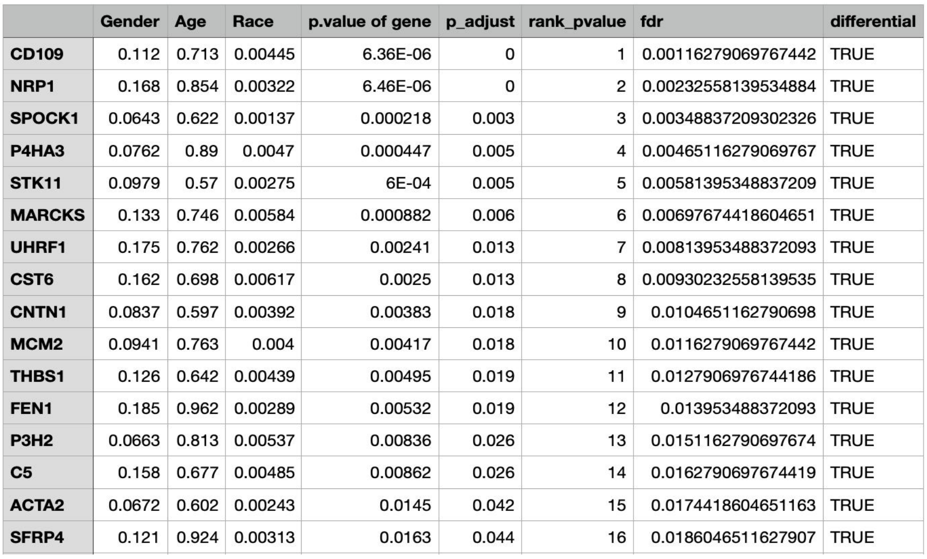
In summary, the 6-gene cox proportional hazards regression model was created using the data from TCGA, compared to the principal component regression model and was shown to be more accurate. Four genes (MARCKS, NRP1, STK11, MCM2) in the 6 gene model supports previous research on their correlation with survival overall survival in GC. Two comparatively novel gene (C5, CTS6) in the field of GC have also been identified. Further experimental studies should continue to investigate the relationship between novel genes and their association with GC and investigate the implication significant gene on potential treatments such as chemotherapy and immune therapy.

Appendix A



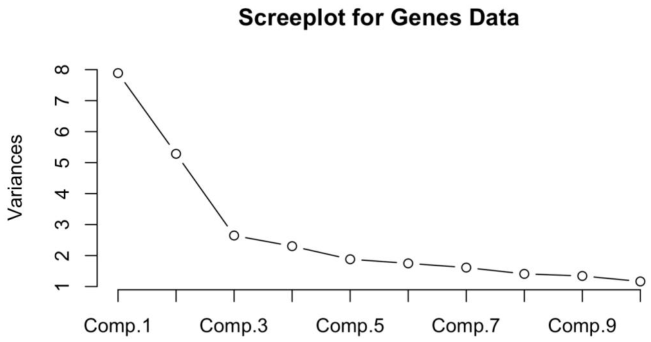


Appendix B



	Gender	Age	Race	p.value of gene	p_adjust	rank_pvalue	fdr	differential
CD109	0.112	0.713	0.00445	6.36E-06	0	1	0.00116279069767442	TRUE
NRP1	0.168	0.854	0.00322	6.46E-06	0	2	0.00232558139534884	TRUE
SPOCK1	0.0643	0.622	0.00137	0.000218	0.003	3	0.00348837209302326	TRUE
P4HA3	0.0762	0.89	0.0047	0.000447	0.005	4	0.00465116279069767	TRUE
STK11	0.0979	0.57	0.00275	6E-04	0.005	5	0.00581395348837209	TRUE
MARCKS	0.133	0.746	0.00584	0.000882	0.006	6	0.00697674418604651	TRUE
UHRF1	0.175	0.762	0.00266	0.00241	0.013	7	0.00813953488372093	TRUE
CST6	0.162	0.698	0.00617	0.0025	0.013	8	0.00930232558139535	TRUE
CNTN1	0.0837	0.597	0.00392	0.00383	0.018	9	0.0104651162790698	TRUE
MCM2	0.0941	0.763	0.004	0.00417	0.018	10	0.0116279069767442	TRUE
THBS1	0.126	0.642	0.00439	0.00495	0.019	11	0.0127906976744186	TRUE
FEN1	0.185	0.962	0.00289	0.00532	0.019	12	0.013953488372093	TRUE
P3H2	0.0663	0.813	0.00537	0.00836	0.026	13	0.0151162790697674	TRUE
C5	0.158	0.677	0.00485	0.00862	0.026	14	0.0162790697674419	TRUE
ACTA2	0.0672	0.602	0.00243	0.0145	0.042	15	0.0174418604651163	TRUE
SFRP4	0.121	0.924	0.00313	0.0163	0.044	16	0.0186046511627907	TRUE

Appendix C



References

[1] Arnold M et al. “Global Burden of 5 Major Types of Gastrointestinal Cancer.” Gastroenterology vol. 159,1 (2020): 335-349.e15.

[2] Bray F et al. “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.” CA: a cancer journal for clinicians vol. 68,6 (2018): 394-424.

[3] Busuttil RA, et al. “SFRP4 Drives Invasion in Gastric Cancer and Is an Early Predictor of Recurrence.” Gastric Cancer, vol. 24, no. 3, 2020, pp. 589–601.

[4] Cai L, et al. “The Better Survival of MSI Subtype Is Associated with the Oxidative Stress Related Pathways in Gastric Cancer.” Frontiers in Oncology, vol. 10, 2020.

[5] Champeris TS et al. “Licensing of DNA replication, cancer, pluripotency and differentiation: an interlinked world?.” Seminars in cell & developmental biology vol. 30 (2014): 174-80.

[6] Chen DH, et al. “SPOCK1 Promotes the Invasion and Metastasis of Gastric Cancer through Slug-Induced Epithelial-Mesenchymal Transition.” Journal of Cellular and Molecular Medicine, 2017.

[7] Cheong JH et al.” Development and validation of a prognostic and predictive 32-gene signature for gastric cancer”. Nature communications, 13, 774. 2022.

[8] Dai J et al. “Whole Genome Messenger RNA Profiling Identifies a Novel Signature to Predict Gastric Cancer Survival.” Clinical and translational gastroenterology vol. 10,1 (2019): e00004.

[9] DeMartino JA, et al. “The amino terminus of the human C5a receptor is required for high affinity C5a binding and for receptor activation by C5a but not C5a analogs.” The Journal of biological chemistry vol. 269,20 (1994): 14446-50.

[10] Eftang L, et al. “Up-regulation of CLDN1 in gastric cancer is correlated with reduced survival”. BMC Cancer 13, 586. 2013.

[11] Ferlay J, et al. “Cancer statistics for the year 2020: An overview.” International journal of cancer, 10.1002/ijc.33588. 5 Apr. 2021.

[12] Fragkos M, et al. “DNA replication origin activation in space and time.” Nature reviews. Molecular cell biology vol. 16,6 (2015): 360-74.

[13] Gu Y, et al. “Contactin 1: An Important and Emerging Oncogenic Protein Promoting Cancer Progression and Metastasis.” Genes, vol. 11, no. 8, 2020, p. 874.

[14] Huang HP, et al. “High expression of COL10A1 is associated with poor prognosis in colorectal cancer.” OncoTargets and therapy vol. 11 1571-1581. 20 Mar. 2018.

[15] Huang K, et al. “Identification of three predictors of gastric cancer progression and prognosis.” FEBS open bio vol. 10,9 (2020): 1891-1899.

[16] Huang YX, et al. “Identification of Novel Susceptible Genes of Gastric Cancer Based on Integrated Omics Data.” Frontiers in cell and developmental biology vol. 9 712020. 20 Jul. 2021.

[17] Ilic M, and Irena I. “Epidemiology of stomach cancer.” World journal of gastroenterology vol. 28,12 (2022): 1187-1203.

[18] Jin XF, et al. “P3H4 Overexpression Serves as a Prognostic Factor in Lung Adenocarcinoma.” Computational and Mathematical Methods in Medicine, vol. 2021, 2021, pp. 1–9.

[19] Jung H, et al. “The expression of claudin-1, claudin-2, claudin-3, and claudin-4 in gastric cancer tissue.” The Journal of surgical research vol. 167,2 (2011): e185-91.

[20] Junnila S, et al. “Gene Expression Analysis Identifies over-Expression ofcxcl1,Sparc,spp1, andsulf1in Gastric Cancer.” Genes, Chromosomes and Cancer, vol. 49, no. 1, 2010, pp. 28–39.

[21] Kim JM, et al. “Identification of Gastric Cancer–Related Genes Using a cDNA Microarray Containing Novel Expressed Sequence Tags Expressed in Gastric Cancer Cells.” Clinical Cancer research Vol. 11,2, 473–482. 15 Jan. 2005.

[22] Kyrlagkitsis I, and Karamanolis DG. “Genes and gastric cancer.” Hepato-gastroenterology vol. 51,55 (2004): 320-7.

[23] Li Q, et al. “Correlation of Cystatin E/M with Clinicopathological Features and Prognosis in Triple-Negative Breast Cancer.” Annals of clinical and laboratory science vol. 48,1 (2018): 40-44.

[24] Li Y, et al. “A Methylation‐Based Mrna Signature Predicts Survival in Patients with Gastric Cancer.” Cancer Cell International, vol. 20, no. 1, 2020.

[25] Li Z, and Xu XZ. “Post-Translational Modifications of the Mini-Chromosome Maintenance Proteins in DNA Replication.” Genes vol. 10,5 331. 30 Apr. 2019.

[26] Liang S et al. “Identification of Skt11-regulated genes in chondrocytes by integrated bioinformatics analysis.” Gene vol. 677 (2018): 340-348.

[27] Liu J, et al. “Identification of Critical Genes in Gastric Cancer to Predict Prognosis Using Bioinformatics Analysis Methods.” Annals of Translational Medicine, vol. 8, no. 14, 2020, pp. 884–884.

[28] Liu X, et al. “Cancer-Associated Fibroblast Infiltration in Gastric Cancer: The Discrepancy in Subtypes Pathways and Immunosuppression.” Journal of Translational Medicine, vol. 19, no. 1, 2021.

[29] Liu ZP, et al. “Prediction and prognostic significance of ALOX12B and PACSIN1 expression in gastric cancer by genome-wide RNA expression and methylation analysis.” Journal of gastrointestinal oncology vol. 12,5 (2021): 2082-2092.

[30] Lu Y, et al. “Diagnostic, Therapeutic, and Prognostic Value of the Thrombospondin Family in Gastric Cancer.” Frontiers in Molecular Biosciences, vol. 8, 2021.

[31] Ma H, et al. “Identifying of Biomarkers Associated with Gastric Cancer Based on 11 Topological Analysis Methods of CytoHubba.” Scientific Reports, vol. 11, no. 1, 2021.

[32] Machlowska J, et al. “Gastric Cancer: Epidemiology, Risk Factors, Classification, Genomic Characteristics and Treatment Strategies.” International journal of molecular sciences vol. 21,11 4012. 4 Jun. 2020.

[33] Machlowska J, et al. “High-Throughput Sequencing of Gastric Cancer Patients: Unravelling Genetic Predispositions towards an Early-Onset Subtype.” Cancers, vol. 12, no. 7, 2020, p. 1981.

[34] Marimuthu A, et al. “Gene Expression Profiling of Gastric Cancer.” Journal of proteomics & bioinformatics vol. 4,4 (2011): 74-82.

[35] Meng Q, et al. “DNA Methylation Regulator-Mediated Modification Patterns and Tumor Microenvironment Characterization in Gastric Cancer.” Molecular Therapy - Nucleic Acids, vol. 24, 2021, pp. 695–710.

[36] Nitta H, et al. “Cancer cells release anaphylatoxin C5a from C5 by serine protease to enhance invasiveness.” Oncology reports vol. 32,4 (2014): 1715-9.

[37] Niu XJ, et al. “Identification of Potential Diagnostic and Prognostic Biomarkers for Gastric Cancer Based on Bioinformatic Analysis.” Frontiers in genetics vol. 13 862105. 16 Mar. 2022.

[38] Quan RL, et al. “Prognostic Value of Upregulation of Myristoylated Alanine-Rich C-Kinase Substrate in Gastric Cancer.” Medical science monitor: international medical journal of experimental and clinical research vol. 25 279-287. 9 Jan. 2019.

[39] Shim HJ, et al. “BRCA1 And XRCC1 Polymorphisms Associated with Survival in Advanced Gastric Cancer Treated with Taxane and Cisplatin.” Cancer Science, vol. 101, no. 5, 2010, pp. 1247–1254.

[40] Slavin T, et al. “Genetics of gastric cancer: what do we know about the genetic risks?.” Translational gastroenterology and hepatology vol. 4 55. 29 Jul. 2019.

[41] “Stomach (Gastric) Cancer Key Statistics.” American Cancer Society, [Internet], Available from: https://www.cancer.org/cancer/stomach-cancer/about/key-statistics.html.

[42] Sun LP, et al. Ai zheng = Aizheng = Chinese journal of cancer vol. 23,1 (2004): 36-9.

[43] Sun MY, et al. “Prognostic Implications of Novel Gene Signatures in Gastric Cancer Microenvironment.” Medical science monitor : international medical journal of experimental and clinical research vol. 26 e924604. 2 Aug. 2020.

[44] Szász AM, et al. “Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients.” Oncotarget vol. 7,31 (2016): 49322-49333.

[45] Vetrivel P, et al. “Investigation on the Cellular Mechanism of Prunetin Evidenced through next Generation Sequencing and Bioinformatic Approaches against Gastric Cancer.” Scientific Reports, vol. 12, no. 1, 2022.

[46] Wang GH, et al. “Hypomethylated gene NRP1 is co-expressed with PDGFRB and associated with poor overall survival in gastric cancer patients.” Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie vol. 111 (2019): 1334-1341.

[47] Wang GG, et al. “The prediction of survival in Gastric Cancer based on a Robust 13-Gene Signature.” Journal of Cancer vol. 12,11 3344-3353. 12 Apr. 2021.

[48] Wang P, et al. “A novel gene expression-based prognostic scoring system to predict survival in gastric cancer.” Oncotarget vol. 7,34 (2016): 55343-55351.

[49] Pan YL et al. “KCNE2, a down-regulated gene identified by in silico analysis, suppressed proliferation of gastric cancer cells.” Cancer letters vol. 246,1-2 (2007): 129-38.

[50] Yuan J, et al. “Multi-Omics Analysis of MCM2 as a Promising Biomarker in Pan-Cancer.” Frontiers in cell and developmental biology vol. 10 852135. 25 May. 2022.

[51] Zeng DQ, et al. “Tumor Microenvironment Characterization in Gastric Cancer Identifies Prognostic and Immunotherapeutically Relevant Gene Signatures.” Cancer Immunology Research, vol. 7, no. 5, 2019, pp. 737–750.

[52] Zeng Z, et al. “Genome‐Wide Identification of CPG Island Methylator Phenotype Related Gene Signature as a Novel Prognostic Biomarker of Gastric Cancer.” PeerJ, vol. 8, 2020.

[53] Zhang L, et al. “Combination of NRP1-mediated iRGD with 5-fluorouracil suppresses proliferation, migration and invasion of gastric cancer cells.” Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie vol. 93 (2017): 1136-1143.

[54] Zhang YW, et al. “Multi-Omics Data Analyses Construct TME and Identify the Immune-Related Prognosis Signatures in Human LUAD.” Molecular Therapy - Nucleic Acids, vol. 21, 2020, pp. 860–873.

[55] Zhen Yuxu et al. ”Gene expression profile towards the prediction of patient survival of gastric cancer, Biomedicine & Pharmacotherapy, Volume 64, Issue 2, 2010, Pages 133-139.

[56] Zhou LQ, et al. “Establishment of a prognostic model of four genes in gastric cancer based on multiple data sets.” Cancer medicine vol. 10,10 (2021): 3309-3322.

[57] Zhuo et al. “Elevated THBS2, COL1A2, and SPP1 Expression Levels as Predictors of Gastric Cancer Prognosis”. Cell Physiol Biochem vol 40, 1316-1324. 2016.

About the author: Zihan Fang (2004.01), female, Han nationality, Beijing native, Student,High school degree, Research area: medical.