TRANSCRIPT VARIANTS OF CELF2 GENE AS UNIQUE PROGNOSTIC INDICATORS IN BREAST CANCER

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

MOLECULAR BIOLOGY AND GENETICS

By

Shila Azizolli

August, 2022

TRANSCRIPT VARIANTS OF CELF2 GENE AS UNIQUE PROGNOSTIC INDICATORS IN BREAST CANCER

By Shila Azizolli

August 2022

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Ali Osmay Güre (Advisor)

Özlen Konu Karakayalı

Onur Emre Onat

Approved for the Graduate School of Engineering and Science:

ORHAN ARIKAN

Director of the Graduate School

ABSTRACT

TRANSCRIPT VARIANTS OF CELF2 GENE AS UNIQUE PROGNOSTIC INDICATORS IN BREAST CANCER

SHILA AZIZOLLI

M.Sc. in Molecular Biology and Genetics

Advisor: Ali Osmay GÜRE

AUGUST 2022

Breast cancer (BC) is the most common malignant tumor in women around the world. Aside from finding a cure for this disease, it is also critical to identify prognostic biomarkers that can help clinicians intervene with the appropriate treatment and prevent BC progression. Current biomarker identification methods rely primarily on multi-gene prognostic signature models. However, due to the tumors' high heterogeneity, the accuracy of these multi-gene signatures is questionable. As a result, our main objective was to conduct a comprehensive analysis to identify a reliable prognostic biomarker in BC. Previously, a group of eight carnitine metabolites and SAH were linked to a poor prognosis in BC (Dr. Waqas Akbar, Unpublished Data). We discovered the genes associated with these metabolites using correlation analysis, and then we identified CELF2 as a good prognostic biomarker in BC. We validated CELF2's prognostic role in RNA-Seq and Microarray datasets in-silico. We demonstrate that the CELF2 1554569_a_at probeset is more consistent in its association with a favorable prognosis direction than the 202157_s_at probeset. When compared to the other probeset, CELF2 – 202157_s_at is expressed at higher levels and in a broader range of tissues.

We were unable to find a clear significant association between CELF2 expression and prognosis during the in-vitro immunohistochemistry validation experiments. We hypothesized that this could be because our polyclonal anti-CELF2 antibody also recognized the less consistent 202157_s_at CELF2 probeset. We discovered that there are probeset-specific CELF2 transcript variants that are associated with different prognosis while testing this hypothesis.

We created a Risk Score model by combining the expression levels of good and less-favorable CELF2 prognostic transcripts to improve prognosis prediction accuracy. The model successfully stratified the patients and predicted a higher overall survival in the Low-Risk group versus the High-Risk group.

Overall, our findings suggest that each unique transcript variant of a gene can be associated with different prognosis directions. Therefore, we propose that studying prognostic associations at a gene transcript level could be a rich resource for the development of more robust biomarkers and therapeutics in cancer in the future.

Keywords: Breast Cancer, Prognosis, Biomarker, Transcript variants.

ÖZET

MEME KANSERİNDE CELF2 GENİNİN ÖZEL PROGNOSTİK GÖSTERGELER OLARAK TRANSKRİP ÇEŞİTLERİ

SHILA AZIZOLLI

MOLEKÜLER BIYOLOJI VE GENETIK, YÜKSEK LISANS

TEZ DANIŞMANI: YARD. DOÇ. DR. ALI OSMAY GÜRE

AĞUSTOS 2022

Meme kanseri (MK) tüm dünyada kadınlarda en sık görülen malign tümördür. Bu hastalık için bir tedavi bulmanın yanı sıra, klinisyenlerin uygun tedaviyle müdahale etmesine ve MK'nin ilerlemesini önlemesine yardımcı olabilecek prognostik biyobelirteçleri belirlemek de önemlidir. Mevcut biyobelirteç tanımlama yöntemleri, esasen çok genli prognostik imzalara dayanır. Fakat tümörlerin yüksek heterojenliği nedeniyle, bu çoklu gen imzalarının tutarlılığı şüphelidir. Sonuç olarak temel amacımız, MK'de güvenilir bir prognostik biyobelirteç belirlemek için kapsamlı bir analiz yapmaktı. Daha önce, sekiz karnitin metabolitinden oluşan bir grup ve SAH, MK'de kötü bir prognozla bağlantılıydı (Dr. Waqas Akbar, Yayınlanmamış Veriler). Korelasyon analizi kullanarak bu metabolitlerle ilişkili genleri keşfettik ve ardından CELF2'yi MK'de iyi bir prognostik biyobelirteç olarak tanımladık. CELF2 geninin prognostik rolunu RNA-Seq ve Microarray veritabanlarinda in-silico doğruladık. CELF2 1554569_a_at probeset'inin, olumlu bir prognoz yönü ile ilişkisi açısından 202157_s_at probset'inden daha tutarlı olduğunu gösteriyoruz. Diğer probset ile karşılaştırıldığında, CELF2 – 202157_s_at daha yüksek seviyelerde ve daha geniş bir doku aralığında ifade edilir.

İn-vitro immünohistokimya doğrulama deneyleri sırasında CELF2 ekspresyonu ile prognoz arasında açık ve anlamlı bir ilişki bulamadık. Bunun, poliklonal anti-CELF2 antikorumuzun daha az tutarlı 202157_s_at CELF2 probset'i tanımasından kaynaklanabileceğini varsaydık. Bu hipotezi test ederken, farklı prognozla ilişkili, probset'e özgü CELF2 transkript varyantları olduğunu keşfettik.

Prognoz tahmini doğruluğunu iyileştirmek için, iyi ve daha az elverişli CELF2 prognostik transkriptlerinin ifade düzeylerini birleştirerek bir Risk Puanı modeli oluşturduk. Model,

hastaları başarılı bir şekilde sınıflandırdı ve Düşük Riskli grupta Yüksek Riskli gruba kıyasla daha yüksek bir genel sağkalım öngördü.

Genel olarak, bulgularımız bir genin her özel transkript varyantının farklı prognoz yönleriyle ilişkili olabileceğini düşündürmektedir. Bu nedenle, bir gen transkript düzeyinde prognostik ilişkilerin incelenmesinin, gelecekte kanserde daha sağlam biyobelirteçlerin ve terapötiklerin geliştirilmesi için zengin bir kaynak olabileceğini öneriyoruz.

Anahtar Kelimeler: Meme Kanseri, Prognoz, Biyobelirteç, Transkript varyantları.

In honor of all the cancer survivors and those who have lost their battle…

ACKNOWLEDGEMENTS

I am pleased to express my deepest gratitude to my supervisor Dr. Ali Osmay Güre for his guidance throughout my Master's studies. His unique ability to communicate his ideas logically and efficiently helped me learn how to organize my thoughts and work better. Besides being an inspiring mentor with a genuine interest to explore critical questions, I had the opportunity to learn something new every time we would have a conversation outside the lab environment as well. In my scientific development, Dr. Ali Osmay Güre will always be a major influence. It is a privilege for me to be a member of his lab.

Secondly, I owe my regards to Dr. Özlen Konu and Dr. Onur Emre Onat for their valuable feedback and insights on the preparation of this thesis and potential future perspectives.

I would like to thank all the members of AOG Lab for their unconditional support and positivity: Ege Dedeoglu, Baris Kucukkaraduman, Farid Ahadli, Marzana Ishraq, Secil Demirkol, and Murat Isbilen. It was always a pleasure to be a part of this team, and work in a very harmonious environment.

I would like to thank especially Dr. Muhammad Waqas Akbar for entrusting me with this project. I learned a lot of new data analysis techniques thanks to his help. In addition to being a great colleague, he is also an excellent camel rider. I believe the balance created between his pessimistic nature versus my optimistic views is the key to a long-lasting friendship.

I would like to thank Noor Niaz and Ronak Naeemaee, who became my dear friends outside of the lab as well. With her kind soul, Noor would always try to brighten my day by leaving small personalized notes on my bench. Half-scientist, half-writer, she has produced some of my all-time favorite hip-hop verses. I would also like to thank her for her adventurous spirit in encouraging us to try new things, such as auditioning for a photo model agency. Ronak, on the other hand, has always been the voice of reason in our friendship dynamics. She was always the one who planned out our trips, even if they were as short as going to the supermarket. Our long conversations are among the most relatable and enjoyable ones I have ever had. I will always treasure the moments that we shared altogether.

I would also like to thank some of my Albanian friends: Herr Aldo Tali, Arli Lika, and Franc Gripshi. They were a constant source of distraction from the daily routine and always made me feel looked after with their spontaneous little gifts. I would like to thank Sezin Unlu, with whom we have shared the same journey for the past six years and have created some beautiful memories I will cherish forever. I also want to thank my friend Mustafa

Akkaya for teaching me his minimalistic approach to solving problems. Despite meeting later in life, he was a big part of my last chapter in Turkey.

I will be forever thankful to Isli Çela, who has been there for me since the beginning. I cannot count the number of times we laughed and cried together over the years. Every time we had to encourage one another to work harder, stay positive, and not give up. All of the times you helped me regain my confidence, and the vicious cycles of late-night coffee we shared. Aside from being a beautiful person, I believe you are one of the finest and most brilliant young scientists I have met. I cannot wait to see what you accomplish in the future!

Last but not least, I would like to thank my wonderful family for their unwavering faith in me. My mother, Rezarta (Sunshine), my greatest blessing in life, has always managed to transmit her warmth to me even at a distance. She has always been willing to listen, give me valuable advice and strength to carry on. Lisjan never fails to remind me that I am capable of accomplishing anything I set my mind to. I could not have asked for a more understanding and supportive brother. I would also like to express my gratitude to my aunt Anilda and cousin Ergys for their kind words and emotional support. I would like to thank Dr. Isuf and Dr. Lindita, who became my second family and always made me feel appreciated.

Grandmother, Shpresa (Hope), I hope I have done you proud!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

1. Introduction

Cancer is a disease characterized by an uncontrollable cell growth. Cancer can develop at different sites in the human body and then spread to nearby tissues or distant places during metastasis [1]. During their growth, cancer cells form lumps of tissue also known as tumors. However, blood cancers would be an exception to that rule since they do not form solid tumor tissues. The multistep process of the transformation of normal cells to cancer cells has been widely studies and several common capabilities have been identified. These characteristics of cancer cells, referred to as "Cancer Hallmarks" originally constituted the ability to sustain proliferative signaling, resisting cell death, replicative immortality, inducing angiogenesis, evading growth suppressors, and activating invasion and metastasis. After conceptual progress, metabolic reprogramming and immune destruction invasion were included in the list. Currently, non-mutational epigenetic reprogramming, senescent cells, polymorphic microbiomes and unlocking phenotypic plasticity are emerging into the field. An essential factor mediating the cancer cells to gain these abilities is the tumor microenvironment, a complex entity of immune cells, stromal cells, blood vessels, and extracellular matrix [2, 3].

Breast Cancer

1.1 Epidemiology

Breast Cancer (BC) is the most commonly diagnosed cancer and the leading cause of cancer death among women globally. Female breast cancer accounts for 1 in 4 cases of cancer and 1 in 6 cancer deaths [4]. Translated into numbers, that means that in 2020 there were 2.3 million women diagnosed with breast cancer followed by 685.000 deaths worldwide [5]. In 2020, 7123 women died of breast cancer in Turkey [4]. Depending on the available diagnostic methods and health care system infrastructure, the 5-year-survival-rate of breast cancer patients varies across the globe from 80% in transitioned countries to less than 40% in transitioning ones. The survival rate of breast cancer patients

can change based on multiple factors including the stage of the tumor at the point of diagnosis [6]. The 5-year survival rate for localized female breast cancer is 99.1% and decreases to 30% for distant breast cancer in America between 2012 and 2018 [7]. Male breast cancer has a low incidence and it accounts for 1% of the overall breast cancer cases. Even though it occurs rarely, due to the late diagnosis, male breast cancer usually has a high mortality rate [8].

1.2 Histological Classification

1.2.1 Invasive Breast Cancer

According to the cell of origin, invasive BC can be classified into ductal and lobular carcinoma. Cancers arising from the ducts are known as ductal carcinoma, while those arising from the lobules are known as lobular carcinoma [9]. Invasive lobular carcinoma (ILC) is associated with a poor prognosis due to the high rate of lymph node metastasis compared to invasive ductal carcinoma (IDC). Upon chemotherapy treatment, ILC and IDC show similar overall survival (OS) in HER2-positive patients however, ILC tends to have a worse prognosis in the HER2-negative group [10].

1.2.2 Noninvasive Breast Cancer

Noninvasive BC can be classified into ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS). Both DCIS and LCIS can be considered risk factors for subsequent invasive breast carcinoma [11]. DCIS is characterized by a continuous strong expression of E-cadherin, while LCIS shows a loss of E-cadherin expression [12]. Therefore, immunohistochemical staining on E-cadherin can be performed to distinguish between the two [9]. DCIS patients have a favorable long-term survival rate of 98% after a follow-up period of 10 years [13]. LCIS patients as well exhibit high survival rates that surpass 90% at 10-year follow-up, irrespective of surgery [14].

1.3 Molecular Classification

According to the molecular profile, Breast Cancer is classified into Luminal A (LumA), Luminal B (LumB), Her2-enriched (HER2), Basal-Like, and Normal-Like subtypes. This molecular classification is based on gene expression differences between the intrinsic subtypes and is widely used in clinics under the name of PAM50 [15, 16]. LumA subtype is estrogen-receptor positive (ER+), progesterone-receptor positive (PR+), and HER2-negative with low levels of Ki-67 protein. LumA generally has the best prognosis. Compared to LumA, LumB tends to have higher levels of Ki-67 indicating faster cell growth and progesterone-receptor expression [17]. HER2 cancers are characterized by high expression levels of Her2 hormone and tend to grow faster than the luminal subtypes. They are associated with higher rates of recurrence and mortality compared to the prior subtypes [18]. Basal-like cancers test negative for all three markers ER/PR/Her2 and show high Ki67 expression levels. They have the worst prognosis for OS, relapse-free survival (RFS), and disease-specific survival (DSS) among all the intrinsic groups [19, 20].

Table 1.1: Molecular Classification of Breast Cancer.

MOLECULAR SUBTYPE	ER/PR	HER2	Ki67	TUMOR GRADE	PROGNOSIS
Luminal A	ER+ PR+	HER2-	Low	I	Good
Luminal B $_{(HER2-)}$	ER+ PR-	HER2-	High	II	Intermediate
Luminal B $_{(HER2+)}$	ER+ PR+/-	HER2+	High	II	Intermediate
HER2-*enriched*	ER- PR-	HER2+	High	III	Poor
Basal-*like*	ER- PR-	HER2-	High	III	Poor

1.4 Risk Factors in Breast Cancer

Breast Cancer development can be influenced by particular characteristics of the patient. The majority of BC cases are caused by environmental and lifestyle factors, which can be potentially modified to reduce the incidence of the disease. The remaining fraction of BC cases are caused due to family history and inherited gene mutations [21].

1.4.1 Lifestyle Factor

Lifestyle behavior and cancer risk have been associated in multiple studies. The guidelines produced by The World Cancer Research Fund leading to a healthy lifestyle have been previously studied and lower rates of BC incidences have been reported among the women who follow these guidelines [22-24]. Weight gain during adulthood has been reported to double the BC risk. Furthermore, weight loss before or after menopause has been associated with reduced BC risk as compared to the women who gained or had a stable weight [25, 26]. Alcohol consumption dosage is another factor that has been associated with higher cell proliferation in ER-positive BC [27, 28]. A high protein diet is also considered to be a BC risk factor. However, there are controversial studies that don't support this claim. A study based on eight cohorts including 7379 women diagnosed with invasive BC found no significant association between BC risk and regular intake of dairy and meat products [29]. Another study however reported that while red meat consumption may be associated with higher risks of BC, poultry consumption is associated with reduced BC risk [30]. The difference between the relation of red meat and poultry to BC risk could be due to the lower fat content of the latter. It has also been shown that poultry consumption promotes lower levels of DNA damage and oxidative stress [31].

1.4.2 Family History

Women with a history of breast cancer in their first-degree family have been shown to have a two-fold increase in their risk of developing BC. The BC risk tends to be higher for those with a first-degree family member diagnosed before the age of 50 [32, 33]. Two of the main cancer susceptibility genes in hereditary BC are Breast Cancer Type I and Type II

(BRCA1 and BRCA2) involved in maintaining genome integrity by partly engaging in DNA repair and cell-cycle checkpoint control [34]. Inherited BRCA1 and BRCA2 gene mutations can lead to BC development in up to 85% of women [23]. Other high-risk hereditary mutations include p53 and PTEN genes. Patients that have inherited a p53 mutation (Li Fraumeni Syndrome) are likely to develop breast cancer before 45 years of age [35]. Those who inherit PTEN mutations, also known as PTEN Hamartoma-Tumor Syndrome have an 85% risk of developing BC throughout their lifetime [36].

1.5 Prognostic Factors

Prognostic factors have wide applications in clinics. They can be used to decide on the right treatment that would increase the patient's benefits compared to the possible risks of the therapy. Prognostic markers can also identify patients with a poor life expectancy that might need a more aggressive treatment [37]. The classical clinical factors include age, tumor size, tumor grade, and nodal status.

1.5.1 Age

Age is a parameter that needs to be considered while evaluating treatment options for BC patients. It has been reported that with increasing age, the risk of BC development will increase as well [38]. BC tumors diagnosed at an age < 35 years old tend to be more aggressive and exhibit a poorer prognosis [39].

1.5.2 Tumor Size

Tumor size is a good predictor of BC behavior and patient survival. Smaller tumor sizes < 1cm usually have been reported to have a better 5-year OS rate than tumors > 3cm [40]. Larger tumors are also associated with shorter periods of time to develop metastasis [41].

1.5.3 Nodal Status

The lymphatic system has been known to facilitate the dissemination of multiple cancers and the axillary lymph node serves as the initial metastatic niche for BC cells [42]. Therefore, axillary lymph node presence is a critical indicator of survival and recurrence

in BC patients. In the 5-years OS, patients classified as node-negative are reported to have around 83% survival rate, those with 4 to 12 positive nodes a 46% survival rate, and around 28% for patients with more than 13 positive nodes [43].

1.5.4 Tumor Grade

The most clearly defined and frequently used grading system is the Nottingham modification of the Bloom and Richardson system [44]. It classifies the tumors by taking into consideration three parameters: tumor grade, tumor size, and lymph node stage. Based on the Nottingham classification system, a tumor can be assigned grades from 1-to-3, where grade 1 indicates a well-differentiated tumor and grade 3 a poorly differentiated one. Grade 1 patients have a survival chance of 85% at 10 years post-diagnosis, while grade 3 patients have a survival chance of 45% [45, 46].

1.6 Breast Cancer Multiple-Gene Prognostic Signatures

Multiple gene signatures in multiple types of cancer have emerged as principal prognostic tools. This is due to their potential to reduce the treatment cost by identifying the patients that wouldn't benefit from it. The main property of multiple gene prognostic signatures is to stratify the patients into high-risk and low-risk groups and predict the survival differences as an outcome of various treatments. Consequently, an increase in the accuracy of these tests through multiple patient cohorts could provide clinicians with a powerful decision-making tool.

1.6.1 MammaPrint

The MammaPrint test is based on the analysis of the 70 most important genes associated with BC recurrence. MammaPrint assay has been cleared by US FDA to be used by women of all ages diagnosed with ER-positive or ER-negative, lymph node-negative BC. The 70 genes are mainly implicated in metastasis, invasion, stromal integrity, and angiogenesis [15, 47, 48]. Based on multivariate analysis, the 70-gene signature was also found to be an independent prognostic factor [49]. The clinical benefit of the MammaPrint test has been validated by MINDACT, a randomized phase 3 trial including 6693 women diagnosed

with early-stage BC. The main aim of the study was to assess the efficacy of the MammaPrint test in selecting patients eligible for adjuvant chemotherapy. Patients were classified as low- or high- genomic risk based on MammaPrint and low- or high- clinical risk based on their histopathological parameters. Patients in the high clinical risk group alone are usually eligible candidates for adjuvant chemotherapy in clinics. However, those who were classified as high clinical risk but low genomic risk did not undergo adjuvant chemotherapy, which resulted in a rate of survival without distant metastasis at an average of 1.5 percentage points lower than the patients who received chemotherapy [50]. These findings indicated that the women diagnosed with BC who are at high clinical risk might not always require chemotherapy. The TRANSBIG Consortium of European Cancer Centers also validated the MammaPrint clinical utility in an independent patient cohort. The 70-gene signature outperformed the risk assessment based on histopathological characteristics of the patients at all endpoints, including distant metastasis and OS [47].

1.6.2 Oncotype DX

Oncotype DX is an assay that uses reverse-transcriptase–polymerase-chain-reaction (RT-PCR) to quantify the expression of 21 genes and compute a recurrence score in patients with ER-positive, node-negative BC who had been treated with tamoxifen. Considering the low recurrence rate at the 10-year survival time point of tamoxifen-treated patients, the initial aim was to identify a gene signature that would help clinicians to not overtreat them with chemotherapy. The 21-gene assay quantifies the likelihood of distant recurrence by comparing the expression of the 16 cancer-related genes to that of the 5 reference genes. The generated Recurrence Score (RS) divides patients into low-, intermediate-, and high-risk categories. The distant recurrence rate at the 10-year time point was shown to be 6.8%, 14.3%, and 30.5% respectively for each category [51]. The Oncotype DX test has been widely validated for its clinical application. The TAILORx trial validated the function of RS in a cohort including 10,253 women with hormone-receptor-positive, HER2-negative, node-negative BC who were eligible for both endocrine therapy and adjuvant chemotherapy based on the histopathological features of the tumors. After

RT-PCR of the 21-gene panel and RS assignment, the distant recurrence and overall survival rate were calculated. The TAILORx indicated that women classified in the low-RS group, and had undergone endocrine therapy only, had highly favorable clinical outcomes with a rate of 5-year freedom from distant recurrence at 99.3%, a rate of invasive DFS at 93.8%, and an OS rate of 98% [52]. ECOG E2197 trial also studies the prognostic utility of Oncotype DX in a sample of 465 patients with HR-positive, 1-3 node-positive BC, irrespectively of recurrence after chemo-hormonal therapy. The patients that were classified as low-RS (<18) had a recurrence rate of only 5% or less for a 5-year time point. The authors have also shown that the 21-gene panel can perform better at predicting relapse compared to the histopathological features [53]. Even though Oncotype DX is not approved by FDA, the American Society of Clinical Oncology (ASCO) has recommended it to be used in adjuvant chemotherapy decision-making for HR-positive, node-negative, and HR-negative BC patients.

1.6.3 PAM50

Another FDA-approved prognostic signature is PAM50 (Prosigna TM) which consists of 50 genes and 5 reference genes to compute the risk of recurrence score (ROR). ROR is used to classify tumors into LumA, LumB, HER2-enriched, and Basal-like subtypes which are then used to assess the clinical risk of BC [54]. The clinical limitation of this classification assay is the fact that depending on the platform, the same patient can be classified into different subtypes. Basal-like cancer is the most homogeneous subtype and can be clearly distinguished from the other ones [55]. PAM50 classifier and ROR score have been validated to have prognostic value in ER-positive post-menopausal women treated with endocrine therapy. Hazard ratios for the BC DSS after 5 years of follow-up indicated that LumA has the lowest risk and Basal-like subtypes have the highest risk [20]. Another study confirmed the independent prognostic ability of the ROR weighted for PAM50 proliferation score in a cohort of 220 patients diagnosed with primary breast cancer. Based on the estimates of post-relapse survival the luminal cases performed better than the non-luminal ones. In the endocrine-treated group of patients, the ones classified as low-risk based on the ROR had a post-relapse survival time of 35 months while the

intermediate-risk and high-risk groups had 18 months and 8 months RFS, respectively [56].

1.7 Alternative Splicing

1.7.1 Alternative Splicing and regulation

Alternative splicing (AS) is a post-transcriptional regulation mechanism that generates multiple transcripts originating from a single parental gene. It has been estimated that in humans 95% of multi-exon genes undergo AS, making it a primary source of protein

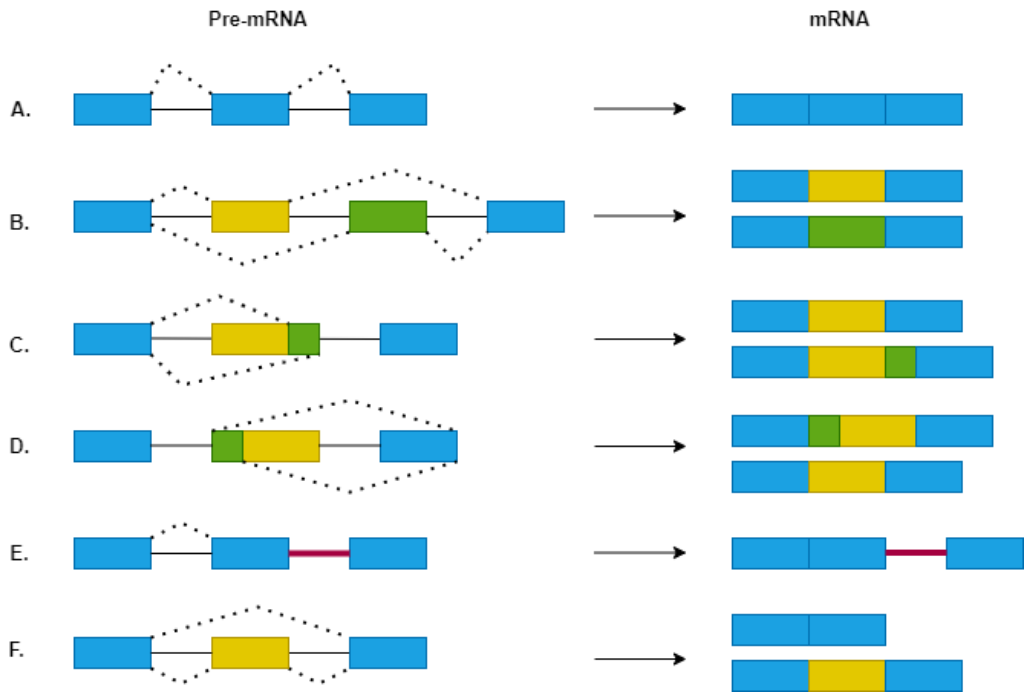


Figure 1.1: Alternative Splicing Types. (A) Constitutive splicing, (B) Mutually Exclusive, (C) Alternative 3' splice site, (D) Alternative 5' splice site, (E) Intron retention, (F) cassette alternative exon.

diversity [57, 58]. Before translation occurs, the mRNA undergoes a splicing reaction, during which the introns from the pre-mRNA are removed and the exons are ligated back together. During AS multiple mRNAs are generated from a number of different combinations of exons [59, 60]. There are five main types of AS which include:

Constitutive splicing, mutually exclusive exons, cassette alternative exon, intron retention, and alternative selection of 5’ or 3’ splice sites [61]. The proteins translated from alternatively spliced mRNAs will often differ in their amino acid sequence and exhibit differences in their biological functions, structures, molecular interactions, enzymatic activities, and localization [62].

The splicing process is carried out by a major complex known as the spliceosome, composed of five small nuclear ribonucleoproteins (snRNPs) and numerous proteins [63]. The conserved regions of an intron, known as 5’ and 3’ termini will constitute the primary recognition binding sites for the spliceosome. The most highly conserved nucleotides at these intronic splice sites are the 5’ terminal GU and 3’ terminal AG. Another conserved sequence is the Branch Site (BS), also known as the lariat formation site, which is followed by a pyrimidine-rich tract at the 3’ end of the intron [64]. However, these conserved regions, considering their short and degenerate nature, are not the sole accountable source for the usually error-free accuracy of the general splicing process nor that of AS.

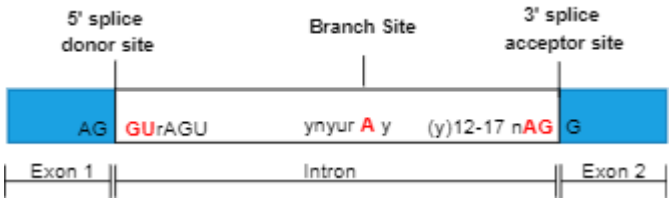


Figure 1.2: Conserved core splicing motifs.

It has been discovered and validated that there are other auxiliary sequences needed for the spliceosomal machinery to distinguish exons from introns during splicing and modulate AS. These auxiliary sequences depending on their location and effect on splicing are referred to as cis-acting elements: exon silencer or enhancer (ESS/ESE), intron silencer or enhancer (ISS/ISE), and trans-acting factors, repressor or activator proteins [65].

1.7.2 Cis-acting Elements

The majority of splicing enhancers, ESE and ISE, contain purine-rich domains and exert their activity via interactions with splicing activators serine/arginine-rich (SR) proteins characterized by RNA Recognition Motifs (RRM) presence [66, 67]. The splicing silencers,

ESS and ISS, inhibit the splice site recognition by interacting with the splicing repressor nuclear ribonucleoproteins (hnRNPs). HnRNPs can then physically block the association of snRNPs, splicing activators, by overlapping with the branch site [68-70]. Generally, the cis-acting elements work additively. The enhancers have more dominant roles in constitutive splicing, while the silencers are more essential in the control of AS [71].

1.7.3 Trans-acting Elements

Alternative splicing is mainly regulated by 4 different types of trans-acting factors. The first group of factors includes U1, U2, U4, U5, and U6 snRNPs spliceosomal components. Initially, U1 recognizes the 5' end of the splice site to be followed by U2 which recognizes the branch site to form a pre-splicing complex [72, 73]. Moreover, U4, U5, and U6 join the pre-splicing complex to create the fully assembled spliceosome [74]. Dynamic conformational changes between these snRNPs will then lead to two transesterification reactions which will result in a fully mature mRNA [75].

The second group is the SR protein family. SR proteins are characterized by an Arginine (R) and Serine (S) rich C-terminal domain and one or two copies of an RNA-recognition motif (RRM) at the N-terminal domain. The RS domain is important in promoting protein-protein interactions during spliceosome recruitment, while the RRM domain provides RNA-binding specificity [76]. Usually, SR protein will bind to ESE to enhance the splicing, while at the same time prevent exon skipping to ensure the correct 5'-to-3' order of exons in mRNA [77].

As mentioned previously, hnRNPs constitute the third group of trans-activating factors. These RNA-binding proteins contain RRM, K homology domain, glycine and tyrosine repeats, zinc-finger, and other protein-specific domains[78]. Besides having an active role in alternative splicing, hnRNPs are also involved in mRNA stabilization, transcriptional and translational regulation [79, 80]. Some of the hnRNP proteins documented to have a role in alternative splicing include hnRNP A1, A2, and H group. HnRNP A1-A2 alter 5' splice site choice and promote exon skipping in alternative splicing in a similar fashion [81-83].

HnRNP H group can also alter the 5’ splice site, inhibit splicing and promote exon skipping [84].

The fourth group is CUGBP and ETR-like factors (CELF) protein family. There are in total six CELF RBPs ranging from CELF1-to-CELF6. Based on phylogenetic studies proteins are divided into two subgroups: CELF1 and CELF2 which share 78% amino acid sequence identity, and CELF3-CELF6 which share 62%-66% amino acid sequence identity [85]. While CELF3-CELF6 have mainly demonstrated tissue-specific expression patterns, CELF1 and CELF2 are widely expressed in human tissues [86, 87]. The proteins share similar structures that include the presence of two RRMs in the N-terminal domain, one RRM, and a lysine/arginine-rich nuclear localization signal in the C-terminal domain, as well as a divergent linker region which is essential for nuclear export and undergoes conformational changes to improve RBP affinity [88, 89]. CELF proteins are involved in global AS regulation in the nucleus by recognizing CUG and UG- repeats. Depending on the expressed protein isoform, the structure, expression, location, or even functional properties of some CELF proteins can be altered [89, 90]. In the cytoplasm, CELF proteins are involved in pre-miRNA maturation, mRNA stability, translation, and alternative polyadenylation (APA) [91-93].

1.8 Aberrant Alternative Splicing in Breast Cancer

There has been increasing evidence over the last years on the effects that alternative splicing has on cancer development and tumorigenesis [94]. Dysregulations of AS can affect several hallmarks of cancer, such as metabolism, invasion, angiogenesis, metastasis, immune escape, and apoptosis signaling [95]. Studies have reported that there are over 15000 cancer-specific splicing variants in 27 types of human cancer, including BC [96]. For example, the BRCA1 gene has three main isoforms that include: BRCA1 full-length, BRCA1-11 complete skipping of exon 11, and BRCA1-11q partial skipping of exon 11. Compared to full-length BRCA1, BRCA1-11q has been shown to promote PARPi and cisplatin resistance both in vitro and in vivo. Furthermore, exon 11 mutation carriers have been reported to have a worse overall survival prognosis

compared to the non-carriers [97]. Another example could be that of tumor growth suppressor gene p53, which can be spliced as smaller isoforms. The splice variant lacking the first 40 amino acids competes with the wild-type p53, affecting its growth-suppressive function [98]. The apoptotic regulatory gene Bcl-x can also be spliced into two isoforms with opposite functions. The long isoform Bcl-x(L) suppresses apoptosis, while the short isoform Bcl-x(S) promotes it. Overexpression of Bcl-x(L) is associated with an increased metastasis risk in BC [99]. BRD4 gene plays a key role in providing epigenetic memory across cell divisions and transcription regulation by maintaining high-order chromatin structure and acetylation status [100]. BRD4 can also be spliced into a longer isoform known as BRD4-L, and a shorter one named BRD4-S. It has been shown that BRD4-L acts as a tumor-suppressive gene, while BRD4-S promotes BC proliferation and metastasis. BRD4-S interacts with the Engrailed-1 (EN1) homeobox transcription factor to activate ECM-associated matrisome network which creates the appropriate tumor microenvironment for cancer progression [101].

1.9 Aim

In this study, we aimed to identify an independent prognostic biomarker in Breast Cancer. To improve the accuracy of our findings, we examined the prognostic role of our gene through multiple microarray and RNA-seq BC datasets. Secondly, we attempted to identify specific associations of the biomarker with other potential interactors or biological pathways in order to investigate the mechanism by which it affects prognosis. Thirdly, we present a novel multiple isoform model for predicting the prognosis of BC patients.

Chapter 2

2. Methods

2.1 Breast Cancer RNA-Seq Datasets

The Swedish Cohort, GSE96058, Illumina HiSeq 2000 and MiSeq 500 expression data of 3273 Breast Cancer samples were downloaded from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). TCGA-BRCA Illumina HiSeq RNA Sequencing Level 2 isoform and gene expression data were downloaded from (firebrowse.org/) using Firehose portal. All the analysis reflect the normalized TPM expression values. The corresponding clinical data, including survival information, were downloaded from the same sources, respectively.

2.2 Breast Cancer Microarray Datasets

GSE1456, GSE2034, GSE2603, GSE3494, GSE4922, GSE6532, GSE7390, GSE11121, GSE12276, GSE19615, GSE20685, GSE21653, GSE25066, GSE37751, GSE58812 microarray datasets were downloaded from Gene Expression Omnibus and ArrayExpress (http://www.ncbi.nlm.nih.gov/geo),(https://www.ebi.ac.uk/arrayexpress/). The Metabric British cohort gene expression data were downloaded from Gene Expression Atlas (https://www.ebi.ac.uk/gxa/home). The datasets contain survival information including the survival status and the survival time until the latest follow-up. The raw expression data were normalized with RMA method using GeneSpring 12.0 (Agilent Technologies).

2.3 Log Rank with Multiple Cut-off (LRMC) Analysis

Each dot in the LRMC graphs represents the data point for one patient/tumor sample. The dots are plotted based on the expression cut-off value on the x-axis, versus its Log-Rank p-value on the y-axis. The red dots indicate an association with a bad prognosis, while the blue ones indicate an association with a good prognosis. The vertical lines represent the 1^{st}, 2^{nd}, and 3^{rd} quartiles, respectively. The dashed horizontal line is the

threshold for the 0.05 p value significance cut-off. The patients can be stratified into High- and Low- groups using the optimal cut-off values that pass the p-value significance threshold.

2.4 Survival Analysis

To test the prognostic association of potential biomarkers in BC datasets, univariate and multiple cox proportional hazard regression analyses, as well as the Kaplan-Meier curve generation were performed using the "survival" package in R programming language (https://cran.r-project.org/web/packages/survival/index.html). The visualization of the graphs was done using the "ggplot2" package in R programming language (https://cran.r-project.org/web/packages/ggplot2/index.html).

2.5 Statistical Analysis

The student T.Test analysis was used to compare the difference between the mean of two groups. Pearson Correlation analysis was used to test the statistical association between variables. Hazard Ratio Forest Plot was used to give a graphical representation of the association of metabolites with prognosis. All the analyses mentioned above were performed in GraphPad Prism v 6.01. ROC Curve was generated in a timely-dependent manner to estimate the prognostic accuracy of the Risk Score model using "survivalROC" package in R (https://cran.r-project.org/web/packages/survivalROC/index.html). Area under the curve (AUC) >= 6 was considered acceptable for prediction. The significant p-value threshold for all the analyses was < 0.05*. A p-value < 0.01 is denoted with two asterisk (**), p-value < 0.001 (***), and p-value < 0.0001 (****).

2.6 Probeset - Isoform alignment

The CELF2 Fasta sequence, and CELF2 transcript and exon position coordinates were downloaded in the GFF file format from the UCSC Genome Browser hg19 (https://genome.ucsc.edu/). The files were manually curated and then mapped against the reference genome. The Probe sequences were downloaded from Affymetrix (https://www.affymetrix.com/site/mainPage.affx). The probes were aligned against

CELF2 Fasta Sequence and plotted on the graph at the genomic position with a perfect match. The script used to build the graph was generated in R Programming Language (https://www.r-project.org/) (Special thank you to Dr. Murat Isbilen).

2.7 Risk Score Prognostic Model

The risk score model assigned to each patient in the dataset was calculated using the mathematical formula:

$$\text{Risk Score} = \sum_{i=1}^{n} coefficient_i * expression_i$$

The $coefficient_i$ represents the multiple cox regression hazard coefficient of LFPI and FPI variables. The $expression_i$ represents the TPM expression levels of the FPI and LFPI for each patient. By multiplying the two, a risk score is calculated and assigned to the patients.

$$\text{Risk Score of Patient X} = (coefficient_{(FPI)} * expression_{(FPI)}) + (coefficient_{(LFPI)} * expression_{(LFPI)})$$

Afterwards, based on the assigned Risk Score, the patients are divided into high-risk and low-risk groups using a suitable cut-off generated by LRMC analysis. The prognostic difference between the groups in OS is assessed by performing the Kaplan-Meier curve and the long-rank test.

2.8 CIBERSORT

CIBERSORT is a deconvolution algorithm used to describe the cell constitution of a tissue based on the gene expression data. Cibersort was used to characterize the TIL cell population content in GSE96058 (Basal) and GSE58812 (TNBC) tumor gene-expression datasets using a signature matrix of 547 genes. The expression data should be in non-log linear space, and contain no negative or missing value. Standard RNA-Seq TPM normalization method was used for the input file of GSE96058 dataset. The Affymetrix microarray dataset GSE58812 was RMA normalized. In both of the cases Cibersort Absolute method was employed. All the samples evaluated with the Cibersort metrics p.value < 0.05 were regarded to be as significant outcomes.

2.9 Immunohistochemistry

The immunohistochemistry analysis was performed at Hacettepe University with the help of Dr. Meral Uner Bugdayci. For the IHC analysis 110 formalin fixed, paraffin-embedded TNBC patients' tumor tissue sections of 4-5 microns were analyzed using Leica BOND-MAX automated IHC/ ISH stainer. The series included the rabbit polyclonal anti-CELF2 antibody HPA038513 and anti-CD3 monoclonal antibody unconjugated LN10. Lymphoid tissue in tonsil and appendix were used as positive control tissues for both CD3 and CELF2. Liver, pancreas, smooth muscle, adipose tissue and peripheral nerve bundles were used as the negative control tissues for CD3 and CELF2. For the detection systems DAB and AEC peroxidase chromogens were used. There were two scoring systems employed in the analysis: Expression Intensity Percentage (%), and Intensity score. An intensity score of 3, represents those tissues with more than two-thirds positive staining, and a score of 1 represents weakly positive staining. The total score of CELF2 staining in the neoplastic cells was calculated by multiplying the *Intensity Percentage * Intensity Score*.

Chapter 3

3. Results

3.1 CELF2 as a Candidate Prognostic Biomarker Gene in Breast Cancer

3.1.1 CELF2 Discovery Cohort

GSE37751 is a microarray dataset that contains metabolome and gene expression information derived from 61 frozen human breast tumor samples. This dataset served as our discovery cohort because it allowed us to make a more comprehensive analysis of a potential prognostic biomarker gene by possibly investigating its association with the metabolome and create hereby a full-scale study. Using a top-down approach, initially the 148 metabolites included in GSE37751 dataset ware screened for a potential association with the OS prognosis in BC. To do so, using Log-Rank Multiple Cut-off (LRMC) analysis, the samples were divided into High- and Low- groups according to the expression values of each of the metabolites separately (Figure 3.1). Using these categorical data, univariate cox regression analysis was performed to identify the corresponding hazard ratios of the metabolites related to the OS event (Figure 3.2). Among the prognostic metabolites group, Pearson Correlation analysis was used to obtain a set of metabolites that could exhibit possible association (Appendix D). The metabolites that were correlated to each-other and associated with the same prognosis direction in BC were selected. This final list contains eight carnitine metabolites and s-adenosylhomocystein (SAH) (Muhammad Waqas Akbar, Unpublished data). The set of nine metabolites have been previously shown to identify the TNBC subtype.

Based on LRMC analysis, the nine metabolites show an association with a bad prognosis in GSE37751 BC samples. SAH is the most highly expressed metabolite compared to the other ones. It displays significant expression cut-off values especially after the 2nd quartile. Regarding the carnitine metabolites, besides being associated with a bad prognosis, they also seem to be able to stratify the patients into High- and Low- groups at multiple significant cut-offs starting generally after the 1st quartile until the 3rd one (Figure 3.1). An

exception to this would be 2-methylbuturoylcarnitine, Hexanoylcarnitine, and Deoxycarnitine. These three metabolites are expressed within the same range and exhibit significant expression cut-offs even below the 1st quartile (Appendix A).

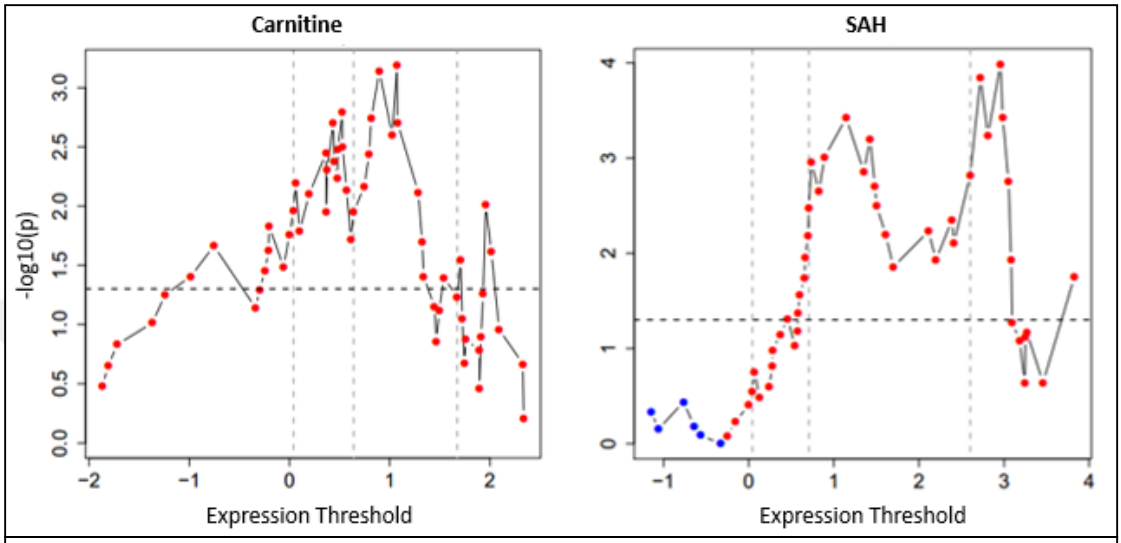


Figure 3.1: LRMC Plot of Carnitine and SAH metabolites in GSE37751 dataset.

Prognostic Metabolites	Exp(HR)	95% CI
butyrylcarnitine	1.58**	1.16-2.16
hexanoylcarnitine	1.56**	1.13-2.16
acetylcarnitine	1.61	0.96-2.7
3-dehydrocarnitine*	1.73**	1.22-2.47
propionylcarnitine	1.34*	1.01-1.78
2-methylbutyroylcarnitine	1.22	0.96-1.55
carnitine	1.94**	1.28-2.93
deoxycarnitine	1.67***	1.23-2.25
S-adenosylhomocysteine (SAH)	1.39**	1.08-1.79

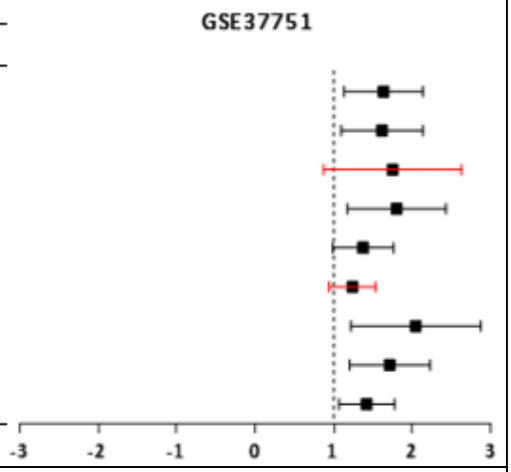


Figure 3.2: Forest Plot of the Univariate Cox Regression analysis of 9 prognostic metabolites.

The hazard ratio, which serves as a measure of the frequency of an event during survival analysis, is >1 for all the metabolites, confirming their association with a bad prognosis (Figure 3.2). According to that, carnitine seems to be associated with a more increased

risk of an event in the patients it is expressed in, as compared to the other metabolites. However, since it also has the largest confidence interval (CI), the precision of the estimate might not be the strongest of the metabolite set. Acetylcarnitine also has a relatively large CI, hence lowering the precision of estimates, which is also reflected in the p-value > 0.05 threshold. The SAH metabolite has the highest precision at predicting estimates with a narrow CI (Figure 3.2).

Table 3.1: Pearson Correlation between CELF2 gene and prognostic metabolites.

Metabolite	CELF2 Correlation Rho Coefficient
SAH	-0.42***
Carnitine	-0.42***
Butyrylcarnitine	-0.39**
2-methylbutyroylcarnitine	-0.33*
Deoxycarnitine	-0.31*
3dehydrocarnitine	-0.29*
Propionylcarnitine	-0.26*
Hexanoylcarnitine	-0.19
Acetylcarnitine	-0.15

The prognostic metabolites were then correlated with the gene probesets present in the GSE37751 discovery dataset. The CELF2 gene was selected based on two criteria. Firstly, compared to the other genes, it shows a significant correlation to 78% (7/9) of the prognostic metabolites, and displays negative direction in all of the cases. CELF2 has the strongest relation with two metabolites: Carnitine and SAH, with an equal rho= -0.42 and a significant adjusted p-value < 0.001 (Table 3.1).

The second reason CELF2 was selected is that its prognosis role has been significantly validated in the most datasets, as will be demonstrated in the following sections. We tested the potential prognostic function of the genes correlated to metabolites via LRMC and survival analysis using the OS information in GSE37751 dataset. The p-value cut-off was chosen in proximity to the 1st quartile based on the significance. According to the Kaplan-Meier (KM) plot, CELF2-High expression group of patients performs better at prognosis than CELF2-Low expression group (Figure 3.3). CELF2 gene is significantly a good prognostic biomarker of the OS in GSE37751 BC dataset.

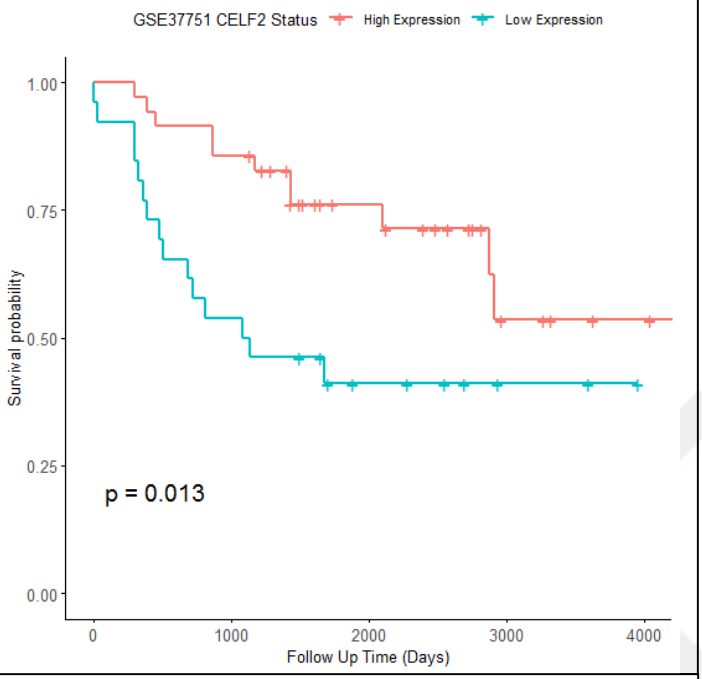


Figure 3.3: Kaplan-Meier OS Curve of CELF2 in GSE37751.

3.1.2 CELF2 Validation in RNA-Sequencing Breast Cancer Datasets

The CELF2 gene was initially validated using two BC RNA-Sequencing datasets. Upon clinical data availability, multiple cox proportional hazard regression was used to determine whether CELF2 can be classified as an independent prognostic factor in the discovery cohort as well as RNASeq cohorts. (Table 3.2). We included other clinically used prognostic factors such as age, node status, tumor grade, and stage in the analysis to test the independence of the CELF2 gene in predicting prognosis.

In the GSE37751 discovery microarray dataset, CELF2 has a significant (p-value = 0.026) HR=0.36 [CI: 0.15 - 0.88], making it an independent prognostic factor. In the Swedish Cohort, the CELF2 gene has a significant HR=0.77 [CI: 0.62-0.96], reinforcing two main concepts: CELF2 is a good prognostic gene (HR<1), and it can be used as an independent factor in predicting OS with precise estimates (narrow CI range, p-value=0.023).

Table 3.2: Multivariate Cox Regression Analysis of CELF2 in GSE37751, and GSE96058.

Multivariate Cox Regression Analysis			
Characteristics	**HR**	**95% CI**	**p-value**
GSE37751 (N = 61)			
CELF2 Gene	0.36	0.15 - 0.88	0.026*
Age	3.49	1.30 - 9.34	0.013*
Node Status	2.95	1.36 - 6.41	0.006**
Tumor Grade	2.10	1.06 - 4.16	0.034*
Stage Status	1.56	0.54 - 4.54	0.415
Swedish Cohort GSE96058 (N = 3273)			
CELF2 Gene	0.77	0.62 - 0.96	0.023*
Age	3.57	2.66 - 4.80	<0.001***
Node Status	1.38	1.10 - 1.72	0.005**
Tumor Grade	1.86	1.55 - 2.24	<0.001***

In the Swedish Cohort, we chose the cut-off expression value close to the sample median, and as expected, the CELF2-High expression group outperforms the CELF2-Low expression group in terms of OS. The same results can be noticed in TCGA-BRCA dataset, where CELF2 is significantly associated with a good prognosis based on the median cut-off expression value. In conclusion, we validated the CELF2 gene as an independent, good prognostic biomarker in the OS of BC samples in the GSE37751 microarray dataset, as well as two RNA-Seq datasets: Swedish Cohort (GSE96058), and TCGA-BRCA.

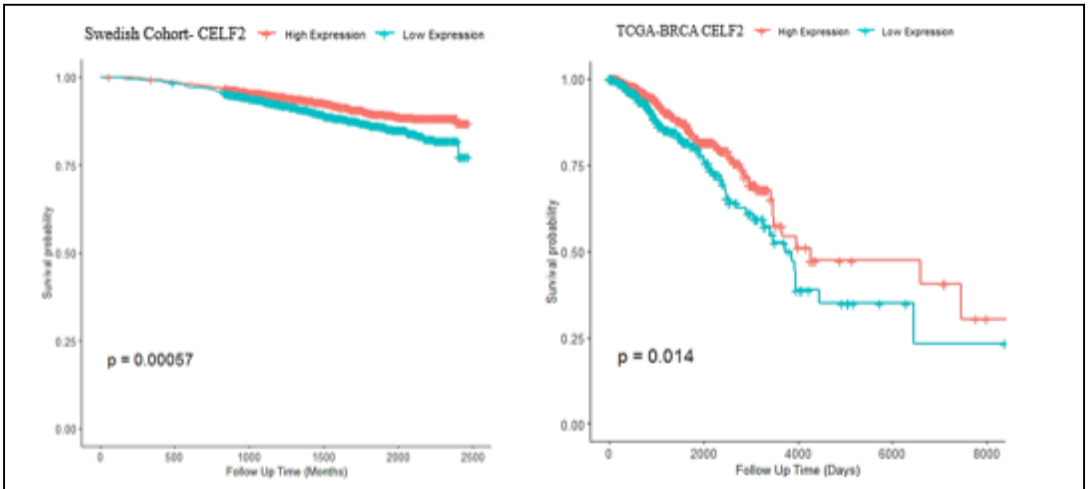


Figure 3.4: Kaplan-Meier OS Curves of CELF2 in the Swedish Cohort, and TCGA-BRCA.

3.1.3 CELF2 Validation in All-Subtypes Breast Cancer Microarray Datasets

We used multiple microarray datasets to assess the consistency of CELF2 gene in association with the direction (good or bad) of prognosis in BC. The microarray datasets contain gene expression information of several CELF2 probesets with the two most frequent ones being: 202157_s_at and 1554569_a_at. We performed LRMC analysis to test the association of CELF2-202157_s_at with prognosis using 13-thirteen microarray datasets from no specific BC subtype. 202157_s_at is significantly prognostic in 86% (19/22) of the cases out of a total of 22 tests. In 73% of cases (14/19) the probeset is associated with a good prognosis. In the remaining cases (27%), it exhibits a relationship with a bad prognosis (Table 3.3).

Table 3.3: The datasets used to test the CELF2-202157_s_at probeset association with prognosis in all BC subtypes. Asterisks (*) indicates significant cut-off p-value < 0.05. Tests performed include: OS=Overall Survival, DSS=Disease Specific Survival, RFS=Relapse Free Survival, MFS=Metastasis Free Survival, LMFS=Lung Metastasis Free Survival, BMFS= Brain Metastasis Free Survival, DFS= Disease Free Survival, DMFS= Distant Metastasis Free Survival, DRFS=Distant Relapse Free Survival.

Dataset All Subtypes	Present Probeset	Type of Event	Prognosis	Nr. of Tests
GSE1456	202157_s_at	OS*/ DSS*/ RFS	Good Prognosis	3
GSE2034	202157_s_at	DMFS	Good Prognosis	1
GSE2603	202157_s_at	MFS*/ LMFS*/ BMFS*	Bad Prognosis	3
GSE3494	202157_s_at	DSS*	Good Prognosis	1
GSE4922	202157_s_at	DFS*	Good Prognosis	1
GSE6532 U133	202157_s_at	DMFS*/ RFS*	Good Prognosis	2
GSE11121	202157_s_at	DMFS*	Bad Prognosis	1
GSE25066	202157_s_at	DRFS*	Bad Prognosis	1

Based on the information available in the microarray datasets platforms, we studied 1554569_a_at probeset in fewer tests than the previous probeset. 1554569_a_at is associated with a good prognosis in 100% (4/4) of the tests and is significantly prognostic in 44% (4/9) of them. Based on these findings, the CELF2 probeset 1554569_a_at is more consistent in its association with a good prognosis across different BC cohorts. In the datasets that included both 202157_s_at and 1554569_a_at, the former is significant in more cases and has higher expression levels (Table 3.4).

Table 3.4: The datasets used to test the association with prognosis of CELF2-202157_s_st and 1554569_a_at probesets in all BC subtypes.

Dataset All Subtypes	Present Probeset	Type of Event	Prognosis	Nr. of Tests
GSE1456	202157_s_at	OS*/ DSS*/ RFS	Good Prognosis	3
GSE2034	202157_s_at	DMFS	Good Prognosis	1
GSE2603	202157_s_at	MFS*/ LMFS*/ BMFS*	Bad Prognosis	3
GSE3494	202157_s_at	DSS*	Good Prognosis	1
GSE4922	202157_s_at	DFS*	Good Prognosis	1
GSE6532 U133	202157_s_at	DMFS*/ RFS*	Good Prognosis	2
GSE11121	202157_s_at	DMFS*	Bad Prognosis	1
GSE25066	202157_s_at	DRFS*	Bad Prognosis	1

As reflected on the x-axis of LRMC plots, the probeset 202157_s_at has an expression range of 6-to-10 log intensity, whereas 1554569_a_at probeset expression values vary between 2.2-to-6.2 (Figure 3.5). In GSE6532-plus2 dataset, 1554569_a_at has no significant cut-off values that pass the p-value < 0.05 threshold, however as the expression increases we can see an association with a good prognosis. The same conclusion can be drawn for GSE12276 dataset in the Bone MFS event. In the case of Brain MFS event of the same dataset, 1554569_a_at is associated with a bad prognosis at most cut-offs until the 3rd quartile, and as expression values increase, the trend shifts to a good prognosis. Probeset 202157_s_at, on the other hand, is expressed at higher levels and has a consistent association with a poor prognosis across all data points. Both probesets in the GSE19615 dataset show a good prognostic association. The 1554569_a_at probeset is significant and can be used to divide samples into High- and Low- groups based on the median, which corresponds to the second quartile (Figure 3.5). The 202157_s_at probeset produces significant cut-offs only at the beginning of the cohort, which may lead to biased survival analysis results due to the considerable difference in the number of participants in the High- and Low- groups.

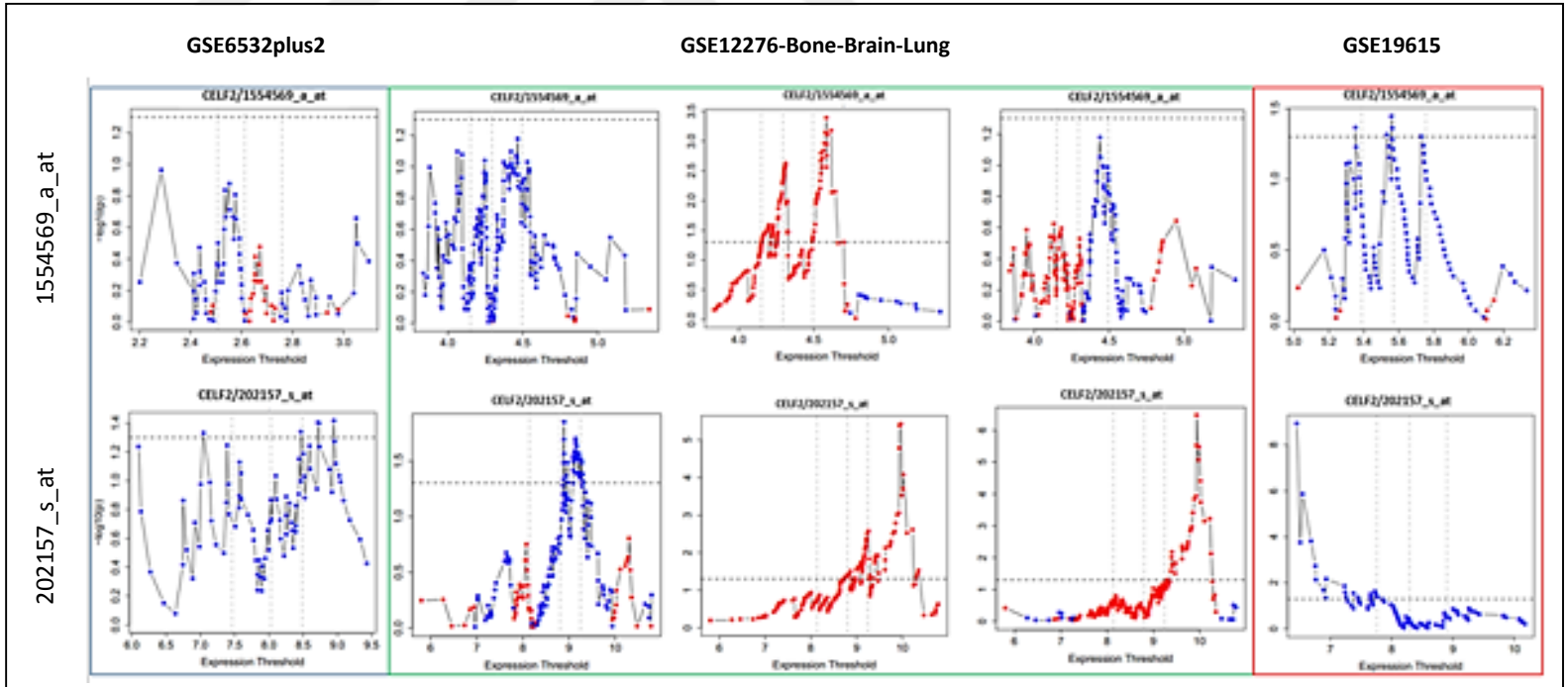


Figure 3.5: LRMC Plot of 1554569_a_at and 202157_s_at CELF2 probesets in GSE6532plus2, GSE12276, and GSE19615 microarray datasets.

3.1.4 CELF2 Validation in TNBC-Subtypes Breast Cancer Microarray Datasets

Because the group of the nine prognostic metabolites shows a significant correlation with CELF2 gene and can identify TNBC subtype, we tested the CELF2 prognostic role in TNBC specific samples (Table 3.5). We validated probeset 202157_s_at in a total of 22 tests, out of which it shows a significant prognostic role in 68% (15/22) of the cases and indicates a good prognosis in 67% (10/15) of them. In the remaining 33%, 202157_s_at is associated with a bad prognosis. On the other hand, we validated 1554569_a_at probeset in a total of 9 tests. It shows a significant association with prognosis in 89% (8/9) of the cases, and a 100% accuracy in the direction of a good prognosis. The expression range of 202157_s_at lies within the range of 7.5-to-12 log intensity ratio, while that of 1554569_a_at varies between 3.5-to-6.6 (Table 3.5).

Table 3.5: The datasets used to test the association with prognosis of CELF2-202157_s_at and 1554569_a_at probesets in TNBC subtype.

TNBC Dataset	Present Probeset	Type of Event	Prognosis	Nr. of Tests
GSE1456	202157_s_at	OS*/ DSS/ RFS	Good Prognosis	3
GSE2034	202157_s_at	DMFS	Bad Prognosis	1
GSE2603	202157_s_at	MFS*/ LMFS/ BMFS*	Good Prognosis	3
GSE3494	202157_s_at	DSS	Bad Prognosis	1
GSE4922	202157_s_at	DFS*	Bad Prognosis	1
GSE6532 U133	202157_s_at	DMFS*	Good Prognosis	1
GSE11121	202157_s_at	TDMFS	Good Prognosis	1
GSE19615	202157_s_at	DRFS*	Good Prognosis	1
GSE21653	202157_s_at	DFS*	Good Prognosis	1
GSE25066	202157_s_at	DRFS	Good Prognosis	1
GSE20685	1554569_a_at	MFS*/ OS*	Good Prognosis	2
GSE58812	202157_s_at	MFS*/ OS*	Good Prognosis	2
GSE58812	1554569_a_at	MFS*/ OS*	Good Prognosis	2
GSE6532 plus 2	202157_s_at	DMFS*/ RFS*	Good Prognosis	2
GSE6532 plus 2	1554569_a_at	DMFS*/ RFS*	Good Prognosis	2
GSE12276	202157_s_at	Bone*-Brain*-Lung*-MFS/ MFS*	Bad-Bad-Bad-Bad Prognosis	4
GSE12276	1554569_a_at	Bone*-Brain-Lung*-MFS	Good-Good-Good	3

In both analyses: All-BC-Subtypes and TNBC, probeset 202157_s_at is expressed at higher levels than 1554569_a_at. 202157_s_at is significant in more tests than 1554569_a_at in all BC-subtypes, but it is less accurate in showing a consistent association with the direction of a good prognosis. In TNBC, 1554569_a_at is significant in more tests and also

has a higher accuracy (100%) of being associated with a good prognosis compared to 202157_s_at.

3.2 CELF2 In-vitro Validation as a Good Prognostic Biomarker in TNBC.

Immunohistochemistry (IHC) experiments were performed to analyze the protein levels of CELF2 gene in 110 TNBC patient samples. Considering that in the in-silico findings TNBC conferred the highest CELF2 accuracy rate in the prediction of a good prognosis, we decided to investigate if CELF2-Protein levels would exhibit the same function in this subtype. Prior to this analysis, using the available information on Human Protein Atlas (HPA), protein expression levels of CELF2 gene in TCGA-BRCA cohort were examined [102]. Based on HPA, CELF2 protein has a low-to-no intensity in BC tissues. On the overall scale however, CELF2 protein was detected to have a nuclear positivity in stromal inflammatory cells. Enrichment of CELF2 in T lymphocytes was also observed in GSE36133 CCLE dataset, where both CELF2 probesets were most highly expressed in T-cell lymphoma and different types of Leukemia (Appendix C).

Considering the fact that tumor microenvironment (TME) infiltrated immune cells have been shown to play major role in cancer, we decided to try and determine if the prognostic effect of CELF2 gene was mediated through T-cell activity. To ensure that CELF2 conveys different prognostic information than other T-cell markers, we observed the overlap between infiltrated T-cells and CELF2 protein expression during IHC by double-staining against the CD3 specific T-cell marker.

3.2.1 CELF2 Immunohistochemistry Analysis

Based on the pathological evaluation, CELF2 protein intensity levels were measured using the numerical grading system ranging from 1-to-3, where 1 indicates a low intensity and 3 a high intensity of expression. Overall, CELF2 protein did not test positive in a high number of neoplasm in the tissue sections. There were in total 60 samples with an interpretable quality on neoplastic cells. CELF2 protein exhibited no intensity signal in

62% (37/60) of these samples. Out of the remaining ones, 31% (19/60) were low intensity, scale 1, and 7% (4/60) depicted a moderate intensity, scale 2.

To assess CELF2 protein expression in the TME inflammatory cells, a percentage % - based scale was used. There are three main divisions which reflect approximately in what percentage of stromal inflammatory cells CELF2 protein is being expressed. The percentage of CELF2 protein infiltration in the stromal inflammatory cells range from [1:25] in 40% of the tissues, [26:75] in 49%, and [>75] in 11% of the tissues. The percentage of CD3-positive lymphocytes in the inflammatory cells of tumor stroma ranges from [1:25], [26:75], [>75], in 52%, 45%, 3% of the tissues, respectively.

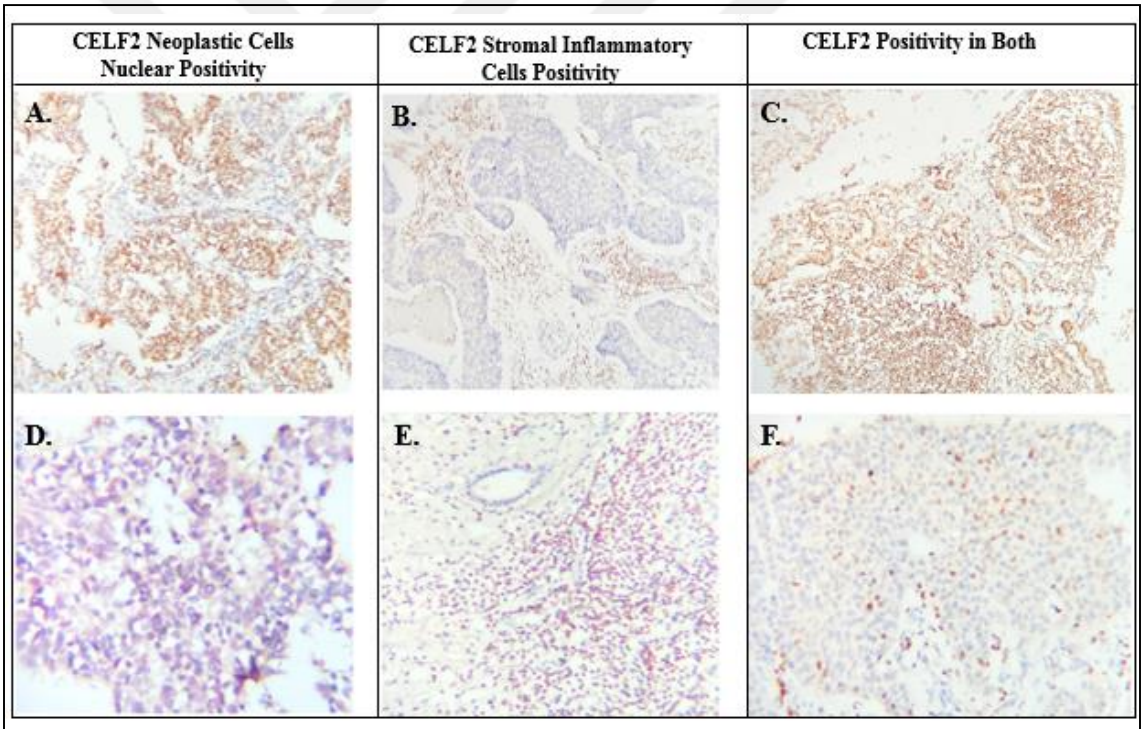


Figure 3.6: CELF2 Immunohistochemistry Results. A & D represent the tissue sections that tested positive for CELF2 in neoplastic/tumor cells; B & E represent the sections that tested positive for CELF2 in the stromal inflammatory cells; C & F represent two of the samples that tested positive for CELF2 in both: tumor cells and stromal inflammatory cells.

We performed LRMC analysis to validate the in-silico findings supporting the good prognostic role of CELF2 gene, using the clinical data of the TNBC patient samples and intensity percentage values. The results did not support the association of any of the tested parameters, CELF2/CD3 stromal infiltration or neoplastic CELF2 intensity, with prognosis. Across all the tests, there is a consistent lack of a significant cut-off value that could be used to stratify patient samples into High- and Low- groups (Fig 3.7 A, B, C). The low number of data points depicted are more heterogeneous in their association with the prognosis, and no clear trend of the relationship between CELF2 and OS can be detected. CELF2 protein levels have a lower intensity-percentage range in neoplastic cells, and are present only at the end of the cohort after the 3rd quartile (Figure 3.7, C). Despite previously being established in the literature, we were unable to validate any prognosis association of the T-cells CD3 marker as well (Figure 3.7, A).

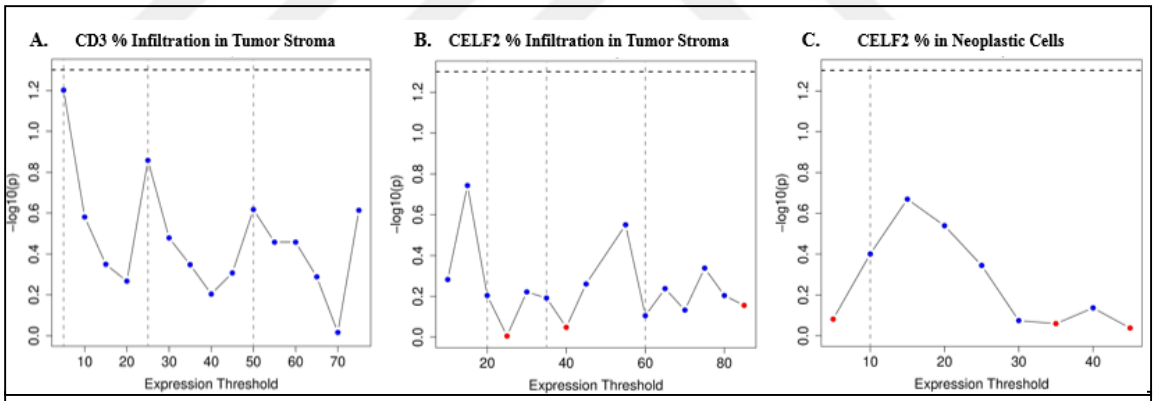


Figure 3.7: LRMC Plot of CELF2 and CD3 intensity percentage values in IHC.

3.2.2 CELF2 and CD3 Dynamics in Immunohistochemistry Validation

To have a better understanding and investigate whether CELF2 expression would consistently overlap with that of CD3 in the tumor stroma tissue sections, paired Pearson Correlation Analysis was performed. As expected, the total score of CELF2-protein expression neoplastic cells does not show a significant correlation with CD3 stromal infiltration. Tissue samples that test negative for neoplasm CELF2 are still noticed to have progressive levels of stromal CD3 (Figure 3.8: A). Instead, there is a significant moderate

correlation (r = 0.58) between CELF2 stromal infiltration and CD3. According to the trend however, there are samples near or on the x-axis that show CELF2 expression but display low or no CD3 expression. The data suggests that CD3 and CELF2 cannot be used interchangeably because they do not convey the same information.

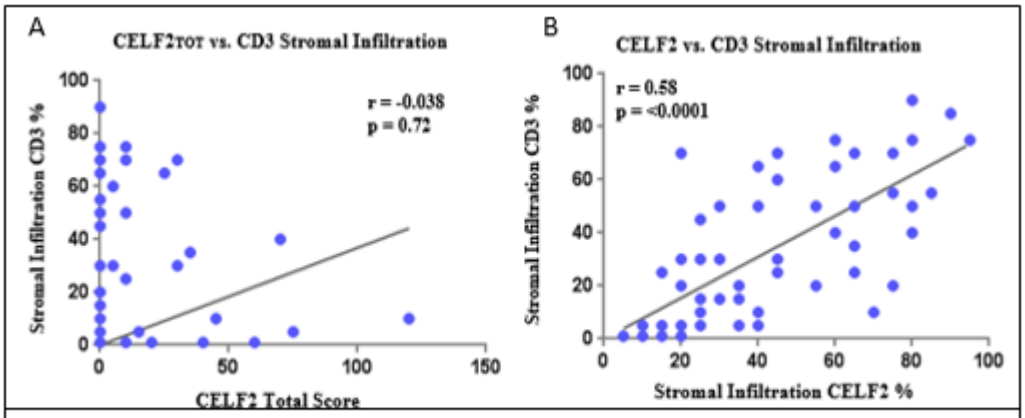


Figure 3.8: Pearson Correlation Analysis of CELF2 and CD3 Intensity Levels.

3.2.3 CELF2 and Infiltrated Immune cells Dynamics in Other Breast Cancer Cohorts

We used previously discussed Swedish cohort-TNBC subtype and GSE58812-TNBC cohort to investigate the extent of behavioral similarities between CELF2 and T-cells *in-silico*. The patients were divided into CELF2-High and CELF2-Low groups based on transcript expression levels, and the difference in CD4 and CD8 T-cell levels, as well as other immune cell types, was assessed between the groups (Appendix B).

In the GSE58812, CD4 naive cells are significantly more abundant in CELF2-High group than they are in CELF2-Low group, p.value < 0.0001 (Figure 3.9: B). In the Swedish Cohort, the same pattern can be observed. However, based on the t-test results, the difference in the CD4 naive cell number between the groups is not significant (Figure 3.10: A). CD8 T-cells abundance shows a significant positive association with higher levels in CELF2-High group in both of the cohorts (Figure 3.9: C&D).

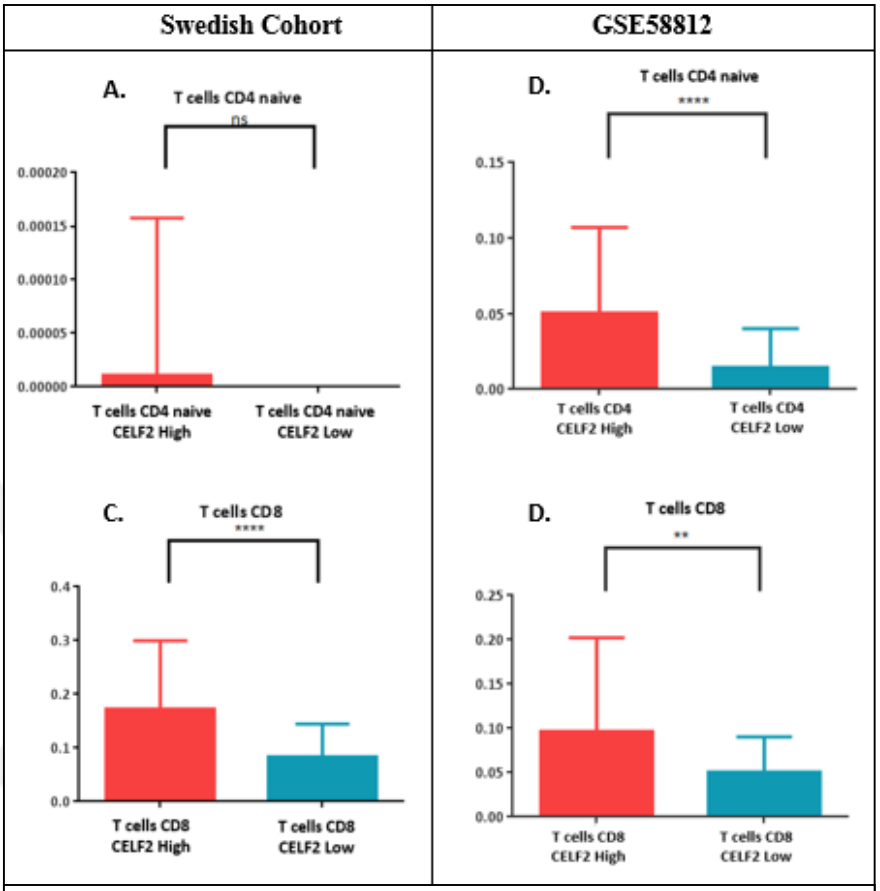


Figure 3.9: Student T-Test results of differences in CD4 and CD8 T cell abundance between the CELF2-High and CELF2-Low groups in the Swedish Cohort (TNBC-Subtype) and GSE58812 (TNBC-Subtype).

Given that the primary goal of this study is to evaluate and validate the role of the CELF2 gene as an independent prognostic biomarker in BC, next we investigated whether the CELF2 prognostic function is dependent on the abundance of CD8+ T-cells in TNBC patients. To do so, we divided the Swedish Cohort – TNBC subtypes (n=339) and GSE58812 – TNBC (n=107) according to the median of the cell abundance into “CD8-High” and “CD8-Low” groups. Afterwards, we assessed the CELF2 prognostic function.

Previously, in Figure 3.9C, we have shown that in the TNBC samples of the Swedish cohort, T cells-CD8 were significantly more abundant in CELF2-High Expression group. Here we show the association of CELF2 with prognosis in CD8-High versus CD8-Low groups in the

same dataset. CELF2 gene is associated with a good prognosis in both of the groups (Figure 3.10).

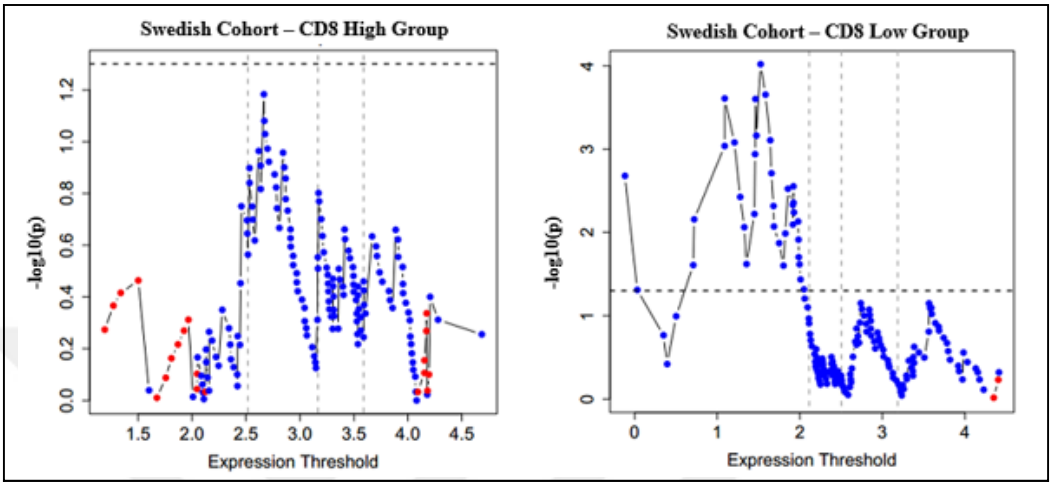


Figure 3.10: LRMC Plot of CELF2 in CD8-High and CD8-Low Groups in Swedish Cohort**.**

In the CD8-High, CELF2 does not present any significant expression cut-off that could be used to stratify the samples for survival analysis. However, in CD8-Low group, not only does CELF2 conserve its association with a good prognosis, but it also displays significant cut-off values (Figure3.10).

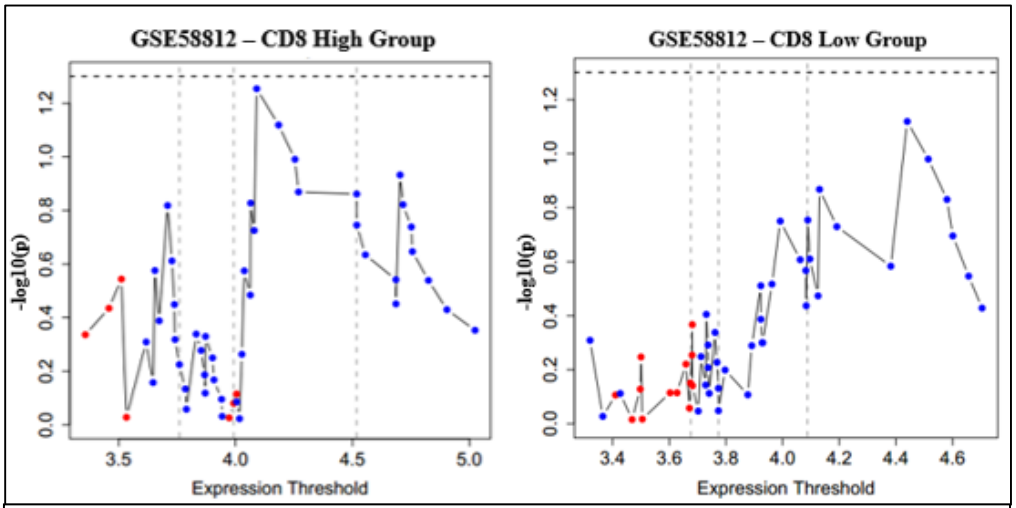


Figure 3.11: LRMC Plot of CELF2 in CD8-High and CD8-Low Groups in GSE58812.

Despite the lack of a significant cut-off expression value in the GSE58812-TNBC dataset, CELF2 is associated with a good prognosis in the majority of the samples, both in CD8-High and CD8-Low. While statistically there is little room for discussion, it can be seen in both cases that as CELF2 expression increases along the x-axis, the association with a good prognosis becomes more prominent (Figure 3.11). In conclusion, while CELF2 expression levels are related to the abundance of CD8 infiltrated immune cells in TME, its prognostic function remains independent.

3.3 CELF2 Isoform Specific Prognostic Signature

One of the reason that motivated us to do an investigation on a CELF2 isoform-specific level was the constant differences regarding the direction of prognosis the CELF2 202157_s_at and 1554569_a_at probesets were associated with. Considering the polyclonal nature of the antibody used during IHC experiment, we also thought that studying CELF2 isoforms individually would be necessary to better understand the in vitro validation results.

3.3.1 Probeset-Specific Isoforms of CELF2 Gene

We have shown that in all-BC-subtypes and TNBC-specific analyses, the CELF2 1554569_s_at probeset is associated with a good prognosis in a higher number of cases and more consistently than 202157_s_at. The HPA03851 Polyclonal Antibody that was used during IHC validation could recognize both of these probesets. Hence, we suspected that a possible underlying cause for the inconclusive CELF2 results during IHC validation, could be the difference probeset 202157_s_at and 1554569_a_at display in their association with the direction of prognosis. We first identified the CELF2 transcripts that these two probesets can recognize. For the Sequence Analysis, we used CELF2 UCSC isoform annotations. As indicated in Figure 3.12, there are in total thirteen (13) CELF2 isoforms. The 202157_s_at probeset can recognize 5/13 CELF2 isoforms that the other probeset, 1554569_a_at, cannot. 1554569_a_at recognizes four isoforms in total: uc001iki.4, uc001ikk.2, uc010qbl.1, and uc010qbp.1. The probeset 202157_s_at can recognize a total of 9 isoforms.

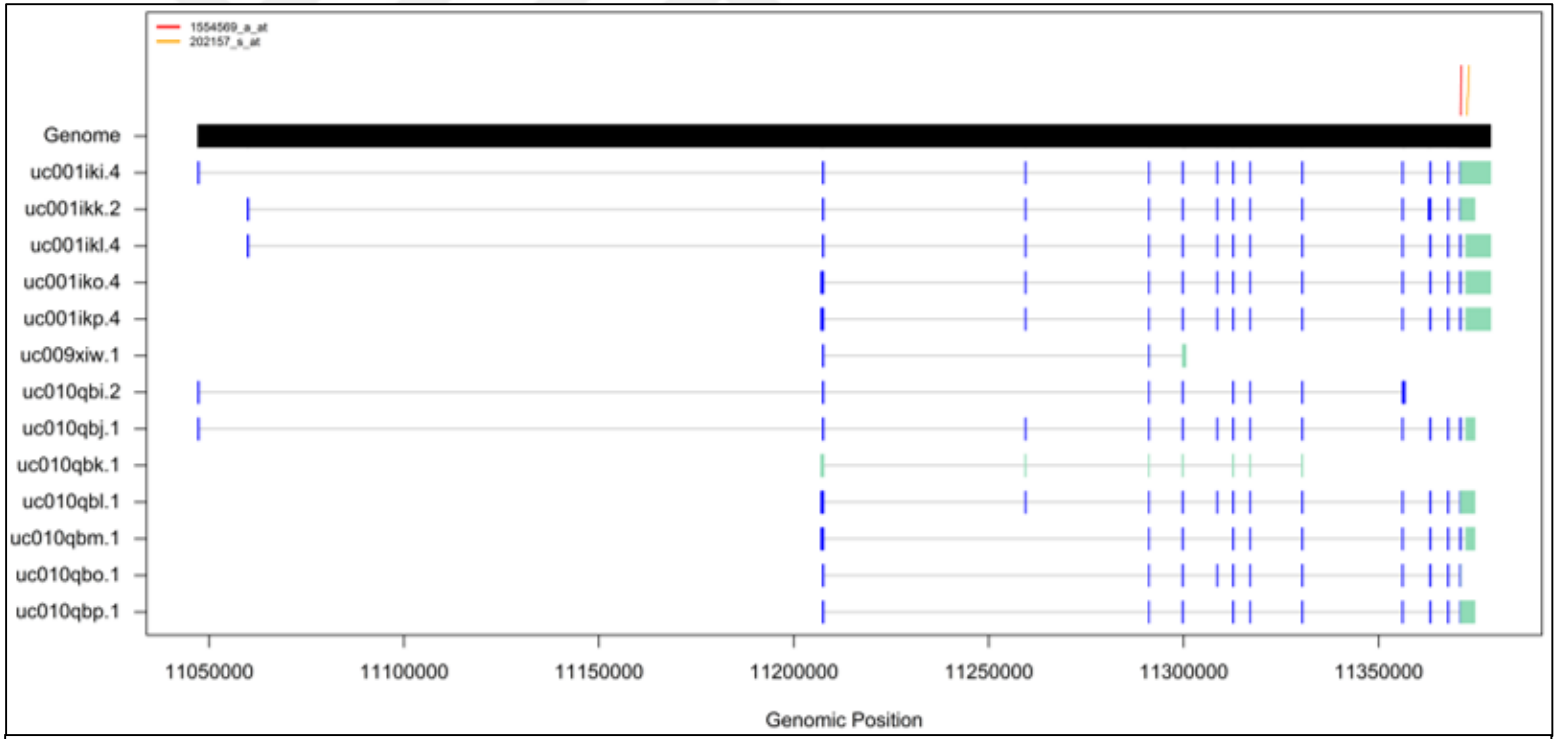


Figure 3.12: The graphs depicts CELF2 isoforms aligned to the reference genome (in black). The exonic regions are depicted in blue color. The intronic regions are depicted in gray, and the 3'-Untranslated Regions in light green. 202157_s_at CELF2 probeset is shown in orange color, while 1554569_a_at is shown in red color.

3.3.2 The prognostic role of CELF2 Isoforms

We have shown that CELF2-202157_s_at and CELF2-1554569_a_at probesets can recognize different isoforms. To test their association with prognosis in BC, we decided to test each of the isoforms individually initially using LRMC analysis. Throughout the study, TCGA-BRCA-Isoform expression and OS data were used. According to the LRMC results, not all of the CELF2 isoforms have a significant association with prognosis. There are in total 4 isoforms that demonstrate a significant association with a good prognosis: Uc001iki, Uc001iko, Uc009xiw, and Uc010qbi. Compared to the others, Uc001iko and Uc001iki are the most highly expressed isoforms in TPM scale (Figure 3.13).

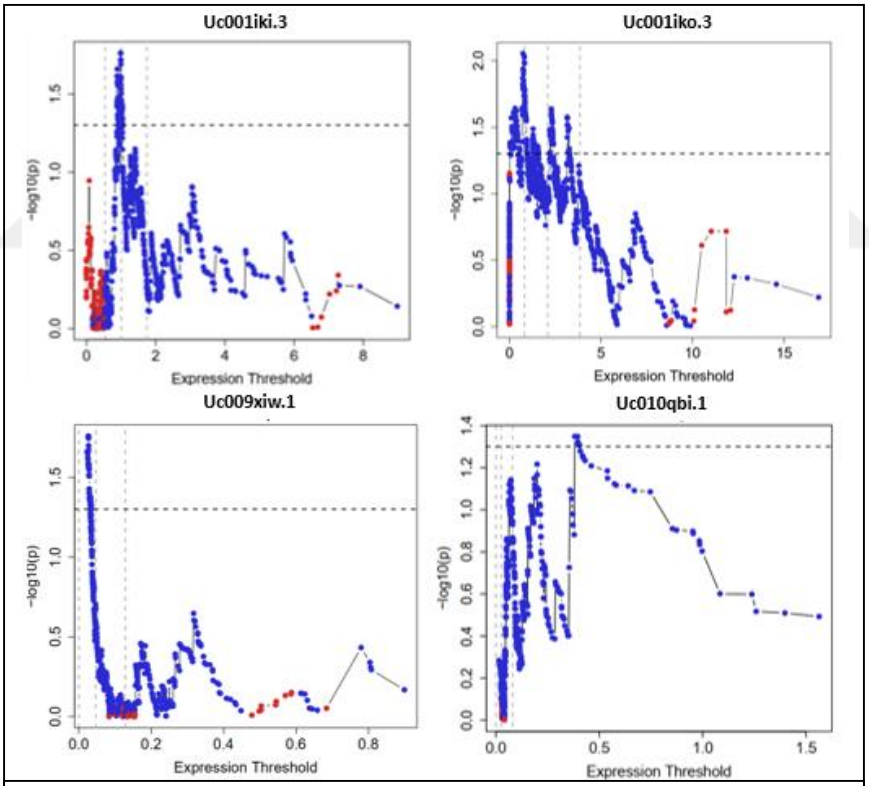


Figure 3.13: LRMC Plot of CELF2 Good Prognostic Isoforms.

On the other hand, three isoforms: Uc001ikl, Uc010qbk.1, and Uc010qbo, display an association with a less-favorable prognosis. This relationship is noticed toward the end of the cohort spanning the 3rd quartile. Compared to the other two isoforms, Uc001ikl, has higher TPM expression levels throughout TCGA-BRCA cohort samples (Figure 3.14).

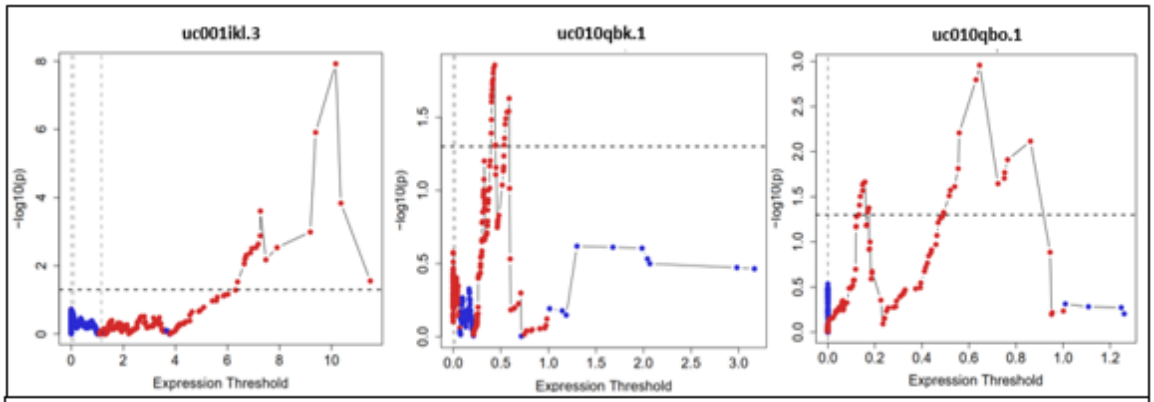


Figure 3.14: LRMC analysis of CELF2 isoforms associated with a less-favorable prognosis.

In the univariate cox regression analysis, the hazard ratio for Uc001iki, Uc001iko, and Uc009xiw isoforms is less than one, indicating an association with a good prognosis. The hazard ratio for Uc001ikl, Uc010qbk, and Uc010qbo is greater than one, suggesting an increased risk for the OS event and a less favorable prognosis (Table 3.6).

Table 3.6: Univariate Cox Regression analysis of CELF2 Isoforms. The probeset that can recognize each isoform is denoted in the last column.

Univariate Cox Regression Analysis						
CELF2 Isoforms	HR	95% CI	p-value	Prognosis	Average Expression(TPM)	Matching Probeset
TCGA-BRCA (N = 992)						
Ucoo1iki	0.61	0.40 - 0.95	0.027	Good Prognosis	1.382	1554569_a_at 202157_s_at
Uc001iko	0.58	0.37 - 0.91	0.018	Good Prognosis	2.673	202157_s_at
Uc009xiw	0.65	0.43 - 0.10	0.049	Good Prognosis	0.090	--
Uc010qbi	0.05	0.00 – 4.47	0.187	Good Prognosis	0.075	--
Uc001ikl	5.39	1.96 - 14.80	0.001	Bad Prognosis	0.885	202157_s_at
Uc010qbk	2.31	1.16 - 4.62	0.018	Bad Prognosis	0.089	--
Uc010qbo	4.42	1.61 - 12.15	0.004	Bad Prognosis	0.036	--

3.4 CELF2 Multiple Isoform Prognostic Signature

We build the Risk Score model based on the most prominent CELF2 isoforms: Uc001iki, Uc001iko, and Uc001ikl. We performed Pearson Correlation analysis to test the independence of the isoforms associated with opposite prognoses. According to the results, the good prognostic isoforms, Uc001iki and Uc001iko, show a moderate degree of correlation to each other (rho = 0.52). However, Uc001ikl, the isoform that relates to a less-favorable prognosis upon increased expression levels, does not show a strong correlation to either of the two good prognostic isoforms (Figure 3.15). Based on the lack of a strong degree association between the good prognostic isoforms Uc001iki and Uc001iko, to the less-favorable prognostic isoform Uc001ikl, we decided to implement them on the prognostic model.

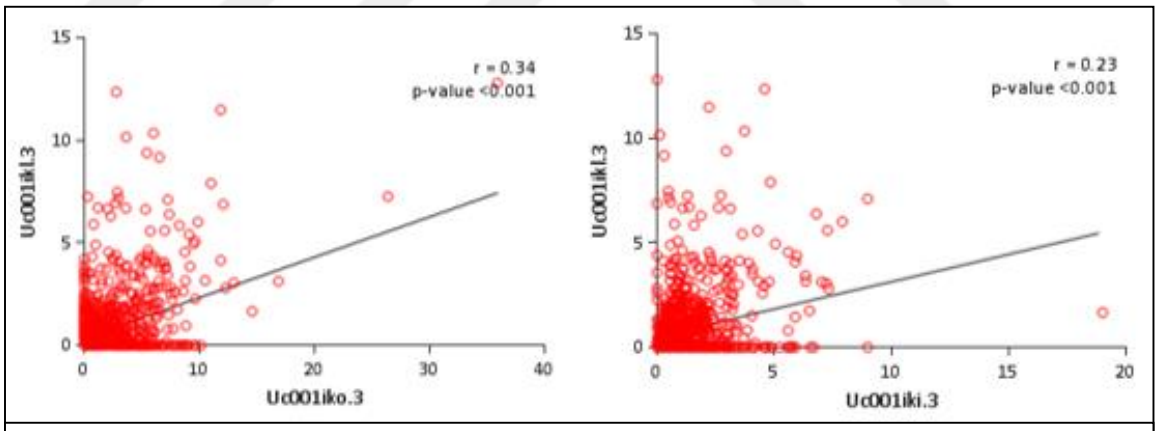


Figure 3.15: Pearson Correlation analysis between Uc001ikl.3 and Uc001iki.3-Uc001iko.3

3.4.1 The Prognostic Risk Score Model

We built the Risk Score prognostic model based on two new variables. **FPI**, **F**avorable-**P**rognostic **I**soforms, is composed of the summation of Uc001iki and Uc001iko good prognostic isoforms. **LFPI**, **L**ess-**F**avorable **P**rognostic **I**soform reflects Uc001ikl isoform

expression levels. Overall, we assigned a FPI and a LFPI value to each patient in TCGA-BRCA cohort (n=991), from which we calculated the Risk Score.

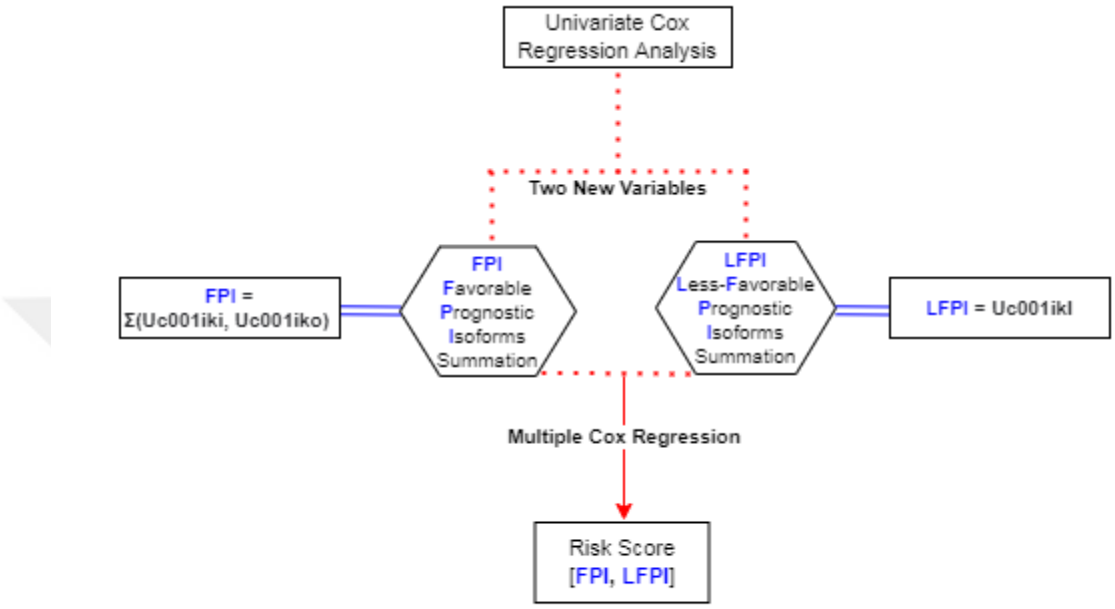


Figure 3.16: The Prognostic Risk Score Model Generation Pipeline.

As displayed on the general pipeline, we tested the FPI and LFPI for their relationship to prognosis using both univariate and multiple cox proportional hazards regression (Figure 3.16). As expected, FPI is related to good prognosis showing significant hazard ratio of 0.59 and 0.55, in univariate and multiple regression analysis, respectively. LFPI is related to less-favorable prognosis as its hazard ratios stand larger than > 1 at 4.70 and 6.52 in univariate and multiple cox regression analysis in that order (Table 3.7).

Table 3.7: Univariate and Multiple Cox Regression of FPI and LFPI in TCGA-BRCA Cohort.

	Univariate Cox Regression			Multiple Cox Regression		
Variables	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
	TCGA-BRCA (N = 992)					
FPI	0.59	0.38 - 0.91	0.017*	0.55	0.35 - 0.85	0.008
LFPI	4.70	1.5 - 15	0.009**	6.52	1.99 - 21.39	0.002

LRMC plots of FPI and LFPI support the previous results. FPI is related to a good prognosis in the majority of the samples, showing a strong cut-off close to the 50th quartile of the entire cohort. LFPI, as it has been shown previously for isoform Uc001ikl, reflects a relationship with bad prognosis after the third quartile of TCGA-BRCA cohort which is represented by the samples that have the highest expression levels of the isoform (Figure 3.17).

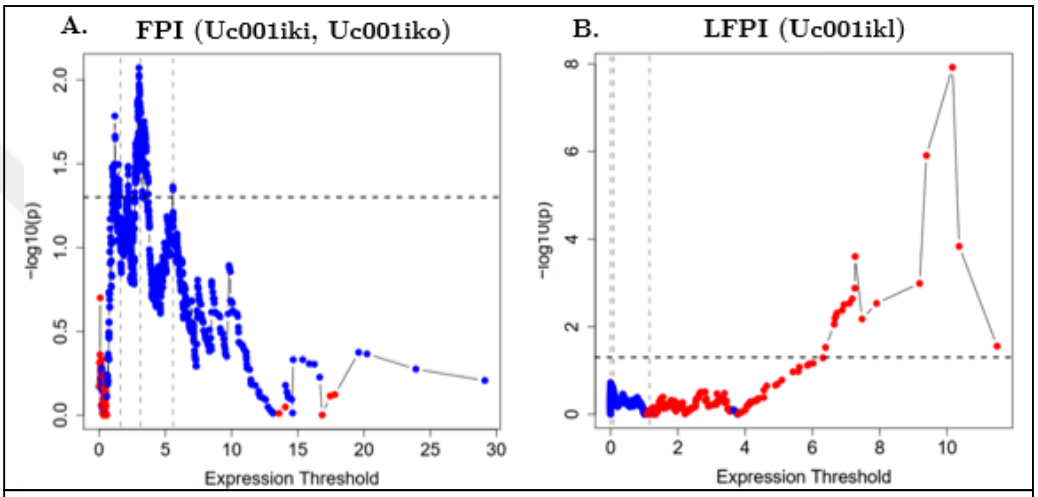


Figure 3.17: LRMC Plot of FPI and LFPI in TCGA-BRCA.

Afterwards, we stratified the patients into High- and Low- FPI and LFPI groups to perform survival analysis. The FPI-High group performs better at prognosis compared to the FPI-Low group. The opposite trend can be noticed in the other case, where LFPI-High group performs significantly worse at prognosis compared to LFPI-Low group. Considering that the less-favorable prognostic effects of Uc001ikl isoform are more pronounced at high expression levels, the stratification cut-off is skewed from the median of the sample to the right (Figure 3.18).

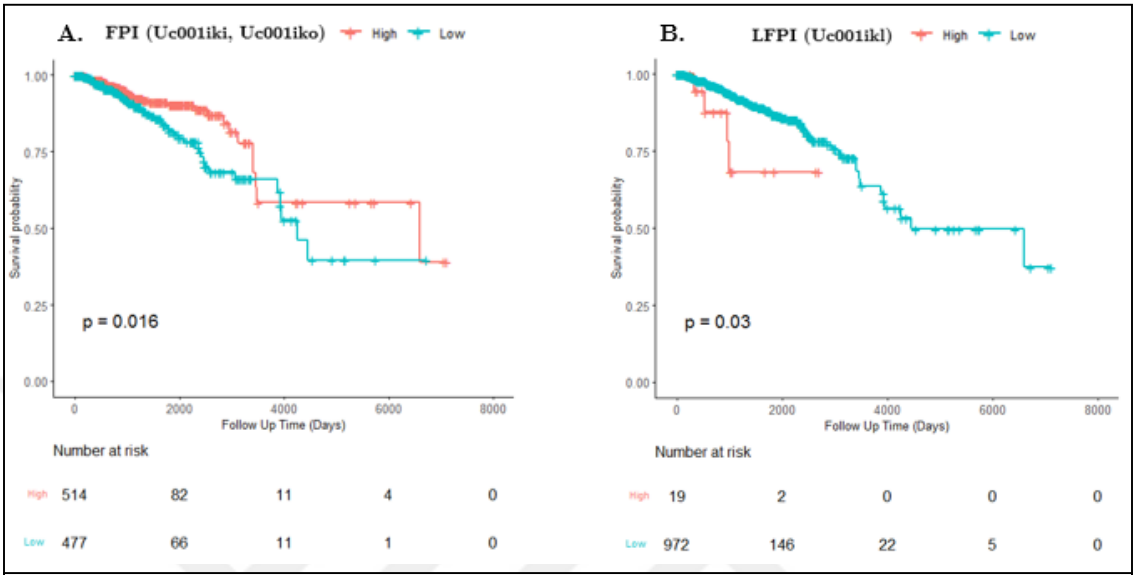


Figure 3.18: Kaplan-Meier OS Curve of FPI and LFPI in TCGA-BRCA.

3.4.2 The Evaluation of the Prognostic Risk Score Model Performance

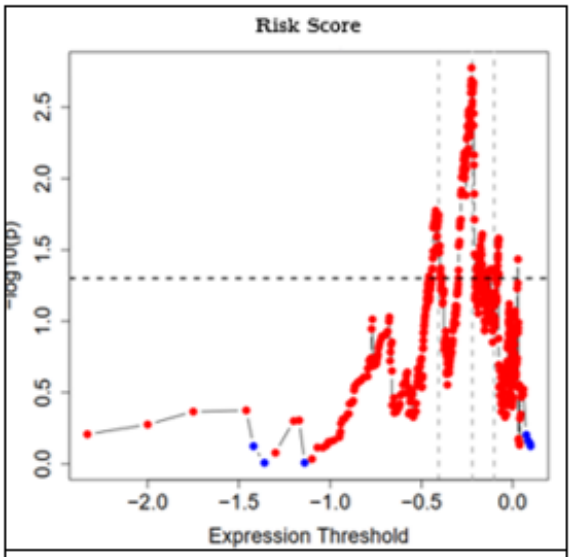


Figure 3.19: LRMC analysis of the Risk Score model.

Having checked the validity of the two created variables in terms of prognosis and significance, we then build the Risk Score model to predict prognosis in TCGA-BRCA cohort. The LRMC plot of the Risk Score model shows a well-distributed pool of significant cut-offs between the first and third quartiles for further stratification of the samples. The increase in Risk Score is related to a bad prognosis in a progressive trend with respect to significance (Figure 3.19)

In survival analysis, we divided the patients at a cut-off close to the overall sample median. The Kaplan Meier plot demonstrates that patients classified in the High-Risk-Score group perform significantly worse at prognosis, compared to those classified in the Low-Risk-Score group. The trend holds true and can be observed throughout different periods in the follow-up time (Figure 3.20A). To measure the validity of the Risk-Score model at predicting the 5-year survival rate in BC patients, we used the Receiver Operating Characteristic (ROC) curve. The AUC measurement of significance, being > 0.5, indicates that the model is able to perform beyond a considerable threshold and that the output is not generated randomly (Figure 3.20B).

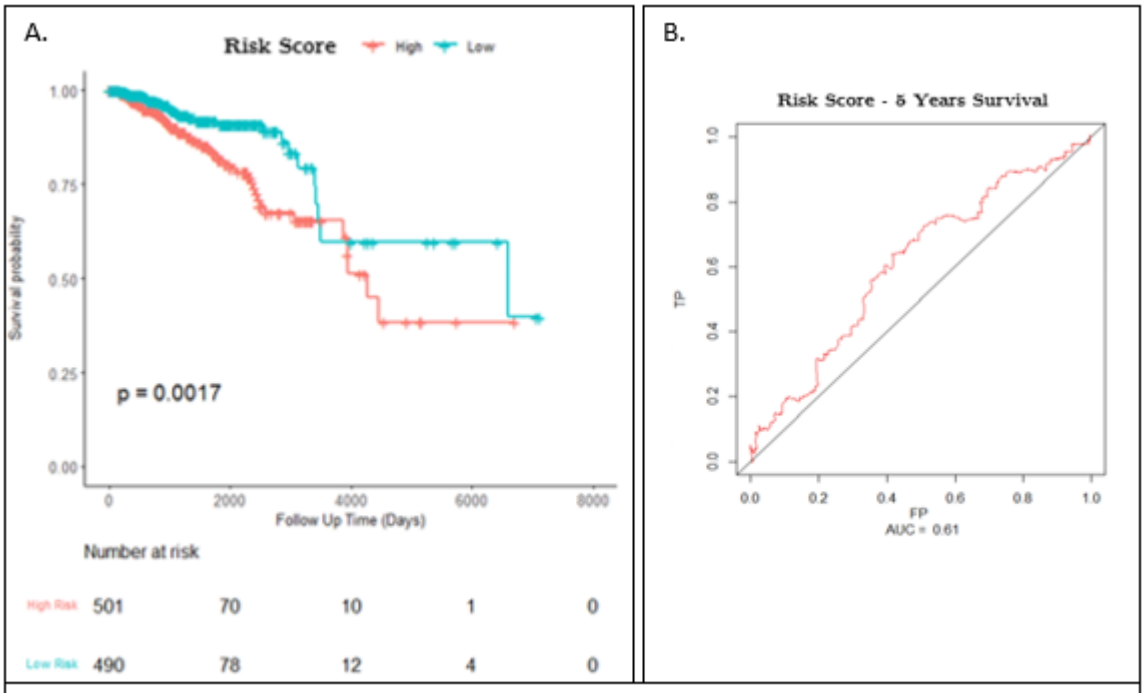


Figure 3.20: Kaplan-Meier plot of the Risk Score model (A). ROC curve of the Risk Score model (B).

3.4.3 The Protein-Isoforms of CELF2 Variants included in the Risk Score Prognostic Model

According to the results, regardless of the prognoses direction association, these three protein isoforms of CELF2 include all the RNA Recognition Motifs (RRM) in their N-terminal and C-domain. Uc001ikl protein isoform has a greater length of 521 amino acids (a.a), followed by Uc001iki and Uc001iko with respective lengths of 508 a.a and 488 a.a.

The linker domain, also known as divergent domain, shows amino acid alterations denoted on the black line. Compared to Uc001iko, the other two protein isoforms include a set of additional a.a on their N-terminal (Figure 3.21).

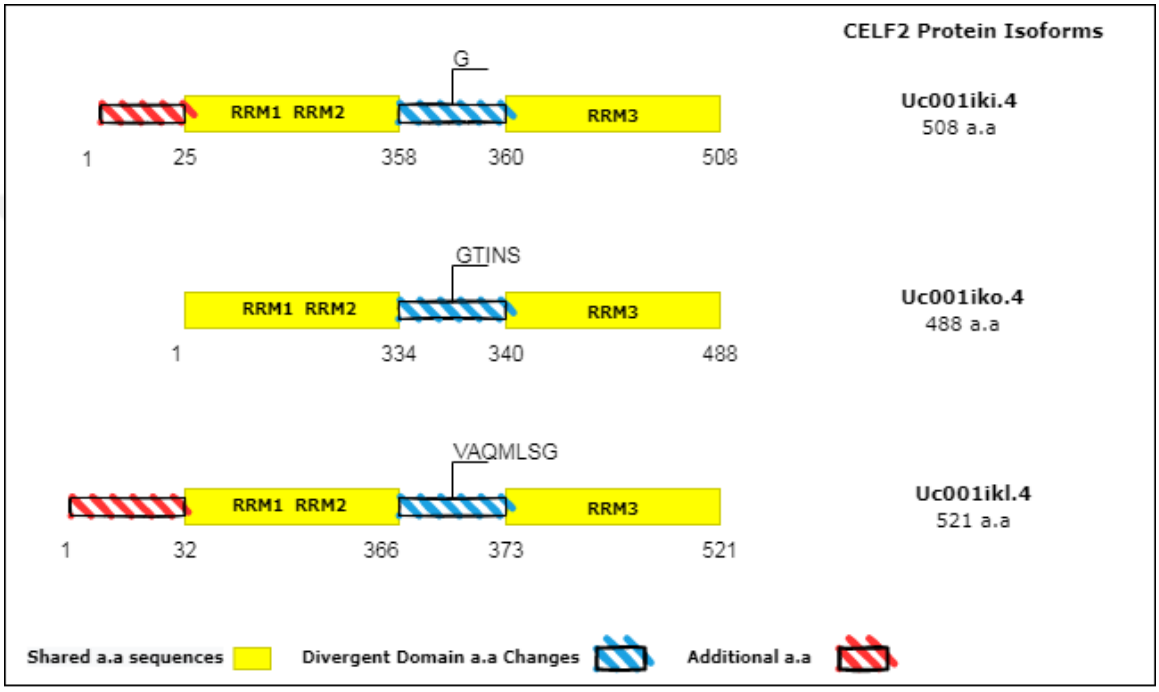


Figure 3.21: Protein Domain Structure of the CELF2 Isoforms included in the Prognostic Model.

Chapter 4

4. Discussion

Prognostic signatures in breast cancer (BC) are used to guide clinicians in their treatment-decision process. These tools provide complementary information to make patient-specific decisions, in addition to the traditional factors such as tumor size, lymph node status, receptor positivity, and so on. A plethora of Multiple Gene Prognostic Signatures has been developed to predict the outcome of BC. Despite this, only a small number of these tests have been approved for use in clinics. We focused on finding more comprehensive BC prognostic biomarkers in this study.

4.1 The Prognostic Function of CELF2 Gene

4.1.1 Discovery Cohort of CELF2 and Prognostic Metabolites

As previously stated, our primary goal was to identify a prognostic biomarker gene in BC with high prediction accuracy, as well as to provide a comprehensive analysis that would elucidate its biological function. In line with this notion, we attempted to identify other co-occurring factors or mechanisms that could aid in the clinical translation of CELF2 as a prognostic biomarker throughout the study.

Based on the previous results in Muhammad Waqas Akbar's unpublished work, several prognostic metabolites in BC were identified using the metabolite expression and clinical data from the microarray dataset GSE37751. The log-rank multiple cut-off (LRMC) method was employed to screen the metabolites in 61 cancer patient samples for any possible association with prognosis. Given the small number of metabolite expression datasets and the low frequency of overlapping between metabolites across cohorts, the prognostic metabolites could not be validated in additional datasets. However, Pearson Correlation analysis was used to investigate the biological relevance underlying the association of the group of nine metabolites with BC prognosis. S-adenosylhomocysteine (SAH) and seven carnitine metabolites were found to have prognostic values and the highest degree of

correlation with one another (Akbar et al, unpublished data). Based on prior research, this group of metabolites was also able to identify the Triple Negative Breast Cancer (TNBC) subtype.

SAH is produced as a byproduct of methionine metabolism. Methionine is one of the essential, sulfur-containing amino acids because it cannot be synthesized in the body. One of the most important functions of methionine metabolism is to maintain the necessary methylation rate in the cell. S-adenosylmethionine (SAM), a universal methyl donor factor for DNA, RNA, and protein methylation, mediates this process, leading to SAH production [103]. Because SAH has been shown to inhibit methyltransferases, the overall methylation rate is dependent on the SAH/SAM ratio. It has been demonstrated that cancer cells and T lymphocytes in the tumor microenvironment (TME) compete for methionine [104]. According to the findings, cancer cells can uptake more methionine due to higher levels of the SLC43A2 transporter system, causing CD8+ T lymphocytes to starve. In such a case, methionine deficiency would result in a significant decrease in SAM levels in CD8+ T cells. Reduced SAM levels have been linked to lower expression of STAT5, the main hub in T cell signaling and survival pathways [105]. High levels of SAH have also been shown to have a negative impact on immunity by limiting or destroying immune cells in another study [106]. We believe it is possible that increased levels of SAH are associated with a poor prognosis of OS in BC patients, because they reflect increased methionine uptake by neoplastic cells and nutrient depletion in the TME, resulting in weakened immunogenicity.

Carnitine is a non-essential amino acid that has been shown to transport long-chain fatty acids to mitochondria for further oxidation reactions that lead to energy production [107]. As a result, carnitine metabolites are critical components which are interconnected in the main pathways that supply energy demands in cancer cells [108]. Higher levels of carnitine metabolites are a direct indicator of increased energy metabolism that can support the abnormal growth of cancer cells. Butyrylcarnitine has been previously related to BC increased odds [109]. Based on previous findings as well, we think one of the

reasons for carnitine's poor prognosis association in BC is the metabolic reprogramming landscape that neoplastic cells go through to meet their energy demands [110].

According to the Pearson Correlation results, SAH and carnitine are the two metabolites that negatively correlate with CELF2 in BC most strongly. According to the survival analysis results, the patient group with higher CELF2 expression levels outperformed the other group in terms of OS. Centered on these data, CELF2 was initially identified as a good prognostic factor in GSE37751.

In summary, we successfully identified two potential prognostic metabolite families: SAH and Carnitine, which are associated with a poor prognosis, as well as a potential good prognostic biomarker gene: CELF2. We also discovered a significant negative relationship between the metabolites, SAH and Carnitine, and CELF2 using Pearson Correlation analysis. Because of the composite nature of such interactivity, metabolite-transcript correlations are challenging to interpret [111]. CELF2 was recently identified as a cancer-specific hypermethylation target that affects both RNA and protein levels in a BC study. The researchers discovered 82 AS events that were specifically mediated by the CELF2 gene [112]. CELF2 epigenetic loss in BC causes an imbalance in isoform content in these AS events, which include well-known oncogenes, metastasis and apoptosis-related genes, and signal transduction activators. According to our findings and previous research, we assume it is probable that cancer cells may increase their uptake of methionine by competing with T lymphocytes in the TME, use it to synthesize SAM, which could act as a methyl donor, and hypermethylate CELF2, lowering its expression levels [113]. To make a more educated assumption, additional data analysis is required to validate the prognostic role of the metabolites and their relationship to CELF2.

4.1.2 CELF2 *in-silico* Validation as a Prognostic Biomarker in All-Subtypes BC and TNBC

The prognostic function of the CELF2 gene has been validated in multiple RNA-Seq and microarray datasets. Using multivariate cox regression and survival analysis, we demonstrated that CELF2 retains its role as an independent good prognostic biomarker

gene. On a clinical perspective, an independent prognostic biomarker could benefit patients by opening up new avenues for individualized treatment strategies.

Moreover, we conducted LRMC analysis to validate the CELF2 prognostic role across other microarray datasets. There are two main CELF2 probesets in these datasets: 202157_s_at and 1554569_a_at. Based on the results, 202157_s_at is significant in a higher number of tests regardless of the prognosis direction (good vs. bad), but 1554569_a_at is associated with a good prognosis in more cases. Overall, these data suggest that the 1554569_a_at probeset is more consistent and compatible with the earlier analysis where CELF2 was demonstrated to have an association with a good prognosis, despite the lesser number of significant tests. We believe that there are two main reasons why 202157_s_at probeset resulted as significant in a higher number of validation tests as compared to 1554569_a_at. Firstly, there was a higher number of datasets that contained information of 202157_s_at probeset and therefore it was validated in a total number of 22 tests, as compared to the 1554569_a_at which was validated in 9 tests only. Secondly, 202157_s_at displayed higher expression levels than the other probeset in the datasets that contained both probesets, which may have impacted the significance of the findings [114]. Additionally, it's important to keep in mind that the biological significance that the 1554569_a_at probeset may have is unrelated to the lack of a significant p-value cut-off.

Previously, we mentioned that SAH and Carnitine metabolites could identify the TNBC subtype, and we showed a possible relation between them and the CELF2 gene via Correlation analysis. To explore this relationship more, and find out whether CELF2 would be able to have a consistent prognostic function in TNBC as well, we used LRMC analysis in patient samples of this subtype specifically. 202157_s_at probeset produced similar results to the previous analysis including all BC subtypes. On the other hand, 1554569_a_at probeset showed 100% accuracy in the prediction of a good prognosis in TNBC samples. Considering that the current prognostic tools are more suitable for early-stage BC, it is very important that other biomarkers be discovered for the more advanced BC stages [115]. Our findings signify that CELF2 gene can be a promising prognostic biomarker at a higher consistency rate in the TNBC subtype specifically.

4.2 Immunohistochemistry *in-vitro* Validation of CELF2

4.2.1 CELF2 Association with Prognosis in Breast Cancer Tissue Samples

Immunohistochemistry (IHC) was used to validate the in-silico identified role of CELF2 as a good prognostic biomarker in TNBC samples. Unfortunately, we were not able to reproduce our observations supporting the prognostic role of CELF2 *in vitro*. We believe this is partly due to the technical issues since a smaller portion of the tissue sections was stained (4-5 microns) and it didn't reflect the entire landscape of the samples. This claim can be supported by previous studies as well during which even a small 2 micron difference in the tissue thickness has been shown to affect the staining intensity [116]. Another factor might have been the long-term storage of the sections which may have interfered with the sample quality [117].

We previously demonstrated that the CELF2 gene has two main probesets, one of which, depending on the dataset, is associated with a poor prognosis. As a result, we show that the two probesets have different expression levels across cohorts and a different distribution across multiple cancer cell lines in a cell type specific fashion. Based on these findings, we hypothesized that the polyclonal antibody employed in the IHC tests, which recognizes both the good and less-consistent probesets, was another factor in our inability to validate CELF2's prognostic role.

4.2.2 CELF2 and Immune Cells Dynamics in IHC

Because data from the Human Protein Atlas (HPA) IHC experiments and available literature indicate that CELF2 protein levels are elevated in CD3+ T-cells (CD8+ T-cells & CD4+ T-cells), double staining for CELF2 and CD3 was performed. We decided to specifically investigate the intensity overlap between CELF2 and CD3 in tumor stroma to find out if CELF2 affects prognosis through a mechanism involving tumor immune infiltrates (TIL). The correlation between the two is moderate (rho = 0.58), and samples with high CELF2 expression levels but low CD3 intensity are found in the IHC data. In the

Swedish and GSE58812 datasets, Cibersort and the Student t-test analysis revealed that CD8+ T-cells are significantly more abundant in the CELF2-High group than in the CELF2-Low group. Similar results were obtained for CD4+ activated T-cells, Regulatory T-cells, Follicular helper T-cells, activated Natural Killer (NK) cells, M1 and M2 Macrophages (Appendix B). These results indicate that there is a potential relation between CELF2 expression levels and T-cell abundance. Based on the literature research, CELF2 expression is required for the regulation of splicing in response to antigens or other signaling pathways in T cells. Out of 200~ AS events that take place upon T cells stimulation, one-third are specifically dependent on CELF2 and include genes involved in cell death, transcription, and phosphorylation [118, 119]. CTTN gene splicing is one of the CELF2 dependent events in T-cells [120]. CTTN encodes for cortactin protein which is directly related to lymphocyte migration [121]. Loss of exon 6 in LEF1 gene due to specific CELF2 knockdown results in reduced expression of TCR-alpha mRNA [122]. It has also been shown that tissue-resident cytotoxic CD8+ T-cells contribute to BC immunosurveillance and improve the prognosis for TNBC [123].

As suggested by our data as well, the CELF2 gene can influence tumorigenesis via T cell regulation. CELF2 may activate signaling pathways associated with T-cells efferent function in response to T-cells stimulation, resulting in an immune response against tumors [120]. However, because CELF2 expression is not restricted to specific cell types, it is possible that it exerts its good prognostic function through other mechanisms. Our findings regarding CELF2 prognostic analysis in CD8-High and CD8-Low groups also indicate that CELF2's function as a prognostic biomarker gene is not dependent on the quantity and activity of immune infiltrates.

4.3 CELF2 Transcript Variants

4.3.1 CELF2 Probeset Sequence Analysis

It is an applied strategy in literature to treat different probesets that target the same gene uniformly and assume that they all represent the same scenario. Therefore, in a number of studies, the expression levels of such probesets are averaged and assigned to the gene

[124]. However, we believe that this is not always the case and that the variances these probeset show might have other underlying causes. We hypothesized that one of the causes might be that, despite the fact that the probesets target the same gene, they may reflect the abundance of different isoforms that are relatively expressed at various amounts depending on specific biological contexts. Therefore, we thought that it would be more informative to highlight the distinctions between these two probesets and to address each of them independently [125].

We confirmed in our sequence analysis that the two probesets do in fact recognize and thus reflect the expression levels of different CELF2 isoforms. The more consistent probeset 1554569_a_at recognizes a total of 4 isoforms, as opposed to 202157_s_at which can recognize a total of 9 isoforms. The isoforms that are specific to 202157_s_at, and cannot be recognized by 1554569_a_at, could offer an explanation on the difference in the association of the two probesets with the direction of prognosis.

4.3.2 CELF2 Transcript Variant Prognostic Analysis

To test our earlier theory that different CELF2 transcript variants might be associated with opposite prognoses, we looked at the prognostic role of each CELF2 isoform independently. We identified six CELF2 prognostic isoforms in the TCGA-BRCA cohort (n = 992); Uc001iki, Uc001iko, and Uc009xiw isoforms are associated with a good prognosis, whereas Uc001ikl, Uc010qbk, and Uc010qbo are associated with a less-favorable prognosis. We refrained from labeling the second group of isoforms as "bad-prognostic" based on the LRMC results because, such an association was only found in a few samples near the end of the cohort when the isoforms' expression levels were significantly higher compared to the general distribution. However, these data imply that the expression levels of both: good and less-favorable prognostic CELF2 isoforms influence the overall prognosis of BC patients. Instances of this include the variations in the LRMC outcomes between the two CELF2 probesets. The different CELF2 transcript variants, associated to different prognosis directions, produce CELF2 protein isoforms. Because of its polyclonal nature, the HPA03813 antibody can recognize these CELF2 protein isoforms. This suggests

that the lack of validation of CELF2 as a good prognostic biomarker during the IHC experiments could be due to the antibody's lack of specificity for the good prognostic isoforms only.

4.4 CELF2 Multiple Isoform Prognostic Signature and Interpretation

The Risk Score Model aims to combine the effects of the most prominently expressed good and less-favorable CELF2 prognostic isoforms. To do so, we created two new variables Favorable Prognostic Isoforms (FPI) to represent the sum of the two constantly good prognostic isoforms, Uc001iki and Uc001iko and "Less-Favorable Prognostic Isoform" (LFPI) for the Uc001ikl expression. After classifying patients into groups based on the risk score, as expected, the low-risk group outperforms the high-risk group in OS. The significance of this model is that we are attempting to reach a more accurate conclusion about the prognostic effects of CELF2 by considering all of its transcript variants. In this manner, during clinical applications, the prognosis will not be determined based on the dominant transcript of CELF2, which may vary from patient to patient, but rather on the overall landscape of all CELF2 isoforms adjusted to their respective weights.

In the available literature, there is a general lack of information on protein isoform characterization to make a thorough interpretation of their distinctive features. CELF2-Uc001iki is the canonical protein according to the protein database UniProtKB [126]. Based on the sequence comparison of Uc001iko and Uc001ikl to Uc001iki, certain characteristics distinguish these proteins from one another. We identified perfect-match amino acid sequence regions, which contain the three RNA Recognition Motifs (RRM), in all isoforms. The amino acid diversification is noticed in the divergent domain, which is the region responsible for the alternative splicing reactions mediated by CELF2. Different amino acids in divergent domain might be an indicator that due to consequential changes in their inter-protein binding conformations, these three CELF2 isoforms localize distinctly and mediate the splicing of separate RNAs. In such a scenario, it might be that the prognostic effects of the isoforms would be mediated via alternative molecular processes. In our sequence analysis, we previously identified that the Uc001iki CELF2 isoform is

characterized by a longer 3'UTR, compared to Uc001iko and Uc001ikl. Differences in the length of 3'UTR of CELF2 isoforms may affect the mRNA half-life, translation efficiency, localization of the protein variants, regulation and the overall function. The longer 3'UTR sequences are more enriched in RBP and microRNA binding sites, as well as secondary structure variations [127-129]. CELF2 has been shown to regulate the 3' UTR intron retention and alternative polyadenylation of its own transcript and other targets [91]. It might be that CELF2, through an auto-feedback mechanism, tightly controls the alternative polyadenylation of its own transcript to give rise to 3'UTR isoforms in a cancer cell specific context. CELF2 might prefer the selective expression of Uc001iko and Uc001ikl isoforms during BC tumorigenesis, because they have shorter 3'UTRs with fewer binding sites for other molecules such as miRNAs to downregulate CELF2. However, considering the several impacts this process can have on CELF2 regulation, further studies are needed to evaluate our hypothesis.

4.5 Conclusion

The goal of this study was to conduct a comprehensive analysis to identify a reliable prognostic biomarker in BC and investigate its biological mechanisms. Previously, a group of eight Carnitine metabolites and SAH were linked to a poor prognosis in BC (Muhammad Waqas Akbar, Unpublished Data). However, in the field of applicable biomarkers, metabolomics is relatively unexplored, and more accurate measurement methods must be developed before they can be used in a clinical setting. As a result, we were eager to identify the genes associated with these metabolites. We discovered that the CELF2 gene is negatively correlated with Carnitine and SAH, and then we identified that it can function as a good prognostic marker of OS in several BC datasets. Based on our findings, we propose a possible mechanism in which neoplastic cells use methionine, a precursor to SAH synthesis, to hypermethylate CELF2 and suppress its expression, resulting in an isoform imbalance of CELF2's target mRNAs.

In the microarray datasets, CELF2 gene was represented by 202157_s_at and 1554569_a_at probesets. The latter probeset is more consistent with the good prognostic

function of CELF2, whereas 202157_s_at shows an association with a bad prognosis in a fraction of validation cohorts. To investigate these variations we performed a sequence analysis, and found out certain CELF2 isoforms that could only be recognized by 202157_s_at probeset. We investigated the prognostic role of each isoform separately and identified a few ones that would relate to a less-favorable prognosis upon increased expression levels. The 1554569_a_at probeset only recognizes good prognostic isoforms, whereas the 202157_s_at probeset also recognizes less-favorable prognostic isoforms. We believe this is one of the major causes of the inconsistencies we encountered during the in-silico validation process.

IHC experiments were carried out to investigate the role of CELF2 as a good prognostic gene *in vitro*. Based on the LRMC analysis, we were unable to identify a clear significant association between CELF2 expression and a good prognosis. Aside from potential technical errors, we believe that the polyclonal antibody used during IHC experiments might have matched both: the good prognostic and the less-favorable prognostic isoforms recognized by the probeset 202157_s_at, leading to the confusing IHC results. We see an association of CELF2 with particularly CD3+ T-cells, in the IHC experiments and in the in-silico validation cohort. CD3+ T cell expression in the tumor stroma has previously been linked to a favorable cancer prognosis. However, by assessing its function in samples with low infiltrated immune cell expression, we demonstrated that CELF2 can function as a good prognostic gene in BC regardless of TIL abundance. Based on literature research, the CELF2 gene is responsible for specific AS splicing events required during T cell maturation and signaling, suggesting a potential role in anti-tumor immunity.

To improve prognosis prediction accuracy, we developed a CELF2 multiple isoform risk score model by combining the expression of good and less-favorable prognostic isoforms. Stratification of patients based on the risk score successfully predicted a longer OS for the Low-Risk group versus the High-Risk group.

In conclusion, we were able to identify CELF2 as a good prognostic gene in the prediction of OS in BC patients and show an interesting association it exhibits with several prognostic

metabolites. Despite the observed association, we also demonstrated that the prognostic function of the CELF2 gene is independent of the T-cell abundance in the TME. We developed a high specificity prognostic model that will consider the predictive power of each CELF2-Isoform, regardless of the direction of prognosis they are associated with, and classify patients into High Risk or Low Risk groups. To the best of our knowledge, there are no transcript variants – specific commercialized prognostic kits. We believe that CELF2 is a promising candidate for clinical application of this strategy. Based on our findings, we strongly advise that prognostic studies be conducted at the transcript level in order to achieve greater accuracy and consistency.

4.6 Future Perspectives

In order for the CELF2 transcript variant specific prognostic model to be implemented in the clinics, several aspect of its mechanism and more extensive validation steps need to be taken.

Initially, the IHC analysis would be more informative if it was to be repeated with isoform specific custom antibodies. In this way, we would be able to confirm or rule out our hypothesis that the presence of both good and bad prognostic CELF2 protein isoforms were an important factor in our lack of validation. After having obtained the custom antibodies we could extend our IHC analysis in a larger sample cohort, and also do in-vivo model validations.

Secondly, using the already available IHC data we could study their significance more extensively. A PCA analysis or other classification algorithms could be performed to see how the samples group and identify main component contributors.

Third, functional analysis of CELF2 at the transcript variant level is required in a Breast Cancer specific context, where we can identify direct mRNA splicing targets for each of the isoforms. This would enable us to elucidate possible mechanisms or pathways in which CELF2 isoform - specific targets are involved and understand their association to prognosis as well. It would also be useful to examine the relative stability of the CELF2

isoforms and their regulation in order to gain a better understanding of the specific role each of them plays on cancer cells.

Aside from the previously mentioned points, the Risk Score Prognostic Model must be validated in other cohorts to see if it performs with the same accuracy. To accomplish this, RNA-Sequencing experiments on BC tissue samples and, preferably, TNBC samples should be performed.

Chapter 5

5. Bibliography

1. Wittekind, C. and M. Neid, *Cancer invasion and metastasis.* Oncology, 2005. **69 Suppl 1**: p. 14-6.
2. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
3. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
4. Sung, H., J. Ferlay, and R.L. Siegel, *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* 2021. **71**(3): p. 209-249.
5. Lei, S. and R. Zheng, *Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020.* 2021. **41**(11): p. 1183-1194.
6. Maajani, K., et al., *The Global and Regional Survival Rate of Women With Breast Cancer: A Systematic Review and Meta-analysis.* Clin Breast Cancer, 2019. **19**(3): p. 165-177.
7. *Surveillance Epidemiology, and end results program*.
8. Nemchek, L., *Male Breast Cancer: Examining Gender Disparity in Diagnosis and Treatment.* Clin J Oncol Nurs, 2018. **22**(5): p. E127-e133.
9. Makki, J., *Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance.* Clin Med Insights Pathol, 2015. **8**: p. 23-31.
10. Yang, C., et al., *Comparison of Overall Survival Between Invasive Lobular Breast Carcinoma and Invasive Ductal Breast Carcinoma: A Propensity Score Matching Study Based on SEER Database.* Front Oncol, 2020. **10**: p. 590643.
11. Cheng, P., et al., *Treatment and survival outcomes of lobular carcinoma in situ of the breast: a SEER population based study.* Oncotarget, 2017. **8**(61): p. 103047-103054.
12. Wen, H.Y. and E. Brogi, *Lobular Carcinoma In Situ.* Surg Pathol Clin, 2018. **11**(1): p. 123-145.
13. van Seijen, M., et al., *Ductal carcinoma in situ: to treat or not to treat, that is the question.* Br J Cancer, 2019. **121**(4): p. 285-292.
14. van Maaren, M.C., et al., *Trends in incidence, treatment, survival and subsequent breast cancer in lobular carcinoma in situ in the Netherlands: A population-based analysis.* Breast, 2021. **59**: p. 376-382.
15. van de Vijver, M.J., et al., *A gene-expression signature as a predictor of survival in breast cancer.* N Engl J Med, 2002. **347**(25): p. 1999-2009.
16. Parker, J.S., et al., *Supervised risk predictor of breast cancer based on intrinsic subtypes.* J Clin Oncol, 2009. **27**(8): p. 1160-7.
17. Prat, A., et al., *Clinical implications of the intrinsic molecular subtypes of breast cancer.* Breast, 2015. **24 Suppl 2**: p. S26-35.
18. Patel, A., N. Unni, and Y. Peng, *The Changing Paradigm for the Treatment of HER2-Positive Breast Cancer.* 2020. **12**(8).
19. Leidy, J., A. Khan, and D. Kandil, *Basal-like breast cancer: update on clinicopathologic, immunohistochemical, and molecular features.* Arch Pathol Lab Med, 2014. **138**(1): p. 37-43.

20. Nielsen, T.O., et al., *A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer.* Clin Cancer Res, 2010. **16**(21): p. 5222-32.
21. Sun, Y.S., et al., *Risk Factors and Preventions of Breast Cancer.* Int J Biol Sci, 2017. **13**(11): p. 1387-1397.
22. Hastert, T.A., et al., *Adherence to WCRF/AICR cancer prevention recommendations and risk of postmenopausal breast cancer.* Cancer Epidemiol Biomarkers Prev, 2013. **22**(9): p. 1498-508.
23. Thomson, C.A., et al., *Nutrition and physical activity cancer prevention guidelines, cancer risk, and mortality in the women's health initiative.* Cancer Prev Res (Phila), 2014. **7**(1): p. 42-53.
24. Catsburg, C., A.B. Miller, and T.E. Rohan, *Adherence to cancer prevention guidelines and risk of breast cancer.* Int J Cancer, 2014. **135**(10): p. 2444-52.
25. Vrieling, A., et al., *Adult weight gain in relation to breast cancer risk by estrogen and progesterone receptor status: a meta-analysis.* Breast Cancer Res Treat, 2010. **123**(3): p. 641-9.
26. Eliassen, A.H., et al., *Adult weight change and risk of postmenopausal breast cancer.* Jama, 2006. **296**(2): p. 193-201.
27. Fan, S., et al., *Alcohol stimulates estrogen receptor signaling in human breast cancer cell lines.* Cancer Res, 2000. **60**(20): p. 5635-9.
28. Li, C.I., et al., *The relationship between alcohol use and risk of breast cancer by histology and hormone receptor status among women 65-79 years of age.* Cancer Epidemiol Biomarkers Prev, 2003. **12**(10): p. 1061-6.
29. Missmer, S.A., et al., *Meat and dairy food consumption and breast cancer: a pooled analysis of cohort studies.* Int J Epidemiol, 2002. **31**(1): p. 78-85.
30. Lo, J.J. and Y.M. Park, *Association between meat consumption and risk of breast cancer: Findings from the Sister Study.* 2020. **146**(8): p. 2156-2165.
31. Tappel, A., *Heme of consumed red meat can act as a catalyst of oxidative damage and could initiate colon, breast and prostate cancers, heart disease and other diseases.* Med Hypotheses, 2007. **68**(3): p. 562-4.
32. Hemminki, K., C. Granström, and K. Czene, *Attributable risks for familial breast cancer by proband status and morphology: a nationwide epidemiologic study from Sweden.* Int J Cancer, 2002. **100**(2): p. 214-9.
33. *Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease.* Lancet, 2001. **358**(9291): p. 1389-99.
34. Simon, K., J.B. Geigl, and G. Pristauz, *[Genes beyond BRCA1 and BRCA2 for hereditary breast cancer].* Wien Med Wochenschr, 2010. **160**(19-20): p. 478-82.
35. Garber, J.E., et al., *Follow-up study of twenty-four families with Li-Fraumeni syndrome.* Cancer Res, 1991. **51**(22): p. 6094-7.
36. Tan, M.H., et al., *Lifetime cancer risks in individuals with germline PTEN mutations.* Clin Cancer Res, 2012. **18**(2): p. 400-7.
37. Kapoor, A. and V.G. Vogel, *Prognostic factors for breast cancer and their use in the clinical setting.* Expert Rev Anticancer Ther, 2005. **5**(2): p. 269-81.
38. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer Statistics, 2017.* CA Cancer J Clin, 2017. **67**(1): p. 7-30.
39. de la Rochefordiere, A., et al., *Age as prognostic factor in premenopausal breast carcinoma.* Lancet, 1993. **341**(8852): p. 1039-43.

40. Carter, C.L., C. Allen, and D.E. Henson, *Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases.* Cancer, 1989. **63**(1): p. 181-7.
41. Koscielny, S., et al., *Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination.* Br J Cancer, 1984. **49**(6): p. 709-15.
42. Li, T., et al., *The Lymph Node Microenvironment May Invigorate Cancer Cells With Enhanced Metastatic Capacities.* Front Oncol, 2022. **12**: p. 816506.
43. Fisher, B., et al., *Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update.* Cancer, 1983. **52**(9): p. 1551-7.
44. Frierson, H.F., Jr., et al., *Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma.* Am J Clin Pathol, 1995. **103**(2): p. 195-8.
45. Elston, C.W. and I.O. Ellis, *Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up.* Histopathology, 1991. **19**(5): p. 403-10.
46. Galea, M.H., et al., *The Nottingham Prognostic Index in primary breast cancer.* Breast Cancer Res Treat, 1992. **22**(3): p. 207-19.
47. Buyse, M., et al., *Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer.* J Natl Cancer Inst, 2006. **98**(17): p. 1183-92.
48. Slodkowska, E.A. and J.S. Ross, *MammaPrint 70-gene signature: another milestone in personalized medical care for breast cancer patients.* Expert Rev Mol Diagn, 2009. **9**(5): p. 417-22.
49. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-536.
50. Cardoso, F., et al., *70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer.* N Engl J Med, 2016. **375**(8): p. 717-29.
51. Paik, S., et al., *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* N Engl J Med, 2004. **351**(27): p. 2817-26.
52. Sparano, J.A., et al., *Prospective Validation of a 21-Gene Expression Assay in Breast Cancer.* N Engl J Med, 2015. **373**(21): p. 2005-14.
53. Goldstein, L.J., et al., *Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features.* J Clin Oncol, 2008. **26**(25): p. 4063-71.
54. Mihály, Z. and B. Győrffy, *Improving Pathological Assessment of Breast Cancer by Employing Array-Based Transcriptome Analysis.* Microarrays (Basel), 2013. **2**(3): p. 228-42.
55. Weigelt, B., et al., *Breast cancer molecular profiling with single sample predictors: a retrospective analysis.* Lancet Oncol, 2010. **11**(4): p. 339-49.
56. Falato, C., et al., *Intrinsic subtypes and genomic signatures of primary breast cancer and prognosis after systemic relapse.* Mol Oncol, 2016. **10**(4): p. 517-25.
57. Johnson, J.M., et al., *Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.* Science, 2003. **302**(5653): p. 2141-4.
58. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.
59. Baralle, F.E. and J. Giudice, *Alternative splicing as a regulator of development and tissue identity.* Nat Rev Mol Cell Biol, 2017. **18**(7): p. 437-451.
60. Zhang, Y. and J. Qian, *Alternative splicing and cancer: a systematic review.* 2021. **6**(1): p. 78.

61. Wang, Y., et al., *Mechanism of alternative splicing and its regulation.* Biomed Rep, 2015. **3**(2): p. 152-158.
62. Liu, Y., et al., *Impact of Alternative Splicing on the Human Proteome.* Cell Rep, 2017. **20**(5): p. 1229-1241.
63. Wahl, M.C., C.L. Will, and R. Lührmann, *The spliceosome: design principles of a dynamic RNP machine.* Cell, 2009. **136**(4): p. 701-18.
64. Lamond, A.I., *The Spliceosome.* BioEssays, 1993. **15**(9): p. 595-603.
65. Pagani, F. and F.E. Baralle, *Genomic variants in exons and introns: identifying the splicing spoilers.* Nat Rev Genet, 2004. **5**(5): p. 389-96.
66. Long, J.C. and J.F. Caceres, *The SR protein family of splicing factors: master regulators of gene expression.* Biochem J, 2009. **417**(1): p. 15-27.
67. Coulter, L.R., M.A. Landree, and T.A. Cooper, *Identification of a new class of exonic splicing enhancers by in vivo selection.* Mol Cell Biol, 1997. **17**(4): p. 2143-50.
68. Wagner, E.J., et al., *Characterization of the intronic splicing silencers flanking FGFR2 exon IIIb.* J Biol Chem, 2005. **280**(14): p. 14017-27.
69. Tange, T.O., et al., *The hnRNP A1 protein regulates HIV-1 tat splicing via a novel intron silencer element.* Embo j, 2001. **20**(20): p. 5748-58.
70. Wang, Y., et al., *A complex network of factors with overlapping affinities represses splicing through intronic elements.* Nat Struct Mol Biol, 2013. **20**(1): p. 36-45.
71. Wang, Z. and C.B. Burge, *Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.* Rna, 2008. **14**(5): p. 802-13.
72. Kondo, Y., et al., *Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition.* Elife, 2015. **4**.
73. Query, C.C., P.S. McCaw, and P.A. Sharp, *A minimal spliceosomal complex A recognizes the branch site and polypyrimidine tract.* Mol Cell Biol, 1997. **17**(5): p. 2944-53.
74. Nguyen, T.H., et al., *The architecture of the spliceosomal U4/U6.U5 tri-snRNP.* Nature, 2015. **523**(7558): p. 47-52.
75. Will, C.L. and R. Lührmann, *Spliceosome structure and function.* Cold Spring Harb Perspect Biol, 2011. **3**(7).
76. Wu, J.Y. and T. Maniatis, *Specific interactions between proteins implicated in splice site selection and regulated alternative splicing.* Cell, 1993. **75**(6): p. 1061-70.
77. Ibrahim, E.C., et al., *Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 5002-7.
78. Lunde, B.M., C. Moore, and G. Varani, *RNA-binding proteins: modular design for efficient function.* Nat Rev Mol Cell Biol, 2007. **8**(6): p. 479-90.
79. He, Y. and R. Smith, *Nuclear functions of heterogeneous nuclear ribonucleoproteins A/B.* Cell Mol Life Sci, 2009. **66**(7): p. 1239-56.
80. Dreyfuss, G., V.N. Kim, and N. Kataoka, *Messenger-RNA-binding proteins and the messages they carry.* Nat Rev Mol Cell Biol, 2002. **3**(3): p. 195-205.
81. Mayeda, A., D.M. Helfman, and A.R. Krainer, *Modulation of exon skipping and inclusion by heterogeneous nuclear ribonucleoprotein A1 and pre-mRNA splicing factor SF2/ASF.* Mol Cell Biol, 1993. **13**(5): p. 2993-3001.
82. Mayeda, A. and A.R. Krainer, *Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2.* Cell, 1992. **68**(2): p. 365-75.
83. Bilodeau, P.S., et al., *RNA splicing at human immunodeficiency virus type 1 3' splice site A2 is regulated by binding of hnRNP A/B proteins to an exonic splicing silencer element.* J Virol, 2001. **75**(18): p. 8487-97.

84. Chen, C.D., R. Kobayashi, and D.M. Helfman, *Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene.* Genes Dev, 1999. **13**(5): p. 593-606.
85. Ladd, A.N., N. Charlet, and T.A. Cooper, *The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing.* Mol Cell Biol, 2001. **21**(4): p. 1285-96.
86. Dasgupta, T. and A.N. Ladd, *The importance of CELF control: molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins.* Wiley Interdiscip Rev RNA, 2012. **3**(1): p. 104-21.
87. Maloney, S.E., E. Khangura, and J.D. Dougherty, *The RNA-binding protein Celf6 is highly expressed in diencephalic nuclei and neuromodulatory cell populations of the mouse brain.* Brain Struct Funct, 2016. **221**(4): p. 1809-31.
88. Ladd, A.N. and T.A. Cooper, *Multiple domains control the subcellular localization and activity of ETR-3, a regulator of nuclear and cytoplasmic RNA processing events.* J Cell Sci, 2004. **117**(Pt 16): p. 3519-29.
89. Vlasova-St Louis, I., et al., *CELFish ways to modulate mRNA decay.* Biochim Biophys Acta, 2013. **1829**(6-7): p. 695-707.
90. Ramalingam, S., et al., *Novel intestinal splice variants of RNA-binding protein CUGBP2: isoform-specific effects on mitotic catastrophe.* Am J Physiol Gastrointest Liver Physiol, 2008. **294**(4): p. G971-81.
91. Chatrikhi, R., et al., *RNA Binding Protein CELF2 Regulates Signal-Induced Alternative Polyadenylation by Competing with Enhancers of the Polyadenylation Machinery.* Cell Rep, 2019. **28**(11): p. 2795-2806.e3.
92. Moraes, K.C., C.J. Wilusz, and J. Wilusz, *CUG-BP binds to RNA substrates and recruits PARN deadenylase.* Rna, 2006. **12**(6): p. 1084-91.
93. Mallory, M.J., et al., *Reciprocal regulation of hnRNP C and CELF2 through translation and transcription tunes splicing activity in T cells.* Nucleic Acids Res, 2020. **48**(10): p. 5710-5719.
94. Singh, B. and E. Eyras, *The role of alternative splicing in cancer.* Transcription, 2017. **8**(2): p. 91-98.
95. Oltean, S. and D.O. Bates, *Hallmarks of alternative splicing in cancer.* Oncogene, 2014. **33**(46): p. 5311-8.
96. He, C., et al., *A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis.* PLoS One, 2009. **4**(3): p. e4732.
97. Wang, Y., et al., *The BRCA1-Δ11q Alternative Splice Isoform Bypasses Germline Mutations and Promotes Therapeutic Resistance to PARP Inhibition and Cisplatin.* Cancer Res, 2016. **76**(9): p. 2778-90.
98. Avery-Kiejda, K.A., et al., *The relative mRNA expression of p53 isoforms in breast cancer is associated with clinical features and outcome.* Carcinogenesis, 2014. **35**(3): p. 586-96.
99. Mercatante, D.R., et al., *Modification of alternative splicing of Bcl-x pre-mRNA in prostate and breast cancer cells. analysis of apoptosis and cell death.* J Biol Chem, 2001. **276**(19): p. 16411-7.
100. Wang, R., et al., *Bromodomain protein Brd4 associated with acetylated chromatin is important for maintenance of higher-order chromatin structure.* J Biol Chem, 2012. **287**(14): p. 10738-52.
101. Wu, S.Y., et al., *Opposing Functions of BRD4 Isoforms in Breast Cancer.* Mol Cell, 2020. **78**(6): p. 1114-1132.e10.

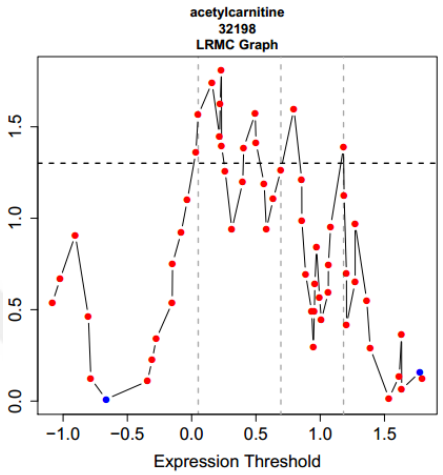
102. Uhlén, M., et al., *Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
103. Martínez, Y., X. Li, and G. Liu, *The role of methionine on metabolism, oxidative stress, and diseases.* 2017. **49**(12): p. 2091-2098.
104. Gubser, P.M. and A. Kallies, *Methio "mine"! Cancer cells steal methionine and impair CD8 T-cell function.* Immunol Cell Biol, 2020. **98**(8): p. 623-625.
105. Bian, Y., W. Li, and D.M. Kremer, *Cancer SLC43A2 alters T cell methionine metabolism and histone methylation.* 2020. **585**(7824): p. 277-282.
106. Kredich, N.M. and M.S. Hershfield, *S-adenosylhomocysteine toxicity in normal and adenosine kinase-deficient lymphoblasts of human origin.* Proc Natl Acad Sci U S A, 1979. **76**(5): p. 2450-4.
107. Flanagan, J.L., et al., *Role of carnitine in disease.* Nutr Metab (Lond), 2010. **7**: p. 30.
108. Melone, M.A.B., et al., *The carnitine system and cancer metabolic plasticity.* Cell Death Dis, 2018. **9**(2): p. 228.
109. Zhang, J., et al., *Circulating Carnitine Levels and Breast Cancer: A Matched Retrospective Case-Control Study.* Front Oncol, 2022. **12**: p. 891619.
110. Sun, C., et al., *Mass spectrometry imaging-based metabolomics to visualize the spatially resolved reprogramming of carnitine metabolism in breast cancer.* Theranostics, 2020. **10**(16): p. 7070-7082.
111. Redestig, H. and I.G. Costa, *Detection and interpretation of metabolite-transcript coresponses using combined profiling data.* Bioinformatics, 2011. **27**(13): p. i357-65.
112. Piqué, L., et al., *Epigenetic inactivation of the splicing RNA-binding protein CELF2 in human breast cancer.* 2019. **38**(45): p. 7106-7112.
113. Zhong, X. and G. Zhong, *Prognostic biomarker identification and tumor classification in breast cancer patients by methylation and transcriptome analysis.* FEBS Open Bio, 2021. **11**(8): p. 2139-51.
114. Sullivan, G.M. and R. Feinn, *Using Effect Size-or Why the P Value Is Not Enough.* J Grad Med Educ, 2012. **4**(3): p. 279-82.
115. Wazir, U. and K. Mokbel, *Emerging gene-based prognostic tools in early breast cancer: First steps to personalised medicine.* World J Clin Oncol, 2014. **5**(5): p. 795-9.
116. Barker, C., et al., *Nondestructive quality control of HER2 control cell line sections: the use of interferometry for measuring section thickness and implications for HER2 interpretation on breast tissue.* Appl Immunohistochem Mol Morphol, 2009. **17**(6): p. 536-42.
117. Fergenbaum, J.H., et al., *Loss of antigenicity in stored sections of breast cancer tissue microarrays.* Cancer Epidemiol Biomarkers Prev, 2004. **13**(4): p. 667-72.
118. Chuvpilo, S., et al., *Multiple NF-ATc isoforms with individual transcriptional properties are synthesized in T lymphocytes.* J Immunol, 1999. **162**(12): p. 7294-301.
119. Zhou, L. and X. Xie, *RNA-binding protein CELF2 inhibits breast cancer cell invasion and angiogenesis by downregulating NFATc1.* Exp Ther Med, 2021. **22**(2): p. 898.
120. Mallory, M.J., et al., *Induced transcription and stability of CELF2 mRNA drives widespread alternative splicing during T-cell signaling.* Proc Natl Acad Sci U S A, 2015. **112**(17): p. E2139-48.
121. Adamson, P., et al., *Lymphocyte migration through brain endothelial cell monolayers involves signaling through endothelial ICAM-1 via a rho-dependent pathway.* J Immunol, 1999. **162**(5): p. 2964-73.
122. Mallory, M.J., et al., *Signal- and development-dependent alternative splicing of LEF1 in T cells is controlled by CELF2.* Mol Cell Biol, 2011. **31**(11): p. 2184-95.

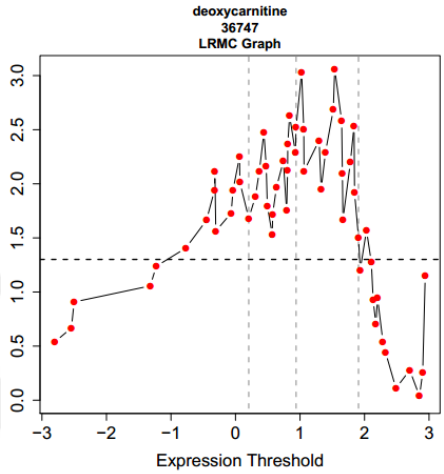
123. Savas, P., et al., *Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis.* 2018. **24**(7): p. 986-993.
124. Dai, M., et al., *Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.* Nucleic Acids Res, 2005. **33**(20): p. e175.
125. Stalteri, M.A. and A.P. Harrison, *Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips.* BMC Bioinformatics, 2007. **8**: p. 13.
126. Consortium, T.U., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Research, 2020. **49**(D1): p. D480-D489.
127. Berkovits, B.D. and C. Mayr, *Alternative 3' UTRs act as scaffolds to regulate membrane protein localization.* Nature, 2015. **522**(7556): p. 363-7.
128. Millevoi, S. and S. Vagner, *Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation.* Nucleic Acids Res, 2010. **38**(9): p. 2757-74.
129. Mayr, C., *Evolution and Biological Roles of Alternative 3'UTRs.* Trends Cell Biol, 2016. **26**(3): p. 227-237.

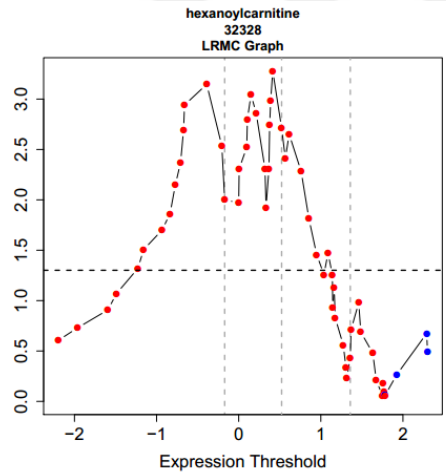
Chapter 6

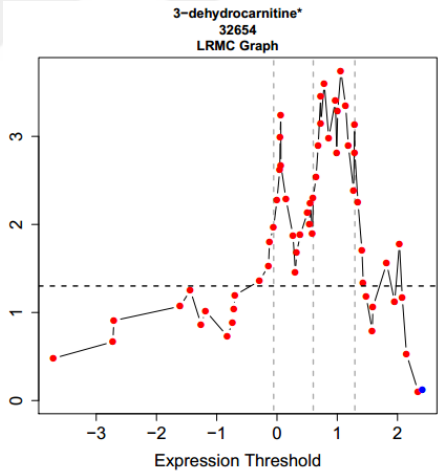
6.1 Appendix A

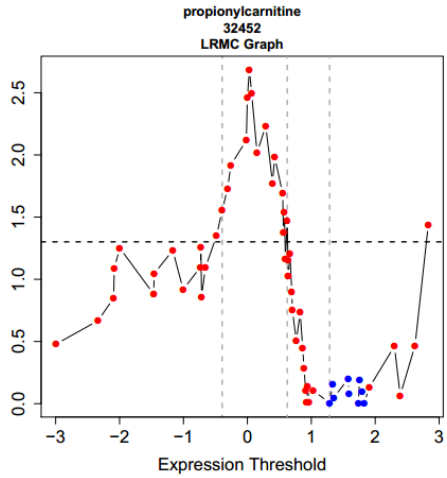
LRMC PLOTS OF THE PROGNOSTIC METABOLITES

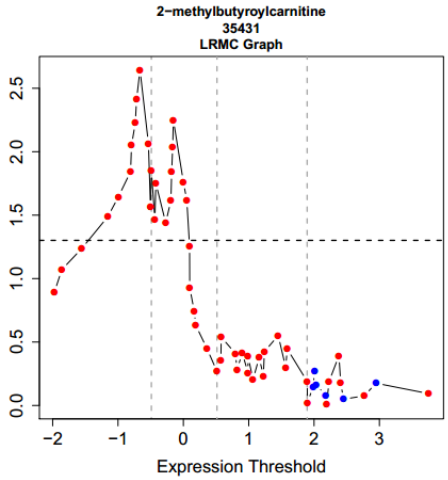


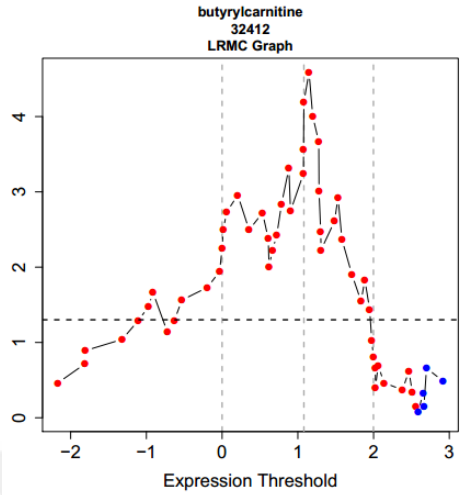








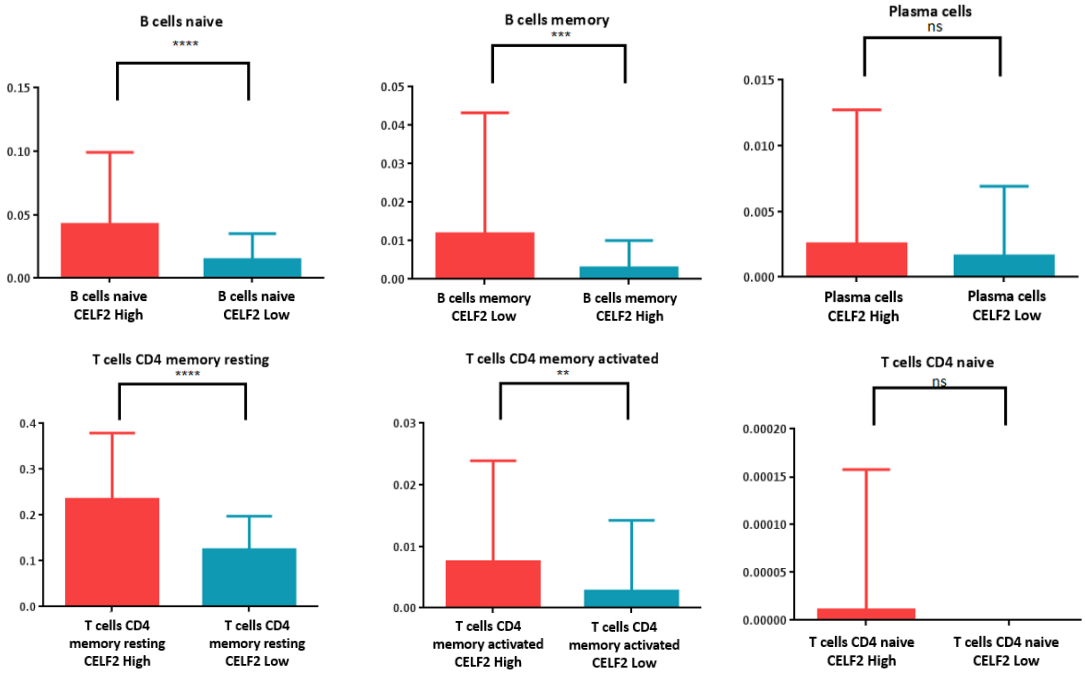


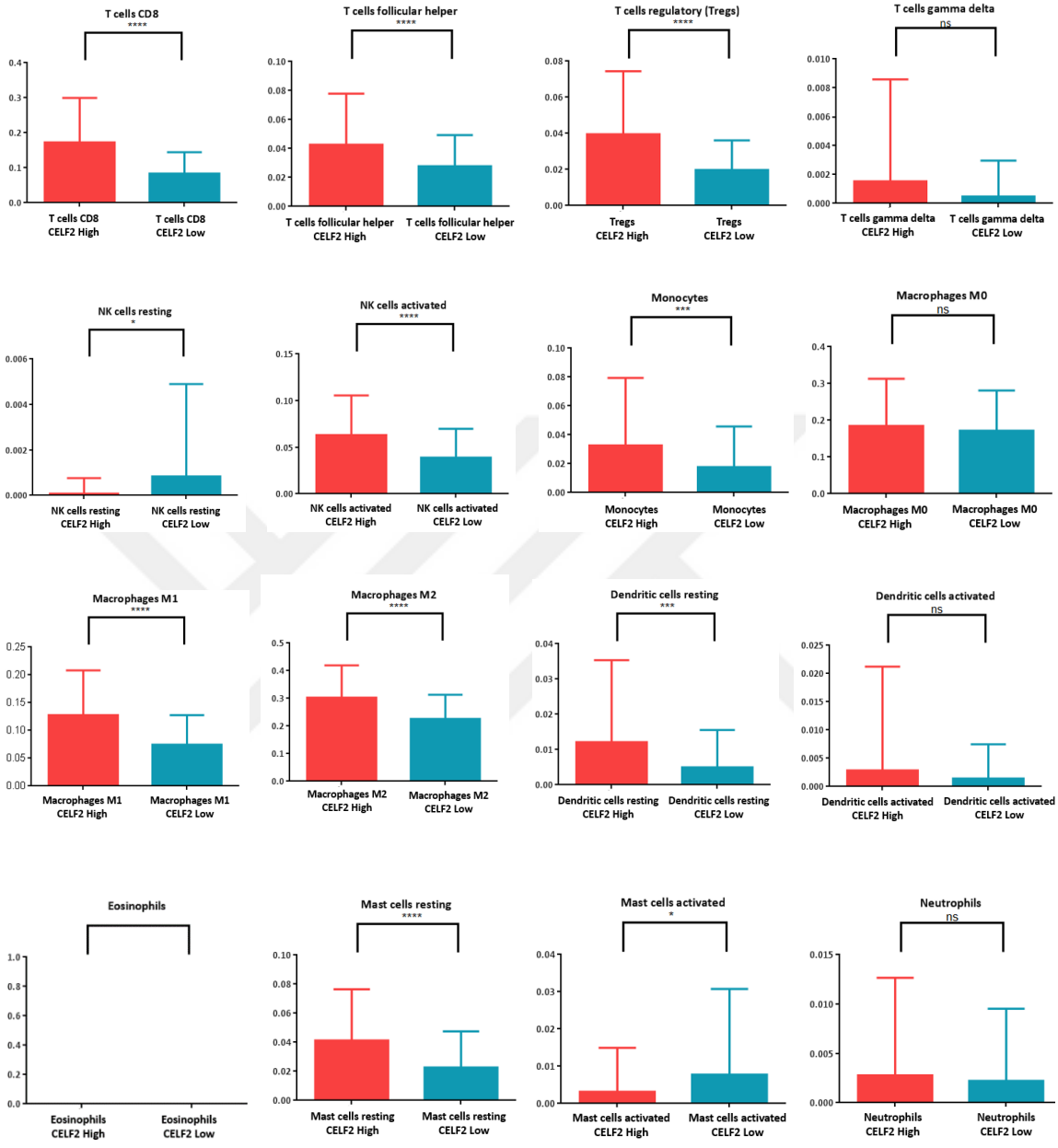


6.2 Appendix B

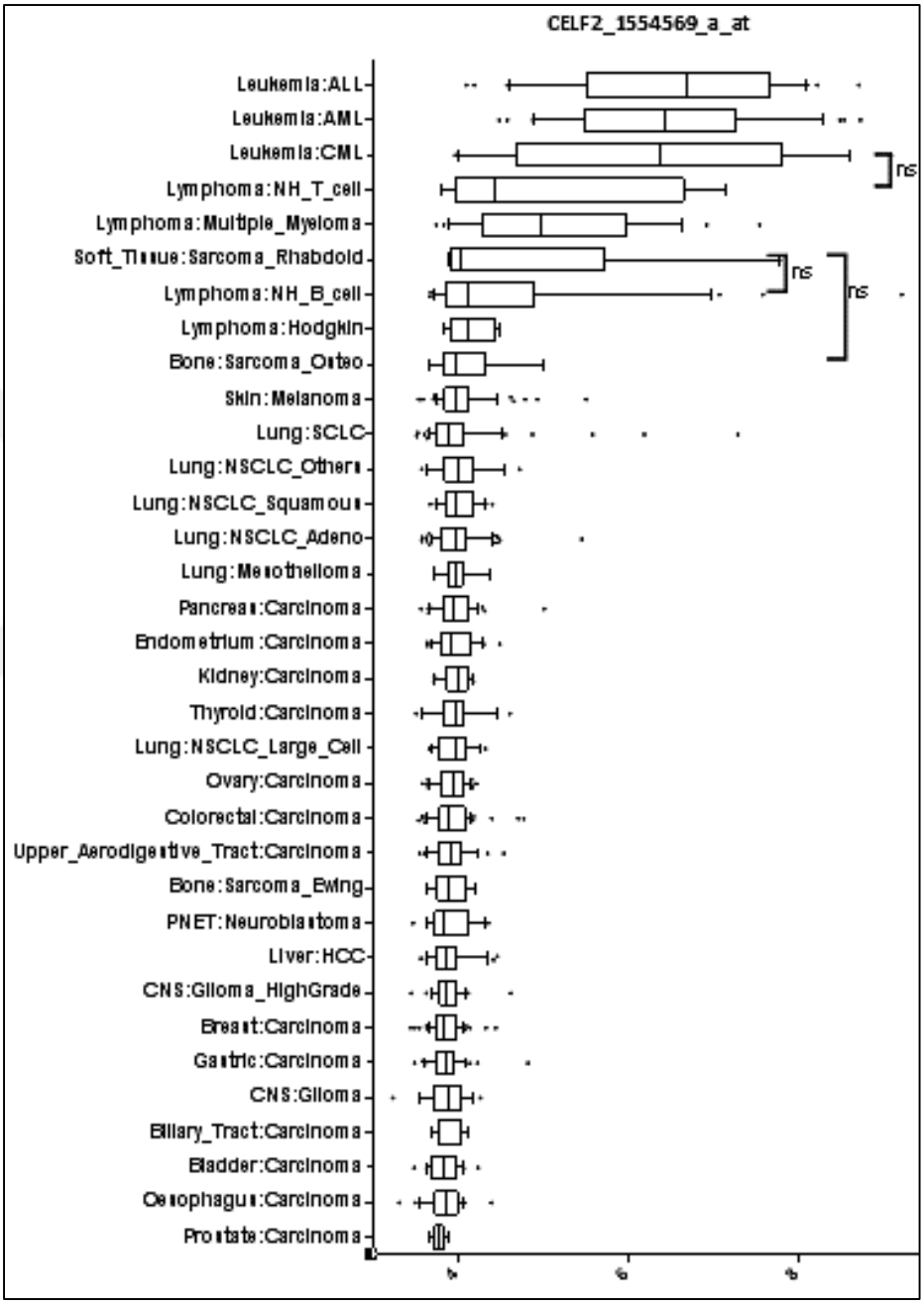
CIBERSORT Analysis

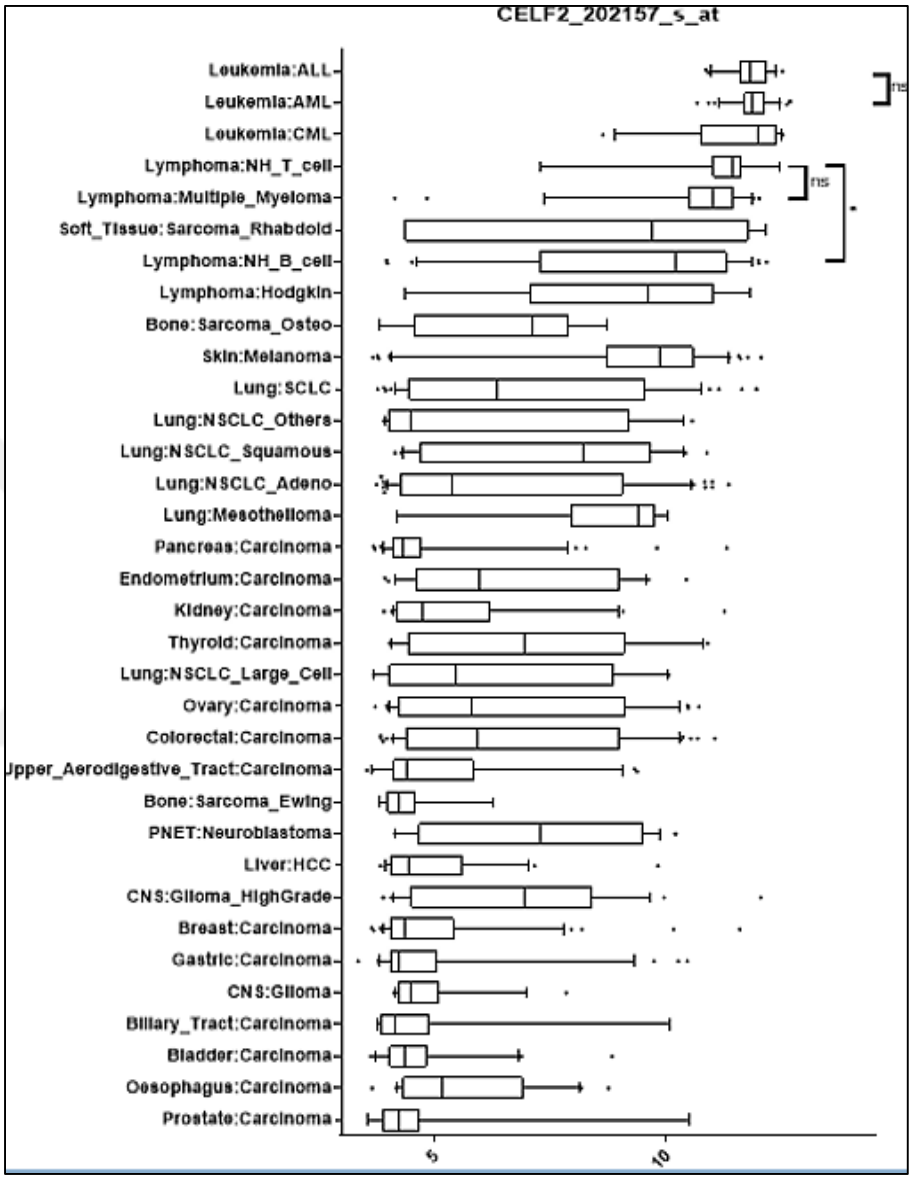
Swedish Cohort



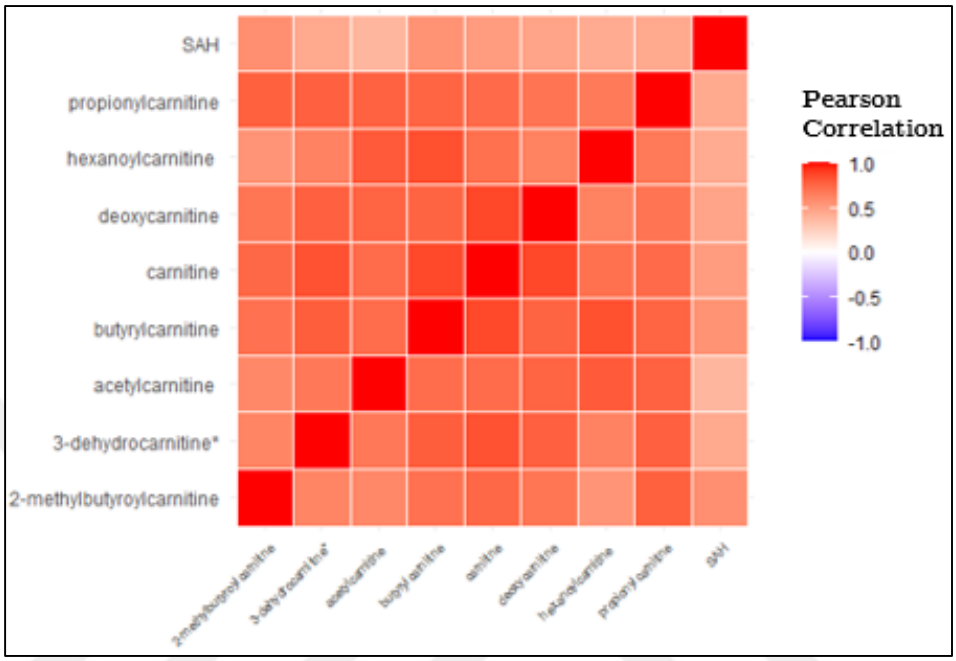

T cells CD8
****
T cells CD8 CELF2 High
T cells CD8 CELF2 Low
T cells follicular helper
****
T cells follicular helper CELF2 High
T cells follicular helper CELF2 Low
T cells regulatory (Tregs)
****
Tregs CELF2 High
Tregs CELF2 Low
T cells gamma delta
ns
T cells gamma delta CELF2 High
T cells gamma delta CELF2 Low
NK cells resting
*
NK cells resting CELF2 High
NK cells resting CELF2 Low
NK cells activated
****
NK cells activated CELF2 High
NK cells activated CELF2 Low
Monocytes
***
Monocytes CELF2 High
Monocytes CELF2 Low
Macrophages M0
ns
Macrophages M0 CELF2 High
Macrophages M0 CELF2 Low
Macrophages M1
****
Macrophages M1 CELF2 High
Macrophages M1 CELF2 Low
Macrophages M2
****
Macrophages M2 CELF2 High
Macrophages M2 CELF2 Low
Dendritic cells resting
***
Dendritic cells resting CELF2 High
Dendritic cells resting CELF2 Low
Dendritic cells activated
ns
Dendritic cells activated CELF2 High
Dendritic cells activated CELF2 Low
Eosinophils
Eosinophils CELF2 High
Eosinophils CELF2 Low
Mast cells resting
****
Mast cells resting CELF2 High
Mast cells resting CELF2 Low
Mast cells activated
*
Mast cells activated CELF2 High
Mast cells activated CELF2 Low
Neutrophils
ns
Neutrophils CELF2 High
Neutrophils CELF2 Low

6.3 Appendix C





6.4 Appendix D



Metabolite	butyrylcarnitine	hexanoylcarnitine	acetylcarnitine	3-dehydrocarnitine*	propionylcarnitine	2-methylbutyroylcarnitine	carnitine	deoxycarnitine	S-adenosylhomocysteine (S
butyrylcarnitine		0.842154	0.725677	0.787087	0.76228	0.708509	0.866244	0.766504	0.554979
hexanoylcarnitine	5.39E-15		0.802748	0.633582	0.66816	0.548592	0.710949	0.635449	0.433843
acetylcarnitine	8.54E-09	1.93E-12		0.675819	0.772794	0.606187	0.730925	0.764808	0.380956
3-dehydrocarnitine*	1.4E-11	8.74E-06	4.95E-07		0.77481	0.621437	0.834454	0.775314	0.44322
propionylcarnitine	2.35E-10	8.62E-07	7.44E-11	5.92E-11		0.778362	0.735123	0.691744	0.444264
2-methylbutyroylcarnitine	3.8E-08	0.000897	4.51E-05	1.84E-05	3.94E-11		0.749757	0.687638	0.572336
carnitine	6.28E-17	3.09E-08	5.29E-09	1.92E-14	3.58E-09	8.62E-10		0.872516	0.50992
deoxycarnitine	1.49E-10	7.77E-06	1.79E-10	5.59E-11	1.48E-07	2.03E-07	1.71E-17		0.471784
S-adenosylhomocysteine (SAH)	0.000662	0.084395	0.4225	0.061663	0.059511	0.00028	0.004992	0.022379	