IN SILICO VALIDATION OF PROGNOSTIC mRNA SIGNATURE IN GASTRIC CANCER AND IDENTIFICATION AND VALIDATION OF NOVEL GASTRIC TISSUE SPECIFIC REFERENCE GENES FOR QUANTITATIVE PCR

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING
AND SCIENCES OF BILKENT UNIVERSITY
IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN
MOLECULAR BIOLOGY AND GENETICS

By

Marzana Ishraq

July 2021

ABSTRACT

***IN SILICO* VALIDATION OF PROGNOSTIC mRNA SIGNATURE IN GASTRIC CANCER AND IDENTIFICATION AND VALIDATION OF NOVEL GASTRIC TISSUE SPECIFIC REFERENCE GENES FOR QUANTITATIVE PCR**

Marzana Ishraq

M.Sc. in Molecular Biology and Genetics
Advisor: Ali Osmay Güre

July 2021

Gastric adenocarcinoma is a molecularly and histologically heterogeneous neoplasm with a predictively disastrous outcome if undiagnosed at early stages. Amidst global declines of gastric cancer rates, it remains the 5th most common malignancy with the 4th worst outcome of all cancers. Predictive and prognostic clinical biomarkers are scarce in gastric cancer due to its immense heterogeneity. Identifying novel biomarkers for gastric cancer is an emerging field. Previously in our lab, a 20 gene mRNA signature was developed which successfully stratified gastric cancer patients into poor and good prognosis. In this thesis, we attempted to shorten this list to a five gene signature which will successfully stratify patients into similar clusters as the 20 genes. The 5 genes being, HEYL, CALD1, ACTA2, TAGLN and TPM2. Moreover, we validated the efficacy of our 5-gene signature to stratify patients based on their prognoses *in silico.*

The second half of the thesis focuses on identifying a set of gastric tissue specific reference genes by analyzing high-throughput gene expression data. Most commonly used reference genes are known to have varying expression in cancer tissue which often leads to an issue with reproducibility in cancer research. We aimed to design an algorithm which identifies a list of stable transcripts within a particular tissue type. We identified 3 genes EWSR1, SF1 and HNRNPK which showed stable expression throughout gastric tissue, both in normal and cancer *in silico. Ex vivo* validation experiments using gastric tumor and adjacent normal RNA show promise for the efficacy of our genes compared to GAPDH and B2M.

Keywords: Gastric cancer, prognosis, biomarker, patient stratification, reference gene

ÖZET

GASTRİK KANSERDE PROGNOSTİK mRNA İMZALARININ IN SILICO OLARAK DOĞRULANMASI VE KANTİTATİF PCR İÇİN GASTRİK DOKU SPESİFİK, ÖZGÜN REFERANS GENLERİN TANIMLANMASI VE DOĞRULANMASI

Tez Danışmanı: Prof. Dr. Ali Osmay Güre

Gastrik adenokarsinom erken evrelerinde teşhis edilmediğinde feci sonuçlara neden olabilen, moleküler ve histolojik olarak heterojen bir neoplazmdır. Gastrik kanser oranlarındaki küresel düşüş içerisinde, gastrik adenokarsinom tüm kanser türleri arasında en yaygın 5.en kötü sonucu olan 4. malignite olmaya devam etmektedir. Muazzam heterojenliği nedeniyle gastrik kanserde öngörücü ve prognostik biyobelirteçler oldukça limitli sayıdadır. Gastrik kanser için özgün biyobelirteçlerin tanımlanması gelişmekte olan bir alandır. Daha önce laboratuvarımızda, gastrik kanser hastalarını kötü ve iyi prognozlar olarak başarıyla sınıflandıran 20 gen mRNA imzası geliştirilmiştir. Bu tez çalışmasında, 20 genden oluşan bu listeyi, hastaları benzer kümelere başarılı bir şekilde sınıflandıracak beş gen imzasına indirgemeye çalıştık. Bu 5 gen HEYL, CALD1, ACTA2, TAGLN ve TPM2 olarak bulgulandı.

Tez çalışmasının ikinci yarısı, gastrik dokuya özgü bir dizi referans genin yüksek verimli gen ifadelenmesi veri analizi ile tanımlanmasına odaklanmaktadır. En yaygın kullanılan referans genler, kanser dokularında değişken ifadeye sahiptir ve bu kanser araştırmalarında sıklıkla tekrarlanabilirlik ile ilgili soruna yol açmaktadır. Belirli bir doku tipindeki stabil transkriptlerin listesini tanımlayan bir algoritma dizayn etmeyi amaçladık. Hem normal hem de kanserli gastrik dokuda stabil ifadelenme gösteren 3 geni, EWSR1, SF1 ve HNRNPK olarak, in silico çalışmalarla belirledik. Gastrik tümöre ve bitişik normal dokuya ait RNA kullanılarak gerçekleştirilen ex vivo doğrulama deneyleri, GAPDH ve B2M ile karşılaştırıldığında bu 3 genin etkinlikleri açısından umut vaad ettiğini göstermektedir.

Anahtar Kelimeler: Gastrik kanser, prognoz, biyobelirteç, hasta sınıflandırması, referans gen

"For those who have inspired me in my toughest times..."

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Dr. Ali Osmay Güre for his invaluable guidance and his endless support. He taught me to be more persistent, and inquisitive; his calm and methodical approach to science taught me lessons which I'll carry for the rest of my academic life. Then, I owe a great deal of gratıtude to Dr. Seçil Demirkol Canlı. I owe her immensely for her unending patience and guidance throughout my master's studies. Moreover, I would like to extend my gratitude to the jury members Dr. Sreeparna Banerjee and Dr. Özlen Konu for their valuable feedback on my work. Their compassionate support and critique have helped me immensely in improving my thesis.

Science is all about teamwork and I was blessed to have an amazing team of young, brilliant, and inspiring scientists to work alongside. I am greatly indebted to my AOG family for all their help and friendship throughout the years. I mean it sincerely when I say, they were the best lab I could ask for; and I truly appreciate all the help they have given over the years. I am particularly grateful to Barış Küçükkaraduman, Ege Dedeoğlu, Noor Niaz, Ronak Naeemaee, Shila Azizoli, Isli cela, Farid Ahadli, Muhammad Waqas Akbar for all their help throughout my master's studies.

I will always be grateful to all my family members who believed in me and lifted my spirits throughout the best and worst of days. I thank my father, Dr. Kamal Ahmed Chowdhury, my mother, Shareefa Yasmeen, and my loving brother Aqiq Ishraq for their encouraging words and support. Most of all, I will always cherish the moments I spent on the phone with my grandmother, Dr. Wazifa Ahmed during my thesis. Her kind voice and loving words helped me through the toughest days during quarantine and if it weren't for her, the past year would be exponentially harder than it was. I feel fortunate to have had family beside me throughout this journey, having Aqiq and his wife, Nazifa Tasnim Arony here with me made my bad days much more bearable.

"If you want to make life easy, make good friends." I can attest to the truth of this statement. The confined walls of Bilkent Molecular Biology lab can be extremely isolating, it can leave the best of us in despair. It's a tough life to lead without good friends; and that's where I was incredibly lucky. I have had the pleasure of meeting some of the most wonderful and inspiring folks one can ever hope to meet. Every group of friends come together around a shared interest and for us it was food and coffee. The long winded discussions we've had about foods, both local and exotic, is what defined us. The memories we've made sharing conversation over coffee, pondering over the absurdities of adult life, sharing our passions for arts and science with one another, are what I'll cherish for the rest of my life. Javaid Jabbar, Beste Uygur, Büşra Korkmaz, Melike Aslan, Hazal

Beril Çatalık, Burçin Arıcı, Noor Niaz, I thank them from the bottom of my heart for being there for me. Each and every one of them profoundly impacted my life and I am a better person today because of them.

I thank Javaid for being the calm, objective voice, who was always honest but never unkind. He always lent an ear to me and always had a good solution to whatever obstacle I faced. During the lockdown era of 2020 where take-away were prohibited, we all had him to thank for preparing delicious meals for us. That's a debt I can't ever repay.

Beste I thank for being a bastion of support and for always believing in me, the little notes of inspiration and the surprise cups of coffee she left around my desk brightened up so many of my days. Not many I know share my dry sense of humor, but Beste did. Our witty backs and forth were often the only bright sides to our days or even weeks; and that's what I'll miss the most about our friendship.

Melike is one of the most helpful, and patient people I know. I learned a lot from her, from coffee beans to photography to art, she's a maestro. And she's an excellent friend to top it all off, our philosophically charged conversation is something I'll take with me wherever I go.

Burçin, I thank for spreading an air of positivity only she can spread. Throughout all the uncertainties of the past year, she stayed as a bright beacon of hope in all our lives. She motivated me to keep working and always helped me when I doubted myself, most importantly, she understood my vague pop-culture references.

Those of us who are most comfortable with themselves are the best company, and Beril knew who she was, and that's why spending time with her was so much fun. We could talk for hours about the most random things without boring each other out, that's the mark of a great friendship. Noor is a very poised, gentle lady who I wish I had started to hang out with sooner. Her great sense of humor, proclivity for all things fiction made her an absolute delight to be around. As Bilkenters we define most of our friendships by the walks we take around the campus. And the walks we took together, stopping for coffee and pictures were just perfect. Talking for hours about arts, culture and just ridiculous life and humans can be, was what made our friendship great.

The best surprise of my master's here was me reuniting with my friend, Buse. It's hard to believe it's almost been 10 years since we first became friends. We've known each other since we were

freshmen and I feel incredibly lucky to have her in my life. Her cheerful, and positive attitude and her propensity to always plan spontaneous visits despite her busy schedule made many of my days.

I am forever grateful to my dearest friend Büşra for being an amazing friend and housemate. She was the absolute best quarantine buddy, all the talks we've had about movies, life and the MBG department are precious memories I'll preserve for the rest of my life. I feel extremely lucky to have such an understanding and compassionate friend, she's been there for me like a sister and I'll always be indebted to her for that.

Lastly, I would like to thank Bilkent University Molecular Biology department for the funding and support it has provided for the duration of my studies.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Abbreviations

TCGA- The Cancer Genome Atlas

CCLE- Cancer Cell Line Encyclopedia

CV- Coefficient of Variance

SVM- Support Vector Machines

BCCP- Bayesian Compound Covariate Predictor

ML- Machine Learning

RG- Reference genes

K-M – Kaplan-Meier

LRMC- Log rank based multiple cut-off

mUAST- Multivariate analysis Unsupervised software analysis tool

qPCR- Quantitative Polymerase Chain Reaction

RT-qPCR- Real Time quantitative Polymerase Chain Reaction

Chapter 1
Introduction

The human stomach is among the most well-studied and heavily debated organs throughout history. Records of speculating and studying this particular organ in the viscera go back to ancient Greece and Egypt. The process of digestion has eluded ancient scholars for many years. Doctors in ancient Greece believed that food was heated or concocted in the body to form chyle, a mixture which was later converted into the four humours, which were blood, phlegm, yellow and black bile. This belief was later incorporated into European medicine; it remained within the academic literature and clinical practice late until the 16th century.

Paracelsus, a Swiss alchemist, was the first to postulate that stomachs of some animals contain acid which aids in digestion. Later on, Jean-Baptiste Van Helmont, known to be the last alchemist and first biochemist, postulated that digestion is aided by stomach acid and a ferment which was later known to be digestive enzymes [18-19].

The human digestive system comprises various glands, which secret digestive juices, and the gastrointestinal tract. The stomach is an indispensable organ in this system and it primarily serves two purposes, aiding digestion and sterilizing the ingested food via stomach acid. The stomach can be divided into five main parts, cardia, body, fundus, pylorus, and antrum [17].

The first three parts make up the proximal stomach. Cardia is at the junction between the esophagus and the stomach; fundus is the upper part of the stomach next to the cardia; body, also known as the corpus, is the main and biggest part between the fundus and antrum. The lower part of the stomach is divided into two parts and is called the distal stomach. The two parts are the antrum, the lower part of the stomach, and the pylorus, which is the junction between the lower end of the stomach and the duodenum.

The stomach wall has five main layers, mucosa, submucosa, Muscularis propria, suberosa and erosa. The mucosa is the innermost layer' this is the layer where the secretory cells are present.

The submucosa is a supporting layer followed by muscularis propria, a thick wall of muscle cells that aid in the churning motions. The outermost layers, suberosa, and Erosa line the inner three layers and wrap the stomach [17]. These layers in the stomach walls are indicative of the cancer stage, the deeper into the layers cancer spreads, the worse the prognostic outcome. Patients with cancers in the deeper layers also require more prolonged and extensive treatment options.

The most common type of cancer in the stomach is gastric adenocarcinoma. Approximately 90-95% of stomach cancer cases globally are stomach adenocarcinomas. Adenocarcinomas originate from the secretory cells from the innermost lining, the mucosa [19].

A much less common variant of stomach cancer is a gastrointestinal stromal tumor (GIST), which also originates from the mucosa. A gain-of-function mutation in c-KIT (a tyrosine receptor kinase) is linked with developing GISTs [21]. GISTS are comparatively rare and less common among gastric cancer patients, with an average of 3% of all gastric cancer incidence globally [20].

Stomach adenocarcinomas are largely variable in response to therapy and prognosis and have variations in their histology and etiology [20]. Identifying an mRNA-based prognostic biomarker list that is effective across all histological subtypes is paramount to prognosticate gastric patients in clinics.

In this work, we have attempted to identify and curate a list of genes to use as a prognosis signature in gastric adenocarcinoma patients. In addition, we attempted to identify a list of stable transcripts in gastric tissue which can be used as reference genes for quantitative PCR.

1.1 Global trends in gastric cancer incidence rates and its statistics

Gastric cancer is the top fifth malignancy globally with the 4th highest mortality [1]. Among cancer-related deaths, gastric cancer remains high despite a global effort to reduce incidents by higher rates of regular screening and eradicating H. pylori infections. According to the World Health Organization, the rate of gastric cancer has decreased significantly over the past decade due to increased screening and primary care. In the span of the last 70 years, the decrease in gastric cancer rates has been quite dramatic. Nevertheless, it remains a frequently occurring cancer with potentially disastrous consequences for the patient [2].

Gastric tumors occur in two primary locations, proximal (cardia) and distal (non-cardia), although the rates of proximal gastric cancers have gone down in the past decades, the rate of distal gastric cancers has been rising steadily [6, 7].

The most well-known histological classification for gastric adenocarcinoma is the Lauren classification which divides the tumors into two groups, intestinal and diffuse. The intestinal-type tumors are known to be well-differentiated whereas the diffuse-type tumors are less differentiated or undifferentiated [8].

Diffuse-type tumors are not known to be specific to any particular demography, however, intestinal tumors are more prevalent in East Asia, South America, and Eastern Europe [9, 10].

The main reason why gastric cancer incidence has gone down over the past years is that intestinal tumors are becoming less common, however, the cases of diffuse-type tumors are on the rise globally [10]. The intestinal type appears to be more common in men than women, with a ratio of 2:1. Diffuse-type tumors however do not show any distinct trend in either sex. However, it does have an earlier onset in patients than the intestinal type [13]. A time-trend analysis performed in Japan 1975-1989 has shown that an older population is more prone to an H. pylori-associated intestinal-type gastric carcinoma than younger populations. The study also found that females are more likely to develop diffuse-type carcinoma than males [14].

The aetiological differences between histological subtypes likely explain the age-sex-specific differences in gastric cancer occurrences.

The survival outcome of gastric patients is on average five years [2]; and since the malignancy does not exhibit any symptom early on, in countries where regular screening is rare, many cases are diagnosed in later stages of the disease. Moreover, it is an increasingly complex and largely heterogeneous disease that has posed a challenge in terms of molecular characterization. Stomach cancers are multifactorial which means their pathogenesis depends on environmental and genetic factors.

1.2 Histological subtypes of stomach cancers

Histologically, gastric tumors have significant heterogeneity in their cellular architecture and cytology, making the classification of tumors in the clinical setting very tenuous for clinicians [3]. As of now, there are two main histological classification systems for gastric cancer, Lauren classification and the more recent system developed by the World Health Organization (WHO). The Lauren classification was developed by Pekka Lauren and was published in 1965.

In this classification, gastric adenocarcinoma is categorized into two main groups, intestinal type, and diffuse type. Besides these two groups, Lauren classified a third group, the indeterminate type which is comparatively rare in occurrence. The intestinal-type tumors are known to be well-differentiated whereas the diffuse-type tumors are less differentiated or undifferentiated [8]. Approximately 54%, 32%, and 15% of gastric tumors fall into the groups intestinal, diffuse, and indeterminate respectively [15].

The World Health Organization in 2010 passed a newer more elaborate histological classification system for gastric carcinoma which has four major groups, tubular, mucinous, papillary, and poorly cohesive. Poorly cohesive tumors include signet ring cell carcinoma. Besides the four major groups, the WHO categorized gastric tumors into several other groups which are less common

histological variants [16, 22]. Among the four major groups, tubular carcinoma and papillary adenocarcinoma are frequently occurring in early gastric carcinoma. Moreover, the papillary type is more common among the aging population. Approximately 10% of gastric adenocarcinoma cases are known to be Mucinous carcinoma, these tumors have characteristically large mucinous pools which can take up to 50% of the tumor volume [22].

Due to histological and molecular diversity in gastric adenocarcinoma, it's very difficult to assess the prognosis of individual patients in a clinical setting by their histological grouping.

Efforts have been taken in molecularly classifying gastric tumors based on mutation profiles and gene expression. Two main studies which undertook these efforts were the ones by The Cancer Genome Atlas and the Asian Cancer research institute [23, 24].

1.3 Molecular subtypes of gastric cancer

In 2014, The Cancer Genome Atlas research group published a paper in which they successfully categorized gastric adenocarcinoma into 4 molecular subgroups. In group molecularly profiled 295 primary cancers with six molecular platforms [23].

The team has categorized the neoplasms into four groups: Epstein-Barr virus-positive, microsatellite unstable tumors, genomically stable tumors, and tumors with chromosomal instability. In order to perform molecular characterization, six molecular platforms were used to analyze a total of 295 tumor samples. The platforms are as follows, array-based DNA methylation profiling, array-based somatic copy number analysis, whole-exome sequencing, microRNA sequencing, messenger RNA sequencing, and reverse-phase protein array. Approximately 77% of the samples were analyzed by all six platforms. Unsupervised clustering analysis was done with the results from these six platforms to categorize the samples into the four aforementioned groups.

Around 9% of gastric cancer cases are EBV positive with its presence in malignant epithelial cells. As previously identified, EBV-positive tumors are more prevalent in males [25]. In the TCGA study, unlike EBV, the presence of H. pylori in the tumors was relatively sporadic, thus the molecular signatures of the tumor could not be classified based on the bacterial infection. The authors concluded that this is possibly a result of increased screening and eradication of H. pylori in recent years. Moreover, they postulated that the bacteria may have been lost from luminal samples during specimen processing.

Unsupervised clustering of CpG methylations by the authors identified that the EBV positive tumors clustered in one group, moreover, they exhibited extreme CpG island methylator phenotype. The authors reported CDKN2A promoter to be heavily methylated in EBV-positive tumors. Prior studies indicated a link between PIK3CA mutation and EBV-positive gastric cancers, TCGA study found non-silent PIK3CA mutations in 80% of the EBV-positive subgroup (P= 9x10-12). Hence, the use of PI (3)-kinase inhibitors in EBV-positive gastric cancer patients should be further evaluated.

The microsatellite instability group was shown to have a high mutational burden and hypermethylation, particularly in the ML1H1 region. The two other groups, genomically stable and chromosomal instability were selected based on the presence or absence of somatic copy number alterations. The genomically stable group heavily coincided with the diffuse histological type whereas the chromosomal instability subgroups heavily aligned with the intestinal histology [23].

An independent study estimated the prognostic benefits of these four molecular subgroups in two separate datasets. The authors successfully demonstrated that the EBV-positive subgroup has the best prognostic outcome and the genomically stable group has the worst. The remaining two groups have a comparatively better prognosis than GS, but worse than EBV-positive [26].

The ACRG study, primarily categorized GC into two groups, microsatellite stable and microsatellite unstable. MSS is further categorized into MSS/EMT, MSS/TP53- and MSS/TP53+. When these subgroups were correlated with clinical cofactors a pattern emerged. The MSS/EMT subtype occurred more frequently in younger patients, moreover, the tumors in this group were predominantly diffuse-type. This group also had the worst prognostic outcome of the four. Like the MSS/EMT subtype, the MSI subtype also was more prevalent in younger patients, however, over 60% of tumors from this subtype were intestinal-type and originated from the antrum. This subtype unsurprisingly had the best prognosis. EBV was more prevalent in the MSS/TP53+ subtype [24].

The ACRG and TCGA subtypes have several features in common, for example, the tumors that are MSI were common in both datasets. The GS, EBV+, and CIN subtypes were enriched in the MSS/EMT, MSS/TP53+, and MSS/TP53- groups respectively. Even then, there are some marked differences between the cohorts. The TCGA cohort had a higher number of CDH1 mutations compared to the ACRG, moreover, the TCGA cohort had a less heterogeneous pool of diffuse-type tumors whereas the diffuse-type in ACRG were significantly more diverse. There have been

other studies published that attempted to molecularly classify gastric cancer, [27-41] however, the ACRG and TCGA studies have been the most comprehensive thus far.

1.4 Gastric cancer risk factors

The key contributors to stomach cancer development are, excess salt intake, smoking, excess alcohol consumption, diet, family history and, Helicobacter pylori and Epstein-Barr virus infections [4].

The pathogenesis after EBV-infection which leads to gastric cancer carcinogenesis is vaguely understood [43]. EBV has been shown to be present in malignant epithelial cells. In a recent study, gastric mucosa collected from healthy individuals has been shown to be free of EBV whereas individuals with gastritis had EBV-RNA in their gastric mucosa. This indicates that EBV infection may be associated with persistent inflammation in the gastric mucosa which consequently results in carcinogenesis [44].

Among all frequently occurring cancers, gastric cancer has the highest occurrence due to infectious agents [45]. The gradual fall in gastric cancer incidences, especially in East Asia is mainly due to an attempt taken to screen individuals for H.pylori infections. Over the years, an improvement in food preservation, hygiene and diet, has led to a decrease in gastric cancer. Despite all efforts, 50% of the global population is affected by H.pylori infections [46]. There is a complex interaction between the host, H. pylori, and the environment at play in the stomach which results in gastric carcinogenesis. This interaction is not fully understood. In some parts of the world like China, higher H.pylori infections have been correlated to higher gastric cancer-related mortality. However, in Africa and South Asia, gastric cancer incidences are relatively low, yet a large part of the populations in these regions are positive for H. pylori infections. This phenomenon is called the African or Asian enigma. This is proof that the host-bacteria-environment interaction is extremely complex and environmental factors such as diet may play a more significant role than we anticipated [47].

Gastric cancer is significantly more likely to affect people older than 50. The median age of gastric cancer diagnosis was 70 in the United States 2005-2009. During this time, only 1% of the population aged between 25 and 34 were diagnosed with gastric cancer, as opposed to 29% of the patients who were aged between 75 and 84 [48]. Besides age, sex seems to be a major factor in the incidence and severity of gastric cancer. Males are affected 2 fold more than females for non-cardia gastric cancer and 5 fold for cardia gastric cancer [49]. The cause of this disparity is still under scrutiny and there is insufficient evidence to draw a conclusion. Historically it was suggested

that the disparity may be due to increased tobacco smoking by males and occupational hazards. However, recent studies which looked at countries where there is no key difference in smoking patterns between men and women, have found that males are more likely to be diagnosed with gastric cancer there as well [50].

In 2002 the International Agency for Research on Cancer declared that smoking is a possible risk factor for cardia and non-cardia gastric cancer [51]. There have been numerous studies that investigated the effects of smoking cigarettes and other tobacco-based products like hookah on developing gastric cancer and the results were not always consistent [52-55]. However, there is enough evidence in the literature to suspect that smoking is in fact a risk factor for gastric carcinogenesis.

Diet plays a crucial role in gastric cancer pathogenesis. In areas of the world where diets are high in salt, contain foods preserved in salt, and use certain cooking techniques, gastric cancer incidences are higher. Foods high in salt have been declared a risk factor for developing stomach cancer by the World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) [56]. Salt may directly damage the gastric mucosa which results in gastric dysplasia [57]. Moreover, the formation of polycyclic aromatic hydrocarbons in smoked meat and fish has been implicated as a risk factor [58].

Cooking techniques such as broiling, roasting, grilling, and baking may cause the formation of N-nitroso compounds which are also associated with gastric cancer pathogenesis [59].

1.5 Family history and gastric cancer

In families where parents or close relatives had gastric cancer, the likelihood of younger members of the family being diagnosed with this cancer is very high; the chances of people being diagnosed with gastric cancer where close relatives fell victim to the disease is 2-3 fold higher than the average person [70-71].

A significant number of gastric cancer cases, 10%, show familial aggregation (meaning it’s highly probable to occur in siblings and offspring) [72]. Inherited gastric cancers with a Mendelian inheritance pattern are very rare, studies show only 3% of total cases are due to Mendelian inheritance [69]. There are three main autosomal dominant gastric cancers, hereditary diffuse gastric cancer (HDGC), gastric adenocarcinoma and proximal polyposis of the stomach, and familial intestinal gastric cancer [72]. These gastric cancer incidences have a strong correlation to hereditary factors. Hereditary diffuse-type gastric cancer is due to a mutation in the gene CDH1

which is an autosomal dominant mutation. HDGC is a highly invasive, diffuse-type gastric cancer with a poor prognosis [64]. The loss of CDH1 causes defective or no E-cadherin to be produced. E-cadherin is a cell adhesion protein and in its absence, the intracellular adhesion is poor. Around 25% of patients with HDGC have inactivating CDH1 germline mutations [65]. Mutations in alpha E-catenin (CTNNA1) were recently shown to be a causative factor in HDGC development [73].

Among the three inherited gastric cancers, only HDGC can be screened in clinics. The genetic information of patients and their families is sequenced in a clinical setting to identify mutations in CDH1 and CTNNA1 to diagnose and assess risk. Currently, screening tests are not available for gastric adenocarcinoma and proximal polyposis of the stomach, familial intestinal gastric cancer [72].

Genetic risk factors associated with stomach cancer are highly debated and currently under extensive scrutiny. Multiple Genome-Wide Association studies performed in recent years have shed some light on possible genetic factors which might lead to stomach carcinogenesis. Several single nucleotide polymorphisms (SNP) were associated significantly with gastric cancer incidence. SNPs in prostate stem cell agent (PSCA) and Mucin 1 (MUC1) came up significantly in several independent GWAS in East Asia. GWAS results from other parts of the world are yet to be published, however, there appears to be a significant correlation between PSCA and MUC1 with certain gastric cancers. The biological mechanisms underlying these associations are still unknown. Nevertheless, once these mechanisms are discovered, they will shed some light on the complex pathogenesis of gastric cancer [66-68].

Earlier studies have found that gastric cancer incidence does not decrease in immigrant families after they move to the host country. The incidence rate remained comparable to the home country's rates until the second generation. One possible explanation for this pattern is that most H. pylori infections take place in early childhood; thus children who move to a new country already have the infection. Moreover, there is a high possibility that the children born after their families immigrate are also likely to pick up the infection from their families [60-63].

Genetic implications which lead to gastric cancer are poorly understood, and no screening is available in clinics except for HDGC. However, there have been several genes implicated in gastric cancer pathogenesis in recent years. These work in concert with environmental factors such as H.pylori infection to promote carcinogenesis. A series of low-penetrance alleles of tumor necrosis factor a (TNF α), interleukin 6 (IL-6), IL-8, IL-10, IL-1b, transforming growth factor b (TGFβ),

and toll-like receptor 4 (TLR4) have been implicated in gastric cancer to have compounding effects in developing a metastatic tumor [74-76].

Even though exact genetic causes of stomach adenocarcinoma are hard to define, numerous studies indicate a strong correlation between certain environmental factors and familial aggregation. For example, alcohol consumption has been shown to significantly increase the risk of gastric cancer in family members who have more than one relative with cancer, whereas alcohol consumption had no significant increase in risk in people with one relative suffering from gastric cancer [77].

Intestinal metaplasia is a condition where the cells in the stomach and esophageal lining take on the morphology of intestinal cells. It is widely accepted that IM is triggered by external stimuli. In many cases, IM is considered precancerous growth which eventually gives rise to dysplasia and later neoplastic growth [78]. However, genetic factors may affect IM as well. A study performed by Kim et al. showed that risk factors for IM include smoking, H.pylori infection, being older than 61, consumption of spicy food, and the SNP IL-10-592 C/A [79]. Large-scale genetic studies are required to explain and define concrete links between gastric cancer and genetic variants. This task proves very challenging due to the complicated intermingling of environmental and genetic factors, however, better clinical data collection protocols and high throughput sequencing of patients can hopefully generate the answers.

1.6 Treatment strategies and Clinical biomarkers of gastric adenocarcinoma

Major strides have been taken in identifying accurate predictive and prognostic biomarkers in recent years. However, there are still many gaps in the treatment plan of GC patients in most clinics. Due to the mutational and molecular heterogeneity of gastric adenocarcinoma it is very hard to predict an accurate treatment plan for individual patients. Moreover, most of these potential therapeutic options do not improve patient outcomes [109].

Very few GC markers have been clinically approved so far and even few have prognostic potential. Gastric cancer biomarkers which are in clinical use right now and those which hold potential are described below.

1.6.1 HER2

Much like many breast cancer cases, there are several human epidermal growth factors receptor 2 (HER2) positive gastric and gastroesophageal junction cancers; approximately 10-25% of gastric cancer cases according to the latest census are HER-2 positive [80]. Recently in vitro studies have shown that HER-2 positive gastric cancer cell lines respond to trastuzumab, an anti-2 HER-2

antibody [80-81]. In 2014, the European Group on Tumor Markers published a guideline urging the use of trastuzumab with chemotherapeutic drugs as a combination therapy in HER-2 positive breast cancer. HER-2 status is checked by immunohistochemistry and/or fluorescent in situ hybridization (FISH) of gastric biopsies in most clinics, and it's one of the few gastric cancer markers routinely used in clinics [81-82].

1.6.2 EGFR

Like HER2, epidermal growth factor receptor (EGFR) is a member of the HER family. Epidermal growth factor (EGF) and transforming growth factor α (TGF α) are two of its ligands and upon binding, induce proliferation and differentiation in healthy cells [97]. EGFR gene amplification has been shown to occur in certain populations of gastric cancer at varying levels. A study by Kyose et al. found 4.9% of GC patients out of 365 had EGFR amplification; the method they used was fluorescent in situ hybridization [98]. Another study, using SNP assay reported EGFR amplification in 7.7% of the GC cases [99]. With a larger patient group, a study found that using IHC, overexpression of EGFR can be detected in 27.4% GC tumors, but for the same group, FISH only identified 2.3% to have overexpression [100]. These studies show significant variation in findings from IHC and FISH, however, there does appear to be EGFR amplification in GC and an investigation to assess the prognostic capabilities of EGFR found that EGFR overexpression is associated with worse prognosis. [102]. Anti-EGFR monoclonal antibodies have been in clinical trials to treat gastric cancer to no avail, currently there is little evidence using anti-EGFR treatment as a first-line treatment of advanced gastric cancer is beneficial [103].

1.6.3 PD-L1

According to the National Cancer Institute in the United States, few markers are in use in clinics for gastric cancer. Besides HER-2, there are other markers in use in clinics , programmed death ligand-1 (PD-L1), dihydropyrimidine dehydrogenase (DPD) gene mutation, and carbohydrate antigen 19-9 (CA19-9) all routinely used for different purposes [84]. Two key immune checkpoint regulators, programmed cell death 1 (PD-1) and programmed cell death-2 are up-regulated in natural killer T-cells, T and B cells, as well as monocytes. When PD-L1 binds to PD-1 in activated T-cells, there is a drop in cytotoxic T-cell activity which consequently weakens the immune response to tumors, generating immune tolerance in the tumor [85]. In 15-70% of GC patients, PD-L1 expression has been detected and PD-L1 expression correlates with poor prognosis [86]. Two monoclonal antibodies, nivolumab and pembrolizumab, target the PD-1 pathway and enable immune checkpoint blockade thus facilitating the immune system to attack tumor cells. In recent

years, GC patients with PD-L1 expression have been shown to have an improved overall response rate (ORR) when treated with pembrolizumab [30].

Tumors which were profiled MSI-High were more responsive to PD-L1 targeted therapy compared to MSI-Low tumors, moreover, most PD-L1 positive tumors have been shown to be EBV-positive as well [30]. PD-L1 status is now checked routinely in clinics as a predictive biomarker for pembrolizumab response in GC.

1.6.4 E-cadherin

The CDH1 gene codes for the cell adhesion protein E-cadherin in epithelial cells. It is a transmembrane protein which is a calcium-dependent cell-cell adhesion molecule [87]. The role of E-cadherin in cancer has been heavily researched and investigated over the years and E-cadherin remains as a hallmark of epithelial-like cancer cells. Loss of E-cadherin expression in tumor cells results in the cells becoming less adhesive to the tissue and is more likely to detach from the mass and enter the bloodstream. This process is of course facilitated by the expression of vimentin, another transmembrane protein that is a hallmark of mesenchymal cells. This transition of tumor cells from a more epithelial state to a more mesenchymal state is known as epithelial-to-mesenchymal transition, better known as EMT [88].

In gastric cancer, E-cadherin plays a major role in carcinogenesis. As discussed beforehand, in HDGC cases, germline mutations in CDH1 are prominent. Apart from HDGC, E-cadherin expression is vastly affected in other gastric cancers. Loss of E-cadherin expression causes epithelial cells to lose their adhesive capacity and it is a prerequisite for a malignant cancer cell. E-cadherin expression is more commonly lost in diffuse-type tumors compared to intestinal-type; moreover, Gabbert et al. studied 413 gastric tumors and found that loss of expression of E-cadherin significantly affects 3- and 5-year survival rates [90]. Studies suggest that this transmembrane protein may be an ideal candidate as a prognostic factor in GC pathogenesis. The mechanisms by which CDH1 expression is minimized in GC cells have been studied extensively. DNA hypermethylation of CDH1 promoters due to H.pylori infection is among the most frequent in GC cases. [89] Furthermore, genetic inactivation of CDH1 has been detected in patients who do not have a germline mutation. Overall, E-cadherin has tumor-suppressor functions in gastric cancer, and its inactivation leads to devastating effects in the stomach tissue. One caveat is, most of the studies conducted on E-cadherin expression in GC tissues were conducted by performing immunohistochemistry experiments. Even though IHC is widely used in clinics and research facilities across the globe, its specificity and replicability have been in question when the tissue

under study is heterogeneous in nature. In order to accurately predict the prognosis of a GC patient by CDH1 expression, other tests such as RT-PCR, blood tests, and gastric juice analysis needs to be developed. There is evidence that somatic DNA methylation of CDH1 is detected by testing blood samples [89], if so, these tests can greatly modify the accuracy of CDH1 as a prognostic biomarker.

1.6.5 CA19-9

Carbohydrate antigen 19-9 is a glycolipid antigen that was primarily identified in colorectal cancer. It is a ligand for E-selectin which is expressed on the cell-surface membrane of endothelial cells. [91] Its a tumor serum marker. In gastric cancer, it has been associated with advanced staging as well as a series of other histopathological features of gastric tumors including a higher level of lymph node metastasis, higher chances of lymphatic and venous invasion, mixed histology, and antral location [95]. Elevated CA19-9 levels are more frequent in female patients as opposed to male patients [94]. Using CA19-9 in combination with CEA, and AFP have proven successful in diagnosing GC [96].

1.6.6 CEA

Carcinoembryonic antigen (CEA) is an excessively used marker for colorectal cancer. In gastric cancer, however, it is not recommended to be used as a diagnostic marker because CEA levels are high in all higher-stage gastric cancers. Be that as it may, it is a very reliable risk factor to predict liver metastasis relapse [91]. Moreover, analysis of the peritoneal lavage fluid after curative resection of GC can accurately predict the likelihood of peritoneal recurrence [92]. A study conducted found that RT-PCR of CEA mRNA from peritoneal cavities can accurately detect micrometastasis. In another study, between 2008 and 2015 that compared CEA levels in serum in early gastric cancer patients found that elevation of CEA levels in early gastric cancer patients leads to a worse prognosis (log-rank p-value= 0.014), making it an independent risk factor in early GC [93].

1.6.7 VEGFA and VEGFR

Vascular endothelial growth factor A (VEGFA) and its receptors, vascular endothelial growth factor receptors (VEGFRs) are part of the angiogenesis pathway. In gastric cancers, these key players are needed to generate new blood vessels and propagate tumorigenesis. In GC, VEGFA was detected in 40% of patients and VEGFRs were detected in 36% of the GC patient group. [104] VEGFA expression in GC tissue and serum has been associated with poor prognosis, lymph node involvement and metastasis [105]. Multiple clinical trials tested the efficacy of inhibiting VEGFA

and VEGFRs in increasing overall survival. However, few have shown promise, a trial where the anti-VEGFR2 monoclonal antibody, ramucirumab has shown elongation of the overall survival in patients [106].

1.6.8 MSI

The mismatch repair machinery is known to be faulty in gastrointestinal cancers which leads to microsatellite instability, in this condition cells accrue multiple genetic errors. In 18-128% of GC, MSI is detected; primarily due to hypermethylation of the MLH1 promoter [107].

In most cases MSI-positive tumors occur in older patients and they generally have a good prognosis [108].

1.6.9 FGFRs and MET

In a recent study using high-resolution SNP arrays, fibroblast growth factors (FGFRs) have been found to have more copy number gain than EGFR, HER2, or MET. FGFRs had approximately 9.3% copy number gain (CNG) and MET had 4.3%. Moreover, FGFRs were associated with lymph node metastasis and shorter overall survival [99]. FGFRs are a potential target for GC treatment in clinics as well as a potential prognostic marker.

MET is a member of the human growth factor receptor family (HGFR) and its amplification and overexpression have a role in GC carcinogenesis thus implying that MET can be a poor prognosis indicator [101]. A recent study found that MET-positive tumors had a significantly poorer prognosis than MET-negative tumors in GC [108].

1.7 Gene expression signatures in gastric cancer diagnosis and prognosis

Since the advent of microarray technology, there has an exponential increase in research on gene expression signatures. By analyzing thousands of genes all at once in different biological states, information on pathway activation, and the relationship between the molecular profiles of tissues to disease state can be studied extensively. There has been several gene expression signatures identified in cancer, however, most of these signatures have not been validated for use in clinical setting. A good gene signature is considered one which has three main characteristics, analytical validity, clinical validity and clinical utility. Analytical validity refers to a signature which is consistent across an array of data types and accurately presents the same results. Clinical validity is the ability of the signature to classify the patient pool into two or more clinically relevant sub-groups. Clinical utility is the signature's ability to help clinicians make informed decisions based

on the results. The results must give a significant difference in outcome in order to be utilized in the clinics. [138]

One less discussed aspect of a gene signature is the cost-to-benefit ratio of using it in clinics across the globe. Most well-known gene expression signatures (GES) consist of a large list of genes, such MammaPrint or Oncotype Dx. MammaPrint assesses the activity of 70 genes whilst Oncotype DX assesse 21 genes [138]. However, in many clinics around the globe there is a shortage of resources and this limits their capacity to routinely test cancer patients for gene signatures. As a result, many patients go undiagnosed in earlier stages of cancer. More often than not, most people in developing nations are cursed with a malignancy due to lack of diagnostic facilities. Moreover, since many of these less fortunate clinics use more archaic methods of cancer diagnosis, the treatment options available to the patients are less efficacious. These treatments are less personalized and is often leads to higher mortality. [139]

Generating gene signatures which are shorter would allow smaller clinics to perform these tests thus potentially reducing the incidence of mortality among patients in less privileged communities.

There are several bioinformatics pipelines by which a prognostic gene expression signature can be generated. The purpose of a prognostic signature is to effectively analyze gene expression data and assess survival outcome.

A meta-analysis which studied 39 gastric cancer prognostic gene signatures identified 3 main methods by which these signatures are generated. Method 1 identified differentially expressed genes to separate good and bad prognostic groups; method 2 relied on identifying genes which are part of an aberrantly active signaling pathway in gastric cancer; and method 3 relied on molecularly classifying gastric cancer based on certain gene expression [141].
In method 1, authors implemented an integrated analysis of gene expression data by the Robust Rank Aggregation method [142]. After selecting a top list of DEGs, the authors performed a univariate cox regression model to identify genes which were significantly associated with survival. The list was shortened by performing a multivariate cox regression model to finalize a gene list. A predictive risk score calculation was generated which allows the authors to rank each individual patient into poor and good survival groups.

A common practice when generating a prognostic gene signature is to perform a univariate analysis to select significant genes, followed by a MVA to further shorten gene list and generate a list with highly significant genes which are independent of stage. Lasso cox regression analysis

has been used between a univariate cox regression and a multivariate cox regression [143] to shrink the genelist. LASSO (Least Absolute Shrinkage and Selection Operator) is a variable selection algorithm which has been used in recent years to select the likeliest genes from Cox regression model. There has been an increase in use of machine learning algorithms in biological sciences in the last few years. Combined with the big data high throughput transcriptomic assays generate, machine learning algorithms have been used to generate gene signatures for prognosis, molecular classification, prediction of treatment options etc. Machine learning algorithms are mainly two types, supervised and unsupervised. In supervised learning a known class or a group of known classes are used to train the algorithm (it's known as labeled training). The other type is unsupervised learning and as the name suggests unknown classes are used to train the algorithm (also known as unlabeled training) [144]. When detecting genomic patterns, particularly in cancer genomics, unsupervised algorithms are used to classify molecular subgroups, generate predictive models for drug efficacy, generate prognostic signatures etc.

One extremely powerful machine learning algorithm is SVM (Support Vector Machine). It has been in use since the early 2000s when microarray technology was first used. It is a highly versatile method and has been used to classify tumor subgroups in many different cancers [144]. It is a supervised learning method which means it undergoes labeled training to classify a certain dataset based on class labels.

1.8 A novel mRNA based prognostic 20-gene signature for gastric adenocarcinoma

Previously in our lab, a 20-gene prognostic signature was generated by Dr. Secil Demirkol using publicly available transcriptomic data from GEO (Gene Expression Omnibus). Two datasets, GSE15459 and GSE62254 were used for the purpose of this study. These datasets contained gene expression data (Affymetrix) for 200 and 300 gastric adenocarcinoma patients respectively. Analyses of these datasets and their respective clinical data was done by the R script mUSAT, the code uses the survival package to calculate the log rank p-value, cox p value and maximally selected rank statistics. The original 20 gene prognostic signature was generated using this script. The key difference between mUSAT and most other prognostic signature selection pipeline us that it performs MVA as well as UVA. The probe-sets with lowest Cox p values in the datasets GSE15459 and GSE62254 were selected and ranked. Afterward the ranks of each top gene in the two datasets were summed up to select the best ranked 20-gene list. [140] Hierarchical clustering analyses shows distinction between Up (poor prognosis) and Down (good prognosis).

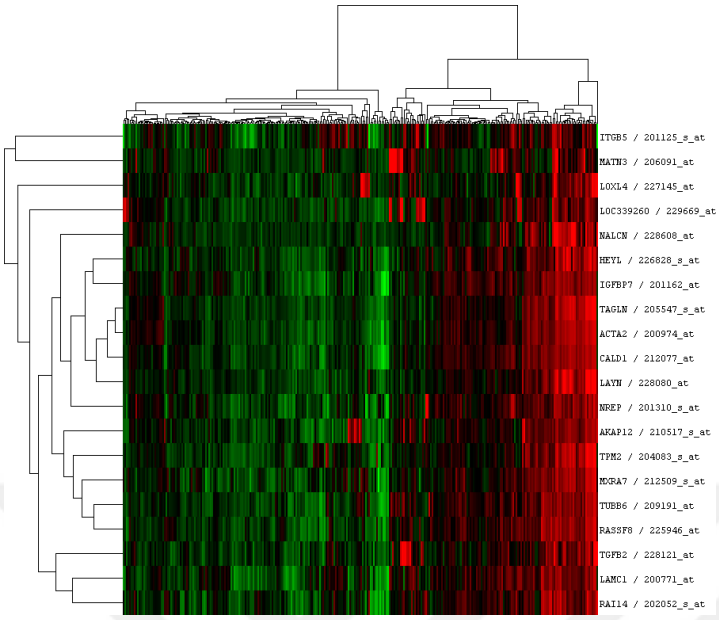


Figure 1.1: Hierarchical clustering analysis showing the two distinction in Up and Down groups in GSE62254.

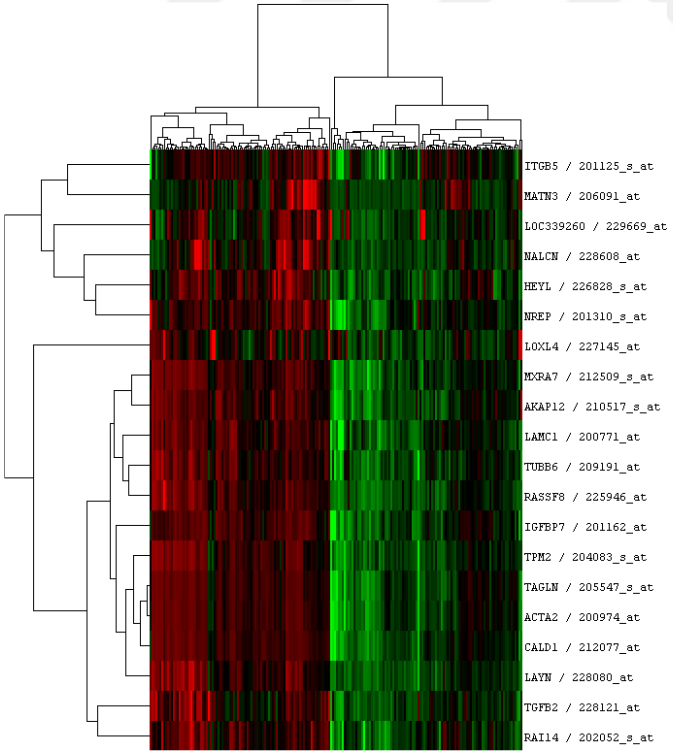


Figure 1.2: Hierarchical clustering analysis showing the two distinction in Up and Down groups in GSE15459.

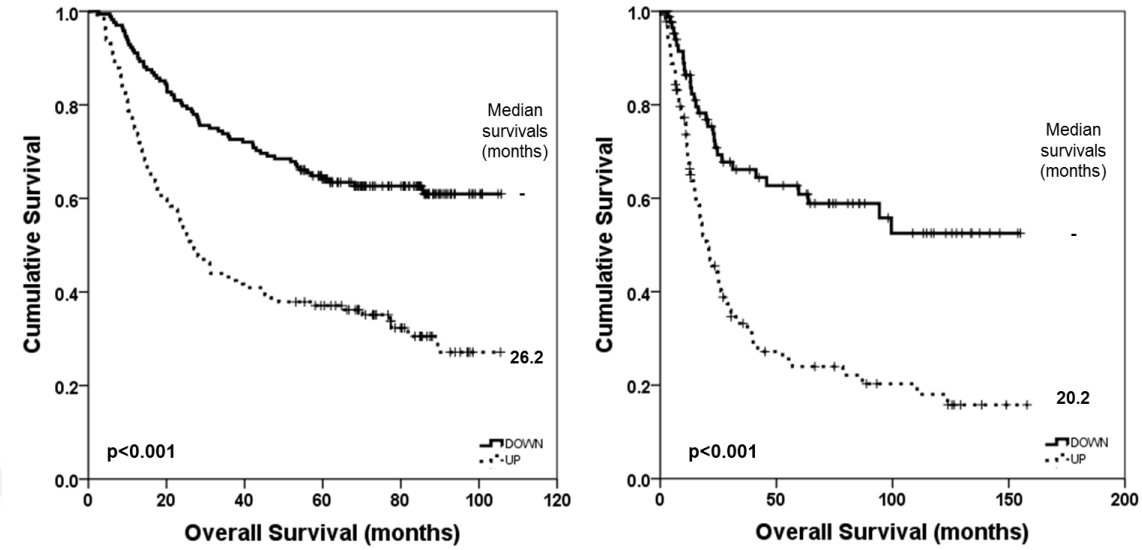


Figure 1.3: Kaplan-Meier plots of the gastric cancers in GSE15459 (left) and GSE62254 (right).

Currently, there aren't any gene signatures which can prognosticate gastric cancer patients into good and bad survival groups. This 20 gene signature can successfully stratify patient groups into good and bad prognosis groups with a significant difference in survival. Moreover, this signature is independent of histological subtypes, intestinal and diffuse which indicate its versatility in stratifying GC patients.

1.9 Gastric tissue specific reference genes

The advent of the quantitative PCR has revolutionized biological research. However, the history of qPCR use in research is plagued with poor experimental designs and lack of transparency regarding the experimental set-up and questionable data analysis [110]. In order to circumvent many of these issues, Minimum Information for Publication of Quantitative Real-time PCR Experiments (MIQE) guidelines were published in 2009 [111].

The use of reference genes in qPCR is a standard practice that is widely accepted in the scientific community. However, the use of conventional reference genes have been brought into question in recent years in cancer research. In cancer biology, there has been an uptick in using mRNA levels of genes of interest as a marker for a myriad of cellular processes including proliferation, apoptosis, metabolic activities, etc. The need for stable reference genes are paramount to quantify the levels of these genes in cancer cells. The use of qPCR is well established in medical research and diagnostics. In a clinical setting, it is of utmost importance that the control genes are stably

expressed in both tumor and normal tissues, otherwise it creates a variability in analyzing results and may give rise to the possibility of improper diagnosis.

The use of housekeeping genes such as Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), Actin beta (ACTB) and beta 2-microglobulin (B2M) is commonplace in biological research. The high expression levels of these genes and their steady expression made them appealing reference genes for many years [112,113]. Recent research however, found expression inconsistencies of these known reference genes in various cancer tissues [114,115]. Furthermore, selecting which genes best serve as internal controls for cancer tissues pose a challenge. A study conducted in 2012 found both GAPDH and ACTB have 67 and 64 pseudogenes respectively. These pseudogenes are intron-less and pose a risk of skewing qPCR data [116].

All things considered, there is a growing body of researchers who agree that it should not be assumed that known reference genes have stable expression in tissues as dynamic and as diverse as cancer. Tissue-specific reference genes are required to accurately assess gene expression levels.

1.10 Aims of the study

The primary aim of the study was to validate the 5-gene mRNA-based prognostic signature in silico using high-throughput gastric cancer datasets. We've utilized publicly available gene expression and clinical data from Gene Expression Omnibus for this purpose. The patient groups were stratified based on our gene signature and their survival difference was assessed.

Lastly, we generated an algorithm which takes normalized transcriptomic data and generate a list of gastric tissue-specific reference genes which has the most stable expression across all genes expressed in the tissue-type. Post-discovery, we attempted to validate our novel reference genes via in silico methods and qPCR.

Chapter 2

Materials and Reagents

2.1 General laboratory reagents

Common laboratory reagents such as ethanol, isopropyl alcohol, and chloroform, EDTA were purchased from SIGMA ALDRICH (St. Louis, MO, USA). Nuclease-free water, DEPC treated water were bought from Ambion (Carlsbard, CA, USA). DMSO was purchased from AppliChem (Darmstadt, Germany). Basic laboratory consumables such as gloves, paper towels, 70% ethanol were all provided by the Molecular Biology and Genetics department of Bilkent University.

2.2 cDNA synthesis of tumor, normal, and cell line mRNA

Total RNA was extracted from cultured cell lines using the PureZOL RNA Isolation Reagent (BIO-RAD, Hercules, CA, USA) following the manufacturer's protocol. Afterward, DNase I treatment was performed using the DNA-freeTM Kit (Ambion, Carlsbard, CA, USA) following manufacturer's protocol. Total RNA concentration and purity was measured using NanoDROP One spectrophotometer (Thermo Scientific, Wilmington, DE, USA).

300 ng of cell lines and 500 ng RNA of tumor was used to perform cDNA synthesis using the Revertaid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, Boston, MA, USA) with oligoDT primers according to the manufacturer's protocol.

2.3 Cell culture materials

A panel of 6 gastric cancer cell line were purchased from the American Type Culture Collection (ATCC). The delivered cell lines were thawed according to ATCC cell thawing protocol. 1% Penicillin, streptomycin was added to RPMI growth media along with filtered FBS (fetal bovine serum).

The cells were cultured in a cell culture incubator accordingly at 37 °C and 5% CO2 according to the ATCC protocol. The cells were passaged according to ATCC protocol.

RPMI with stable Glutmine was used for cell culture. (Biowest, Nuaille, FR). Trypsin-EDTA (0.25%) was purchased from SIGMA-ALDRICH (St. Louis, MO, USA). Fetal Bovine serum was bought from Biowest (Nuaille, FR). 1% Penicillin-streptomycin was used in complete media (Hyclone, Rockford, USA).

For freezing media, ATCC protocol was followed and the media consisted of complete growth media with 5% (v/v) DMSO. Additionally, vials with FBS and 5% (v/v) DMSO was preserved according to our lab's protocol.

Cell line	Type	Culture condition	Classification
AGS	Adhesion	1% pen-strep, 10% FBS, RPMI	Epithelial
KATO-III	Adhesion + Suspension	1% pen-strep, 10% FBS, RPMI	Uncategorized
NCI-N87	Adhesion	1% pen-strep, 10% FBS, RPMI	Epithelial
SNU-1	Suspension	1% pen-strep, 10% FBS, RPMI	Mesenchymal
SNU-16	Suspension	1% pen-strep, 10% FBS, RPMI	Epithelial
SNU-5	Suspension	1% pen-strep, 20% FBS, RPMI	Intermediate/ Mesenchymal

Table 2-1: Cell line characteristics and growth conditions.

2.4 Quantitative Real-time PCR assay

Real-time quantitative (qPCR) was performed using the Roche LightCycler 480 type II (Roche diagnostics, Nederland, BV) and iTaqTM Universal SYBR® Green Supermix (BIO-RAD, Hercules, CA, USA) according to the manufacturer's specifications. Amplification was performed under the following conditions: 95 °C for 30 seconds, followed by 40 cycles of 95° for 15s and of 60 °C for 60s. The genes, EWSR1, SF1 and HNRNPK were selected as candidate reference genes and GAPDH and PGK1/B2M were selected for comparison.

2.4.1 Primers

SF1

Forward: 5-ACCATGGCCCTCCTCCAAT-3

Reverse: 5-GGTGGCGGCATCATACCTTT-3

EWSR1

Forward: 5-GAGAGCGAGGTGGCTTCAAT-3

Reverse: 5-GGATCTACAGGTGGGCCTAGA-3

B2M

Forward: 5-TGGAGCTCCTGGAAGGTAAAG-3

Reverse: 5-AGTTGACTTAGGGGCTGTGC-3

HNRNPK

Forward: 5- CGGGGAAGAGGTGGTTTTGA-3

Reverse: 5-TGAGGTCTCCCCCTCTAGGT-3

GEO accession code	**No. of samples**	**Platform**	**Survival information**	
GSE15459 [117]	192	Affymetrix HG-U133_Plus_2	Yes	Discovery
GSE62254 [118]	300	Affymetrix HG-U133_Plus_2	Yes	Discovery
GSE54129	111	Affymetrix HG-U133_Plus_2	No	Hierarchical clustering
GSE51105 [119]	94	Affymetrix HG-U133_Plus_2	No	Hierarchical clustering

GSE100935 [120]	66	Affymetrix HG-U133_Plus_2	No	Hierarchical clustering
GSE57303 [121]	70	Affymetrix HG-U133_Plus_2	No	Hierarchical clustering
GSE13861 [122,123]	90	Illumina HumanWG-6 v.3.0	Yes	Survival Data Validation
GSE26253 [122,124]	432	Illumina HumanRef-8 WG-DASL v.3.0	Yes	Survival Data Validation
GSE26899 [122]	108	Illumina HumanHT-12 V3.0	Yes	Survival data Validation
GSE26901 [122]	109	Illumina HumanHT-12 V3.0	Yes	Survival data Validation
GSE28541 [122]	40	Illumina HumanWG-6 v2.0	Yes	Survival data Validation

Table 2-2: List of datasets used to validate the prognostic gene list.

2.5 *Ex vivo* validation group

Tumor tissues and adjacent normal tissues were collected from 10 individuals who were diagnosed with gastric adenocarcinoma. Prior to collecting the samples, the patients signed an informed consent form. All patients were above the age of 20. The samples underwent routine analysis by the clinical pathologists after acquisition.

Chapter 3

Methods

3.1 Statistical analyses

We used R version 3.5.3 and 4.0.3 [125] respectively throughout our work. We used RStudio to run our scripts. For calculating average gene expression, standard deviation of gene expression and log fold change we've used Microsoft excel. Variation was calculated using =var() function in Excel.

Pearson correlation and p-value calculations were done in RStudio using the package Hmisc [126]; we've implemented a p-value cut-off of <0.01.

Graphpad Prism version 8 for Windows was used to generate expression plots.

3.2 Microarray data Normalization

Raw data for each dataset was downloaded from NCBI GEO. For the Affymetrix data-sets, .CEL files were used to perform RMA (Robust Multiarray Average) normalization by using BRB array tool [127] which is an Excel Add-in. The Illumina datasets were quantile normalized by using an R package PreProcessCore [128] and then they were log2 transformed in R version 3.5.3.

3.3 Hierarchical Clustering Analysis

For hierarchical clustering we used Cluster 3.0 and Java Treeview. These desktop applications were downloaded from http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm.

In Cluster 3.0 we clustered our dataset by measuring Euclidean distance and complete linkage for both genes and arrays. The software generated a .CDT file with the clustered data, each data point was then standardized using the formula (Value-mean)/ standard deviation. The standardized dataset was then visualized by Java Treeview.

3.4 Survival Analysis

For Survival analysis we used the Survminer [129] and Survival [130] packages in R. We classified the patients into good and bad groups based on low and high expression of our 20 genes respectively. We plotted Kaplan-Meiers for our discovery and validation datasets after classifying patients into good and bad groups based on the 5 gene prognostic signature.

We calculated the cox regression and hazards ratio for the 20 gene signature in our 5 survival validation datasets by using an R script, mUSAT which was written by one of our previous lab member; for this script we used the R packages Survival and Maxstat [131].

All survival analysis was performed with overall survival (in months) information.

3.5 Correlation plots

We used the package Corrplot [132] in R to generate correlation plots. We generated correlation plots for the discovery and validation datasets for the 20 genes to see which genes are highly correlated with each other. The function was written to eliminate the genes which did not have a significant correlation; thus the diagrams display genes that are only significantly correlated with a p-value cutoff of 0.05.

3.6 RNA-seq data and normalization

We downloaded raw count data from TCGA (link: https://portal.gdc.cancer.gov/) for gastric cancer and performed De-seq2 [133] normalization in Rstudio. The sample size was 413 and post-normalization we log2 transformed the data. Besides the De-seq2 normalized TCGA data we downloaded RSEM [134] normalized TCGA data from FireBrowse (link: http://firebrowse.org/) and log2 transformed the data as well. The RSEM dataset had 450 samples in total and 35 of these were from normal tissue, and 415 were stomach adenocarcinoma tissue. We performed our gene stability analysis with the 415 samples. The two samples in the RSEM data which were not present in the HT-seq count data were eliminated. We downloaded RSEM normalized data from CCLE and log2 transformed them as well.

3.7 Assessment of gene stability

We develop an in-house algorithm to identify the genes that are the most stably expressed in gastric cancer tissue. The statistical method we chose was coefficient of variance calculation. The standard deviation of each gene is calculated and then divided by its average expression to generate its coefficient of variance. A smaller coefficient of variance (CV) indicated less variation around the mean among gastric tumors.

$$CV = s/x$$

(s = sample standard deviation, x = sample mean)

3.8 Random repeated sub-sampling method

The code we wrote takes a normalized, log2 transformed RNA-seq data as an input where the columns correspond to the samples and the rows correspond to the genes.

In the first step of the algorithm, the samples were shuffled and then divided into two random two and matrices were generated with their expression data (n=207, n=206). For both matrices, the code calculates the CV for each gene, ranks the genes based on the coefficient of variances (higher rank means lower CV). The ranks were stored in two matrices titled “ranks 1” “ranks 2”.

The code then repeats these steps for a thousand iterations. After we generated the two rank tables, for a thousand iterations, we calculated the average rank for each gene (average of 2000 ranks) and called that “rank_mean”. We used rank_mean 200 as a cutoff to generate a table of top candidate reference genes. We ran the code on the RSEM normalized data as well as the Deseq2 data for TCGA and for RSEM normalized CCLE data.

3.9 Multivariate Unsupervised software analysis tool (mUSAT)

The mUSAT code was written by a former member of our lab, Murat İşbilen using R (version 3.1.2). The code uses the survival package to calculate the log rank p-value, cox p value as well as perform maximally selected rank statistics. The original 20 gene prognostic signature was generated using mUSAT. The probe-sets with lowest Cox p values in the datasets GSE15459 and GSE62254 were selected and ranked. Afterward the ranks of each top gene in the two datasets were summed up to select the best ranked 20-gene list.

Chapter 4

Results

4.1 Selection of a shortened gene-list from the 20-gene mRNA signature for further analysis

In a clinical setting, analyzing the gene expression of each gastric cancer patient for a 20 gene signature is tedious. Pathology and diagnostics labs are inundated with tests on a daily basis. Unless the facility has abundant resources and targeting sequencing facilities, assessing gene expression of 20 genes via qPCR seemed daunting. Thus we, attempted to shorten our gene-list, in order to make them more suitable for clinical applications. Furthermore, we postulated a two-step verification method to classify patients into good or poor prognostic group would increase the efficacy of our signature. Thus we aimed to find a list of genes which stratify patients into prognostic groups both at a transcriptomic and at a proteomic level. To generate a smaller gene signature based on transcriptomic data, we attempted two separate strategies; correlation matrix to find highly correlated genes and find genes with the biggest difference in log fold change between the good and bad prognostic groups, classified by our initial 20 gene-signature. Our aim was to find a group of genes which show up on both lists to work on further. As for the proteomic approach, we checked for the protein level expression of our genes from the Human Protein Atlas. A list of genes with high expression in gastric tumor tissues, which also have readily available commercial antibodies would be selected and compared with the gene-list generated from transcriptomic data to find a list of genes which are present in both lists. The flow chart below depicts the steps we took to finalize our gene list.

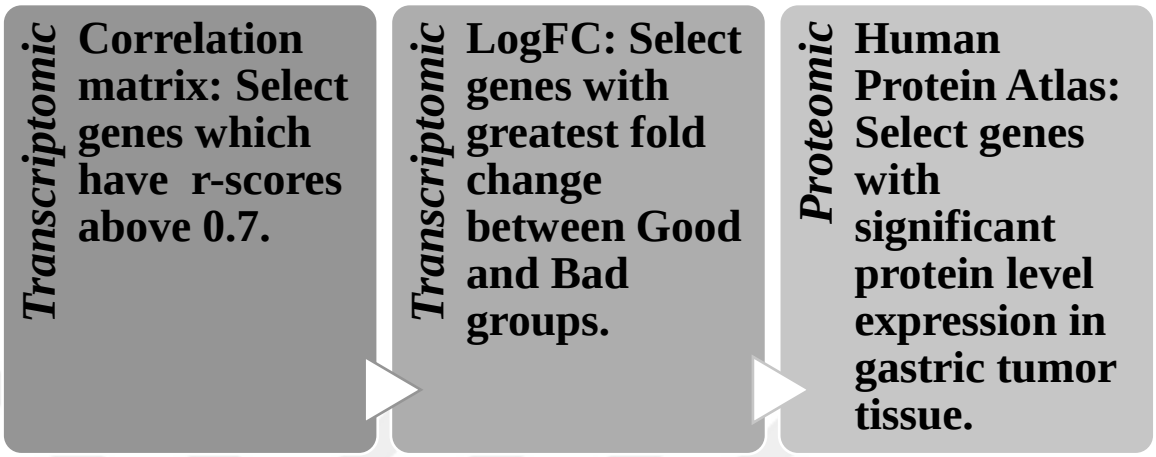


Figure 4.1: Flow chart depicting the series of decision making to select the final 5 genes for our signature.

4.1.1 Generating correlation plots for the 20-gene list to identify a strongly correlated subgroup

Primarily, we attempted to generate a strongly correlated group of genes which will stratify the patients in the discovery datasets into good and bad prognosis groups as the 20-gene signature did. We postulated that a strongly correlated set of genes would allow us to shorten our list to the point where it would be more manageable in a clinical setting. We generated correlation matrices with the 20 genes to identify genes with a higher correlation than 0.7 in our discovery cohorts, GSE62254 and GSE15459.

	ITGB5	MATN3	LOXL4	LOC339260	NALCN	HEYL	IGFBP7	TAGLN	ACTA2	CALD1	LAYN	NREP	AKAP12	TPM2	MXRA7	TUBB6	RASSF8	TGFB2	LAMC1	RAI14
ITGB5	1																			
MATN3	0.32852	1																		
LOXL4	0.453639	0.322356	1																	
LOC339260	0.367031	0.434187	0.387597	1																
NALCN	0.399743	0.408492	0.509943	0.501387	1															
HEYL	0.45621	0.369416	0.454943	0.483082	0.683166	1														
IGFBP7	0.665627	0.339526	0.50304	0.46212	0.608579	0.789274	1													
TAGLN	0.460957	0.327324	0.507198	0.524645	0.767596	0.787075	0.748121	1												
ACTA2	0.586169	0.3313	0.525364	0.526401	0.722914	0.798895	0.875956	0.935084	1											
CALD1	0.653378	0.360185	0.529424	0.511684	0.685123	0.797012	0.925895	0.858759	0.927568	1										
LAYN	0.394605	0.32681	0.579145	0.486529	0.734486	0.807672	0.749582	0.877166	0.832059	0.801449	1									
NREP	0.520799	0.38548	0.510754	0.497284	0.658301	0.726121	0.770054	0.747188	0.803353	0.805741	0.703132	1								
AKAP12	0.401668	0.324057	0.481515	0.456418	0.621023	0.650397	0.611385	0.730864	0.726786	0.701141	0.703063	0.602728	1							
TPM2	0.515661	0.35429	0.525555	0.470711	0.681403	0.720523	0.677009	0.867163	0.817591	0.768148	0.772552	0.682992	0.677245	1						
MXRA7	0.618391	0.36265	0.638706	0.491381	0.670864	0.710342	0.782718	0.812852	0.824798	0.841619	0.760733	0.688451	0.666255	0.775776	1					
TUBB6	0.550611	0.431913	0.605152	0.481779	0.618913	0.730715	0.790185	0.766902	0.802756	0.841938	0.777547	0.705845	0.720473	0.733752	0.802629	1				
RASSF8	0.435052	0.405803	0.547433	0.536397	0.663393	0.742732	0.70884	0.787366	0.750218	0.78868	0.80624	0.69442	0.728694	0.714164	0.760456	0.819727	1			
TGFB2	0.457037	0.430083	0.495965	0.513855	0.532454	0.646729	0.613137	0.600411	0.622319	0.666289	0.628106	0.58734	0.608268	0.620968	0.627746	0.688707	0.656045	1		
LAMC1	0.73584	0.423404	0.579238	0.476484	0.511913	0.711796	0.8239	0.64708	0.746038	0.818239	0.63562	0.70461	0.552611	0.658923	0.745409	0.758677	0.679652	0.657224	1	
RAI14	0.616924	0.398087	0.542941	0.447935	0.572358	0.690335	0.789545	0.665377	0.741594	0.807428	0.70584	0.71274	0.629308	0.654345	0.758157	0.797089	0.737842	0.650021	0.791081	1

Figure 4.2: Correlation matrix of the 20 gene signature in discovery dataset 1 (GSE66254). All correlations higher than 0.7 is highlighted in red and Genes with a significant correlation is highlighted yellow.

A total of 14 genes had a strong positive correlation. The genes were, HEYL, IGFBP7, CALD1, ACTA2, TAGLN, LAYN, NREP, LAMC1, RAI14, AKAP12, MXRA7, RASSF8, TUBB6, and TPM2.

Next we generated a correlation plot with our second discovery dataset, GSE15459.

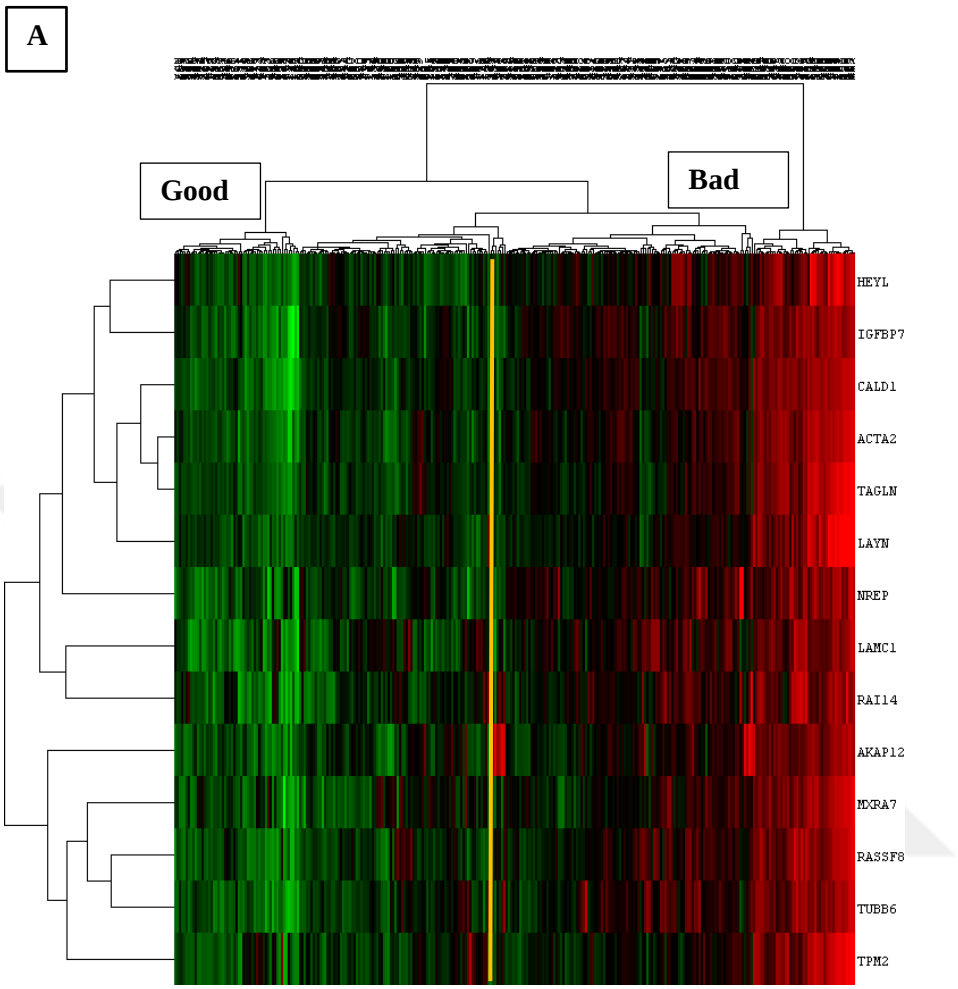
	ITGB5	MATN3	LOXL4	LOC339260	NALCN	HEYL	IGFBP7	TAGLN	ACTA2	CALD1	LAYN	NREP	AKAP12	TPM2	MXRA7	TUBB6	RASSF8	TGFB2	LAMC1	RAI14
ITGB5	1																			
MATN3	0.472422	1																		
LOXL4	0.217728	0.187858	1																	
LOC33926	0.3037	0.41626	0.066803	1																
NALCN	0.483315	0.546197	0.256915	0.511806	1															
HEYL	0.313678	0.398179	0.223662	0.407946	0.58603	1														
IGFBP7	0.485551	0.402918	0.256353	0.285928	0.581743	0.773077	1													
TAGLN	0.409081	0.256163	0.305977	0.268	0.473418	0.656861	0.844449	1												
ACTA2	0.440368	0.281764	0.304214	0.257001	0.493498	0.665342	0.87549	0.983078	1											
CALD1	0.496724	0.343732	0.304332	0.288697	0.522118	0.673792	0.913391	0.938661	0.951855	1										
LAYN	0.338496	0.275498	0.399865	0.316584	0.479735	0.665461	0.78536	0.84596	0.850138	0.868407	1									
NREP	0.478138	0.460787	0.234231	0.389141	0.577259	0.652829	0.763432	0.686294	0.715311	0.741313	0.610092	1								
AKAP12	0.428473	0.332639	0.410683	0.282876	0.463183	0.61112	0.779617	0.822399	0.827577	0.845335	0.773147	0.668705	1							
TPM2	0.40671	0.237562	0.362166	0.213664	0.438461	0.629415	0.786341	0.928362	0.926677	0.897474	0.88656	0.634471	0.802335	1						
MXRA7	0.428739	0.300669	0.362248	0.324491	0.524033	0.644959	0.817264	0.873547	0.874386	0.895132	0.841472	0.687866	0.882487	0.855167	1					
TUBB6	0.414812	0.203059	0.409447	0.218144	0.388777	0.508333	0.74499	0.848154	0.836177	0.869449	0.821414	0.587996	0.814161	0.842687	0.838068	1				
RASSF8	0.449958	0.312432	0.413438	0.315387	0.500589	0.602893	0.792028	0.843462	0.833767	0.87374	0.860308	0.655783	0.861869	0.851702	0.872943	0.868549	1			
TGFB2	0.377812	0.313326	0.379982	0.288892	0.436075	0.496576	0.603945	0.59959	0.609431	0.660506	0.68198	0.641402	0.642908	0.633974	0.691073	0.656423	0.722379	1		
LAMC1	0.517826	0.295949	0.282088	0.228668	0.433345	0.567489	0.799977	0.766757	0.794223	0.859441	0.764433	0.716231	0.774037	0.795425	0.807485	0.848062	0.82765	0.673947	1	
RAI14	0.499399	0.37882	0.277157	0.257211	0.428451	0.508931	0.678311	0.647036	0.650541	0.726862	0.731575	0.615183	0.610674	0.690314	0.682282	0.697391	0.698591	0.666598	0.755349	1

Figure 4.3: Correlation matrix of the 20 gene signature in discovery dataset 2 (GSE15459). All correlations higher than 0.7 is highlighted in red and Genes with a significant correlation is highlighted yellow.

A total of 15 genes had a strong positive correlation. The genes were, HEYL, IGFBP7, TAGLN, ACTA2, CALD1, LAYN, NREP, LAMC1, RAI14, AKAP12, MXRA7, RASSF8, TUBB6, TGFB2, and TPM2.

4.1.2 Hierarchical clustering of correlation-based shortened gene-lists in the two discovery datasets

Next, we attempted to see whether the shortened gene-lists stratify patients into good and bad groups based on prognosis. Gene expression data from the two discovery cohorts were hierarchically clustered to see whether there is significant overlap between the patient groups stratified by the 20 gene signature to the shorter gene signature.


A
Good
Bad
HEYL
IGFBP7
CALD1
ACTA2
TAGLN
LAYN
NREP
LAMC1
RAI14
AKAP12
MXRA7
RASSF8
TUBB6
TPM2

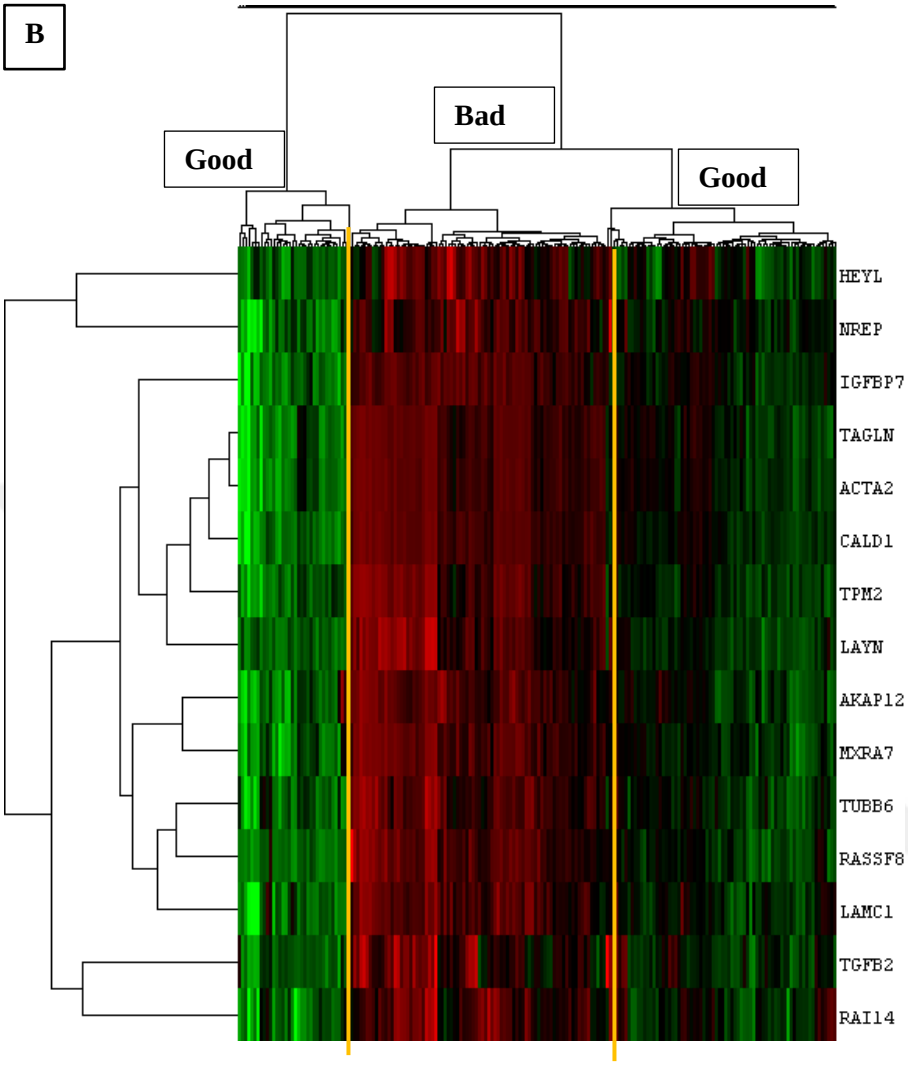


Figure 4.4: Hierarchical clustering of discovery cohorts with the selected prognostic genes. 14 and 15 genes that were highly correlated were hierarchically clustered in the datasets GSE62254 and GSE1545, respectively. The discernible patient groups were labelled Good and Bad based on their gene expression. Orange lines separate the Good and Bad groups. Scale: Green clusters represent lower expression and red represents overexpression.

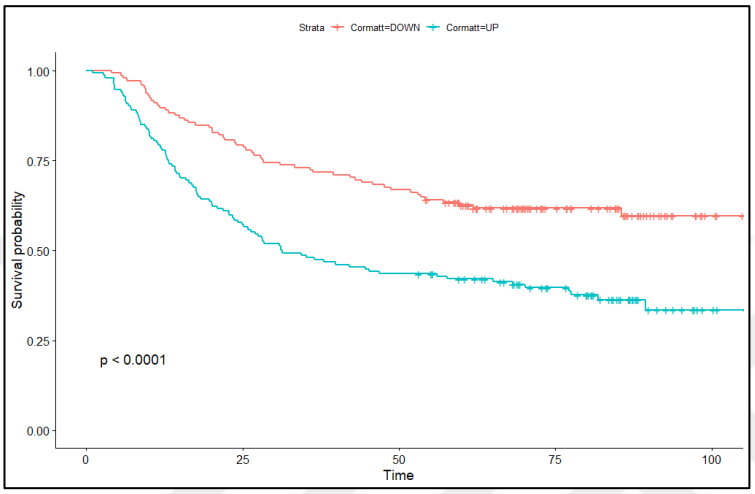
Dataset	Total samples	GOOD	BAD	Mismatch	%overlap
GSE62554	300	136	122	42	86
GSE15459	192	99	82	11	94

Table 4-1: Overlap of patient groups between the 20 gene signature and the shorter gene signatures.

From table 1, it is clear that the shorter gene-lists are capable of classifying the patients into similar groups as the larger signature.

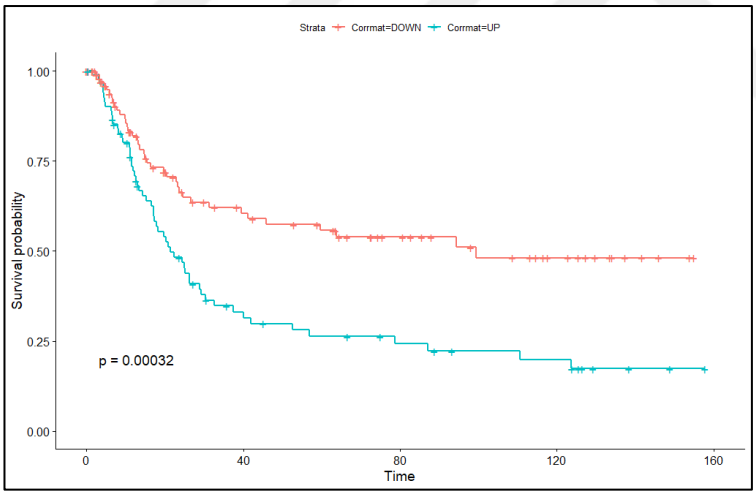
Since, there was significant patient overlap between the grouping done by the two genelists, we generated Kaplan-Meier plots to see if our shorter gene-lists had any prognostic advantage.

A



GSE15459

B



GSE62254

	n	events	median	0.95LCL	0.95UCL
Corrmat=DOWN	110	38	99.4	41.2	NA
Corrmat=UP	82	57	21.4	17.1	30.4

GSE15459

	n	events	median	0.95LCL	0.95UCL
Cormatt=DOWN	145	56	NA	NA	NA
Cormatt=UP	154	96	31.2	24.9	65

GSE62254

Figure 4.5: Kaplan-Meier plots for the two potential prognostic groups. A and B shows the survival plots for the two discovery cohorts, GSE15459, and GSE62254 respectively. Red graph represents the good group, labelled Down, and the blue graph represents bad group, labelled Down. Median survival is depicted below the graphs.

4.1.3 Log-fold change based gene-list generation

We next attempted to generate a gene-list based on log-fold change values. The hierarchical clustering in figure 1.1 shows subgrouping based on low and high gene expression. We postulated that, this might result in significant fold change between the good and bad prognostic groups. Hence, we calculated the log2 fold change of the 20 genes, based on the stratification by the 20 gene signature. We ranked the 20 genes based on their log fold change. The gene with the greatest fold change was ranked lowest. We then took the average of the ranks to generate a ranksum for each gene.

GSE15459				GSE62254			
	Gene Sym	LogFC	RANK	Gene nam	LogFC	RANK	RANKSUM
	AKAP12	-0.39676	2	AKAP12	-1.06843	13	15
	LAYN	-0.26813	5	LAYN	-1.15744	11	16
	LOC33926	-0.0977	13	LOC33926	-1.53482	4	17
	CALD1	-0.24302	8	CALD1	-1.16008	10	18
	HEYL	-0.13034	12	HEYL	-1.40337	6	18
	ITGB5	-0.031	17	ITGB5	-1.85498	1	18
	RASSF8	-0.50384	1	RASSF8	-0.91596	17	18
	TAGLN	-0.2188	10	TAGLN	-1.31754	8	18
	TPM2	-0.31684	4	TPM2	-0.97156	14	18
	TUBB6	-0.39496	3	TUBB6	-0.91817	16	19
	ACTA2	-0.14785	11	ACTA2	-1.18743	9	20
	LOXL4	-0.01912	18	LOXL4	-1.58562	3	21
	MATN3	-0.00157	19	MATN3	-1.76108	2	21
	MXRA7	-0.25219	6	MXRA7	-0.95868	15	21
	IGFBP7	-0.06513	15	IGFBP7	-1.39769	7	22
	NALCN	0.06581	20	NALCN	-1.49381	5	25
	LAMC1	-0.24333	7	LAMC1	-0.44048	19	26
	NREP	-0.0655	14	NREP	-1.14983	12	26
	TGFB2	-0.22494	9	TGFB2	-0.77724	18	27
	RAI14	-0.04749	16	RAI14	-0.37741	20	36

Table 4-2: Log2 fold change values and the ranks for 20 genes in the two discovery datasets. The genes highlighted in mustard yellow are the ones which were selected for hierarchical clustering analysis. The genes were selected based on a ranksum cutoff of 20.

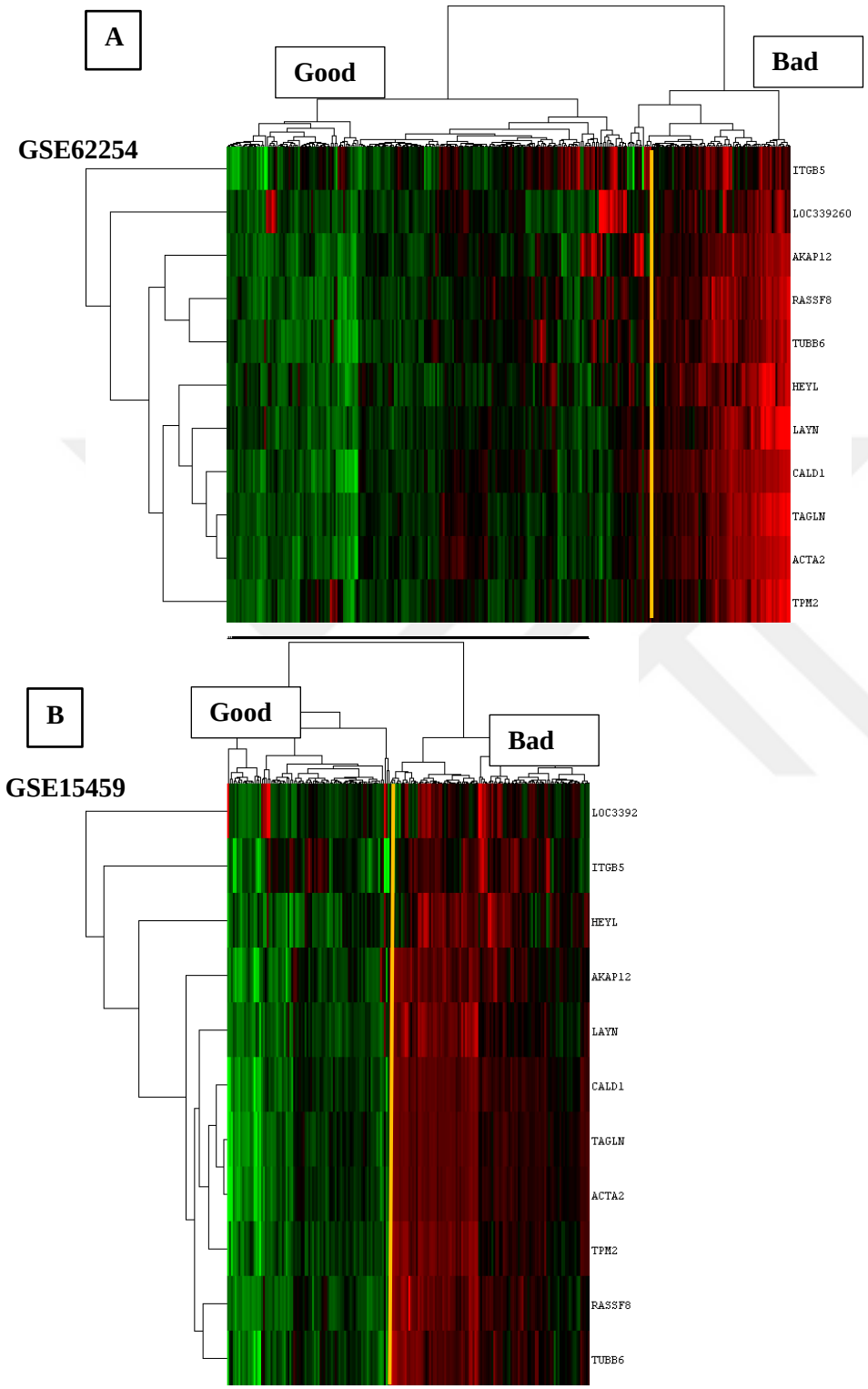
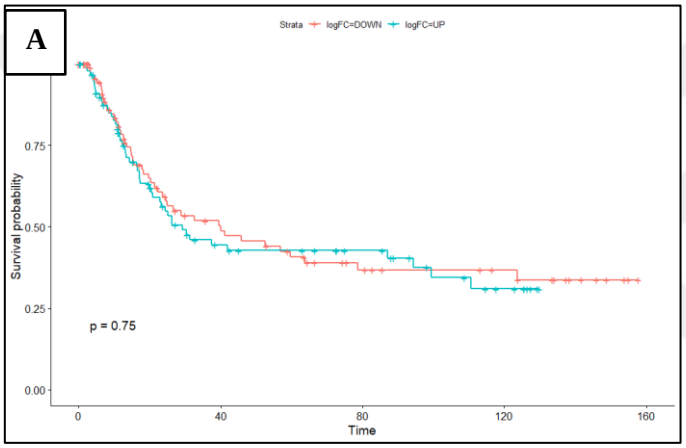
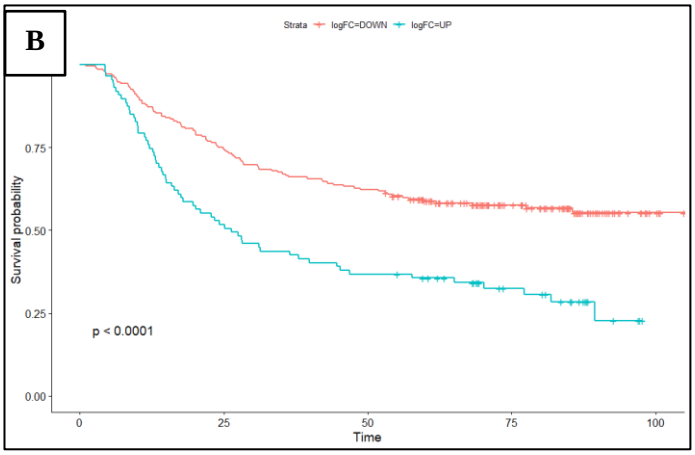


Figure 4.6: Hierarchical clustering analysis of the candidate genes picked based on log-FC. Discovery datasets GSE162254 and GSE15459 are A and B respectively. The orange line represent the separation of the good and bad groups.

We generated Kaplan-Meier plots for the genes selected by log-FC as well. However, one important point to note is that the logFC values in the cohort, GSE15459 were not significant whereas, the logFC values in the dataset, GSE62254 were mostly less than -1, meaning there is a significant fold change between the good and bad groups. We reasoned that this might be due to the heterogeneity in gene expression between gastric cancer patients. Moreover, there might be technical artefacts in the experiment design which gave rise to this inconsistency.



GSE15459



GSE62254

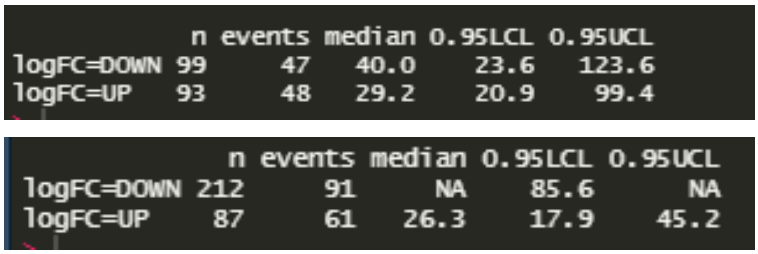


Figure 4.7: Kaplan-Meier plots for the two potential prognostic groups. A and B shows the survival plots for the two discovery cohorts, GSE15459, and GSE62254 respectively. Red graph represents the good group, labelled Down, and the blue graph represents bad group, labelled Down. Median survival is depicted below the graphs.

As expected, the survival plot for the dataset GSE15459 did not have a significant log-rank p-value.

4.1.4 Selecting a shorter gene-list based on protein level expression

After thorough research, we selected the genes ACTA2, TPM2, TAGLN, CALD1 and HEYL from our list of 20 genes. These 5 genes were present in the gene selected from the highly correlated list as well as the logFC-based list. Moreover, these 5 genes had high protein level expression in gastric tumour tissue according to the Human Protein Atlas. Four of these genes were associated with the extracellular matrix and HEYL was a transmembrane protein.
Our aim was to identify a prognostic signature which will be effective both at an mRNA level and protein level. If we can validate our genes, immunohistochemistry of these 5 genes in gastric cancer biopsies can be routinely performed in clinics alongside qPCR to have a more accurate two-fold diagnostic system.

4.2 Validation of the 5 gene signature by hierarchical clustering and survival analysis

In order to validate our gene-list *in silico*, we utilized two main approaches. Hierarchical clustering and survival analysis.
The discovery cohorts were used to first validate the efficacy of the 5-genelists by clustering and survival plots. We had 7 additional microarray datasets, all of which were used for validation via hierarchical clustering and 3 were used to generate additional survival plots.

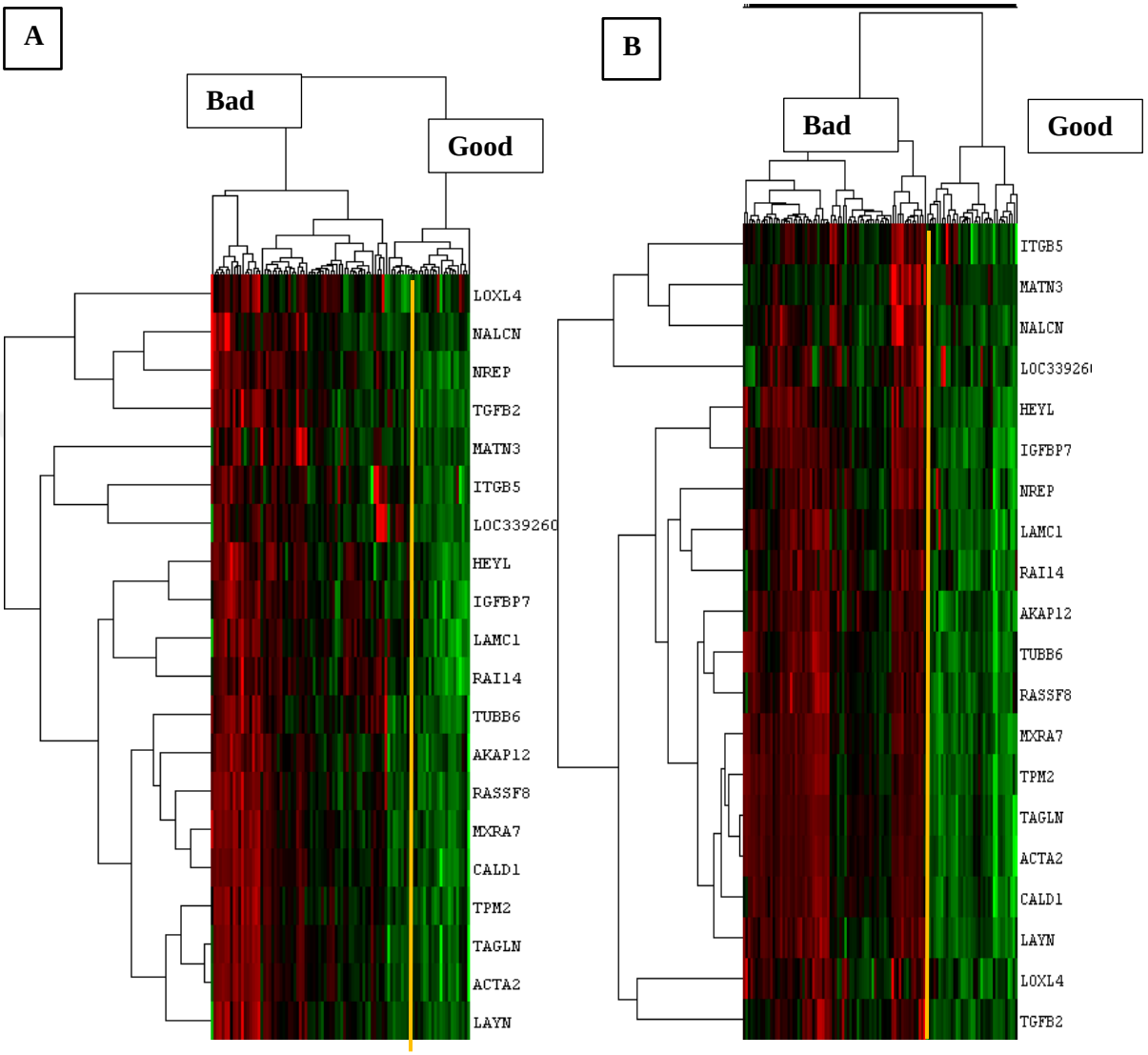
4.2.1 Patient stratification based on 5-gene signature

Before moving forward with the 5 genes, we needed to make sure that these genes hold similar capacity to discernibly categorize patients into comparable groups as the 20-gene mRNA signature. In order to do that, we selected 4 affymetrix datasets from Gene Expression Omnibus (GEO). All four datasets were from the platform, HGU-133 Plus 2. All datasets were RMA normalized in R and standardized before analyses. For simplicity these datasets, GSE51105,

GSE54129, GSE100935, and GSE 37303 will be referred to as cohorts-3, 4, 5, and 6 respectively from here on out.

First wanted to see if our 20 gene signature can stratify patients in independent datasets. Hierarchical clustering analysis was performed on cohorts-3, 4, 5, and 6.

Cohort-3 **Cohort-4**



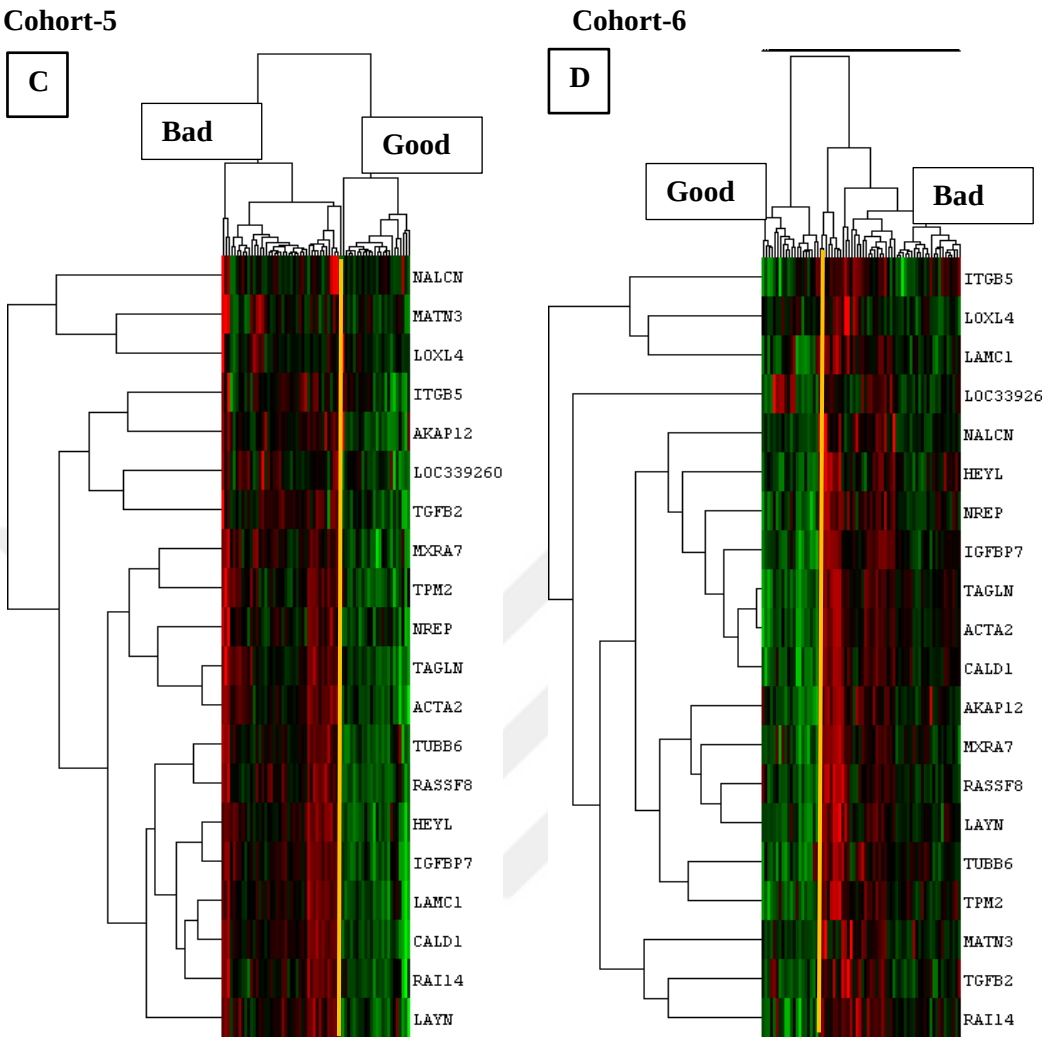


Figure 4.7 Results of hierarchical clustering analyses for 4 validation datasets with the 20 genes signature. A-D are cohorts 3-6 respectively. Orange lines separate the good and bad groups.

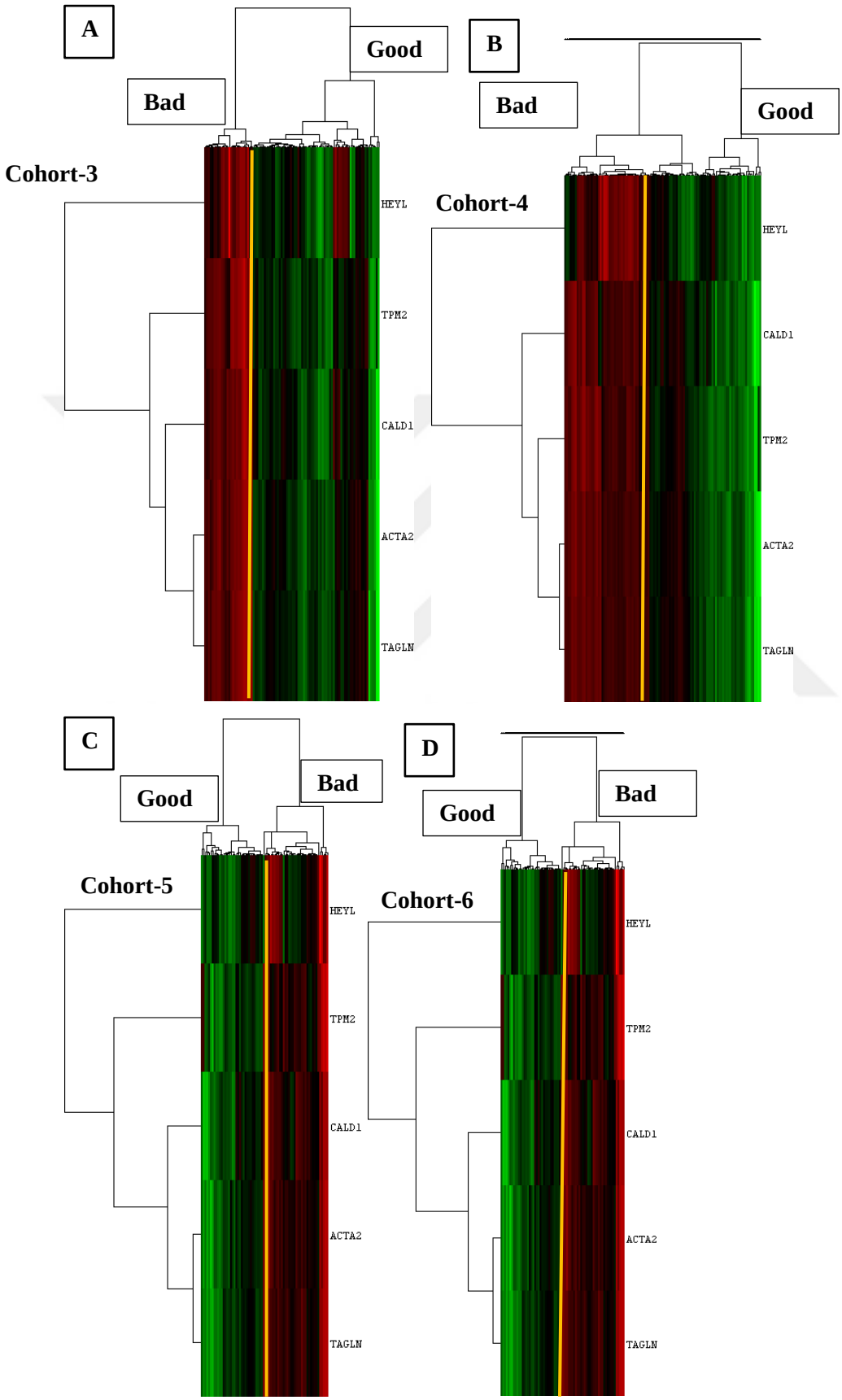

A
Good
Bad
Cohort-3
HEYL
TPM2
CALD1
ACTA2
TAGLN
B
Bad
Good
Cohort-4
HEYL
CALD1
TPM2
ACTA2
TAGLN
C
Good
Bad
Cohort-5
HEYL
TPM2
CALD1
ACTA2
TAGLN
D
Good
Bad
Cohort-6
HEYL
TPM2
CALD1
ACTA2
TAGLN

Figure 4.8: Results of hierarchical clustering analyses for 4 validation datasets with the 5-gene signature. A-D are cohorts 3-6 respectively. Orange lines separate the good and bad groups.

Dataset	Total Sample	Good-Good	Bad-Bad	Mismatch	Total Overlap	% Overlap
Cohort-3	94	57	24	13	81	87
Cohort-4	111	36	67	8	103	93
Cohort-5	66	23	40	3	63	95
Cohort-6	70	20	36	14	56	80

Table 4-3: Overlap between the 20-gene signature and the 5-gene signature in cohorts 3-6.

4.2.2 Validation of 5-gene signature in discovery cohorts

The two discovery cohorts GSE15459 and GSE62254, will henceforth be referred to as cohorts 1 and 2 respectively.

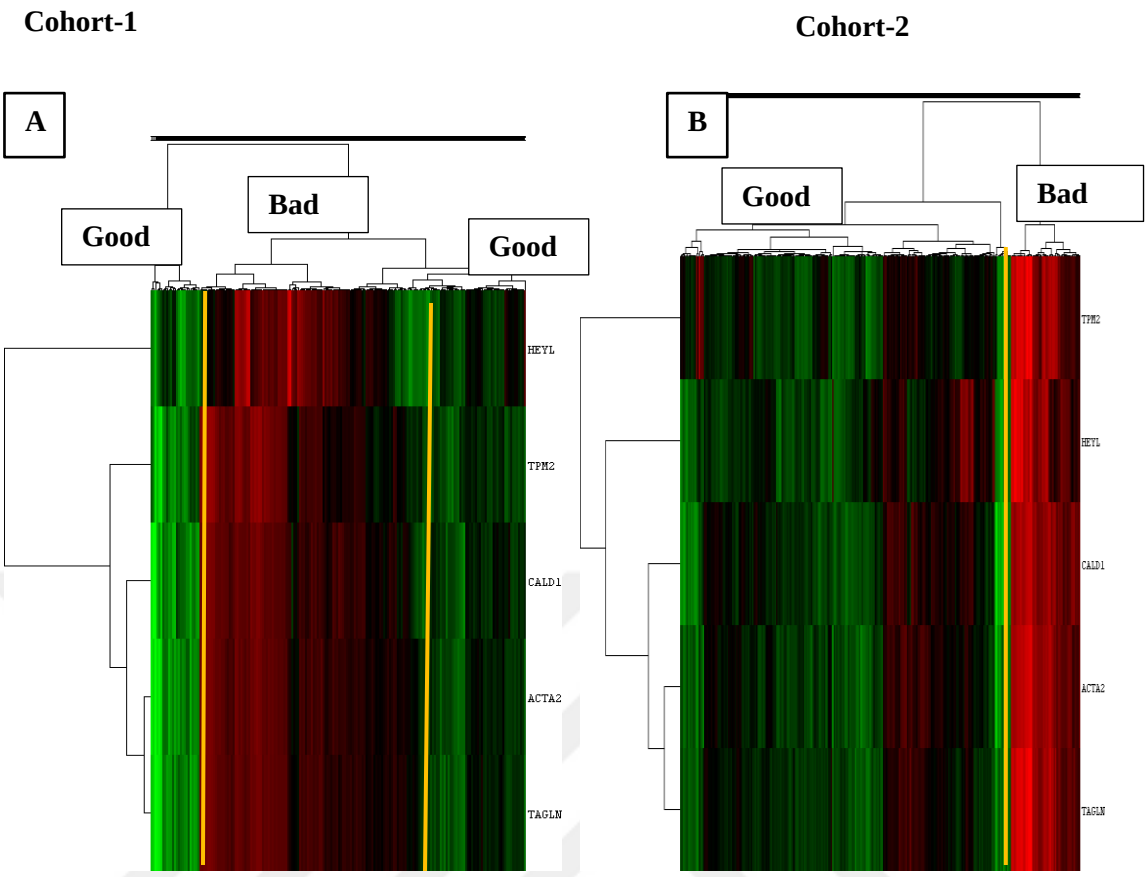
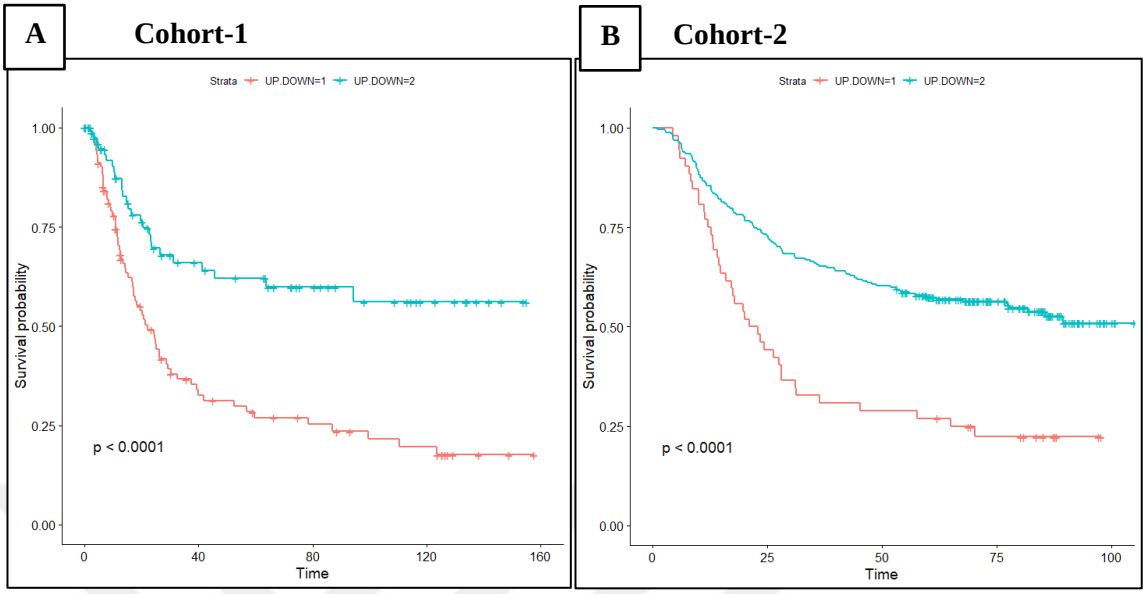


Figure 4.9: Hierarchical clustering analysis of cohorts 1 and 2 with the 5-gene signature. A and B represent cohorts 1 and 2 respectively. Orange lines represent where the Good and Bad groups were separated.

Dataset	Total Sample	Good-Good	Bad-Bad	Mismatch	Total Overlap	% Overlap
Cohort-1	192	82	89	21	171	89.1
Cohort-2	300	168	52	80	220	73.3

Table 4-4: Overlap between the patent groups stratified by the 20 gene-list and the 5-gene signature.




```
             n events median 0.95LCL 0.95UCL
UP.DOWN=1  106     69   22.3    17.1    30.4
UP.DOWN=2   86     26     NA    63.7      NA
```

Cohort-1

```
             n events median 0.95LCL 0.95UCL
UP.DOWN=1   52     40   21.9    15.9    31.3
UP.DOWN=2  247    112     NA    77.2      NA
```

Cohort-2

Figure 4.10: Survival plots for the 5-gene signature in cohorts 1 and 2.
A and B represent the K-M plots for cohorts 1 and 2 respectively. Red graph represents the good group, labelled Down, and the blue graph represents bad group, labelled Down. Log-rank p-value is shown in the graph, the median survival for each group is given below the graph.

4.2.3 Validation of the 5-gene signature by survival analysis

Three validation datasets were acquired from GEO which contained clinical data. The datasets were generated using the Illumina beadchip technology and the normalization process used was quantile normalization. The GEO accession number of the datasets were, GSE26901, GSE13861, and GSE26899, which from now on will be referred to as cohorts 7-9 respectively. The 5-genes were used to hierarchically cluster the cohorts into Good and Bad groups.

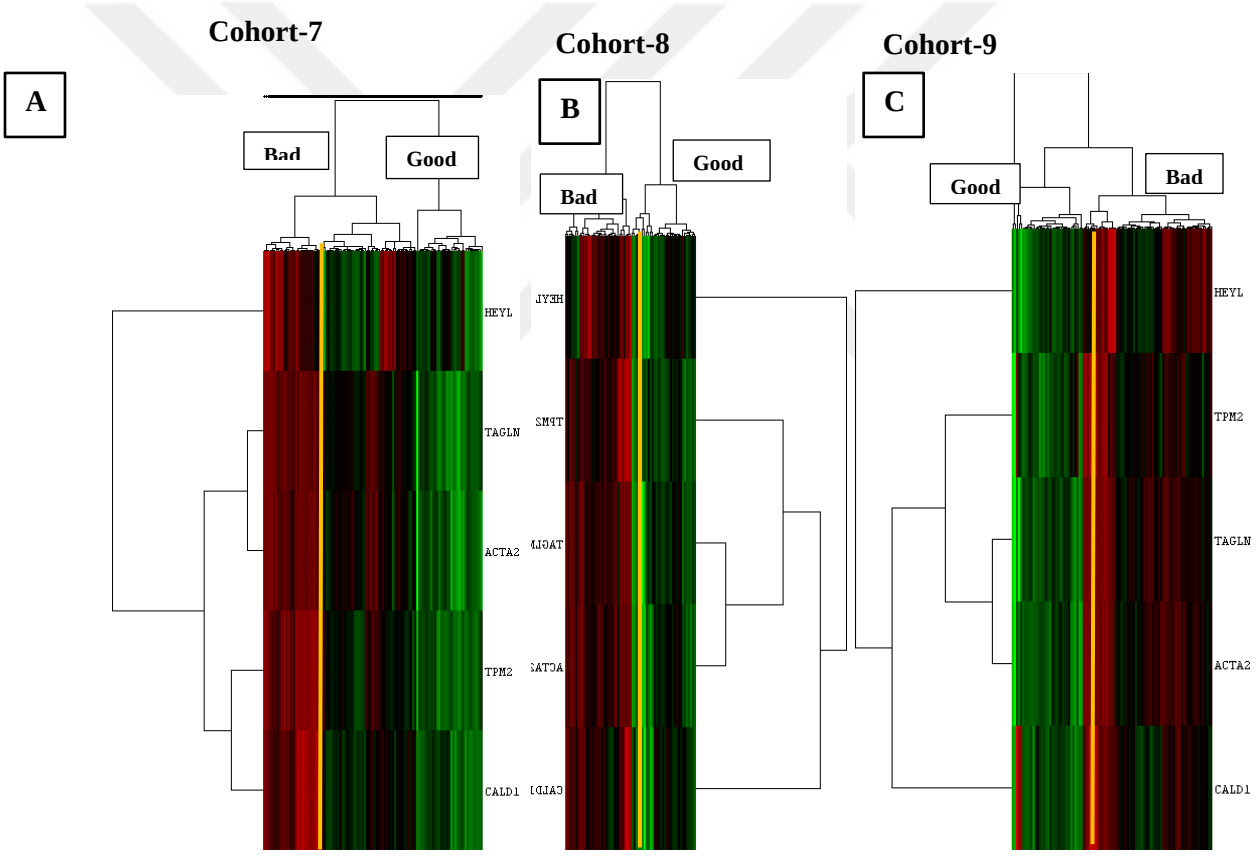


Figure 4.11: Hierarchical clustering results for the survival datasets.

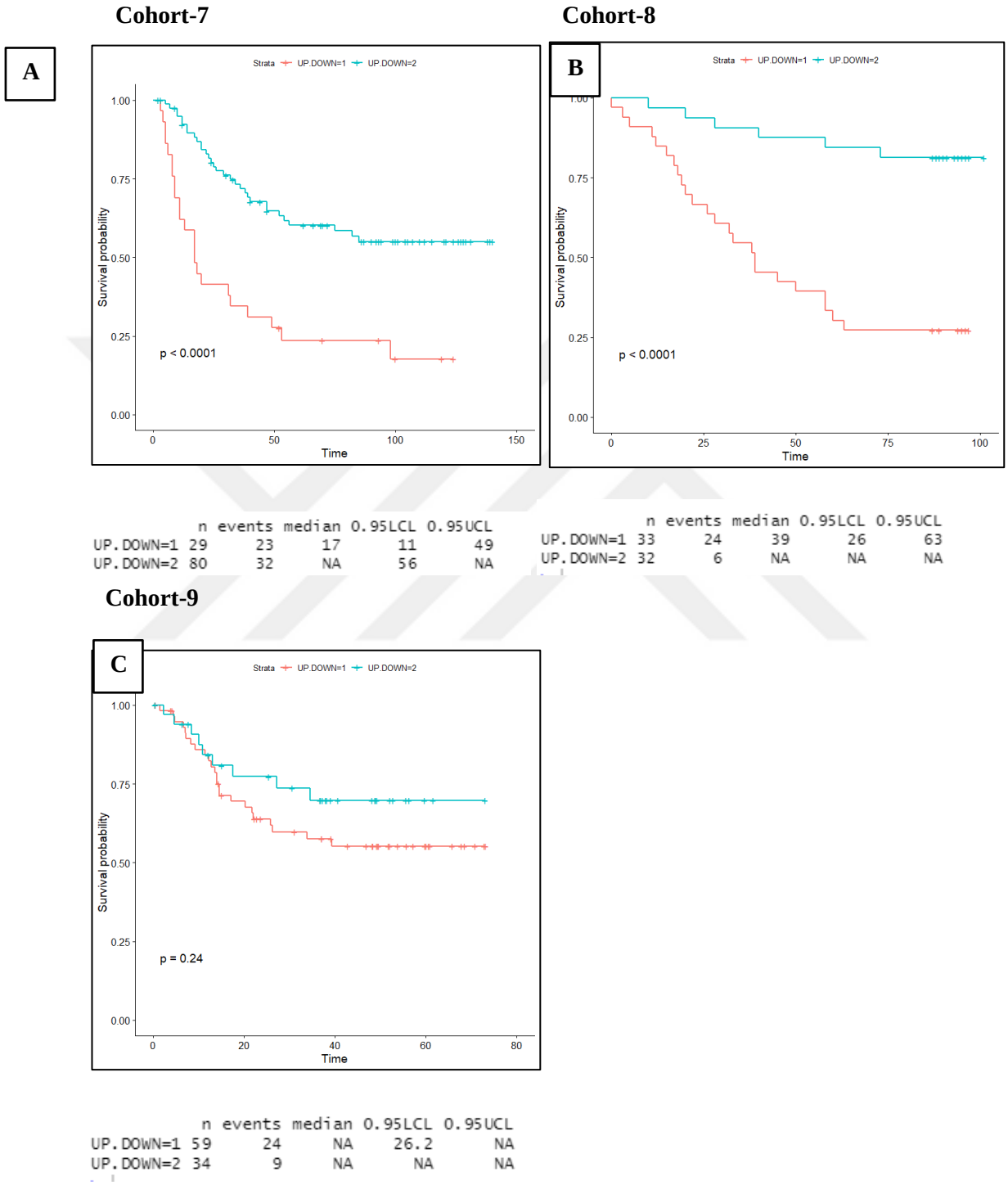


Figure 4.12: Survival plots for the 3 validation cohorts. A-C are Cohorts 7-9 respectively. Red graph represents the good group, labelled Down, and the blue graph represents bad group, labelled Down. Log-rank p-value is shown in the graph, the median survival for each group is given below the graph. Log-rank based multiple cutoff curves were plotted for the three datasets in RStudio.

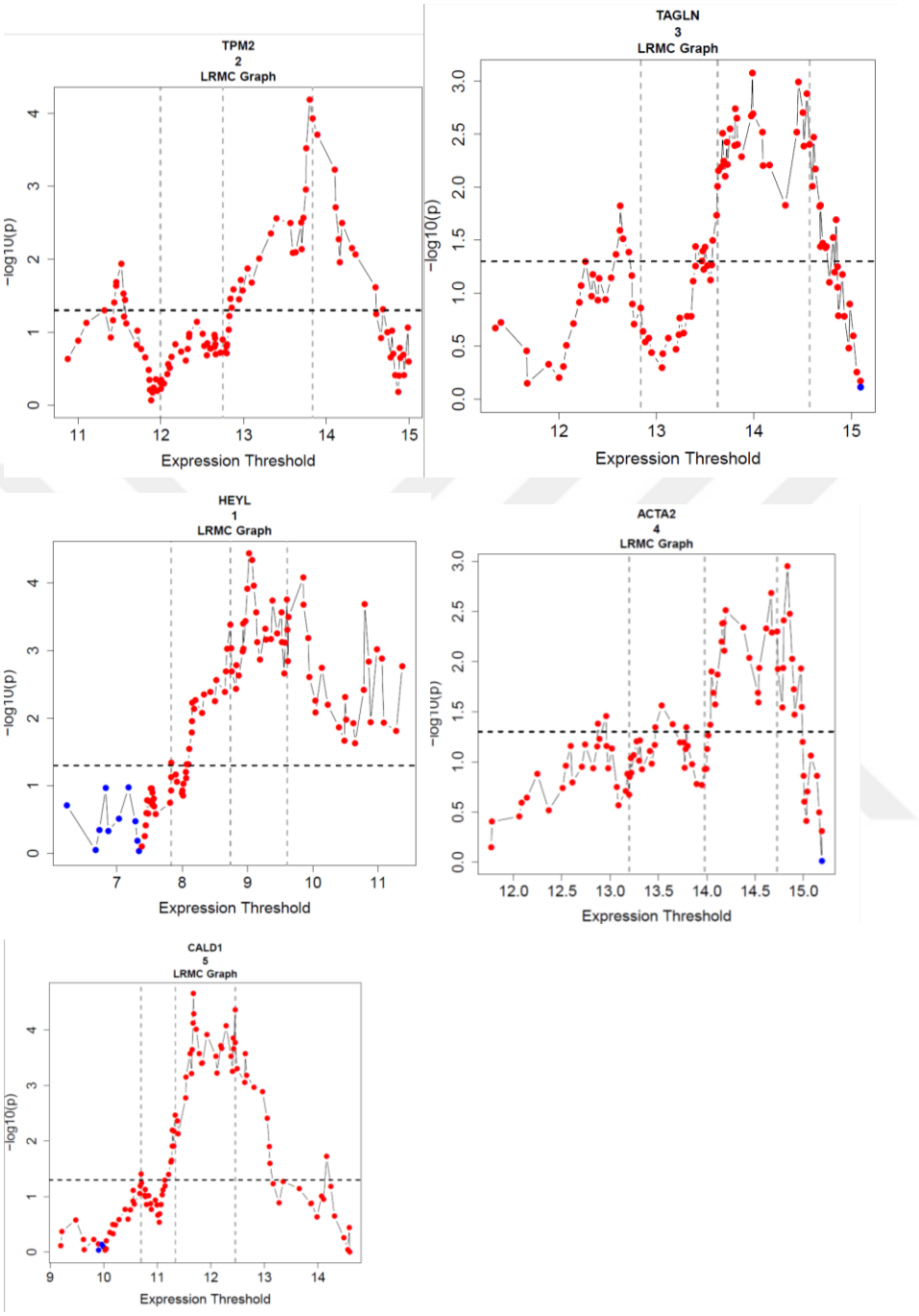


Figure 4.13: LRMCs for cohort 7 with the 5 genes TPM2, TAGLN, HEYL, ACTA2, and CALD1.

The horizontal dotted line is the threshold for log-rank p-value <0.05; the blue dots correspond to gene expression values with a good prognosis whereas the red dots correspond to worse prognosis. The vertical dotted lines represent the 25th, 50th and 75th percentile expression value.

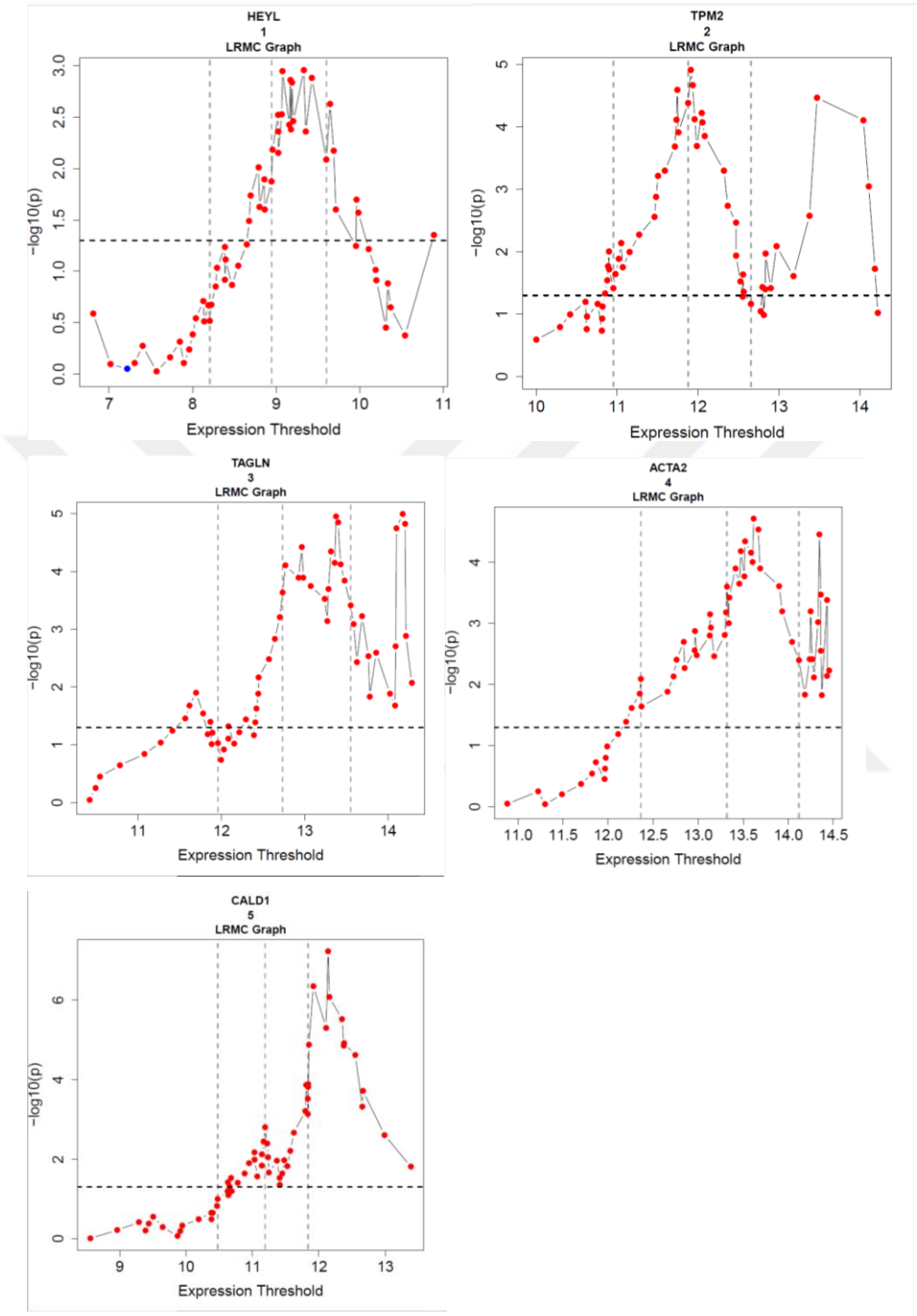


Figure 4.14: LRMCs for cohort 8 with the 5 genes TPM2, TAGLN, HEYL, ACTA2, and CALD1. The horizontal dotted line is the threshold for log-rank p-value <0.05; the blue dots correspond to gene expression values with a good prognosis whereas the red dots correspond to worse prognosis. The vertical dotted lines represent the 25th, 50th and 75th percentile expression value.

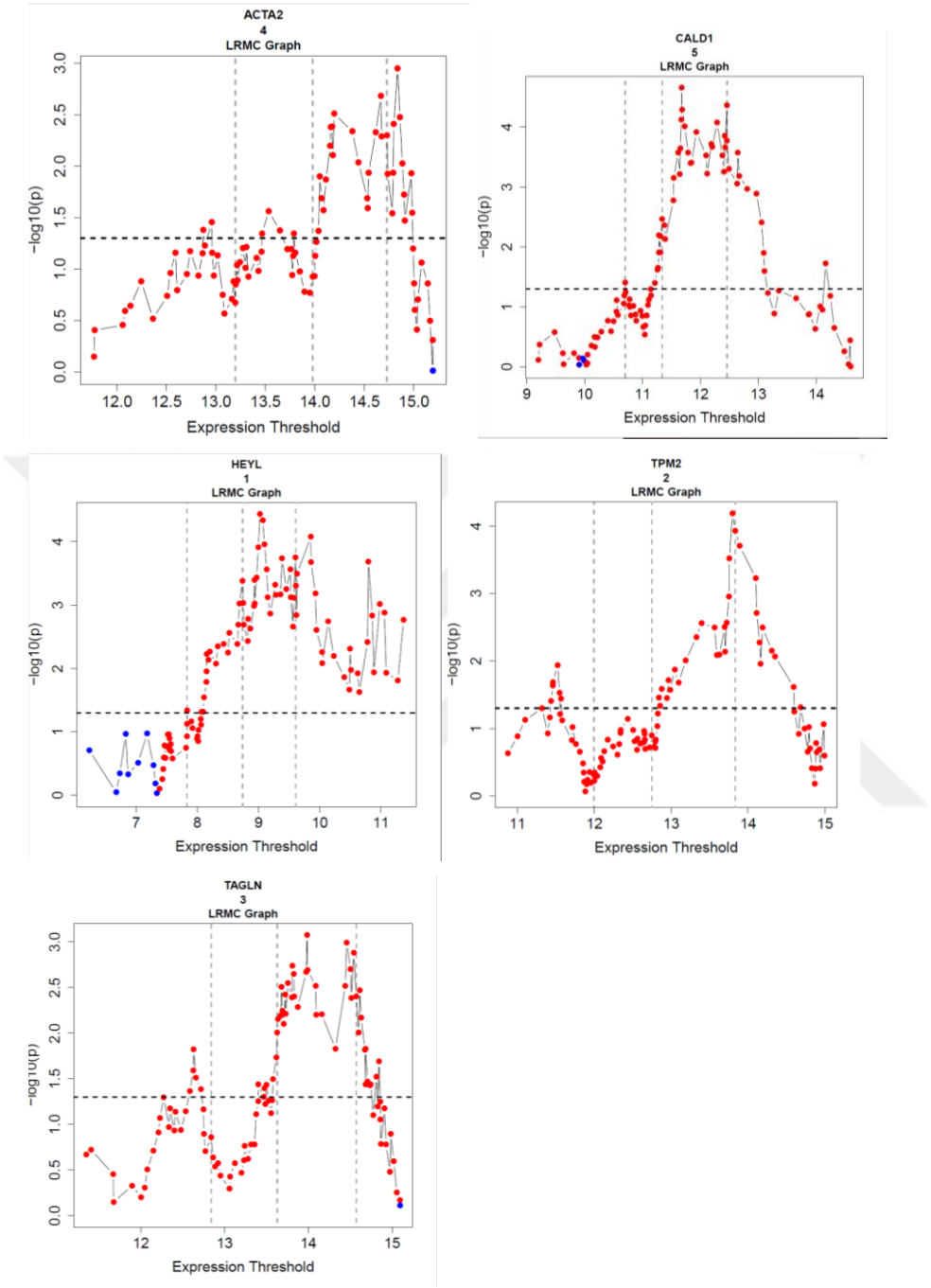


Figure 4.15: LRMCs for cohort 9 with the 5 genes TPM2, TAGLN, HEYL, ACTA2, and CALD1. The horizontal dotted line is the threshold for log-rank p-value <0.05; the blue dots correspond to gene expression values with a good prognosis whereas the red dots correspond to worse prognosis. The vertical dotted lines represent the 25th, 50th and 75th percentile expression value.

4.3 Discovering gastric tissue-specific reference genes using high-throughput transcriptomic data

We developed an in-house algorithm to identify the genes that are the most stably expressed in gastric cancer tissue. The statistical method we chose was coefficient of variance calculation. We would calculate the standard deviation of each gene and then divide by its average expression to generate its coefficient of variance. Our hypothesis was, lower the coefficient of variance (CV) the more stably the gene will be expressed in gastric cancer tissue. CV is a good indicator of the dispersion of expression values in a data set around the mean. The implementation of this method allows us to compare genes with drastically different average expression levels across a date set.

As a sample set we chose The Cancer Genome Atlas (TCGA) data for our analysis. TCGA has RNA-seq data for 413 gastric cancer patients available publicly. For RNA-seq data normalization we utilized two different normalization methods, DeSeq2 and RSEM.

We acquired the RSEM normalized data-set from FireBrowse, which is curated by Broad Institute. For the Deseq2 normalization we downloaded the HT-seq files from the Genomic Data Commons (GDC). The RSEM software package takes in raw sequencing data and quantify both paired-end and single-end reads; the software has a normalization technique which accounts for fragment length and thus the final count is not affected by the length of the transcript. RSEM normalization is a very flexible normalization technique in terms of its applications, thus we chose to use it for our analysis.

Deseq2 is an R package which specializes on differential expression analysis. It takes a matrix of HT-seq counts as an input. The package allows downstream analysis of the gene list to identify differentially expressed genes. However, for our purpose, we did not perform any downstream analysis. Both the Deseq2 normalized and the RSEM normalized data were log2 transformed.

The code we wrote takes a normalized, log2 transformed RNA-seq data as an input where the columns correspond to the samples and the rows correspond to the genes. We named it Random, repeated sub-sampling method for stable gene identification.

In the first step, the code creates two empty matrices and then randomly assigns samples from the input into these two matrices. After generating the matrices, the code calculates the CV for each gene and stores these values into two new tables. Then the code ranks the genes based on the coefficient of variance calculated and stores the ranks in two matrices titled “ranks 1” “ranks

2”. The genes are ranked by ascending order of coefficient of variance. Meaning the gene with the lowest coefficient of variance would be ranked 1.

The code then repeats these steps for a thousand iterations. The reason we divided the matrix into two groups is primarily because TCGA has over 400 gastric cancer samples giving the two groups over 200 samples each, which is a statistically significant number of samples to make an observation. By dividing the data into two groups we could have a discovery cohort and a validation cohort. So if the same genes are ranked top in both groups, we can consider it for further analysis.

We performed a thousand iterations of the code to randomly shuffle the 400 samples into two groups and rank the genes according to their CV in every iteration. After a thousand iterations, the genes that consistently appear in to be best ranked would indicate that they are stably expressed. Since ranks correspond to CV and lower CV indicate a more stable expression across all the samples. After we generated the two rank tables, discovery and validation, for a thousand iterations, we calculated the average rank for each gene.

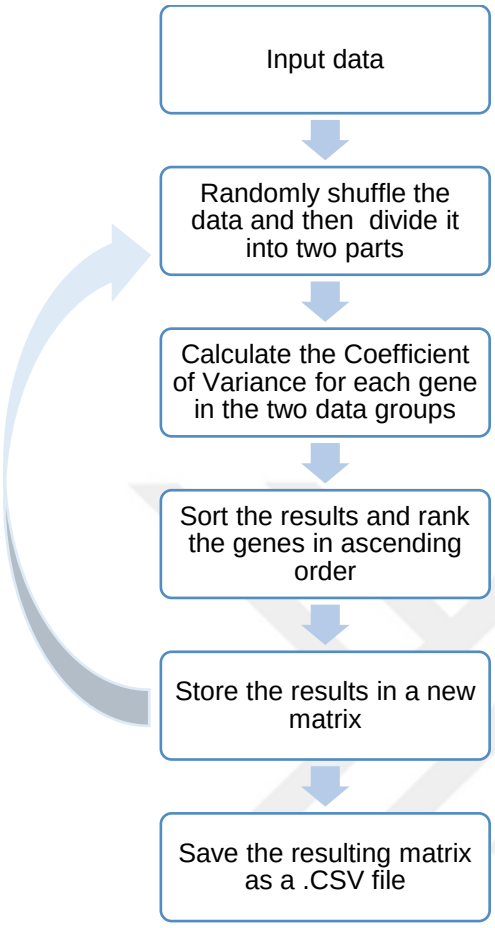


Figure 4.16: A visual representation of the code to discover novel reference genes.

4.3.1 The results from TCGA and CCLE

The best ranked candidate reference genes were plotted against the more conventionally known reference genes to visualize the variation difference between the genes.

The rationale behind using two types of normalization to analyze TCGA data was to observe any perceived difference between coefficients of variance. The objective was to ensure the code is applicable to all sorts of high-throughput gene expression data, regardless of the normalization used. The top ranked genes from the De-Seq2 results and the RSEM results were compared; the ranks were comparable and the top list of genes were similar. Thus, for the final selection of reference genes, RSEM data from TCGA and CCLE were considered.

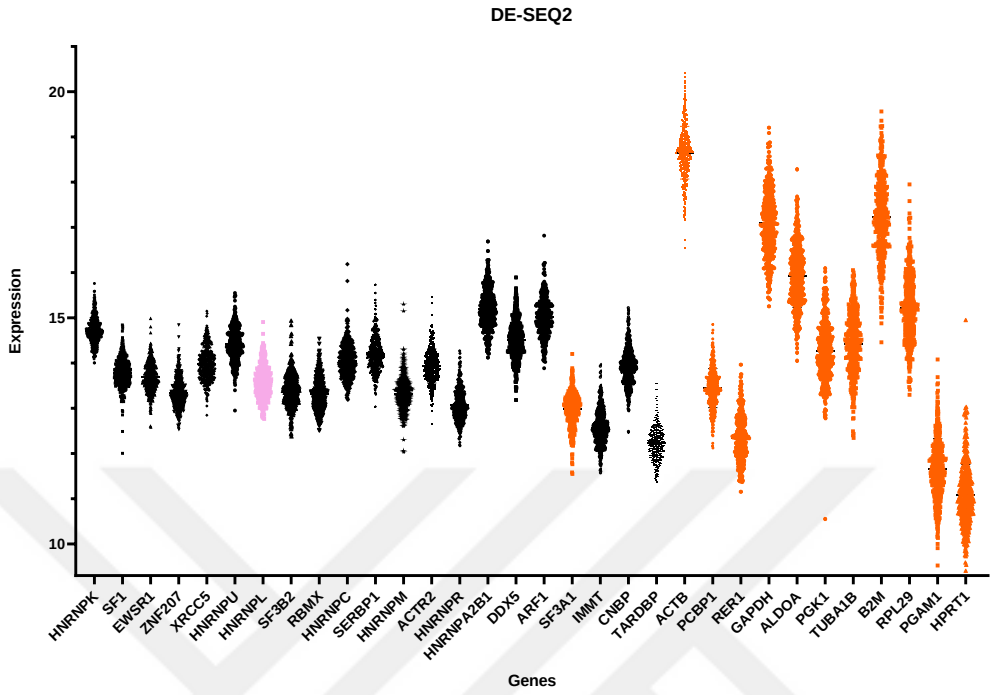


Figure 4.17: Expression pattern of top ranked reference genes (black) against the more well-known reference genes (orange) in TCGA data. Data was normalized by the De-seq2 package. Pink represents genes that were discovered to be a good reference gene in literature.

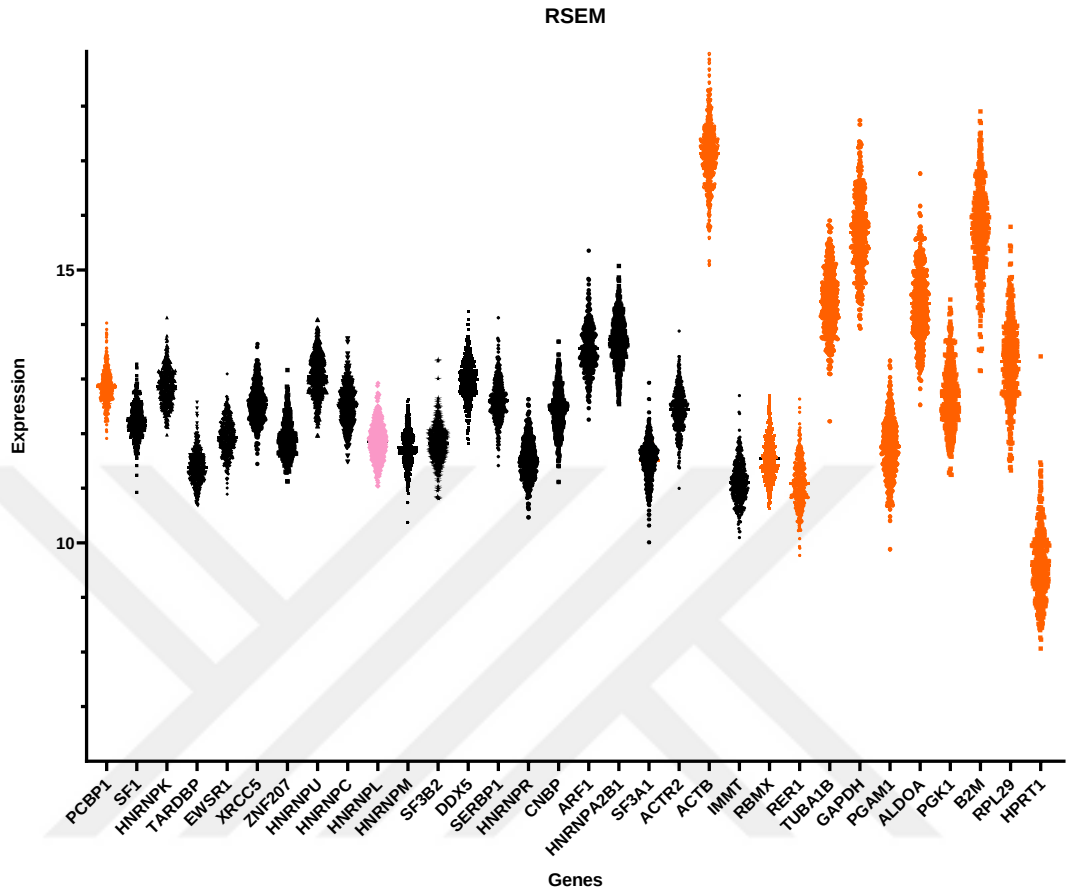


Figure 4.18: Expression pattern of top ranked reference genes (black) against the more well-known reference genes (orange) in TCGA data. Data was normalized by RSEM. Pink represents genes that were discovered to be a good reference gene in literature.

Instead of identifying suitable reference genes in gastric cancer tumors only, we aimed to use the RNA-seq data from Cancer Cell Line Encyclopedia (CCLE) to identify a list of reference genes which would act as control genes for both gastric tissue and gastric cancer cells lines. CCLE has RNA-seq RSEM normalized data for 37 gastric cancer cell lines. We ran our novel reference gene code with this data to identify a list of top reference gene candidates. 150 of the top genes from CCLE and TCGA were compared to generate a list of common genes. The resulting list is displayed in table 5.

In figures 4.17-4.19, the expression patterns of the top 20 reference gene candidates (black) are presented alongside 12 known reference genes (orange). The one consistent pattern that emerges is that our method identifies reference genes which less varied expression than the well-known reference genes, particularly GAPDH, B2M and ACTB.

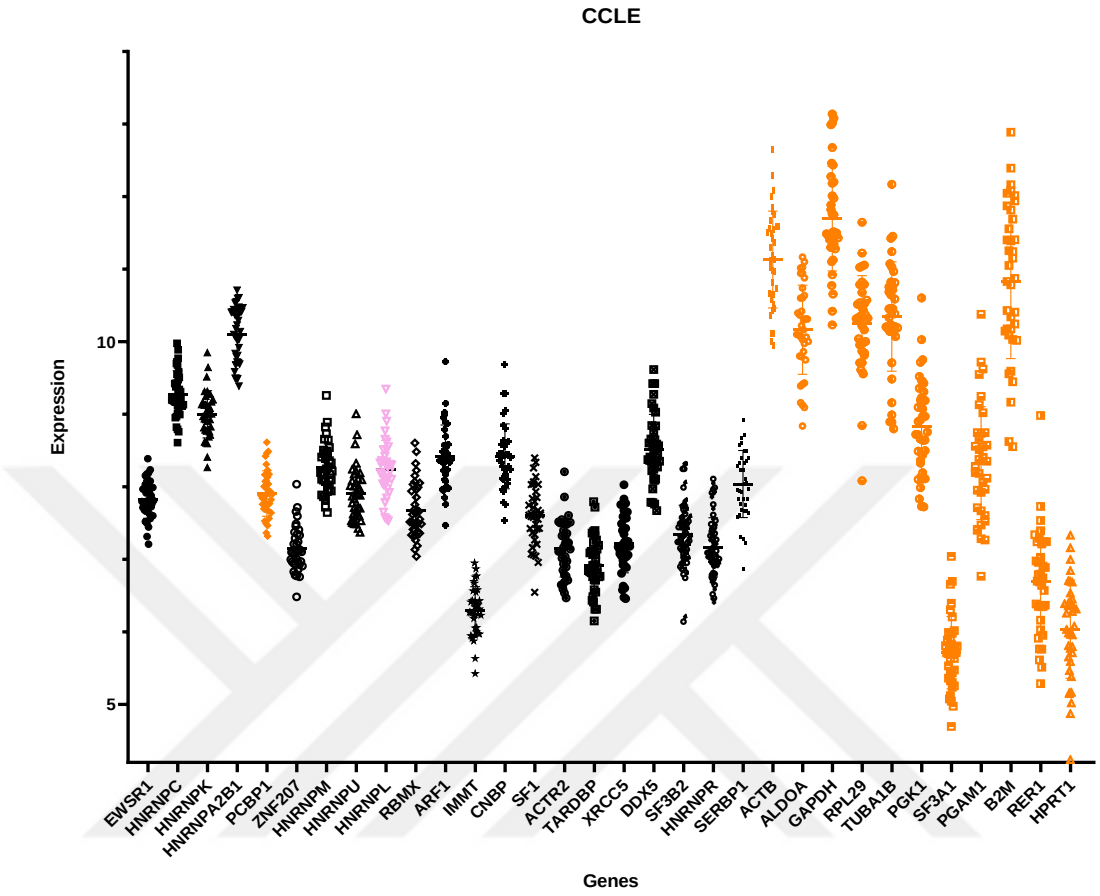


Figure 4.19: Expression pattern of top ranked reference genes (black) against the more well-known reference genes (orange) in CCLE data which was normalized by RSEM. Pink represents genes that were discovered to be a good reference gene in literature.

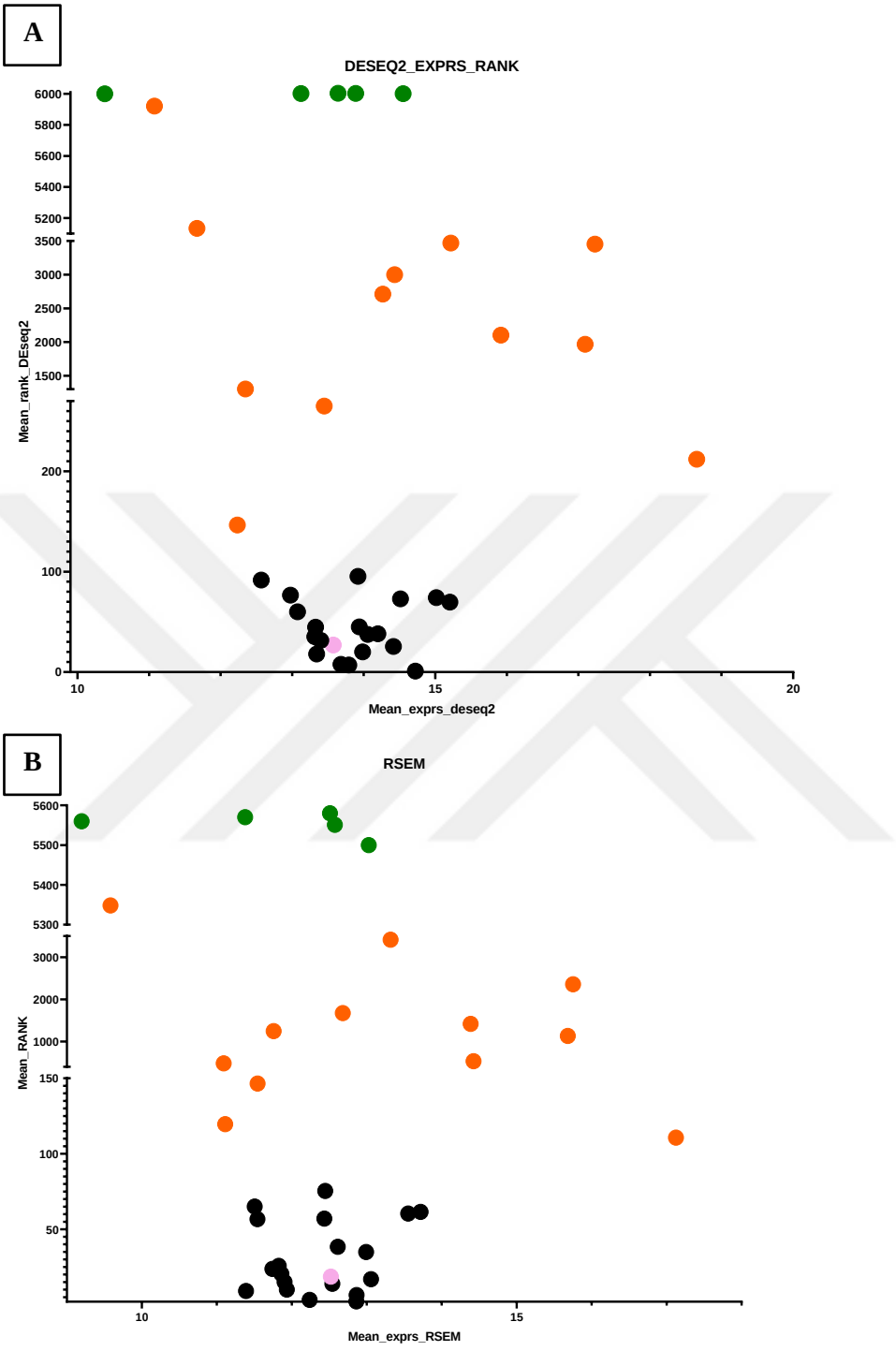


Figure 4.20: Mean_rank vs. expression plots. A) TCGA data with De-Seq2 normalization. B) TCGA data with RSEM normalization. Green- prognostic gene signature, Orange- known reference genes, Pink- not novel, Black- Candidate Reference genes

The purpose of identifying gastric tissue specific reference genes was to perform qPCR with our 5-gene prognostic signature to accurately assess the mRNA levels of our genes in our gastric cancer cohort. Thus, we aimed to see if the expression levels of our reference genes were comparable to our prognostic genes; the objective was to select 3 reference genes whose

expression covered the range of expression for our 5 prognostic genes. However, as can be seen in figure 4.20, this proved difficult as the expression range of our candidate reference genes were less variable than that of the prognostic genes. Thus, we selected the reference genes to test with qPCR based on their ranks.

Genes	*Mean_RANK_RSEM*	*Mean_Rank_CCLE*	*Rank_Mean*
EWSR1	**10.0**	**8.0**	**9.0**
HNRNPK	**6.5**	**14.5**	**10.5**
HNRNPC	18.6	10.2	14.4
ZNF207	15.1	38.2	26.6
HNRNPM	23.7	40.3	32.0
HNRNPU	16.9	59.9	38.4
HNRNPA2B1	61.5	17.5	39.5
HNRNPL	20.5	77.6	49.0
SF1	**3.2**	**152.9**	**78.1**
TARDBP	9.1	155.0	82.0
XRCC5	13.9	176.0	95.0
ARF1	60.4	139.5	99.9
CNBP	57.0	143.5	100.3
SF3B2	25.7	191.0	108.4
DDX5	34.9	187.9	111.4
ACTR2	75.4	154.4	114.9
RBMX	146.5	89.1	117.8
HNRNPR	56.7	195.0	125.8
SERBP1	38.4	213.3	125.9
IMMT	119.6	139.8	129.7

Table 4-5: Top ranked candidate reference genes. The genes highlighted red are the genes we've selected for validation via qPCR.

The sum of the mean ranks for TCGA_RSEM and CCLE_RSEM was calculated for each gene. Then, the top two were selected, EWSR1 and HNRNPK. SF1 was selected as it had the highest rank in tumor tissue.

Figure 4.21 depicts the expression patterns or our candidate reference genes compared to 12 known reference genes for TCGA (right) and CCLE (left).

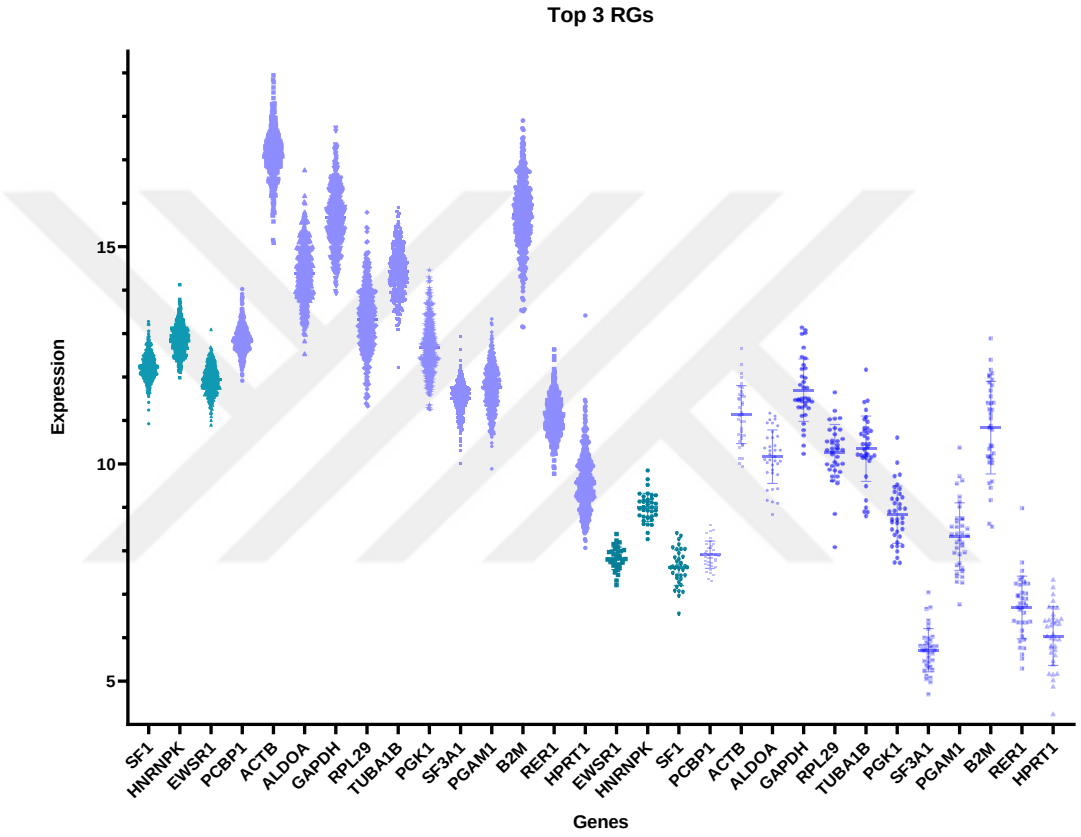


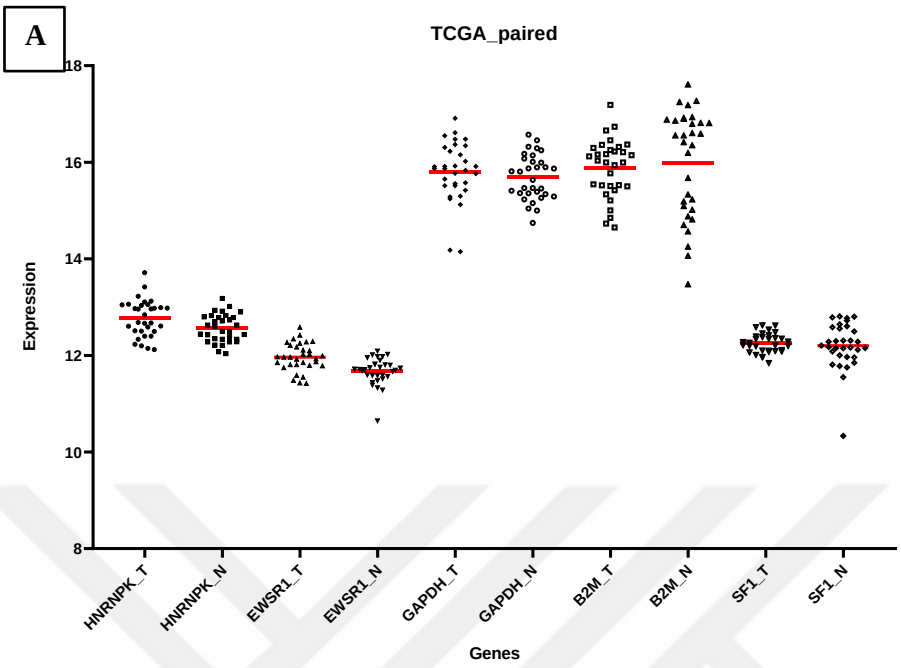
Figure 4.21: The gene expression pattern of our 3 candidate reference genes compared to the 12 known reference genes. Teal- candidate reference genes, lilac- 12 known reference genes. TCGA (on the left) and CCLE (to the right).

4.3.2 Comparing expression of our candidate reference genes in normal gastric tissue

Upon selecting a list of reference genes for further validation in cancer tissue and cell lines, we set forth to analyze the expression levels of our candidate reference genes in normal gastric tissue. Our aim was to observe whether our candidate genes are expressed stably in normal tissues or not. For this analyses, we ultimately selected two microarray datasets, 2 and 3 from table 4.6; as well as TCGA, which has tumor adjacent normal tissue expression data for 35 patients; out of which 32 patient barcodes could be matched to their tumor RNA-seq data.

GEO accession ID		Platform	Tumor sample	Normal Samples	Description	Location of study
GSE54129	Dataset 1	HG-U133_Plus_2	111	21	Normal gastric tissue were collected from healthy individuals undergoing heath examination	China
GSE66229	Dataset 2	HG-U133_Plus_2	300	100	Normal tissues were collected from 100 patient-matched gastric cancer patients.	Korea
GSE79973	Dataset 3	HG-U133_Plus_2	10	10	Paired data, tumor tissue and normal gastric mucosa were collected from 10 patients	China

Table 4-6: Datasets selected to analyze expression levels in tumor adjacent normal tissues for the 3 candidate reference genes.



B

Genes	P-value
B2M	0.615412
EWSR1	**0.000206**
GAPDH	0.363175
HNRNPK	**0.006563**
SF1	0.568597

Figure 4.22: Comparison of candidate control gene expression between normal and tumor tissue in TCGA. A) Expression patterns of paired tumor and normal tissue in TCGA. N_gene-name denotes gene expression in normal tissue and T_gene-name denotes expression in tumor. B) Table representing p-values from a paired t-test between gene expression in tumor tissue and normal tissue.

To compare the expression levels of our candidate genes, EWSR1, SF1 and HNRNPK, we chose to use GAPDH and B2M due to their popularity as control genes. In figures 4.22 (A), 4.23 and 4.24 (A), it appears that the expression levels of the 5 genes are comparable between tumor and normal gastric tissue. However, in 4.22 (B) p-values for the paired t-test shows significance difference between tumor and normal for the genes, EWSR1 (p-value= 0.000206) and HNRNPK (p-value= 0.006563). Moreover, in 4.24 (B), paired t-test returned significant results for GAPDH (p-value= 0.000599), HNRNPK (p-value= 0.004182) and SF1 (p-value= 0.044789).

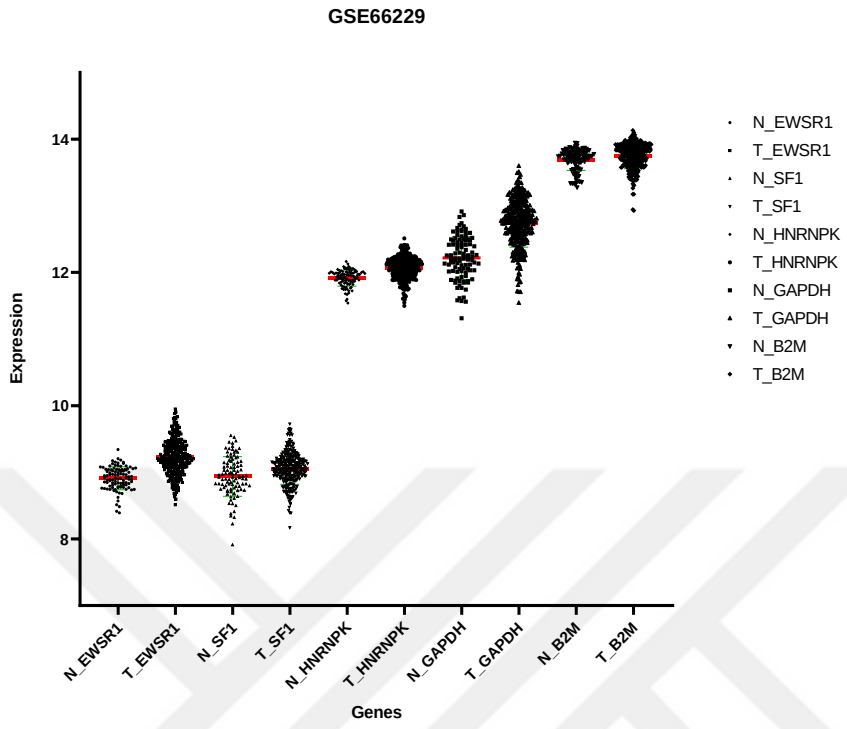


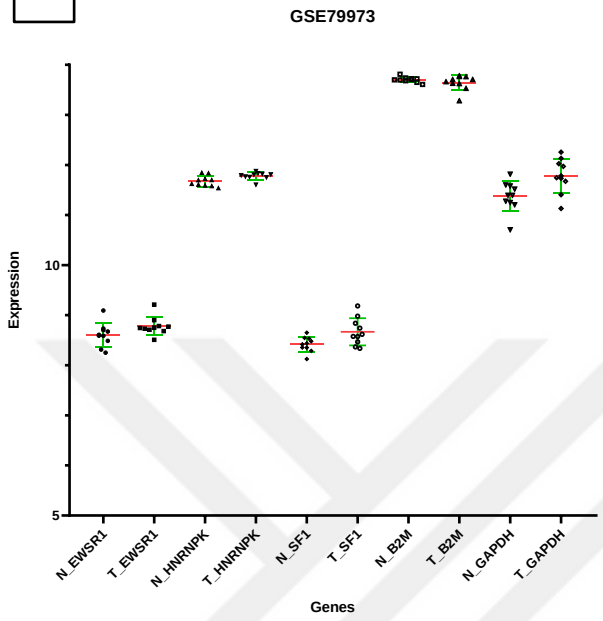
Figure 4.23: 23 Comparison of candidate control gene expression between normal and tumor tissue in dataset_2.

Expression patterns of 300 tumor and 100 normal tissue in dataset 2. N_gene-name denotes gene expression in normal tissue and T_gene-name denotes expression in tumor tissue.

Although, HNRNPK is differentially expressed in tumor and normal gastric tissues two datasets, the other genes only are significant in either dataset with paired data. Besides t-test, we aimed to verify whether there is potential differential expression of the 5 genes by calculating their log fold change (logFC) using average expression value.

In figure 4.25, the logFC in each dataset is displayed. Here we see that there is no discernible or significant difference in gene expression between normal and tumor tissues. Due to the inconsistency between the t-test results in the two paired datasets and he logFC values, we decided to move forward with the logFC values as reference.

A



B

Genes	p-values
B2M	0.242083
EWSR1	0.05017
GAPDH	**0.000599**
HNRNPK	**0.004182**
SF1	**0.044789**

Figure 4.24: Comparison of candidate control gene expression between paired normal and tumor tissue in dataset_3. A) Expression patterns of paired tumor and normal tissue in dataset 3. N_gene name denotes gene expression in normal tissue and T_gene name denotes expression in tumor. B) Table representing p-values from a paired t-test between gene expression in tumor tissue and normal tissue.

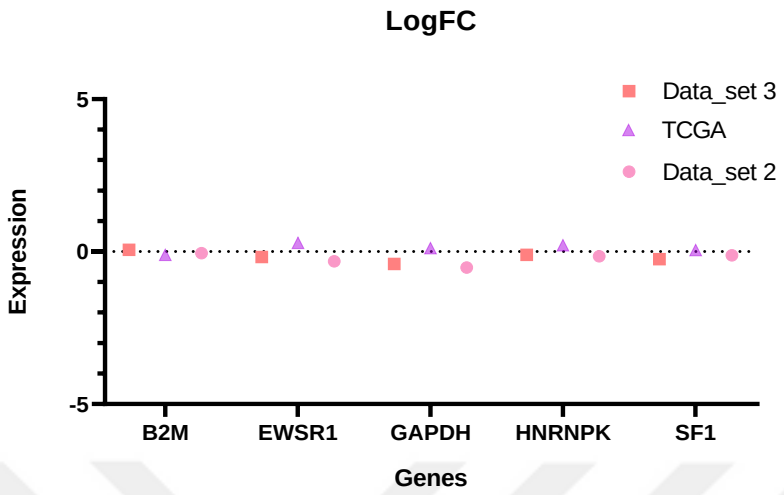


Figure 4.25: A scatter plot showing the lack of variation in the log fold change value of the reference genes between tumor and normal tissue.

4.3.3 Assessing the efficacy of our novel reference genes compared to known reference genes

In order to estimate whether our candidate reference genes have an added advantage to be used as control genes for gastric tissue, both normal and tumor, as well as in stomach cancer cell lines we developed a simple mathematical concept. We postulated, if the ratios of mean gene expression between each of the candidate reference genes have a lower variation than GAPDH and B2M, then we can claim that they are better suited as control genes in gastric cancer tissues and cell lines.

To verify our claim, we calculated the log2-difference between all 5 reference genes with one another to generate a simple matrix which holds the ratios between each of the genes. For this analysis, we utilized datasets 2 and 3 from table 4.6 as well as TCGA gastric data.

A

		EWSR1	SF1	HNRNPK	GAPDH	B2M
		8.90957	8.93775	11.9141	12.2176	13.6922
EWSR1	8.90957	0.0	0.0	3.0	3.3	4.8
SF1	8.93775	0.0	0.0	3.0	3.3	4.8
HNRNPK	11.9141	-3.0	-3.0	0.0	0.3	1.8
GAPDH	12.2176	-3.3	-3.3	-0.3	0.0	1.5
B2M	13.6922	-4.8	-4.8	-1.8	-1.5	0.0

B

		EWSR1	SF1	HNRNPK	GAPDH	B2M
		8.60508	8.41859	11.6752	11.3773	13.7037
EWSR1	8.60508	0.0	-0.2	3.1	2.8	5.1
SF1	8.41859	0.2	0.0	3.3	3.0	5.3
HNRNPK	11.6752	-3.1	-3.3	0.0	-0.3	2.0
GAPDH	11.3773	-2.8	-3.0	0.3	0.0	2.3
B2M	13.7037	-5.1	-5.3	-2.0	-2.3	0.0

C

		EWSR1	SF1	HNRNPK	GAPDH	B2M
		11.6749	12.2017	12.5645	15.6892	15.9915
EWSR1	11.6749	0.0	0.5	0.9	4.0	4.3
SF1	12.2017	-0.5	0.0	0.4	3.5	3.8
HNRNPK	12.5645	-0.9	-0.4	0.0	3.1	3.4
GAPDH	15.6892	-4.0	-3.5	-3.1	0.0	0.3
B2M	15.9915	-4.3	-3.8	-3.4	-0.3	0.0

Figure 4.26: Expression ratios between each gene in the 3 datasets for normal tissues. Ratios between average expressions in A) dataset_2, B) dataset_3, C) TCGA.

In figure 4.26, the ratios between the genes is consistent throughout all three datasets which suggests that the datasets are reliable. Moreover, the consistency suggests that our method will most likely generate reliable results.

We generated similar ratio tables with tumor expression from the same three datasets.

A

		B2M	EWSR1	GAPDH	HNRNPK	SF1
		13.75	9.22	12.74	12.07	9.06
B2M	13.75	0.0	-4.5	-1.0	-1.7	-4.7
EWSR1	9.22	4.5	0.0	3.5	2.8	-0.2
GAPDH	12.74	1.0	-3.5	0.0	-0.7	-3.7
HNRNPK	12.07	1.7	-2.8	0.7	0.0	-3.0
SF1	9.06	4.7	0.2	3.7	3.0	0.0

B

		B2M	EWSR1	GAPDH	HNRNPK	SF1
		13.64	8.78	11.79	11.78	8.67
B2M	13.64	0.0	-4.9	-1.9	-1.9	-5.0
EWSR1	8.78	4.9	0.0	3.0	3.0	-0.1
GAPDH	11.79	1.9	-3.0	0.0	0.0	-3.1
HNRNPK	11.78	1.9	-3.0	0.0	0.0	-3.1
SF1	8.67	5.0	0.1	3.1	3.1	0.0

C

		B2M	EWSR1	GAPDH	HNRNPK	SF1
		15.88	11.96	15.80	12.77	12.26
B2M	15.88	0.0	-3.9	-0.1	-3.1	-3.6
EWSR1	11.96	3.9	0.0	3.8	0.8	0.3
GAPDH	15.80	0.1	-3.8	0.0	-3.0	-3.5
HNRNPK	12.77	3.1	-0.8	3.0	0.0	-0.5
SF1	12.26	3.6	-0.3	3.5	0.5	0.0

Figure 4.27: Expression ratios between each gene in the 3 datasets for tumor tissues. Ratios between average expressions in A) dataset_2, B) dataset_3, C) TCGA.

The ratios in both tumors and normal are consistent albeit different from each other. A slight difference is expected as cancer tissues often have uncontrolled growth due to aberrant expression of a multitude of genes.

These ratios were then used to calculate a variation between the genes, for this purpose datasets 2 and 3 were used, TCGA data was abandoned since it is RNA-seq data whereas the other 2 datasets were microarray, as can be seen from figures 4.26-4.27, even though the ratios have comparable patterns, there is significant numerical difference which is expected as the technologies used to generate these data are different. And a cross-platform analyses will skew the results inaccurately.

Hence we used the datasets 2 and 3, both of which are affymetrix HG-U133-Plus-2 to calculate the variation between the ratios.

A

	HNRNPK	EWSR1	GAPDH	B2M	SF1
HNRNPK					
EWSR1	**0.002**				
GAPDH	0.181	0.144			
B2M	0.031	0.050	0.363		
SF1	**0.039**	**0.023**	0.052	0.141	

B

	HNRNPK	EWSR1	GAPDH	B2M	SF1
HNRNPK					
EWSR1	0.013				
GAPDH	0.224	0.129			
B2M	0.016	0.059	0.361		
SF1	0.005	0.002	0.160	0.041	

Figure 4.28: Variation between each gene for datasets 2 and 3. A) Variation in normal tissues, B) variation in tumor tissue. Values in black represent variation between candidate control genes, and values in blue represent the remaining.

Figure 4.28 demonstrates an interesting point, our candidate reference genes have less variation in both tumor and normal tissue. The combination of GAPDH and B2M shows most variation, which supports our hypothesis that conventional reference genes are not as reliable in quantifying mRNA levels in tumor and normal tissues.

4.3.4 Quantitative PCR of gastric cancer cell lines and gastric tumor tissue with the three candidate reference genes

In order to assess the efficacy of the three candidate RGs in vitro, we performed a qPCR with 6 gastric cancer cell lines and gastric tumor and normal tissues. The two known RGs we selected to test our genes against were B2M and GAPDH.

We generated ratios of the log2 transformed Ct values between samples as we generated ratios between log2 expression values of transcriptomic data beforehand. One thing to note is that our sample size was significantly smaller than the *in silico* datasets we've used. We had 10 tumor and adjacent normal tissue pairs and 6 gastric cancer cell lines. Due to this our results were inconclusive and further work needs to be done before reaching any conclusion. The method by which we generated the ratios were similar between our *in silico* and *ex vivo* validation data. We first separated the tumors into two groups of 5, log2 transformed the Ct values and took an average of the log2 transformed Ct value for each group. Then we calculated the ratio between the groups as we did with the affymetrix data in figure 4.27. After generating the ratios, we calculated the variation between these ratios in these two groups. This left us with a small matrix like the ones in figure 4.28 which indicates the level of variation in these datasets. Since the variation is a direct cause of the ratios of gene expression changing between datasets, we've postulated that gene pairs with lowest variation is the most stable. Thus these pairs would be the best candidates to be qPCR control genes.

For cell lines we've also implemented a similar approach, however, we treated each cell line as one group, giving us 6 groups in total. The rationale was, gastric cancer cell lines vary in gene expression and molecular subtype and since our cell lines are a mix of epithelial and mesenchymal subtypes, taking an average of expression for all cell lines would be eliminate the differences in gene expression these cell lines have. Thus, we made 6 groups, one for each cell line and calculated the average Ct values for each gene in these groups. Then the mean Ct values were all log2 transformed and ratio tables were generated for each group. Figure 4.29 depicts the variation tables for normal, tumor and cell line qPCR data.

We can see a much lower variation in the known control gene pairs B2M and GAPDH in figure 4.29. Contrary to what we observe in the computational data, the experimental results are least to say conflicting. However, these differences are likely due to the technical artefacts as well as a low sample size. A larger group of samples need to be analyzed in order to draw a conclusion. One important aspect to note is that there are multiple factors that can affect qPCR results, which are discussed in detail in the discussion chapter. As for a conclusion based on our results, the data analysis we've performed shows promise, we've successfully identified a list of candidate reference genes which have very low coefficient of variance and are likely better reference genes than the known reference genes. However, more experiments need to be done to verify this claim.

A

Tumor	GAPDH	B2M	EWSR1	SF1	HNRNPK
GAPDH					
B2M	0.00116				
EWSR1	0.00882	0.00358			
SF1	0.00014	0.00213	0.01122		
HNRNPK	0.00122	0.00476	0.01659	0.00052	

B

Normal	EWSR1	GAPDH	HNRNPK	B2M	SF1
EWSR1					
GAPDH	0.00052				
HNRNPK	0.00034	2.0126E-05			
B2M	0.00055	2.602E-07	2.4963E-05		
SF1	0.00031	0.00163952	0.00129634	0.00168	

C

Cell_line	GAPDH	B2M	EWSR1	SF1	HNRNPK
GAPDH					
B2M	0.0046				
EWSR1	0.0184	0.0075			
SF1	0.0277	0.0143	0.0029		
HNRNPK	0.0087	0.0035	0.0024	0.0064	

Figure 4.29: Variation between each gene for tumor, normal and cell line qPCR data.
A) Variation in tumor tissues, B) variation in normal tissue, C) variation in cell lines. Values in black represent variation between candidate control genes, and values in red represent the remaining.

4.3.5 A novel quality control approach for high-throughput transcriptomic data

The work involving analysis of normal and tumor data to identify stable control genes has helped us develop a novel method to identify data quality of high-throughput transcriptomic data. Sometime in the last ten years, the field of biological research was inundated with microarray data. However, these datasets, which are mostly accessible through GEO do not always face scrutiny with regard to the quality of the data. As RNA-seq becomes gradually more accessible, we need to be more vigilant over the quality of sequencing data as well.

In order to evaluate the quality of a dataset, we've developed a simple pipeline which involves using a list of stable control genes, in our case we used EWSR1, SF1, HNRNPK, GAPDH and B2M. First, we generated expression plots for all available probe-sets in datasets 1, 2 and 3 (table 6).

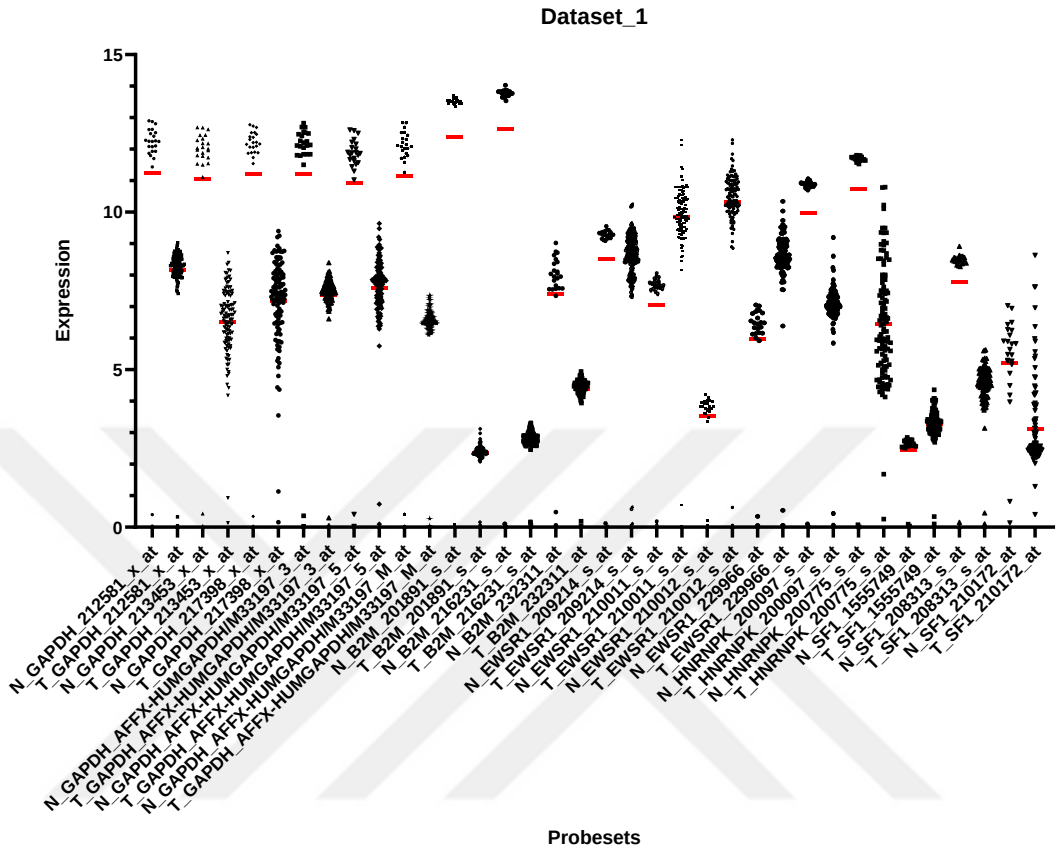


Figure 4.30: Expression plot of Dataset_1 with all probe-sets from both tumor and normal tissue. Identical probe-sets tumor and normal and tumor are arranged side-by-side for comparison. Tissue of origin, gene symbol and probe-set ID was merged to make a unique identifier (UID).

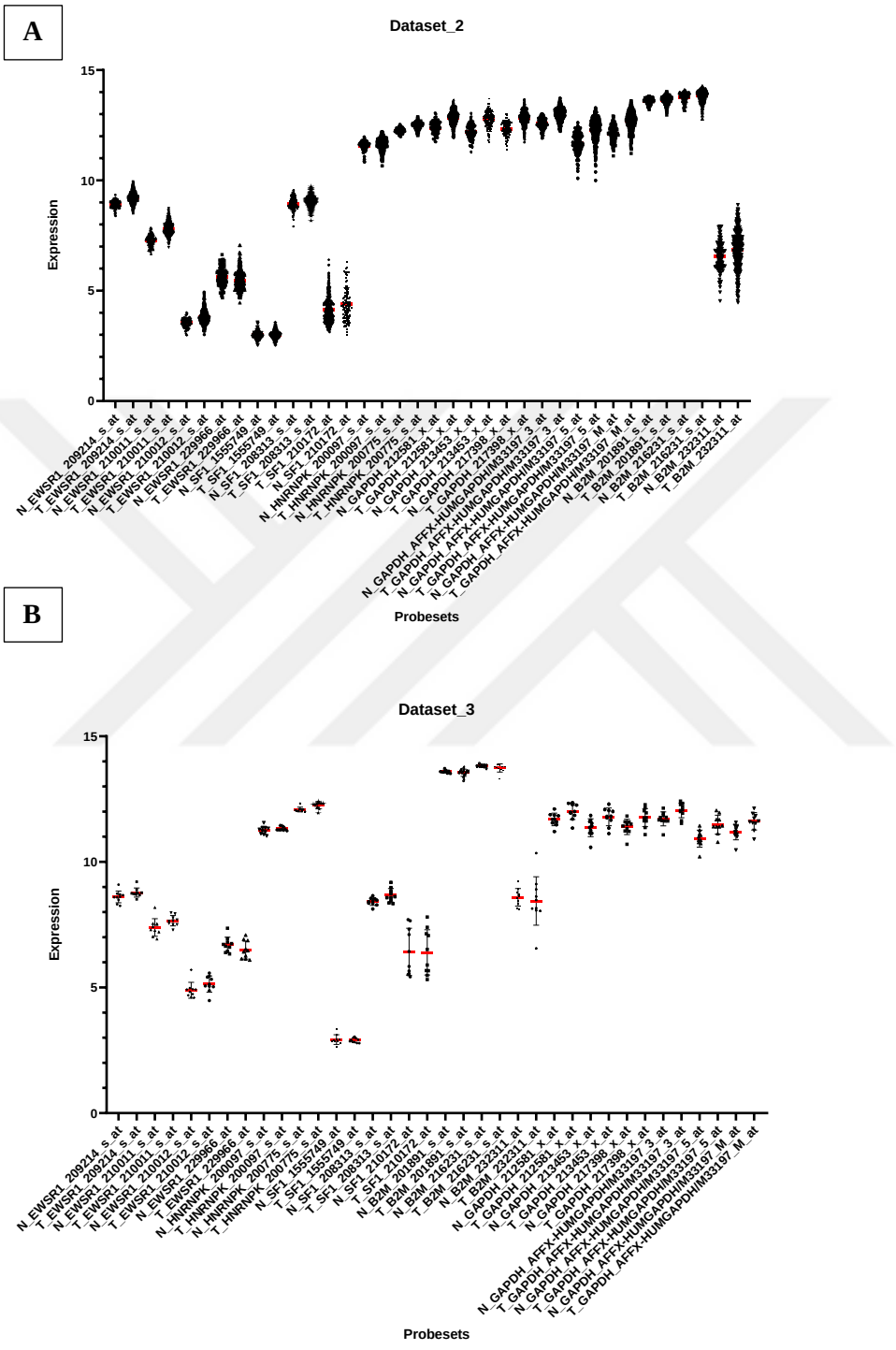


Figure 4.31: Expression plots of Datasets 2 and 3 with all probe-sets from both tumor and normal tissue. A) dataset 2, B) dataset 3. Identical probe-sets tumor and normal and tumor are arranged side-by-side for comparison. Tissue of origin, gene symbol and probe-set ID was merged to make a unique identifier (UID).

When figures 4.30 is compared to the plots in figure 4.31, one key observation is in the difference in expression range between tumor and normal in dataset 1 compared to datasets 2 and 3. Moreover, there is greater deviation around the mean in dataset 1 which implies that the dataset is most likely going to skew the results of an analysis due to what appears to be technical error. The expression patterns in datasets 2 and 3 are consistent. These plots also helps to weed out probe-sets which have inconsistent expression patterns or has very low expression.
One interesting observation is that the tumor probe-sets in figure 4.30 act completely unlike the tumor probe-sets in figure 4.31 A and B. Where the expression range of the normal probe-sets are comparable to the plots in fig.4.31, which lead us to believe that the normal group can be used to generate ratio tables as in figure 4.26. It is quite possible that there is significant batch effect between the tumor group and patient group in dataset 1 hence the variation. However, when we calculated the variation of the reference genes by considering all three datasets for the normal samples, we got very mixed results prompting us to discard dataset 1 from our analyses. For the sake of identifying datasets with poor quality, we argue that generating a scatter plot with the expression levels of all available probe-sets of a list of reference genes (tissue-specific) along with calculating the gene expression ratios between said genes will help biologists immensely. Gene expression ratio tables can be generated for normal tissue and tumor separately to assess the quality of each data compared to verified datasets which show consistent expression pattern.

Chapter 5
Discussion

5.1 5-gene prognostic signature

As seen in figure 7.1 (Appendix), our 20-gene signature appears to be a highly correlated list of genes in gastric cancer transcriptomic datasets of multiple platforms. Moreover, this gene-list clusters gastric cancer patients into two groups consistently. As mentioned thoroughly in the introduction, gastric adenocarcinomas are extremely heterogeneous, with extreme variability in mutations which make it extremely difficult to classify into subgroups. Yet our signature is consistent in stratifying patients into good and bad prognostic groups. While selecting a prognostic gene list by mUSAT, an MVA is performed for each gene while running the script which allowed us to select genes which are already independent of stage. Consequently, the 20 gene signature allows us to stratify patients within GC histological subtypes [140].
We intended to shorten our gene list for two primary reasons, one was to reduce burden of analyzing a large gene signature but more importantly, we wanted to generate a gene signature which can be assessed on a transcriptomic and a proteomic level. By selecting a group of genes which have high mRNA and protein level expression in gastric tumor tissue, we could make a two tier prognostic system. We ultimately selected a list of genes which have strong correlation with other genes, a significant logFC value between good and bad prognostic groups and has commercially available antibodies. RT-qPCR and IHC are commonly used methods in clinics and thus our signature can be easily adapted into most clinics if proven effective in *ex vivo* studies.

When discovering the 20-gene signature, an unsupervised hierarchical clustering analysis was used to cluster the GC patients in cohort 1 and 2 into good and bad prognostic groups [140]. In this, the same unsupervised method has been independently used to validate the efficacy of our shortened gene list in several datasets. Validation was performed in two steps, in first step the datasets were clustered by the 20 gene and the 5 gene signature to see if the subgroups generated by these two signatures were similar. Secondly, datasets with survival data was used to perform survival analysis to see if the subgroups generated by the 5-gene signature is capable of predicting prognostic outcome.

The five genes in the prognostic signature have been implicated in various cancer in recent literature. ACTA2 (alpha actin) is also known as smooth muscle actin, and it is a suspected EMT marker. HEYL is a member of the division-related family of transcription factors. It not only

regulates the differentiation, self-renewal and proliferation of cancer cells, but also promotes tumor angiogenesis, so it plays an important role in tumor progression.

Transgelin (TAGLN) and Caldesmon (CALD1) are cytoskeleton associated genes, as well as ACTA2 and beta-Tropomysin, (TPM2) which explains their co-expression in gastric cancer tissue.

Recently, HEYL has been found to be a prognostic factor in gastric cancer by Zhou *et al* [143]. The authors presented HEYL as a member of their prognostic gene signature which was a hub gene in their analysis. Their study have also concluded that HEYL is associated with poor prognosis.

In a recent study published by Zhao *et al*. ACTA2, CALD1 and TPM2 were found to be major hub proteins in a PPI network in colorectal cancer, with ACTA2 showing the highest degree of connectivity [145]. TAGLN along with four other genes have successfully performed risk stratification in CRC with high risk group facing worse prognosis [146]. CALD1 has also been shown to be a potential biomarker in bladder cancer, with high expression leading to poor survival. Overall, the 5 genes in our prognostic signature appear to be highly associated with multiple cancer types suggesting that our signature may have a biological role in gastric cancer and their overexpression may result in aberrant activation of metastatic pathways in gastric cancers.

In this thesis, the biological role of these 5 genes were not studied, however, in order to move forward with this genelist, a functional enrichment analysis with the good and bad prognosis groups need to be completed. This will allow us to biologically define the two groups.

By generating differentially expressed genes in gastric cancer datasets, we could generate Protein-protein interaction networks to see if our genes come up as hub proteins, which would help elucidate their biological role in GC progression.

Our *in silico* methods show a remarkable capacity of our genelist to subtype gastric tumor patients into good and bad groups in multiple cohorts across Illumina and affymetrix datasets. The patient overlap between good groups with the 5 gene signature and the 20 gene signature is significantly high in the Affymetrix datasets (tables 4-3 & 4-4). This indicate that our subgrouping is successful. In figure 4.12, the survival plots show significant log-rank p-values for 2 of the 3 cohorts indicating our signature has prognostic potential. The largest validation dataset however, was not successfully clustered by either the 20 gene or the 5 gene signature (figure 7.2, appendix). Since hierarchical clustering failed, we could not perform a survival

analysis with this dataset. One explanation is that our unsupervised method failed to recognize a pattern. A supervised method may have had better outcome in this particular dataset.
A similar study where two GC subgroups were identified by hierarchical clustering, later performed SVM and BCCP (Bayesian Compound Covariate Predictor) with their two defined subgroups to stratify patients in the validation datasets to either of those groups [122].
We can implement a supervised learning algorithm of our own, using SVM to first classify the good and bad prognostic groups based on the 5 gene signature. This could be an additional method to validate our signature gene list.

The discovery datasets cohorts 1 and 2 are both demographically East Asian. Whereas, two of the validation cohorts with survival data, cohorts 7&8 are from the MD Anderson Medical center. The authors of these two datasets disclosed that the patients are from White and Hispanic backgrounds [122]. As explained in detail in the introduction for this thesis, gastric cancer occurrence, mortality and molecular and histological class is greatly dependent on the demography of the patients, thus we can definitively say our signature will be effective across several demography.
The 5 genes in our signature are heavily correlated with each other, which can be considered a limitation of our study. The strong correlation between the genes makes it difficult to assess the effects of these individual genes on the patient groups; moreover, the strong correlation limits us in trying to understand a direct biological role these genes have on GC tumorigenesis. Moreover, the correlations are stronger in between genes in the Affymetrix datasets whereas the correlations are weaker in the Illumina datasets (figure 7.1). This indicate that this strong correlation is likely an artefact of the platform type. In figure 7.2, the lack of strong correlations may be why the gene signatures failed to cluster the patients into ant discernable groups. Moreover, we've run mUSAT on the validation datasets (Appendix). The Cox p values were significantly more variable in these datasets compared to the discovery datasets [140]. Overall, there is a likelihood and that stronger correlation between the genes in the Affymetrix datasets is the reason why our signatures are more successful. However, our signature is still very efficient at stratifying tumor groups irrespective of platform type, which suggests it works in most scenarios. Further work need to be done to validate the efficacy of the signature. The TCGA gastric cohort can be used to see if our signature stratify the patients into prognostic groups. Or other RNA-seq gastric cancer data from GEO could be analyzed to assess the cross-platform efficacy or our gene-list.

5.2 Discovery of novel reference genes in gastric tissue

There has been many conflicting literature regarding the use of GAPDH or other well know internal control genes in cancer research in recent years. Which was the main reason we wanted to identify a gastric tissue specific reference gene list. Due to technical limitations, we could fully validate our *in silico* findings by *ex vivo* means. However, our results suggest that further experiment is required to draw a conclusion.
One aspect we plan to look into is the expression level of our candidate genes. In our analysis, we planned to identify stables genes with lowest CV (Appendix) but we did not take into consideration an expression threshold. However, an expression threshold should be written into the code to prevent genes with low expression levels from being ranked. Internal control genes need to be stable however, the main reason GAPDH and ACTB are so widely used is because they have high expression levels, which is a requirement to be a qPCR reference gene.

A key issue as to why our experimental results and our computational results are not matching is the primer design. In later experiments we plan to take into consideration the specific probesets in the microarray data which have stable expression and design primers specific to said probesets.

Another issue with our primers was that we tried to design primers which target all transcriptional variants of each gene as our primary data was from bulk RNA-seq and we did not look into which specific transcripts were identified in our analysis. In order to get reliable results we plan to first identify which specific stable transcript of genes are expressed in gastric tissue, whether these transcripts are the same in tumor, normal and cell line is something we also need to look into.

Our algorithm successfully generated a list of genes with very low variation, compared to the more well-known reference genes. Our data supports our hypothesis that the candidate genes are more promising than known reference genes as internal controls in qPCR experiments. Gene stability analysis such as GENORM and Bestkeeper has been used for a long time to estimate the stability of control genes. However, in recent years there has been a trend in using high-throughput transcriptomic data to estimate gene stability in tissue-specific and pan-cancer analyses.

Our candidate reference genes have shown promise *in silico* to be better reference genes than the well-known genes. In the qPCR results, HNRNPK and SF1 showed promise a good reference gene pair. However, the expression levels of these genes are comparatively lower than B2M-

GAPDH. In qPCR, higher Ct values indicate lower levels of amplicon and our genes consistently had higher Ct values compared to GAPDH-B2M. In conclusion, in order to validate the efficacy of our code to identify stable variants, we need to add an expression threshold and we need to select specific transcripts and design transcript specific primers in the future.

In conclusion, our code successfully identified the most stable variants in TCGA RNA-seq data (Appendix). The coefficient of variance is much lower for our genes than the better known reference genes (Appendix). If we implement the changes discussed in this thesis, we will likely be successful at finding tissue specific reference genes for qPCR.

Chapter 6
Future Perspectives

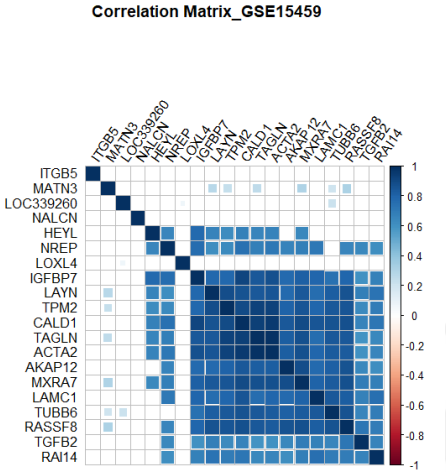
Whilst the 5-gene signature shows promise in stratifying patients into good and bad prognostic groups (figures 4.9-4.12) in this thesis, I did not look into whether the 5-gene signature stratify GC histological subtypes, intestinal and diffuse, into good and bad prognosis. In order to show that the signature is independent of histological subtype, survival analysis needs to be performed for both intestinal and diffuse type tumors separately.

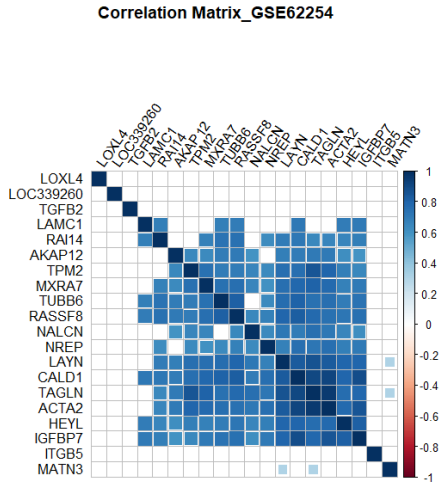
In order to validate our prognostic gene signature, RT-qPCR has to be performed with gastric tumor RNA for our 5 genes. Looking at our qPCR data from the reference genes identification part in this thesis, I would ensure our primers are specific to the probesets we've identified by *in silico* analysis. Different probesets don't always target the same transcripts and it is imperative we design probeset specific primers for reliable results.

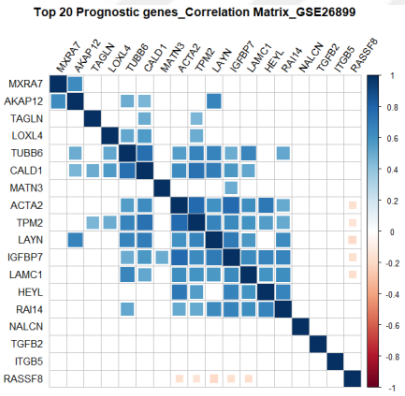
Moreover, in a clinical setting, our prognostic gene list we would have to identify prognosis on a patient to patient basis. For which reason it is imperative we design a prognostic scoring system which incorporates the expression levels of each gene and generates a prognostic risk score. Furthermore, immunohistochemistry validations of the 5 proteins in gastric tumor need to be performed.

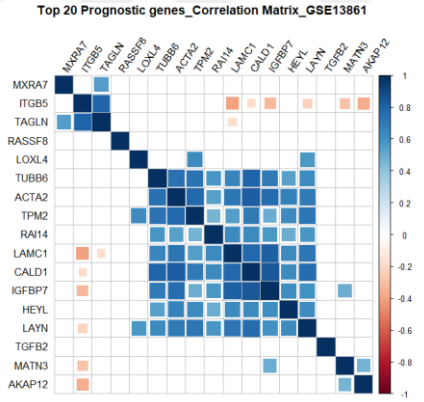
The reference gene discovery project has many questions unanswered which is beyond the scope of this thesis. Our next step would be to design transcript specific primers for the 3 candidate reference genes, EWSR1, HNRNPK, and SF1. And perform qPCR with them to assess stability compared to known reference genes B2M and GAPDH. One more area we need to look into is to increase our sample size in order to make statistically significant inferences about the efficacy of our control genes. We can perform a gene stability analysis with GeNorm and BestKeeper to see if our genes are more stably expressed than GAPDH and B2M in gastric tumor and normal tissue. Lastly, one criticism of the genes was, their expression is significantly lower than B2M and GAPDH in TCGA data (figures 4.17-4.18). One potential solution is to add an expression threshold into the code, which will then select genes based on coefficient of variance as well as average expression levels before ranking them in a descending order.

Chapter 7
Appendix









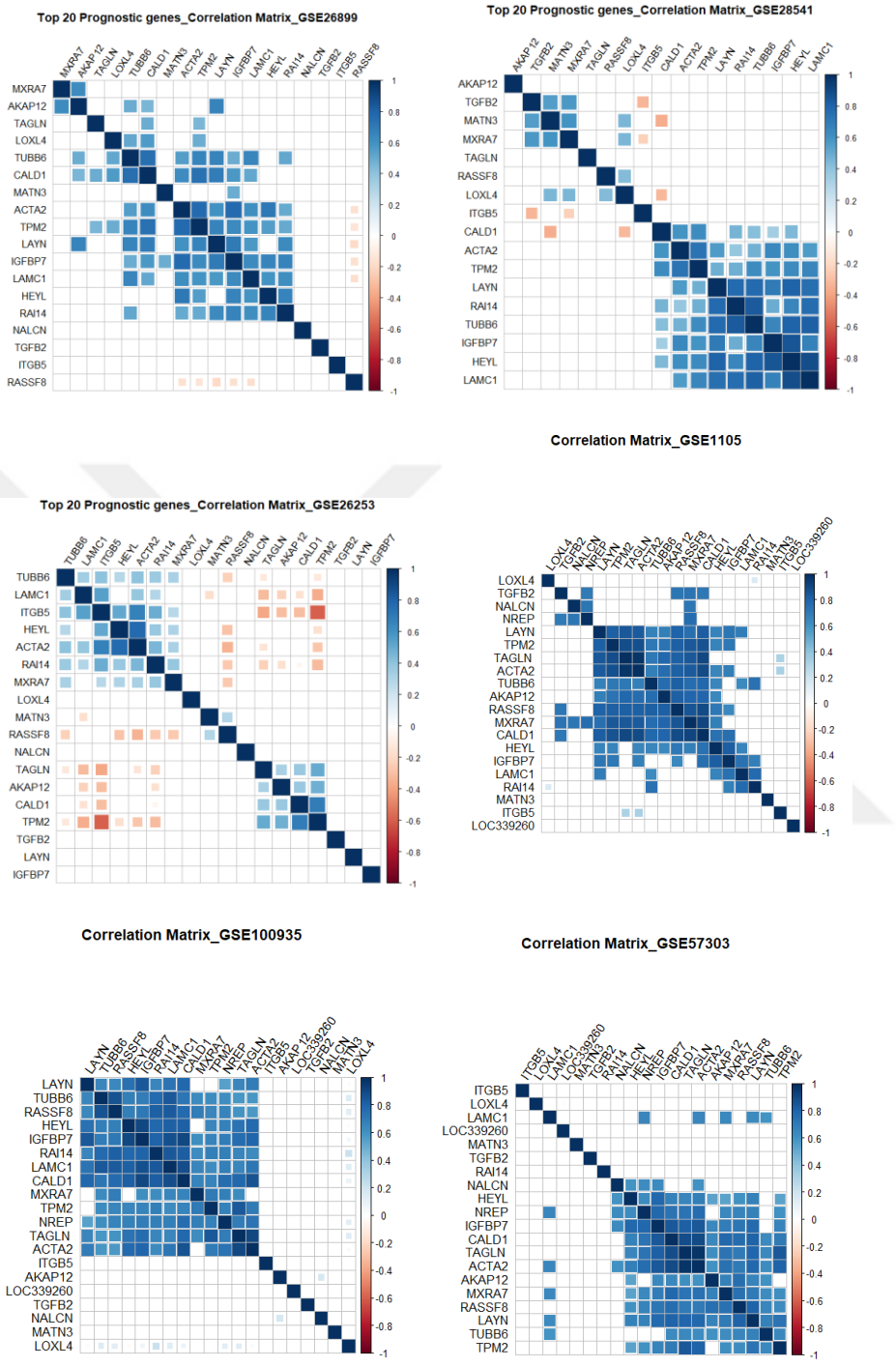


Figure 7.1: Correlation matrices for the 20 genes signature in all discovery and validation datasets.

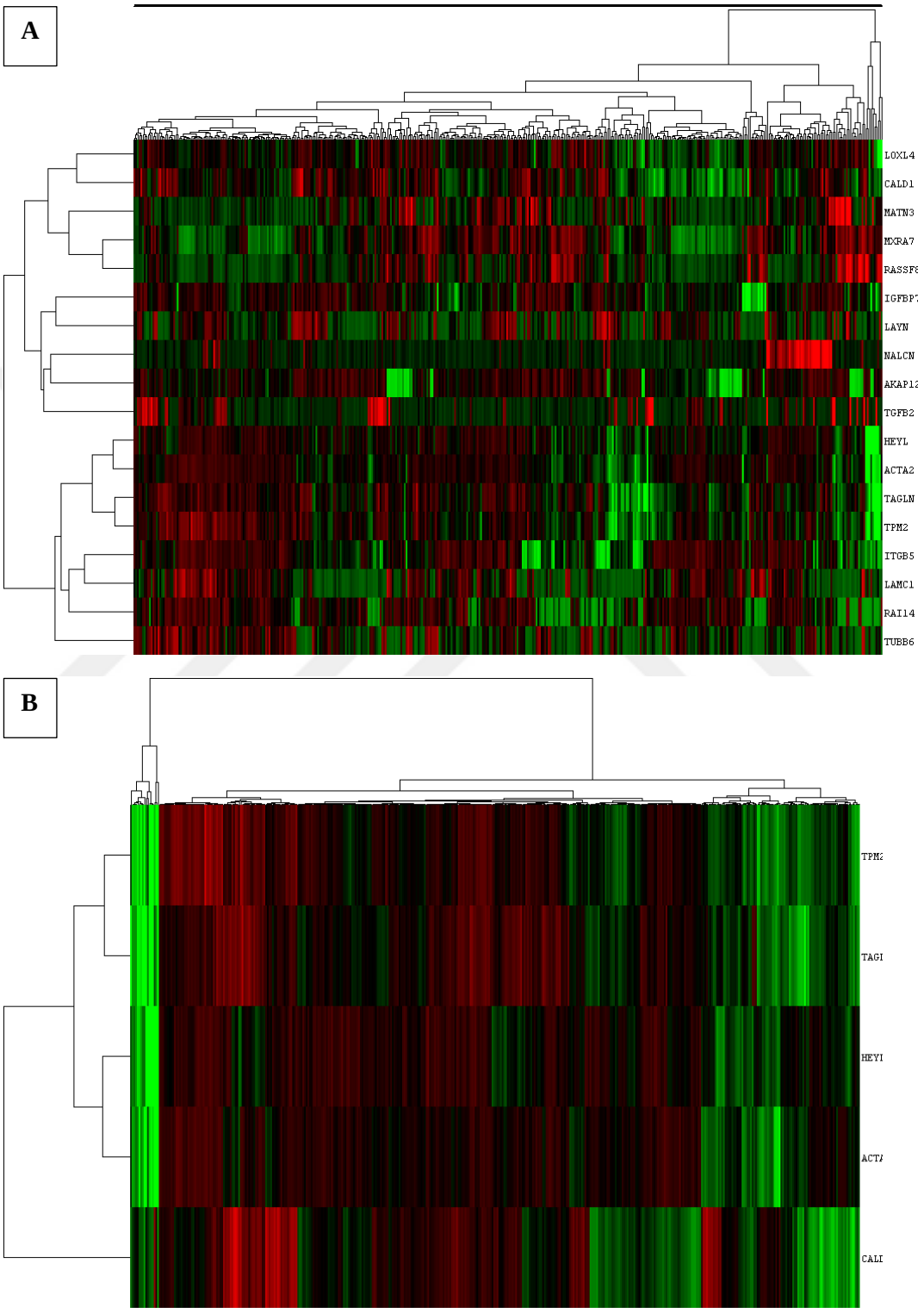


Figure 7.2: Results for hierarchical clustering analysis with the 20 gene (A) and 5 gene (B) prognosis signature in the dataset GSE26253. A) 2 genes were not available in this platform thus the analysis was done with 18 genes.

Gene	Probeset	HR	Cox p	Threshold	Maxstat p	Median Surv High Expr	Median Surv Low Expr	Log-Rank p	Censored/High Expr	Censored/Low Expr
MATN3	ILMN_1663171	1.368483	0.001905	7.175001	6.97E-05	98.94509	133.5877	2.51E-06	151/284	114/148
NALCN	ILMN_1736317	1.363568	0.002054	7.691844	0.005061	72.74924	103.7267	5.84E-05	22/56	243/376
IGFBP7	ILMN_2062468	1.451641	0.004997	11.75705	0.002381	95.68948	122.2205	7.28E-05	97/194	168/238
HEYL	ILMN_1654324	1.445391	0.00954	12.39659	0.0001	97.95089	130.1034	2.86E-06	133/258	132/174
TUBB6	ILMN_1699489	1.14965	0.02285	10.37692	0.036177	90.79421	116.0448	0.00166	38/86	227/346
MXRA7	ILMN_1692077	1.10145	0.036993	7.399358	0.068498	105.2127	125.4289	0.004883	178/313	87/119
ITGB5	ILMN_2311166	1.103765	0.187505	10.54466	0.311784	102.5026	116.6297	0.030186	123/220	142/212
TPM2	ILMN_1789196	1.57531	0.244412	13.90468	0.518947	103.2002	115.8583	0.074021	98/177	167/255
TAGLN	ILMN_1706783	0.9336	0.25344	9.801012	0.140199	120.0646	104.2816	0.011	116/169	149/263
RAI14	ILMN_1682139	1.063971	0.310921	7.059097	0.659559	108.6636	129.5407	0.10139	232/388	33/44
AKAP12	ILMN_1686846	0.936588	0.491652	9.059013	0.549373	114.7108	104.0744	0.07242	120/184	145/248
ACTA2	ILMN_1671703	1.088261	0.497422	13.93515	0.311265	97.6173	113.4285	0.030598	58/112	207/320
CALD1	ILMN_1717990	1.142233	0.64685	12.80159	0.867744	108.3422	122.4653	0.187479	226/377	39/55
LAYN	ILMN_1716397	0.972324	0.692652	9.764802	0.326799	130.1962	108.4038	0.041112	34/45	231/387
LOXL4	ILMN_2179083	1.060873	0.697619	11.54837	0.239001	106.8136	130.7369	0.023435	217/369	48/63
RASSF8	ILMN_1736741	0.948041	0.701364	7.303875	0.339961	111.429	94.94773	0.028386	212/332	53/100
LAMC1	ILMN_1810852	0.972173	0.744599	6.988722	0.521683	112.0631	88.88017	0.053723	236/375	29/57
TGFB2	ILMN_1812526	1.005029	0.951487	7.648537	0.813834	99.64075	110.2165	0.143764	61/109	204/323

Table 7-1: mUSAT results for the dataset GSE26253

Genes	Probeset	HR	Cox p	Threshold	Maxstat p	Median Surv High Expr	Median Surv Low Expr	Log-Rank p	Censored/High Expr	Censored/Low Expr
MXRA7	ILMN_1743836	2.117521	2.67E-05	8.015093	6E-05	39.5	83.89024	9.42E-08	43945	31/41
TPM2	ILMN_1789196	1.948226	7.86E-05	11.91033	0.000597	48.96774	84.91176	1.22E-05	44074	27/34
ITGB5	ILMN_1668374	3.615471	8.14E-05	10.55052	0.001166	53.25	90.3	8.9E-05	14/40	21/25
CALD1	ILMN_2324002	1.643904	8.98E-05	8.616613	0.000676	41.04762	80.52273	3.83E-06	43942	31/44
AKAP12	ILMN_2308950	1.940236	0.000804	8.341034	0.006459	45	80.2381	9.26E-05	44005	29/42
ACTA2	ILMN_1671703	2.137553	0.001081	13.61554	0.002753	42.63636	80.62791	1.91E-05	43973	30/43
TAGLN	ILMN_1778668	1.979653	0.001085	12.96503	0.00147	48.6129	85.23529	3.45E-05	11567	26/34
LAYN	ILMN_1716397	1.859579	0.002517	6.592137	0.023678	45.33333	78.07955	0.000489	44003	29/44
TUBB6	ILMN_1699489	2.169972	0.00385	8.68109	0.017777	43.94444	76.52128	0.000294	43939	31/47
LAMC1	ILMN_1810852	1.944805	0.007842	11.21011	0.002478	44.66667	78.39773	3.79E-05	43942	31/44
HEYL	ILMN_1654324	1.598928	0.011266	9.075549	0.026717	50.38462	79.35897	0.001129	44069	27/39
RAI14	ILMN_1682139	1.737276	0.013429	9.354107	0.010101	51.74194	82.38235	0.000457	44135	25/34
IGFBP7	ILMN_2062468	1.684873	0.01575	11.57416	0.023586	55.18919	84.39286	0.00187	14/37	21/28
LOXL4	ILMN_2179083	1.84644	0.036615	5.999356	0.029067	61.13462	94.30769	0.004953	23/52	44178
MATN3	ILMN_1663171	2.345735	0.071048	5.295777	0.170353	62.29245	90.5	0.029233	25/53	44116
RASSF8	ILMN_1736741	2.852761	0.353118	5.767769	0.460163	38	70.75862	0.020158	43868	33/58
TGFB2	ILMN_1812526	1.383239	0.611853	5.355395	0.039913	54.69697	81.25	0.002775	12389	23/32

Table 7-2: mUSAT results table for GSE13861

Genes	Probeset	HR	Cox p	Threshold	Maxstat p	Median Surv High Expr	Median Surv Low Expr	Log-Rank p	Censored/High Expr	Censored/Low Expr	
MATN3	ILMN_1663171	1.398346	0.263259	6.215658	0.3449	43.11821	55.87188	0.039032	3/42	37/51	
NALCN	ILMN_1736317	4.058382	0.246541	5.689597	0.448158	44.60703	61.815	0.077367	50/82		44115
IGFBP7	ILMN_2062468	1.412112	0.056156	11.9597	0.061613	45.38486	67.05295	0.007032	40/71	20/22	
HEYL	ILMN_1654324	1.305251	0.10066	8.257496	0.230772	47.7987	73.115	0.032655	50/83		44114
TUBB6	ILMN_1702636	1.22362	0.280517	6.884311	0.468042	45.26892	57.07922	0.06738	30/53	30/40	
MXRA7	ILMN_1796126	1.997375	0.416252	6.352562	0.545638	40.29947	53.36697	0.07383		44125	50/72
ITGB5	ILMN_2311166	1.245246	0.529127	10.54109	0.890021	47.75588	55.76628	0.239913	40/65	20/28	
TPM2	ILMN_1789196	1.440108	0.021082	12.69228	0.033455	38.59151	58.06	0.001762	16/36	44/57	
TAGLN	ILMN_1778668	1.637002	0.00824	13.48285	0.055242	40.36535	58.42562	0.003439	21/42	39/51	
RAI14	ILMN_1682139	1.039551	0.894627	10.33266	0.755782	45.60834	54.62833	0.147557	26/45	34/48	
AKAP12	ILMN_1684836	1.150562	0.498907	6.708396	0.439564	45.83463	58.18533	0.067361	36/61	24/32	
ACTA2	ILMN_1671703	1.662338	0.017147	14.13801	0.042185	37.8271	57.21144	0.001914	15/33	45/60	
CALD1	ILMN_1730487	1.368507	0.048465	12.66656	0.111341	35.31026	54.50132	0.005539		44063	52/73
LAYN	ILMN_1716397	1.206434	0.321039	7.579312	0.306729	38.58909	53.26011	0.024943		44158	49/70
LOXL4	ILMN_2179083	1.180089	0.377776	6.500645	0.071987	46.53554	69.58077	0.013855	47/79	13/14	
RASSF8	ILMN_1736741	0.556958	0.585345	5.974715	0.576028	54.49914	43.60701	0.084961	39/55	21/38	
LAMC1	ILMN_1810852	1.165241	0.481717	11.10472	0.224875	47.34015	67.71958	0.035347	48/80		44178
TGFB2	ILMN_1812526	1.863458	0.462346	5.826175	0.600124	47.53864	59.9503	0.115151	45/74	15/19	

Table 7-3: mUSAT results table for GSE26899.

Gene	Probeset	HR	Cox p	Threshold	Maxstat p	Median Surv High Expr	Median Surv Low Expr	Log-Rank p	Censored/High Expr	Censored/Low Expr	
MATN3	ILMN_1663171	3.91055	0.001021	6.3603	0.152246	51.61728	83.93622	0.008912	44071	46/81	
NALCN	ILMN_1736317	1.248682	0.77525	6.053144	0.485638	73.16611	99.44119	0.073363	38/86	16/23	
IGFBP7	ILMN_2062468	1.396195	0.002235	12.24024	0.001271	42.16402	88.24359	1.14E-05	12236	47/76	
HEYL	ILMN_1654324	1.466026	0.000566	9.0229	0.002099	50.42632	92.54434	3.64E-05	16041	42/66	
TUBB6	ILMN_1702636	1.265977	0.062733	8.019349	0.126296	49.52381	82.06082	0.005574	43972	49/88	
MXRA7	ILMN_1796126	2.106782	0.105058	6.31229	0.05998	57.84117	88.67271	0.003404	13/42	41/67	
ITGB5	ILMN_2311166	2.536364	0.000168	10.62191	0.000878	39.41667	83.78115	7.21E-06	43981	49/79	
TPM2	ILMN_1789196	1.381367	0.004065	13.79828	0.006301	44.90714	86.68184	6.46E-05	44010	48/81	
TAGLN	ILMN_1778668	1.454705	0.007763	13.98464	0.024321	54.41617	88.94255	0.000834	14580	42/70	
RAI14	ILMN_1682139	1.929027	0.000199	10.06512	0.000744	38.35897	86.94216	2.65E-06	43947	50/83	
AKAP12	ILMN_1684836	1.552335	0.00184	8.163666	0.028882	40.47368	81.68972	0.000228	43940	50/90	
ACTA2	ILMN_1671703	1.511282	0.007358	14.83655	0.037711	46.4	83.57054	0.001118	43943	50/87	
CALD1	ILMN_1730487	1.324758	0.001949	11.6774	0.001392	50.22633	93.87638	2.18E-05	16407	42/65	
LAYN	ILMN_1716397	1.68673	0.000194	7.287088	0.000235	44.42643	92.35388	1.83E-06	13728	46/72	
LOXL4	ILMN_2179083	1.855907	0.001468	7.258483	0.013233	59.45175	95.6002	0.000742	1/60	33/49	
RASSF8	ILMN_1736741	0.387155	0.288471	5.957703	0.500756	69.02333	39.54	0.044692	52/99		43871
LAMC1	ILMN_1810852	1.693451	0.000279	11.43697	0.00132	50.69362	94.25228	2.32E-05	13/46	41/63	
TGFB2	ILMN_1812526	5.481082	0.001631	6.264753	0.119436	57.05208	88.07001	0.006428	11963	44/77	

Table 7-4: mUSAT results table for GSE26901.

Gene	Probeset	HR	Cox p	Threshold	Maxstat p	Median Surv High Expr	Median Surv Low Expr	Log-Rank p	Censored/High Expr	Censored/Low Expr
MATN3	ILMN_1663171	0.632088	0.647676	6.35671	0.223512	29.43398	12.007	0.005031	44041	43841
IGFBP7	ILMN_1660390	2.274296	0.00305	13.06791	0.003906	14.73	41.39857	0.000178	43886	43997
HEYL	ILMN_1654324	1.720763	0.007615	7.807067	0.014364	22.06662	66.468	0.007203	12510	43927
TUBB6	ILMN_1697132	1.577908	0.035628	7.990816	0.124733	24.29446	58.7025	0.041449	12905	43895
MXRA7	ILMN_1692077	1.640572	0.585598	6.368855	0.428186	17.95667	32.18819	0.044195	0/12	44071
ITGB5	ILMN_1761674	18.44676	0.178857	6.27039	0.48419	12.962	26.70107	0.027366	43840	44042
TPM2	ILMN_1789196	1.251829	0.307834	11.68579	0.437739	19.55756	33.76688	0.064279	43883	44000
TAGLN	ILMN_1706783	0.810449	0.915347	5.895338	0.802867	26.43652	14.18	0.11673	12997	0/5
RAI14	ILMN_1682139	1.870636	0.006177	10.29575	0.062226	12.34143	32.28622	0.001021	0/14	44069
AKAP12	ILMN_1770149	0.279032	0.511925	6.166593	0.415954	27.56626	12.80556	0.021924	44042	43840
ACTA2	ILMN_1671703	1.489709	0.169741	12.91029	0.356928	22.11731	40.37722	0.065545	43918	43963
CALD1	ILMN_1730487	1.718681	0.042035	11.17153	0.302865	16.80125	31.00284	0.031487	0/16	44067
LAYN	ILMN_1716397	2.769816	0.006413	6.775158	0.017782	19.81821	50.13629	0.004587	43920	43961
LOXL4	ILMN_1754174	1.402506	0.692939	6.413723	0.84139	31.49733	19.87909	0.184238	44041	43841
RASSF8	ILMN_1736741	0.322569	0.334657	6.309247	0.209206	28.68317	10.35125	0.002288	11902	0/8
LAMC1	ILMN_1810852	1.92634	0.022492	11.33052	0.122844	21.4173	46.01889	0.022365	43920	43961
TGFB2	ILMN_1812526	5.999661	0.081142	6.052895	0.082576	22.50998	51.60333	0.027677	12540	43896

Table 7-5: mUSAT results table for GSE28541.

GSE13861			GSE26899		GSE26253		GSE26901		GSE28541	
Genes	HR	Cox p	HR	Cox p	HR	Cox p	HR	Cox p	HR	Cox p
CALD1	1.64	8.98E-05	1.37	0.04	1.14	0.6	1.32	0.001	1.72	0.04
HEYL	1.59	0.011	1.31	0.1	1.45	0.009	1.47	0.0005	1.72	0.007
ACTA2	2.14	0.001	1.66	0.01	1.09	0.4	1.51	0.007	1.49	0.16
TPM2	1.95	7.86E-05	1.44	0.02	1.58	0.2	1.38	0.004	1.25	0.3
TAGLN	1.98	0.001	1.64	0.008	0.93	0.2	1.45	0.007	0.81	0.91
AKAP12	1.94	0.0008	1.15	0.4	0.94	0.4	1.55	0.001	0.28	0.51
IGFBP7	1.68	0.01	1.41	0.05	1.45	0.004	1.40	0.002	2.27	0.003
ITGB5	3.62	8.14E-05	1.25	0.5	1.10	0.1	2.54	0.0001	18.45	0.1
LAMC1	1.94	0.007	1.17	0.4	0.97	0.74	1.69	0.0002	1.93	0.02
LAYN	1.86	0.002	1.21	0.3	0.97	0.69	1.69	0.0001	2.77	0.006
LOXL4	1.85	0.03	1.18	0.3	1.06	0.69	1.86	0.001	1.40	0.6
MATN3	2.35	0.071	1.40	0.2	1.37	0.001	3.91	0.001	0.63	0.6
MXRA7	2.12	0.00002	2.00	0.4	1.10	0.03	2.11	0.1	1.64	0.5
NALCN			4.06	0.2	1.36	0.002	1.25	0.7		
RAI14	1.74	0.013	1.04	0.8	1.06	0.31	1.93	0.0001	1.87	0.006
RASSF8	2.85	0.35	0.56	0.5	0.95	0.7	0.39	0.28	0.32	0.33
TGFB2	1.38	0.6	1.86	0.4	1.01	0.95	5.48	0.001	6.00	0.08
TUBB6	2.17	0.003	1.22	0.2	1.15	0.02	1.27	0.06	1.58	0.03

Table 7-6: mUSAT results summary of HR and Cox p value for the five validation datasets.

Genes	Functions
SF1	Splicing factor 1
HNRNPK	Binds to pre-mRNAs, likely plays a role in metabolising hnRNAs
TARDBP	RNA binding protein that has various processes steps in RNA biogenesis and processing
EWSR1	Various cellular processes, including cell signalling, RNA processing, and transport
XRCC5	Single-stranded DNA dependent, ATP dependent helicase
ZNF207	Binds to pre-mRNAs, likely plays a role in metabolising hnRNAs
HNRNPU	Many cellular functions, binds to DNA and RNA.
HNRNPC	Binds pre-mRNA and nucleates the assembly of 40S hnRNP particles
HNRNPL	Splicing factor, involved with intronic/exonic inclusion
HNRNPM	Pre-mRNA binding protein, possibly involved in splicing
SF3B2	Splicing factor
DDX5	Involved in alternative regulation of pre-mRNA splicing, has RNA helicase activity
SERBP1	May play a role in mRNA stability, binds to mRNA
HNRNPR	Component of ribo-nucleosomes, plays a role in pre-mRNA processing
CNBP	ssDNA binding protein, with specificity to the SRE, involved in sterol-mediated repression
ARF1	GTP-binding protein involved in protein transport between different components
HNRNPA2B1	Packages nascent pre-mRNAs into hnRN particles
ACTR2	ATP-binding component of the Arp2/3 complex. It is involved in Actin polymerization
IMMT	Inner membrane mitochondrial protein
RBMX	RNA-binding protein with multiple pre- and post-translational functions

Table 7-7: Biological roles of the top candidate reference genes.

	CCLE		TCGA_DeSeq2		TCGA_RSEM	
Genes	Mean_RANK	Mean_CV	Mean_RANK	Mean_CV	Mean_RANK	Mean_CV
EWSR1	8	3.58	8	1.93	10	2.59
HNRNPC	10	5.37	37	2.34	19	2.46
HNRNPK	15	3.31	1	2.38	6	2.66
HNRNPA2B1	17	4.27	70	2.46	62	2.74
ZNF207	38	4.59	18	2.55	15	2.77
HNRNPM	40	5.52	45	2.57	24	2.84
HNRNPU	60	5.52	26	2.59	17	2.72
HNRNPL	78	4.77	27	2.6	20	2.81
RBMX	89	4.88	35	2.6	146	3.37
ARF1	139	5.71	74	2.63	60	2.95
IMMT	140	3.45	92	2.63	120	2.78
CNBP	144	4.12	95	2.68	57	2.84
SF1	153	5.17	7	2.7	3	3.07
ACTR2	154	5.44	45	2.74	75	3.14
TARDBP	155	5.64	147	2.78	9	3.06
XRCC5	176	5.59	20	2.82	14	2.93
DDX5	188	3.73	73	2.83	35	3.08
SF3B2	191	5.26	32	2.88	26	3.06
HNRNPR	195	5.26	60	2.93	57	3.29
SERBP1	213	5.43	38	2.94	38	2.64

Table 7-8 Coefficient variants as calculated by our code for the candidate reference genes

	CCLE		TCGA_DeSeq2		TCGA_RSEM	
Genes	Mean_RANK	Mean_CV	Mean_RANK	Mean_CV	Mean_RANK	Mean_CV
RBMX	89	4.88	35	2.6	146	3.37
RER1	2155	10.62	1303	3.92	483	3.86
TUBA1B	612	7.22	3000	4.72	533	3.91
GAPDH	296	6.12	1968	4.34	1133	4.38
PGAM1	1472	9.26	5134	5.65	1246	4.45
ALDOA	268	6.01	2103	4.24	1419	4.55
PGK1	675	7.41	2712	4.47	1674	4.7
B2M	1723	9.79	3454	4.91	2357	5.07
RPL29	340	6.25	3469	4.61	3414	5.62
HPRT1	2413	11.13	5921	6.19	5348	6.74
PCBP1	27	3.94	265	3.23	2	2.39
ACTB	265	5.97	212	3.12	111	3.26

Table 7-9: Coefficient of variance values of known reference genes according to our code

Random repeated subsampling code

```
library(tidyr)
library(dplyr)
library(tidyverse)
library(stats)
library(plyr)

datax <- read.csv("trial.csv",header = TRUE, row.names=1)

datax<-data.matrix(datax)

repeats <- 1000 #number of times coefficient of variance will be calculated for each gene

output1 <- matrix(ncol = repeats, nrow = nrow(datax))

output2 <- matrix(ncol = repeats, nrow = nrow(datax))

rownum<- nrow(datax)

#the loop divides datax into two parts, then shuffles the columns in datax and assigns them to two datasets randomly

for (j in 1:repeats) {

 if (j%%(repeats/10)==0) print(paste(j*100/repeats,"% completed"))

testss= round(ncol(datax)/2)
trainindex = sample(1:ncol(datax), testss, replace=FALSE)
testindex = c(1:ncol(datax))[-trainindex]

data1 <- datax[,trainindex]
data2 <- datax[,testindex]

 cv1_vec<-c()

 cv2_vec<-c()

 for(i in 1:rownum) #the loop calculates the coefficient of variance for each iteration of data1 and data2

 {   cv1=(sd(data1[i,], na.rm=TRUE)/

          mean(data1[i,], na.rm=TRUE))

 cv2=(sd(data2[i,], na.rm=TRUE)/

       mean(data2[i,], na.rm=TRUE))

 cv1_vec<-c(cv1_vec,c(cv1))

 cv2_vec<-c(cv2_vec,c(cv2))

 }

 output1[,j]<-cv1_vec

 output2[,j]<-cv2_vec

 ## Setting row names of output matrices once out of the for loop for 1000 iterations is enough. So I put them out of the loop. This also works fine, but it will slow down the running process of the script for bigger datasets.

 ##rownames(output2)<-rownames(datax)
```

```
  ##rownames(output1)<-rownames(datax)

  }

rownames(output2)<-rownames(datax)

rownames(output1)<-rownames(datax)

write.csv(output1, file = "trial_RESULTS1.csv", row.names = TRUE)
write.csv(output2, file = "trial_RESULTS2.csv", row.names = TRUE)

ranked1 <- matrix(nrow=nrow(output1),ncol = repeats)
ranked2 <- matrix(nrow=nrow(output2),ncol = repeats)

v<-ncol(output1)


for(i in 1:v){ #creates two matrices with the ranks of CV for each gene
  ranked_output1<-rank(output1[,i])
  ranked_output2<-rank(output2[,i])
  ranked1[,i]<-ranked_output1
  ranked2[,i]<-ranked_output2

}

## For the ranked outputs, alternatively, you can run the following code instead of a for loop, for a better performance in terms of processing time. Your code above also works fine. Do not forget to remove comment signs "#" to run the codes below.

## ranked11 = apply(output1,2,rank)
## ranked22 = apply(output2,2,rank)

rownames(ranked1)<-rownames(datax)
rownames(ranked2)<-rownames(datax)
write.csv(ranked1, file = "trial_RANKS1.csv", row.names = TRUE)
write.csv(ranked2, file = "trial_RANKS2.csv", row.names = TRUE)
```

Chapter 8
Bibliography

1. H. Sung, J. Ferlay, R. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal and F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", CA: A Cancer Journal for Clinicians, vol. 71, no. 3, pp. 209-249, 2021 [Online]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/epdf/10.3322/caac.21660

2. M. Balakrishnan, R. George, A. Sharma and D. Graham, "Changing Trends in Stomach Cancer Throughout the World", Current Gastroenterology Reports, vol. 19, no. 8, 2017.

3. M. Feldman, L. Friedman and M. Sleisenger, Sleisenger & Fordtran's gastrointestinal and liver disease. (Vol. 1 of 2 Volume Set). Philadelphia: Saunders, 2002.

4. J. Machlowska, J. Baj, M. Sitarz, R. Maciejewski and R. Sitarz, "Gastric Cancer: Epidemiology, Risk Factors, Classification, Genomic Characteristics and Treatment Strategies", International Journal of Molecular Sciences, vol. 21, no. 11, p. 4012, 2020.

5. Parkin DM, Pisani P, Ferlay J. Estimates of the worldwide incidence of eighteen major cancers in 1985. Int J Cancer 1993;54: 594-606

6. Blot WJ, Devesa SS, Kneller RW, Fraumeni JF. Rising incidence of adenocarcinoma of the esophagus and gastric cardia. JAMA 1991; 265: 1287-1289

7. Brown LM, Devesa SS. Epidemiologic trends in esophageal and gastric cancer in the United States. Surg Oncol Clin N Am 2002; 11: 235-256

8. Lauren P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification. Acta Pathol Microbiol Scand 1965; 64:31-49

9. Munoz N. Gastric Carcinogenesis. In: Reed PI, Hill MJ, eds. Gastric carcinogenesis: proceedings of the 6th Annual Symposium of the European Organization for Cooperation in Cancer Prevention Studies (ECP). Amsterdam: Elsevier Science, 1988: 51-69

10. Muñoz N, Correa P, Cuello C, Duque E. Histologic types of gastric carcinoma in high- and low-risk areas. Int J Cancer 1968; 3: 809-818

11. Lauwers GY, Carneiro F, Graham DY. Gastric carcinoma. In: Bowman FT, Carneiro F, Hruban RH, eds. Classification of Tumours of the Digestive System. Lyon:IARC;2010. In press.

12. B. Hu, N. El Hajj, S. Sittler, N. Lammert, R. Barnes and A. Meloni-Ehrig, "Gastric cancer: Classification, histology and application of molecular pathology", Journal of Gastrointestinal Oncology, vol. 3, no. 3, pp. 251-261, 2012 [Online]. Available: https://jgo.amegroups.com/article/view/427.

13. Henson DE, Dittus C, Younes M, Nguyen H, Albores-Saavedra J. Differential trends in the intestinal and diffuse types of gastric carcinoma in the United States, 1973-2000: increase in the signet ring cell type. Arch Pathol Lab Med. 2004 Jul;128(7):765-70. doi: 10.5858/2004-128-765-DTITIA. PMID: 15214826.

14. S. Kaneko and T. Yoshimura, "Time trend analysis of gastric cancer incidence in Japan by histological types, 1975–1989", British Journal of Cancer, vol. 84, no. 3, pp. 400-405, 2001 [Online]. Available:

https://www.nature.com/articles/6691602.

15. Polkowski W, van Sandick JW, Offerhaus GJ, et al.Prognostic value of Laurén classification and c-erbB-2 oncogene overexpression in adenocarcinoma of the esophagus and gastroesophageal junction. Ann Surg Oncol 1999;6:290-7.

16. I. Nagtegaal, R. Odze, D. Klimstra, V. Paradis, M. Rugge, P. Schirmacher, K. Washington, F. Carneiro and I. Cree, "The 2019 WHO classification of tumours of the digestive system", Histopathology, vol. 76, no. 2, pp. 182-188, 2019.

17. D. Soybel, "Anatomy and Physiology of the Stomach", Surgical Clinics of North America, vol. 85, no. 5, pp. 875-894, 2005.

18. L. Rosenfeld, "The last alchemist--the first biochemist: J.B. van Helmont (1577-1644).", Clinical Chemistry, vol. 31, no. 10, pp. 1755-1760, 1985.

19. J. Baron, "The Discovery of Gastric Acid", Gastroenterology, vol. 76, no. 5, pp. 1056-1064, 1979.

20. H. Hoshi, "Management of Gastric Adenocarcinoma for General Surgeons", Surgical Clinics of North America, vol. 100, no. 3, pp. 523-534, 2020.

21. S. Hirota, K. Isozaki, Y. Moriyama, K. Hashimoto, T. Nishida, S. Ishiguro, K. Kawano, M. Hanada, A. Kurata, M. Takeda, G. Tunio, Y. Matsuzawa, Y. Kanakura, Y. Shinomura and Y. Kitamura, "Gain-of-Function Mutations of c-kit in Human Gastrointestinal Stromal Tumors", Science, vol. 279, no. 5350, pp. 577-580, 1998.

22. F. Bosman, F. Carneiro, R. Hruban and N. Theise, WHO classification of tumours of the digestive system. Lyon: International agency for research on cancer, 2010.

23. "Comprehensive molecular characterization of gastric adenocarcinoma", Nature, vol. 513, no. 7517, pp. 202-209, 2014.

24. R. Cristescu, J. Lee, M. Nebozhyn, K. Kim, J. Ting, S. Wong, J. Liu, Y. Yue, J. Wang, K. Yu, X. Ye, I. Do, S. Liu, L. Gong, J. Fu, J. Jin, M. Choi, T. Sohn, J. Lee, J. Bae, S. Kim, S. Park, I. Sohn, S. Jung, P. Tan, R. Chen, J. Hardwick, W. Kang, M. Ayers, D. Hongyue, C. Reinhard, A. Loboda, S. Kim and A. Aggarwal, "Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes", Nature Medicine, vol. 21, no. 5, pp. 449-456, 2015.

25. G. Murphy, R. Pfeiffer, M. Camargo and C. Rabkin, "Meta-analysis Shows That Prevalence of Epstein–Barr Virus-Positive Gastric Cancer Differs Based on Sex and Anatomic Location", Gastroenterology, vol. 137, no. 3, pp. 824-833, 2009.

26. B. Sohn, J. Hwang, H. Jang, H. Lee, S. Oh, J. Shim, K. Lee, E. Kim, S. Yim, S. Lee, J. Cheong, W. Jeong, J. Cho, J. Kim, J. Chae, J. Lee, W. Kang, S. Kim, S. Noh, J. Ajani and J. Lee, "Clinical Significance of Four Molecular Subtypes of Gastric Cancer Identified by The Cancer Genome Atlas Project", Clinical Cancer Research, vol. 23, no. 15, pp. 4441-4449, 2017

27. Wang, G. et al. Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in China. PLoS ONE 8, e63826 (2013).

28. Shah, M.A. et al. Molecular classification of gastric cancer: a new paradigm. Clin. Cancer Res. 17, 2693–2701 (2011).

29. Tay, S.T. et al. A combined comparative genomic hybridization and expression microarray analysis of gastric cancer reveals novel molecular subtypes. Cancer Res. 63, 3309–3316 (2003).

30. Cancer Genome Atlas Research Network. et al. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513, 202–209 (2014).

31. Chen, X. et al. Variation in gene expression patterns in human gastric cancers. Mol. Biol. Cell 14, 3208–3215 (2003).

32. Tan, I.B. et al. Intrinsic subtypes of gastric cancer, based on gene expression pattern, predict survival and respond differently to chemotherapy. Gastroenterology 141, 476–485 (2011).

33. Cho, J.Y. et al. Gene expression signature-based prognostic risk score in gastric cancer. Clin. Cancer Res. 17, 1850–1857 (2011).

34. Deng, N. et al. A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets. Gut 61, 673–684 (2012).

35. An, C. et al. Prognostic significance of CpG island methylator phenotype and microsatellite instability in gastric carcinoma. Clin. Cancer Res. 11, 656–663 (2005).

36. Liu, Z. et al. Large-scale characterization of DNA methylation changes in human gastric carcinomas with and without metastasis. Clin. Cancer Res. 20, 4598–4612 (2014).

37. Zouridis, H. et al. Methylation subtypes and large-scale epigenetic alterations in gastric cancer. Sci. Transl. Med. 4, 156ra140 (2012).

38. Wang, K. et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat. Genet. 46, 573–582 (2014).

39. Kakiuchi, M. et al. Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. Nat. Genet. 46, 583–587 (2014).

40. Liu, J. et al. Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. Nat. Commun. 5, 3830 (2014).

41. Ooi, C.H. et al. Oncogenic pathway combinations predict clinical prognosis in gastric cancer. PLoS Genet. 5, e1000676 (2009).

42. Wu, Y. et al. Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. Gut 62, 1100–1111 (2013).

43. X. Chen, H. Chen, F. Castro, J. Hu and H. Brenner, "Epstein–Barr Virus Infection and Gastric Cancer", Medicine, vol. 94, no. 20, p. e792, 2015.

44. Ryan JL, Shen YJ, Morgan DR, et al. Epstein-Barr virus infection is common in inflamed gastrointestinal mucosa. Dig Dis Sci 2012; 57:1887–1898.

45. M. Venerito, A. Link, T. Rokkas and P. Malfertheiner, "Gastric cancer - clinical and epidemiological aspects", Helicobacter, vol. 21, pp. 39-44, 2016.

46. T. Ang and K. Fock, "Clinical epidemiology of gastric cancer", Singapore Medical Journal, vol. 55, no. 12, pp. 621-628, 2014.

47. Fock KM, Ang TL. Epidemiology of Helicobacter pylori infection and gastric cancer in Asia. J Gastroenterol Hepatol 2010; 25:479-86

48. P. Karimi, F. Islami, S. Anandasabapathy, N. Freedman and F. Kamangar, "Gastric Cancer: Descriptive Epidemiology, Risk Factors, Screening, and Prevention", Cancer Epidemiology Biomarkers & Prevention, vol. 23, no. 5, pp. 700-713, 2014.

49. J. Marqués-Lespier, M. González-Pons and M. Cruz-Correa, "Current Perspectives on Gastric Cancer", Gastroenterology Clinics of North America, vol. 45, no. 3, pp. 413-428, 2016.

50. Freedman N, Derakhshan M, Abnet C, Schatzkin A, Hollenbeck A, McColl K. Male predominance of upper gastrointestinal adenocarcinoma cannot be explained by differences in tobacco smoking in men versus women. Eur J Cancer 2010;46:2473–8.

51. IARC monographs on the evaluation of carcinogenic risks to humans. Ingested nitrate and nitrite, and cyanobacterial peptide toxins. IARC Monogr Eval Carcinog Risks Hum 2010;94:1–412.

52. Ladeiras-Lopes R, Pereira AK, Nogueira A, Pinheiro-Torres T, Pinto I, Santos-Pereira R, et al. Smoking and gastric cancer: systematic review and meta-analysis of cohort studies. Cancer Causes Control 2008;19:689–701.

53. Cook MB, Kamangar F, Whiteman DC, Freedman ND, Gammon MD, Bernstein L, et al. Cigarette smoking and adenocarcinomas of the esophagus and esophagogastric junction: a pooled analysis from the international BEACON consortium. J Natl Cancer Inst 2010;102:1344–53.

54. Freedman ND, Abnet CC, Leitzmann MF, Mouw T, Subar AF, Hollenbeck AR, et al. A prospective study of tobacco, alcohol, and the risk of esophageal and gastric cancer subtypes. Am J Epidemiol 2007;165:1424–33.

55. Neglected role of hookah and opium in gastric carcinogenesis: a cohort study on risk factors and attributable fractions. Int J Cancer 2014;134:181–8.

56. Wiseman M. The second World Cancer Research Fund/American Institute for Cancer Research expert report. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. Proc Nutr Soc 2008;67:253–6.

57. Tsugane S, Sasazuki S, Kobayashi M, Sasaki S. Salt and salted food intake and subsequent risk of gastric cancer among middle-aged Japanese men and women. Br J Cancer 2004;90:128–34.

58. Nagini S. Carcinoma of the stomach: a review of epidemiology, pathogenesis, molecular genetics and chemoprevention. World J Gastroint Oncol 2012;4:156–69.

59. Wogan GN, Hecht SS, Felton JS, Conney AH, Loeb LA. Environmental and chemical carcinogenesis. Semin Cancer Biol 2004;14:473–86.

60. Coggon D, Osmond C, Barker DJ. Stomach cancer and migration within England and Wales. Br J Cancer 1990;61:573–4.

61. Malaty HM, El-Kasabany A, Graham DY, Miller CC, Reddy SG, Srinivasan SR, et al. Age at acquisition of Helicobacter pylori infection: a follow-up study from infancy to adulthood. Lancet 2002;359:931–5.

62. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of human migrations in Helicobacter pylori populations. Science 2003;299:1582–5.

63. Lynch HT, Grady W, Suriano G, Huntsman D. Gastric cancer: new genetic developments. J Surg Oncol 2005;90:114–33.

64. Kaurah P, MacMillan A, Boyd N, Senz J, De Luca A, Chun N, et al. Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer. JAMA 2007;297:2360–72.

65. Oliveira C, Senz J, Kaurah P, Pinheiro H, Sanges R, Haegert A, et al. Germline CDH1 deletions in hereditary diffuse gastric cancer families. Hum Mol Genet 2009;18:1545–55.

66. Sakamoto H, Yoshimura K, Saeki N, Katai H, Shimoda T, Matsuno Y, et al. Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. Nat Genet 2008;40:730–40.

67. Shi Y, Hu Z, Wu C, Dai J, Li H, Dong J, et al. A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. Nat Genet 2011;43:1215–8

68. Abnet CC, Freedman ND, Hu N, Wang Z, Yu K, Shu XO, et al. A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. Nat Genet 2010;42:764–7.

69. Y. Choi and N. Kim, "Gastric cancer and family history", The Korean Journal of Internal Medicine, vol. 31, no. 6, pp. 1042-1053, 2016.

70. Zanghieri G, Di Gregorio C, Sacchetti C, et al. Familial occurrence of gastric cancer in the 2-year experience of a population-based registry. Cancer 1990;66:2047–2051.

71. Rokkas T, Sechopoulos P, Pistiolas D, Margantinis G, Koukoulis G. Helicobacter pylori infection and gastric histology in first-degree relatives of gastric cancer patients: a meta-analysis. Eur J Gastroenterol Hepatol 2010;22:1128–1133.

72. Oliveira C, Pinheiro H, Figueiredo J, Seruca R, Carneiro F. Familial gastric cancer: genetic susceptibility, pathology, and implications for management. Lancet Oncol 2015;16:e60–e70.

73. IJ Majewski, I Kluijt, A Cats, et al. An alpha-E-catenin (CTNNA1) mutation in hereditary diffuse gastric cancer J Pathol, 229 (2013), pp. 621-629

74. El-Omar EM, Carrington M, Chow WH, et al. Interleukin-1 polymorphisms associated with increased risk of gastric cancer. Nature 2000;404:398–402.

75. El-Omar EM, Rabkin CS, Gammon MD, et al. Increased risk of noncardia gastric cancer associated with proinflammatory cytokine gene polymorphisms. Gastroenterology 2003;124:1193–1201.

76. Canedo P, Figueiredo C, Machado JC. After Helicobacter pylori, genetic susceptibility to gastric carcinoma revisited. Helicobacter 2007;12 Suppl 2:45–49.

77. Choi YJ, Kim N, Jang W, et al. Familial clustering of gastric cancer: a retrospective study based on the number of first-degree relatives. Medicine (Baltimore) 2016;95:e3606.

78. [9]V. Giroux and A. Rustgi, "Metaplasia: tissue injury adaptation and a precursor to the dysplasia–cancer sequence", Nature Reviews Cancer, vol. 17, no. 10, pp. 594-604, 2017.

79. Kim N, Park YS, Cho SI, et al. Prevalence and risk factors of atrophic gastritis and intestinal metaplasia in a Korean population without significant gastroduodenal disease. Helicobacter 2008;13:245–255.

80. Smyth EC, Cunningham D. targeted therapy for gastric cancer. Curr Treat Options Oncol 2012;13:377–89.

81. Fujimoto-Ouchi K, Sekiguchi F, Yasuno H, et al. Antitumor activity of trastuzumab in combination with chemotherapy in human gastric cancer xenograft models. Cancer Chemother Pharmacol 2007;59:795–805.

82. Matsui Y, Inomata M, Tojigamori M, et al. Suppression of tumor growth in human gastric cancer with HER2 overexpression by an anti-HER2 antibody in a murine model. Int J Oncol 2005;27:681–5.

83. M. Duffy, R. Lamerz, C. Haglund, A. Nicolini, M. Kalousová, L. Holubec and C. Sturgeon, "Tumor markers in colorectal cancer, gastric cancer and gastrointestinal stromal cancers: European group on tumor markers 2014 guidelines update", International Journal of Cancer, vol. 134, no. 11, pp. 2513-2522, 2013.

84. "Tumor Markers in Common Use", National Cancer Institute, 2021. [Online]. Available: https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-markers-list. [Accessed: 02- Jul- 2021]

85. Sharpe AH, Wherry EJ, Ahmed R, Freeman GJ. The function of programmed cell death 1 and its ligands in regulating autoimmunity and infection. Nat Immunol. 2007;8:239-245.

86. Gu L, Chen M, Guo D, Zhu H, Zhang W, Pan J, Zhong X, Li X, Qian H, Wang X. PD-L1 and gastric cancer prognosis: A systematic review and meta-analysis. PLoS One. 2017;12:e0182692.

87. F. van Roy and G. Berx, "The cell-cell adhesion molecule E-cadherin", Cellular and Molecular Life Sciences, vol. 65, no. 23, pp. 3756-3788, 2008.

88. Y. Zhang and R. Weinberg, "Epithelial-to-mesenchymal transition in cancer: complexity and opportunities", Frontiers of Medicine, vol. 12, no. 4, pp. 361-373, 2018.

89. A. Chan, "E-cadherin in gastric cancer", World Journal of Gastroenterology, vol. 12, no. 2, p. 199, 2006.

90. Gabbert HE, Mueller W, Schneiders A, Meier S, Moll R, Birchmeier W, Hommel G. Prognostic value of E-cadherin expression in 413 gastric carcinomas. Int J Cancer. 1996;69:184-189.

91. Shimada H, Noie T, Ohashi M, Oba K, Takahashi Y. Clinical significance of serum tumor markers for gastric cancer: a systematic review of literature by the Task Force of the Japanese Gastric Cancer Association. Gastric Cancer. 2014;17:26-33.

92. Asao T, Fukuda T, Yazawa S, Nagamachi Y. Carcinoembryonic antigen levels in peritoneal washings can predict peritoneal recurrence after curative resection of gastric cancer. Cancer. 1991;68:44-47.

93. Zhang YS, Xu J, Luo GH, Wang RC, Zhu J, Zhang XY, Nilsson-Ehle P, Xu N. Detection of carcinoembryonic antigen mRNA in peritoneal washes from gastric cancer patients and its clinical significance. World J Gastroenterol. 2006;12:1408-1411.

94. Feng, F., Tian, Y., Xu, G. et al. Diagnostic and prognostic value of CEA, CA19–9, AFP and CA125 for early gastric cancer. BMC Cancer 17, 737 (2017). https://doi.org/10.1186/s12885-017-3738-y

95. Kannagi R, Yin J, Miyazaki K, Izawa M. Current relevance of incomplete synthesis and neo-synthesis for cancer-associated alteration of carbohydrate determinants--Hakomori's concepts revisited. Biochim Biophys Acta. 2008;1780:525-531

96. T. Matsuoka and M. Yashiro, "Biomarkers of gastric cancer: Current topics and future perspective", World Journal of Gastroenterology, vol. 24, no. 26, pp. 2818-2832, 2018.

97. B. Rude Voldborg, L. Damstrup, M. Spang-Thomsen and H. Skovgaard Poulsen, "Epidermal growth factor receptor (EGFR) and EGFR mutations, function and possible role in clinical trials", Annals of Oncology, vol. 8, no. 12, pp. 1197-1206, 1997.

98. Kiyose S, Nagura K, Tao H, Igarashi H, Yamada H, Goto M, Maeda M, Kurabe N, Suzuki M, Tsuboi M, Kahyo T, Shinmura K, Hattori N, Sugimura H (2012) Detection of kinase amplifications in gastric cancer archives using fluorescence in situ hybridization. Pathol Int 62(7):477–484. doi:10.1111/j.1440-1827.2012.02832.x

99. Deng N, Goh LK, Wang H, Das K, Tao J, Tan IB, Zhang S, Lee M, Wu J, Lim KH, Lei Z, Goh G, Lim QY, Tan AL, Sin Poh DY, Riahi S, Bell S, Shi MM, Linnartz R, Zhu F, Yeoh KG, Toh HC, Yong WP, Cheong HC, Rha SY, Boussioutas A, Grabsch H, Rozen S, Tan P (2012) A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets. Gut 61(5):673–684. doi:10.1136/gutjnl-2011-301839

100. Kim MA, Lee HS, Lee HE, Jeon YK, Yang HK, Kim WH (2008) EGFR in gastric carcinomas: prognostic significance of protein overexpression and high gene copy number. Histopathology 52(6):738–746. doi:10.1111/j.1365-2559.2008.03021.x

101. Graziano F, Galluccio N, Lorenzini P, Ruzzo A, Canestrari E, D'Emidio S, Catalano V, Sisti V, Ligorio C, Andreoni F, Rulli E, Di Oto E, Fiorentini G, Zingaretti C, De Nictolis M, Cappuzzo F, Magnani M (2011) Genetic activation of the MET pathway and prognosis of patients with high-risk, radically resected gastric cancer. J Clin Oncol 29(36):4789–4795. doi:10.1200/JCO.2011.36.7706

102. Lieto E, Ferraraccio F, Orditura M, Castellano P, Mura AL, Pinto M, Zamboli A, De Vita F, Galizia G (2008) Expression of vascular endothelial growth factor (VEGF) and epidermal growth factor receptor (EGFR) is an independent prognostic indicator of worse outcome in gastric cancer patients. Ann Surg Oncol 15(1):69–79. doi:10.1245/s10434-007-9596-0

103. B. Kim, J. Kim, H. Jang and H. Kim, "The role of anti-EGFR agents in the first-line treatment of advanced esophago-gastric adenocarcinoma: a meta-analysis", Oncotarget, vol. 8, no. 58, pp. 99033-99040, 2017.

104. Yoshikawa T, Tsuburaya A, Kobayashi O, Sairenji M, Motohashi H, Yanoma S, Noguchi Y (2000) Plasma concentrations of VEGF and bFGF in patients with gastric carcinoma. Cancer Lett 153(1–2):7–12

105. Fuchs CS, Tomasek J, Yong CJ, Dumitru F, Passalacqua R, Goswami C, Safran H, Dos Santos LV, Aprile G, Ferry DR, Melichar B, Tehfe M, Topuzov E, Zalcberg JR, Chau I, Campbell W, Sivanandan C, Pikiel J, Koshiji M, Hsu Y, Liepa AM, Gao L, Schwartz JD, Tabernero J, for the RTI (2013) Ramucirumab monotherapy for previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (REGARD): an international, randomised, multicentre, placebo-controlled, phase 3 trial. Lancet. doi:10.1016/S0140-6736(13)61719-5

106. Pinto M, Oliveira C, Machado JC, Cirnes L, Tavares J, Carneiro F, Hamelin R, Hofstra R, Seruca R, Sobrinho-Simoes M (2000) MSI-L gastric carcinomas share the hMLH1 methylation status of MSI-H carcinomas but not their clinicopathological profile. Lab Invest 80(12):1915–1923

107. Corso G, Velho S, Paredes J, Pedrazzani C, Martins D, Milanezi F, Pascale V, Vindigni C, Pinheiro H, Leite M, Marrelli D, Sousa S, Carneiro F, Oliveira C, Roviello F, Seruca R (2011) Oncogenic mutations in gastric cancer with microsatellite instability. Eur J Cancer 47(3):443–451. doi:10.1016/j.ejca.2010.09.008

108. Sehdev A, Catenacci DV (2013) Gastroesophageal cancer: focus on epidemiology, classification, and staging. Discov Med 16(87):103–111

109. Durães, C., Almeida, G.M., Seruca, R. et al. Biomarkers for gastric cancer: prognostic, predictive or targets of therapy?. Virchows Arch 464, 367–378 (2014). https://doi.org/10.1007/s00428-013-1533-y

110. S. Bustin and C. Wittwer, "MIQE: A Step Toward More Robust and Reproducible Quantitative PCR", Clinical Chemistry, vol. 63, no. 9, pp. 1537-1538, 2017.

111. S. Bustin, V. Benes, J. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. Pfaffl, G. Shipley, J. Vandesompele and C. Wittwer, "The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments", Clinical Chemistry, vol. 55, no. 4, pp. 611-622, 2009.

112. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet. 2013;29(10):569–74.

113. Zhu J, He F, Hu S, Yu J. On the nature of human housekeeping genes.Trends Genet. 2008;24(10):481–4.

114. Sharan RN, Vaiphei ST, Nongrum S, Keppen J, Ksoo M. Consensus reference gene(s) for gene expression studies in human cancers: end of the tunnel visible? Cell Oncol (Dordr). 2015;38(6):419–31.

115. Jacob F, Guertler R, Naim S, Nixdorf S, Fedier A, Hacker NF, Heinzelmann- Schwarz V. Careful selection of reference genes is required for reliable performance of RT-qPCR in human normal and cancer cell lines. PLoS One. 2013;8(3):e59180.

116. Y. Sun, Y. Li, D. Luo and D. Liao, "Pseudogenes as Weaknesses of ACTB (Actb) and GAPDH (Gapdh) Used as Reference Genes in Reverse Transcription and Polymerase Chain Reactions", PLoS ONE, vol. 7, no. 8, p. e41659, 2012.

117. Ooi CH, Ivanova T, Wu J, Lee M et al. Oncogenic pathway combinations predict clinical prognosis in gastric cancer. PLoS Genet 2009 Oct;5(10):e1000676. PMID: 19798449

118. Cristescu R, Lee J, Nebozhyn M, Kim KM et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med 2015 May;21(5):449-56. PMID: 25894828

119. Busuttil RA, George J, Tothill RW, Ioculano K et al. A signature predicting poor prognosis in gastric and ovarian cancer represents a coordinated macrophage and stromal response. Clin Cancer Res 2014 May 15;20(10):2761-72. PMID: 24658156

120. Yong WP, Rha SY, Tan IB, Choo SP et al. Real-Time Tumor Gene Expression Profiling to Direct Gastric Cancer Chemotherapy: Proof-of-Concept "3G" Trial. Clin Cancer Res 2018 Nov 1;24(21):5272-5281. PMID: 30045931

121. Qian Z, Zhu G, Tang L, Wang M et al. Whole genome gene copy number profiling of gastric cancer identifies PAK1 and KRAS gene amplification as therapy targets. Genes Chromosomes Cancer 2014 Nov;53(11):883-94. PMID: 24935174

122. Oh SC, Sohn BH, Cheong JH, Kim SB et al. Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. Nat Commun 2018 May 3;9(1):1777. PMID: 29725014

123. Cho JY, Lim JY, Cheong JH, Park YY et al. Gene expression signature-based prognostic risk score in gastric cancer. Clin Cancer Res 2011 Apr 1;17(7):1850-7. PMID: 21447720

124. Lee J, Sohn I, Do IG, Kim KM et al. Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. PLoS One 2014;9(3):e90133. PMID: 24598828

125. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

126. Hmisc

Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2020). Hmisc: Harrell Miscellaneous. R package version 4.4-1.

https://CRAN.R-project.org/package=Hmisc

127. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, et al. (2006) Design and Analysis of DNA Microarray Investigations: Springer New York.

128. Bolstad B (2020). preprocessCore: A collection of pre-processing functions. R package version 1.50.0, https://github.com/bmbolstad/preprocessCore.

129. Alboukadel Kassambara, Marcin Kosinski and Przemyslaw Biecek (2020). survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.8. https://CRAN.R-project.org/package=survminer

130. Terry M. Therneau, Patricia M. Grambsch (2000). _Modeling Survival Data: Extending the Cox Model_. Springer, New York. ISBN 0-387-98784-3.

131. Torsten Hothorn (2017). maxstat: Maximally Selected Rank Statistics. R package version 0.7-25. https://CRAN.R-project.org/package=maxstat

132. Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

133. Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, 550. doi: 10.1186/s13059-014-0550-8.

134. Li, B., Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011). https://doi.org/10.1186/1471-2105-12-323

135. Rees, M., Seashore-Ludlow, B., Cheah, J. et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol 12, 109–116 (2016). https://doi.org/10.1038/nchembio.1986

136. Iorio, F., Knijnenburg, T., Vis, D., Bignell, G., Menden, M., & Schubert, M. et al. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell, 166(3), 740-754. doi: 10.1016/j.cell.2016.06.017

137. Ghandi, M., Huang, F.W., Jané-Valbuena, J. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature 569, 503–508 (2019). https://doi.org/10.1038/s41586-019-1186-3

138. M. Kwa, A. Makris and F. Esteva, "Clinical utility of gene-expression signatures in early stage breast cancer", *Nature Reviews Clinical Oncology*, vol. 14, no. 10, pp. 595-610, 2017.

139. R. Vieira, G. Biller, G. Uemura, C. Ruiz and M. Curado, "Breast cancer screening in developing countries", Clinics, vol. 72, no. 4, pp. 244-253, 2017.

140. S. Demirkol Canli, “Prediction of Prognosis and Chemosensitivity in Gastrointestinal Cancers,” Ph.D. dissertation, Dept. of Molecular Biology and Genetics, Bilkent Univ., Ankara, 2017. [Online]. Available: http://hdl.handle.net/11693/35716

141. L. Xie, L. Cai, F. Wang, L. Zhang, Q. Wang and X. Guo, "Systematic Review of Prognostic Gene Signature in Gastric Cancer Patients", Frontiers in Bioengineering and Biotechnology, vol. 8, 2020.

142. L. X., Wu, J., Zhang, D., Bing, Z., Tian, J., Ni, M., et al. (2018). Identification of potential key genes associated with the pathogenesis and prognosis of gastric cancer based on integrated bioinformatics analysis. Front. Genet. 9:265. doi: 10.3389/fgene.2018.00265

143. L. Zhou, H. Lu, F. Zeng, Q. Zhou, S. Li, Y. Wu, Y. Yuan and L. Xin, "Constructing a new prognostic signature of gastric cancer based on multiple data sets", Bioengineered, vol. 12, no. 1, pp. 2820-2835, 2021.

144. S. Huang, Y. Wang and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics", Cancer Genomics & Proteomics, vol. 15, no. 1, 2018.

145. B. Zhao, Z. Baloch, Y. Ma, Z. Wan, Y. Huo, F. Li and Y. Zhao, "Identification of Potential Key Genes and Pathways in Early-Onset Colorectal Cancer Through Bioinformatics Analysis", Cancer Control, vol. 26, no. 1, p. 107327481983126, 2019.

146. Y. Liu, J. Wu, W. Huang, S. Weng, B. Wang, Y. Chen and H. Wang, "Development and validation of a hypoxia-immune-based microenvironment gene signature for risk stratification in gastric cancer", Journal of Translational Medicine, vol. 18, no. 1, 2020.