IDENTIFICATION OF THERANOSTIC GENE MARKERS IN CANCERS AND PROGNOSTIC VALIDATION IN COLORECTAL 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

Murat İşbilen

January, 2015

IDENTIFICATION OF THERANOSTIC GENE MARKERS IN CANCERS AND PROGNOSTIC VALIDATION IN COLORECTAL CANCER
By Murat İşbilen
January, 2015

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.

Assist. Prof. Dr. Ali OSMAY GÜRE

Assist. Prof. Dr. Özlen KONU

Assist. Prof. Dr. Aybar Can ACAR

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

IDENTIFICATION OF THERANOSTIC GENE MARKERS IN CANCERS AND PROGNOSTIC VALIDATION IN COLORECRAL CANCER

MURAT İŞBİLEN

M.Sc. in Molecular Biology and Genetics

Advisor: Assist. Prof. Dr. Ali Osmay GÜRE

January 2015

Colorectal cancer (CRC) is the fourth most prevalent cancer type worldwide. Although the 5-year survival rate of CRC is higher than many cancer types, prediction of prognosis and identification of accurate biomarkers still maintain their importance for chemotherapy benefits, thus survival of the patients. Current techniques to identify biomarkers for clinical use are based on building models with multi-gene signatures. However, the accuracy rates of such signatures are not high enough due to heterogeneity of the tumors and low sensitivity of gene expression measurement techniques, although cell lines can be predicted very well with such signatures. There has also been sufficient evidence that multi-gene signatures may not be better predictors than random signatures with the same size. Therefore, in this study, we aimed to develop two R-based statistical analysis tools, SSAT and USAT, to identify single-gene expression markers for prognosis with chemotherapy benefit prediction power. We identified two genes, ULBP2 and SEMA5A, with SSAT and 6 genes, PTRF, TGFB1I1, DUSP10, KLF9, CLCN7 and CLDN3, with USAT for colon cancer and CRC, respectively. We were able to validate independent prognostic power of ULBP2 and SEMA5A in an independent cohort. However, we could only validate CLCN7 among 6 genes that we identified by USAT. Those results showed that SSAT may be a better tool to identify prognostic gene markers and USAT needs to

be improved to identify better candidate genes. We could also reveal the chemotherapy benefit prediction power of ULBP2 and SEMA5A in CCLE and CGP drug databases, although these in silico results should be validated by in vitro experiments. We believe that the approach that we used in this study may pioneer the studies to develop commercial theranostic tools for clinical use in various types of cancer.

Keywords: CRC, Prognosis, Chemotherapy, Biomarkers

ÖZET

KANSERDE PROGNOZ VE KEMOTERAPİ FAYDASI TAHMİNİ YAPABİLEN GEN BELİRTEÇLERİNİN BELİRLENMESİ VE KOLOREKTAL KANSERDE PROGNOZ BELİRTEÇLERİNİN DOĞRULAMA ÇALIŞMASI

MURAT İŞBİLEN

Moleküler Biyoloji ve Genetik, Yüksek Lisans

Tez Danışmanı: Yard. Doç. Dr. Ali Osmay GÜRE

Ocak 2014

Kolorektal kanser (KRK) dünyada en yaygın olan dördüncü kanser türüdür. KRK için 5 yıllık hayatta kalma oranı diğer kanser türlerinden daha yüksek olmasına rağmen prognoz tahmini ve yüksek tahmin gücü olan biyobelirteçlerin belirlenmesi hastaların tedavilerden faydalanabilmeleri ve hayatta kalabilmeleri için hala önemini korumaktadır. Klinikte kullanılacak biyobelirteçlerin belirlenmesi için günümüzde kullanılan teknikler çok genli listeler ile model oluşturmak üzerine kuruludur. Fakat, bu tür modeler hücre hatlarını güzelce ayırabilirken, tümördeki çeşitlilik ve gen ifadesini ölçen tekniklerin düşük duyarlılığı yüzünden tümör örneklerinde doğru tahmin oranları yeteri kadar yüksek değildir. Daha önceki yayınlanmış çalışmalar da gösteriyor ki, çok genli listeler eşit sayıda gen içeren rastgele gen listelerinden daha iyi ayrımlar yapamıyorlar. Bu sebeple, biz bu çalışmada, kemoterapi faydasını tespit edebilen tek gen prognoz belirteçleri belirlemek için, iki adet R tabanlı istatistiksel analiz programı (SSAT ve USAT) geliştirmeyi hedefledik. SSAT ile ULBP2 ve SEMA5A genlerini kolon kanseri için ve USAT ile PTRF, TGFB1I1, DUSP10, KLF9, CLCN7 ve CLDN3 genlerini KRK için aday prognoz belirteçleri olarak tespit ettik. ULBP2 ve SEMA5A genlerini bağımsız bir hasta grubunda diğer klinik parametrelerden bağımsız prognoz belirteçleri olarak doğruladık. Fakat,

USAT ile belirlediğimiz aday genlerden sadece CLCN7 genini doğrulayabildik. Bu sonuçlar bize SSAT'ın aday prognoz belirteçleri belirlemek için daha iyi bir program olabileceğini ve USAT'ın daha iyi prognoz belirteçlerini tespit edebilmesi için geliştirilmesi gerektiğini gösterdi. Ayrıca CCLE ve CGP ilaç veritabanlarında ULBP2 ve SEMA5A genlerinin kemoterapi faydası tahmini yapabildiklerini de gösterdik. Fakat bu sonuçlar in vitro deneyler ile doğrulanmalıdır. Bu çalışmada kullandığımız yöntemlerin, klinikte kullanılabilecek, prognoz ve kemoterapi faydası tahmini yapabilen ürünlerin geliştirilmesi için yapılacak çalışmalara katkı sağlayacağına inanıyoruz.

Anahtar Kelimeler: KRK, Prognoz, Kemoterapi, Biyobelirteç

ACKNOWLEDGEMENTS

I am pleased to express my sincere gratitude to my tolerant and wise supervisor Dr. Ali Osmay Güre for his unique and persistent guidance and motivating and encouraging support throughout my M.Sc. project. I am very grateful to him that he provided me with many opportunities, by which I could improve my experimental, computational and analytical skills. I got experienced a lot under his precious supervision. It is an honor for me to be a member of his lab.

Secondly, I cannot thank enough Dr. Mithat Gönen, Dr. Cemal Deniz Yenigün and Dr. Koray Doğan Kaya. They taught me a lot about biostatistics and bioinformatics. They shared their precious time and experiences with me to contribute to my project.

I would like to extend my gratitude to Dr. Özlen Konu and Dr. Aybar C. Acar for their valuable feedbacks and guidance for the project. They provided me with new insights for the development of the project.

I also feel grateful to al AOG lab members for their friendly, kind and motivating behaviors in the course of the project. I would like to thank specifically Kerem Mert Senses for his informative and helpful attitudes when I got stuck in my project. He was like my elder brother. I also thank Secil Demirkol and Dr. Emre Dayanç very much. They did all the validation experiments for SSAT approach.

I would like to express my greatest and profound thanks to my Pakistani friends Muhammad Waqas Akbar and Umar Raza. Indeed, they were more like brothers to me rather than friends. They never made me feel alone in my life in Bilkent University.

I am also very thankful to Gökhan Şentürk, Sevi Durdu and Ali Cihan. Gökhan was Dr. Gure's senior student. He got a funding from TUBITAK and supported my project financially. Sevi started this project and Ali developed it. They also helped me a lot to improve and finalize the project.

Finally, I declare that I owe all my achievements to my beloved parents Hüseyin and Sevda İşbilen, my dear brothers Sezer and Kadircan İşbilen and my beautiful fiancée Melike Öndeş. My parents worked hard throughout their lives to build a bright future for me and my brothers. They supported every decision I made and made me feel powerful against every difficulty I encountered. I felt my family's prayers and motivational supports in every situation.

Murat İşbilen

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Colorectal cancer (CRC) is one of the most prevalent cancer types worldwide, comprising around 10% of all cancer cases [1]. Although CRC is one of the most preventable cancer types as inappropriate eating habit, smoking and body fatness are the main risk factors for CRC, early diagnosis is still crucial for the survival of CRC patients [2, 3]. The patients who are diagnosed with TNM Stage I CRC have 90-95% 5-year survival rate after surgery, although the patients who are diagnosed with TNM Stage IV CRC have 5-10% 5-year survival rate [3]. Distant metastasis and local recurrence are the main reasons why CRC patients die due to cancer within 5 years after surgery.

The main treatment after surgery for CRC patients is adjuvant chemotherapy and radiotherapy to prevent recurrence. Whether CRC patients will be treated with adjuvant chemotherapy is determined according to histopathological characteristics of the resected tumors (infiltration to inner layers, metastasis to lymph nodes etc.) [4] and single molecule markers like p53/p21 expression, KRAS/BRAF mutations MSI etc. [5]. However, those characteristics do not provide opportunity to prediction of chemotherapy outcome, although they give insights into the prognosis of CRC patients. It has been shown that most stage III CRC patients benefit from chemotherapy, although the benefit of adjuvant chemotherapy in stage II CRC patients is still debatable [6]. In other words, many stage II and III CRC patients may be exposed to adverse effects of chemotherapy from which they would not benefit. Therefore, it is crucial to determine which stage II and III CRC patients to treat and not to treat with adjuvant chemotherapy in order to advance the treatment of CRC.

Recently, the common tendency in molecular biology to determine the benefit of any clinical use is to identify genomic signatures containing many parameters, like gene

expression, mutations etc. Identification of such markers requires high-throughput data, which has been published recently on databases, like Gene Expression Omnibus (GEO), in an increasing manner. Studies based on gene expression have been popular in the field of identification of biomarkers due to the advances in high-throughput gene expression measurement techniques, like microarray, next generation sequencing etc. Microarray data is one of the easiest high-throughput data to acquire, because of the publication rate and ease of handling. Therefore, a high amount of research have been conducted based on gene expression data to generate hypotheses or validate results. It has also been of great interest in identification of gene expression markers for clinical parameters, like prognosis, benefit of chemotreatment etc.

1.1. Prognostic Signatures

To date, many prognostic gene expression biomarkers have been identified for CRC. As reviewed by Schaeybroeck et al. [7], the first prognostic gene expression signature was a 23-gene signature, which was described by Wang et al. [8] in 2004 using two different approaches to distinguish between recurrent and relapse-free patients. They used split-sample method as their first approach by separating 74 stage II CRC patients into 36-patient training set and 38-patient test set. They selected 60 genes using training set and built a Cox model to predict recurrence in patients in the test set. In the second approach, they clustered all 74 patients by hierarchical clustering according to the expression of 17,616 genes and identified two distinct clusters. They assigned each patient sample into either cluster by the expression of most differentially expressed gene (Cadherin 17) between those clusters. Analyzing each cluster separately by split-sample method again, they selected 23 genes including Cadherin 17 to build a Cox model to predict recurrence. They report that the second approach successfully identified a gene expression signature with a prognostic value but the first approach did not [8].

Barrier et al. [9] also described a 30-gene tumor and 70-gene non-neoplastic mucosa prognostic signatures in 2005 using microarray expression data of 18 stage II and III (9 recurrent, 9 non-recurrent) CRC patient samples to predict recurrence. Their method consisted of two steps, one of which is the selection of the differentially expressed genes between relapse-free and recurrent samples and the other one is the recurrence prediction through k-nearest neighbor method. They determined the number of genes and the nearest neighbors by six-fold cross-validation. Barrier et al. [10] also used 50 stage II CRC patient samples to validate Wang et al.'s 23-gene prognostic signature using the same split-sample method. It was the first validation of a prognostic gene signature for CRC that was described by a completely independent group [7, 10].

O'Connell et al. [11, 12] developed Oncotype DX® as 12-gene prognostic signature to predict 3-year recurrence risk and 11-gene predictive signature to predict chemotherapy benefit in stage II CRC patients through recurrence and treatment scores using FFPE tissues [12]. They performed multivariate Cox Proportional Hazard Regression to determine the genes associated with recurrence and treatment benefit. Then, they generated a 4-step selection procedure to build models for recurrence risk and treatment benefit. They used 1,851 CRC patients from National Surgical Adjuvant Breast and Bowel Project (NSABP) and Cleveland Clinic cohorts [12]. They also validated their results by 1,436 stage II CRC patients from the QUASAR trial [13].

Salazar et al. developed ColoPrint as an 18-gene prognostic signature to predict recurrence in early stage CRC patients using fresh frozen tumor samples [14]. They used 188- and 206-patient CRC cohorts as discovery and validation sets. They used Kaplan-Meier method and Log-Rank test to determine the variables (e.g. gene expression, clinical parameters) associated with prognosis. They built also multivariate Cox models with the significant variables to find independent variables.

ColDx is another 634-gene signature to predict recurrence risk of stage II CRC patients and was developed by Kennedy et al. using FFPE tissues from 73 and 142 patients with recurrent and non-recurrent CRC patients [15]. They identified classifiers based on recursive feature elimination and partial least square classification. As a result, 10 repeats of five-fold cross validation showed that the 643-gene signature was the optimal for prognostic separation.

OncotypeDX®, ColoPrint and ColDx were three of prognostic gene expression signatures developed in recent years by large number of patients. OncotypeDX® is the only one which is commercially available. The others are currently undergoing clinical validation research. One of the disadvantages of these signatures is that they cannot calculate a treatment score for chemotherapy benefit. In fact, OncotypeDX® underwent a failed retrospective validation study for chemotherapy benefit [16]. Therefore, these signatures have no chemotherapy benefit prediction power even for the patients predicted in high risk group.

1.2. Predictive Signatures

There is no commercially available chemotherapy benefit signatures due to many reasons but one of the main reason is that prediction signatures developed by tumor samples have some bottlenecks like small clinical trial sample size and heterogeneity of tumors. It is also not possible to study many drugs at the same time and to determine the best drug for each individual in those trials. Therefore, high-throughput research settings were required to determine the chemotherapy treatment from which each patient can benefit most. For this reason, several cell line drug response databases were established; Cancer Cell Line Encyclopedia (CCLE) [17] and Cancer Genome Project (CGP) [18] are two of them. One of the main disadvantages of cell line drug databases are that the findings were evaluated under the assumption that cancer cell lines may represent tumors in terms of

chemotherapy response. Nevertheless, CCLE provides evidence that tumor samples resemble cell lines in terms of gene expression profile [17].

CCLE includes 947 cell lines originated from many known cancer types and screened for 24 drugs with different action of mechanisms, as CGP includes 789 cell lines screened for 138 drugs. Mutation data for specific genes, gene expression and chromosomal copy number data are also available in those databases. Analyses of combination of those data may reveal many molecular level interactions and provide with the identification of general biomarkers for chemotherapeutic response in cancers, as Barretina et al. [17] and Garnett et al. [18] published some gene expression and single-molecule markers for the benefit of some drugs. Therefore, these databases have the utmost importance for the analyses of predictive signatures and remain to be analyzed further.

1.3. Aim

In this study, we aimed to develop algorithms to identify independent prognostic markers with chemotherapy benefit prediction power. For that reason, we first used cell line databases to find subtypes of different cancer types with differential drug responses. Secondly, we tried to identify prognostic subtypes in tumors with gene signatures and single-gene lists using approaches both similar with and different than the ones that we used for cell lines. Thirdly, we tried to assess the chemotherapy benefit powers of identified prognostic genes.

2. Materials and Methods

2.1. Cohorts and Datasets

2.1.1. Tumor Microarray Datasets

GSE17536, GSE17537 and GSE41258 microarray datasets were downloaded from GEO ("http://www.ncbi.nlm.nih.gov/geo") and normalized with GC-RMA method using GeneSpring 12.0. Corresponding clinical data were downloaded from ArrayExpress ("http://www.ebi.ac.uk/arrayexpress"). GSE17536 and GSE17537 include the gene expression profiles of 177 and 55 fresh frozen colon cancer tumor tissues, respectively. GSE41258 includes gene expression profiles of 390 samples, 182 of which are colorectal cancer samples. It also includes 54 normal colon samples, 49 polyp samples and many normal and metastatic samples from liver and lung.

2.1.2. Ankara Cohort

Ankara cohort that includes fresh frozen tumor tissues of 47 colon and 37 rectal cancer patients were used as validation cohort. The samples were collected from the patients in ... hospital and validated as tumor tissues by pathologists. The patients were followed for around 4 years and various clinical data were recorded.

2.1.3. Cell Line Microarray Datasets

GSE36133 [17] (CCLE database) and E-MTAB-783 [18] (CGP database) datasets were downloaded from ArrayExpress ("http://www.ebi.ac.uk/arrayexpress") and normalized with RMA method using GeneSpring 12.0.

2.2. Experimental Protocols

2.2.1. Tumor Homogenization and RNA Extraction

The tumor samples were cut into small pieces on dry ice and 100 mg of each tumor sample were homogenized in 1 ml of TRIzol reagent. After homogenization, the samples were frozen and TRIzol reagent RNA extraction protocol (catalog number: 15596-018) was applied for each sample after all the samples were homogenized. The samples were centrifuged at 12,000 g for 10 minutes at 4°C to remove the insoluble content of the samples. After 5 minutes incubation at room temperature, 0.2 ml of chloroform was added to the samples. The samples were shaken vigorously for 15 seconds and incubated at room temperature for 3 minutes. They were centrifuged at 12,000 g for 15 minutes at 4°C and upper aqueous phase was removed and placed into a new tube. 0.5 ml of 100% isopropanol was added onto aqueous phase and incubated at room temperature for 10 minutes. The resulting mixture was centrifuged at 12,000 g for 10 minutes at 4°C. The supernatant was removed and the pellet was washed with 1 ml of 75% ethanol by vortexing briefly. The samples were centrifuged at 7,500 g for 5 minutes at 4°C and the wash was discarded. The RNA pellet was dried for around 10 minutes and resuspended in 100 µl nuclease-free water and incubated at 55°C for 10 minutes. The concentration of the samples were calculated by NanoDrop 1000 Spectrophotometer.

2.2.2. DNase Treatment and cDNA Synthesis

DNase treatment was applied to RNA samples by Life Technologies' Ambion DNA-free Kit (catalog number: AM1906). The RNA samples were diluted so that the concentrations were 200 ng/µl and the experiment was performed with 60 µl reaction setup. 6 µl of 10X DNase buffer and 1.2 µl of rDNase I enzyme were added into 52.8 µl of RNA samples. After 20 minutes incubation at 37°C, 6 µl of resuspended DNase inactivation reagent was

added into the samples and incubated in room temperature for 2 minutes by mixing occasionally. The samples were centrifuged at 10,000 g for 1.5 minutes and the RNA was transferred to a new tube. The concentration of the samples were calculated by NanoDrop 1000 Spectrophotometer.

cDNA synthesis was performed by Thermo Scientific Fermentas RevertAid First Strand cDNA Synthesis Kit (catalog number: K1622). The RNA samples were diluted so that the concentrations were 500 ng/µl in the final reaction and the experiment was performed with 100 µl reaction setup. Master Mix was prepared for all samples at once with 20 µl of 5X reaction buffer, 5 µl of Ribolock RNase inhibitor, 10 µl of 10 mM dNTP mix and 5 µl of RevertAid M-MuLV Reverse Transcriptase enzyme for each sample. The reaction mixture included 55 µl of diluted RNA samples, 5 µl of random hexamer primer and 40 µl of Master Mix for each sample. Before addition of Master Mix, the samples with primers were incubated at 65°C for 5 minutes. The reaction took place under the conditions at 25°C for 5 minutes, 42°C for 60 minutes and 70°C for 5 minutes, in order.

2.2.3. qPCR Experiments

All qPCR reactions were run using TaqMan gene expression assays with 96-well plate format in 7500 Real-time PCR Systems. The best coverage MGB primer-probes with catalog number 4331182 were used for PTRF (Assay ID: Hs00396859_m1), TGFB1I1 (Assay ID: Hs00210887_m1), DUSP10 (Assay ID: Hs00200527_m1), KLF9 (Assay ID: Hs00230918_m1), CLCN7 (Assay ID: Hs01126462_m1) and CLDN3 (Assay ID: Hs00265816_s1). The best coverage MGB primer-probes with catalog number 4448892 were used for PCSK5 (Assay ID: Hs00196400_m1) and PLS3 (Assay ID: Hs00192406_m1).

The reaction mixture was comprised of 10 µl of TaqMan gene expression Master Mix, 1 µl of primer-probe, 7 µl of nuclease-free water and 2 µl of cDNA for each sample. The

initial reaction (holding stages) took place under the condition at 50°C for 2 minutes and 95°C for 10 minutes. The cycling stage reaction took place for 45 cycles under the condition at 95°C for 15 seconds and 60° for 1 minute.

Three experimental replicas were used for each gene of each sample. The mean cycle threshold (CT) values of three replicas were used as actual CT values for relative gene expression calculations. Relative gene expression values were calculated using $2^{-\Delta\Delta CT}$ calculation, where

$$\Delta\Delta CT = \left(CT_{Target} - CT_{GAPDH}\right)_{Sample} - \left(CT_{Target} - CT_{GAPDH}\right)_{Reference}$$

2.3. Statistical Analyses

2.3.1. Hierarchical Clustering Analyses

The hierarchical clustering analyses were performed with Cluster 3.0 with Euclidean distance and complete linkage method and visualized as heat-maps with dendograms with JTreeView. Standardized gene expression values were used for hierarchical clustering analyses. Whole Transcriptome Hierarchical analyses were performed by "hclust" function in R [19] with Euclidean distance and Ward method.

2.3.2. Global Effectiveness Profiles

The global effectiveness distributions of the drugs were visualized with boxplots for each cancer type included in CCLE and CGP databases using "boxplot" function in R. Normalized activity area and IC50 were used as drug sensitivity parameters in CCLE and CGP, respectively.

2.3.3. Comparison of Drug Response between Subtypes

The drug responses of two groups and more than two groups were compared student's t-test and ANOVA, respectively, using "t.test" and "aov" functions in R. The drug response distributions of the groups were visualized as jitter plots using "stripchart" function in R.

2.3.4. Survival Analyses

All univariate survival analyses including Cox proportional hazard regression, maximally selected rank statistics and Log-Rank tests and Kaplan-Meier method were performed using "survival" package [20] in R. Stepwise multivariate Cox proportional hazard regression analyses were performed using IBM SPSS Statistics 19.

2.3.4.1. Discretization of Expression Values

Three types of expression values were used in survival analyses. Continuous expression values were used as they were generated by GC-RMA normalization of datasets using GeneSpring 12.0. Two types of categorical expression values were used in analyses. First, logarithmic gene expression values were divided into 8 different categories so that the expression values between 0 and 2 were category 1, the ones between 2 and 4 were category 2 and so on. The expression values above 14 were considered as category 8. The second categorical expression values contained two categories as high and low expression categories determined by Maxstat threshold values.

2.3.4.2. SSAT

After probeset normalization, all but one of the probesets that hit the same gene were eliminated so that the one with the highest coefficient of variance were used in survival analyses as the representative of the gene. The gene expression values were categorized into 8 groups, as explained in Discretization of Expression Values section. For each gene,

7 sequential expression thresholds were determined to compare the samples in the categories 1 vs. 2, 3, 4, 5, 6, 7 and 8; 1, 2 vs. 3, 4, 5, 6, 7 and 8 and so on with Log-Rank test. The genes with a p-values less than 0.05 were considered as significant for the given threshold.

2.3.4.3. USAT

All the probesets were analyzed with CoxPH, Maxstat and Log-Rank tests, with continuous and categorical expression values. In each test, the p-value threshold was 0.05. 12 different models were generated using combinations of the statistical tests and the expression types. The probesets that were significant in all the statistical tests at least in one model were considered as significant.

2.3.4.3.1. *Model Generation*

The models were generated by combining CoxPH, Maxstat and Log-Rank tests with continuous and categorical expression values. The explanation of generated models are as follows:

1. CC-MC-LC: Univariate CoxPH regression with continuous expression values, Maxstat with continuous expression values, Log-Rank with 2-category expression values.
2. CCS-MC-LC: Bivariate CoxPH regression with continuous expression values and stage, Maxstat with continuous expression values, Log-Rank with 2-category expression values.
3. CG-MC-LC: Univariate CoxPH regression with 8-category expression values, Maxstat with continuous expression values, Log-Rank with 2-category expression values.

4. CGS-MC-LC: Bivariate CoxPH regression with 8-category expression values and stage, Maxstat with continuous expression values, Log-Rank with 2-category expression values.
5. CC-MG-LG: Univariate CoxPH regression with continuous expression values, Maxstat with 8-category expression values, Log-Rank with 2-category expression values.
6. CCS-MG-LG: Bivariate CoxPH regression with continuous expression values and stage, Maxstat with 8-category expression values, Log-Rank with 2-category expression values.
7. CG-MG-LG: Univariate CoxPH regression with 8-category expression values, Maxstat with 8-category expression values, Log-Rank with 2-category expression values.
8. CGS-MG-LG: Bivariate CoxPH regression with 8-category expression values and stage, Maxstat with 8-category expression values, Log-Rank with 2-category expression values.
9. MC-C12-LC: Maxstat with continuous expression values, univariate CoxPH regression with 2-category expression values, Log-Rank with 2-category expression values.
10. MC-C12S-LC: Maxstat with continuous expression values, bivariate CoxPH regression with 2-category expression values and stage, Log-Rank with 2-category expression values.
11. MG-C12-LG: Maxstat with 8-category expression values, univariate CoxPH regression with 2-category expression values, Log-Rank with 2-category expression values.

12. MG-C12S-LG: Maxstat with 8-category expression values, bivariate CoxPH regression with 2-category expression values and stage, Log-Rank with 2-category expression values.

2.3.4.4. Log-Rank with Multiple Cut-offs

All the possible expression threshold values were determined for each probeset and Log-Rank tests were performed for all these expression thresholds. For each probeset, the Log-Rank p-values were plotted for corresponding expression thresholds. Blue and red colors were used to express the direction of the association of the high expression of the genes with good and poor prognosis, respectively, based on the given thresholds.

2.3.5. Gene Expression Profile Comparison of Cohorts

Pearson's correlation between the expression of the genes (probesets for microarray datasets and qPCR expression for Ankara cohort) were calculated within cohorts and shown with heat-maps.

3. Results

In this study, we first identified gene signatures that could differentiate between distinct cancer cell lines subtypes, which were sensitive and resistant to different drugs, using commonly used supervised and unsupervised clustering approaches. However, those approaches, by which gene signatures could be defined, did not work for tumors to find prognostic subtypes. Therefore, we decided to develop algorithms that could identify single gene markers, rather than signatures, which could define distinct prognostic subtypes. We developed Semi-supervised Survival Analysis Tool (SSAT) and Unsupervised Survival Analysis Tool (USAT) for the identification of prognostic gene markers in any cohort with high-throughput gene expression and clinical data. We performed SSAT and USAT analyses with three different CRC microarray datasets and tried to validate their results with a unique CRC cohort by qPCR.

3.1. Identification of Chemotherapeutic Subtypes in Cancers Cell Lines

3.1.1. Hematological Cancer Cell Lines

We first aimed to identify a gene signature to find new genetic subtypes in hematological cancer cell lines, which exhibited differential drug response [21], rather than clinically-defined classification. We used gene expression and drug response data for hematological cancer cell lines from Cancer Cell Line Encyclopedia (CCLE) [17] and Cancer Genome Project (CGP) [18] to identify gene signatures for subtypes and drug response. According to hierarchical clustering with the expression of 1,000 most variant genes, we identified 8 distinct subtypes of hematological cancer cell lines in CCLE database and selected 400 of 1,000 genes as a signature, all of which were either positively or negatively correlated with any one of 8 clusters with a Pearson r more than 0.5 (Figure 3.1A-B). We also validated the expression pattern of each group identified in CCLE using CGP database with the

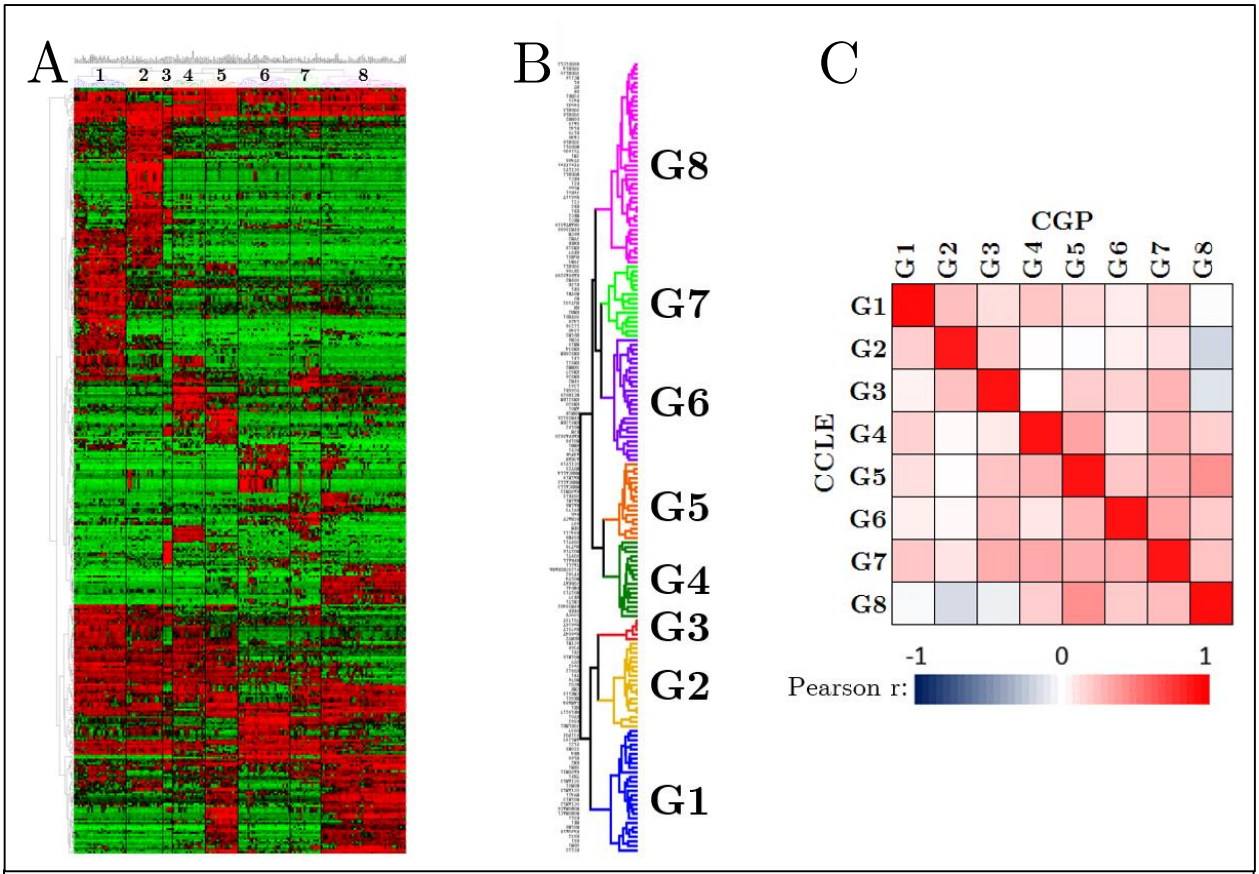

Figure 3.1: Classification of hematological cancer cell lines based on gene expression profile.

(A) Heat-map shows the expression profiles of hematological cancer cell lines clustered as shown in clustering tree **(B)** based on the standardized expression of 400 genes in CCLE database. In the heat-map, red, green and black colors show high, low and intermediate gene expression, respectively. Each cluster in the three is defined by a color. **(C)** The heat-map shows the concordance of the clusters based on the expression profiles of 400 genes in CCLE and CGP databases. Red, blue and white colors present positive, negative and no correlation, respectively.

common cell lines in both databases. The heat-map in Figure 3.1C shows that the expression pattern of each cluster in CCLE was highly correlated with the counterpart in CGP according to our signature.

After we had showed that there had been similar clusters in both databases, we checked whether those clusters could define sensitive subgroups to drugs better than classical classification. We used other cancer types as reference to determine whether hematological

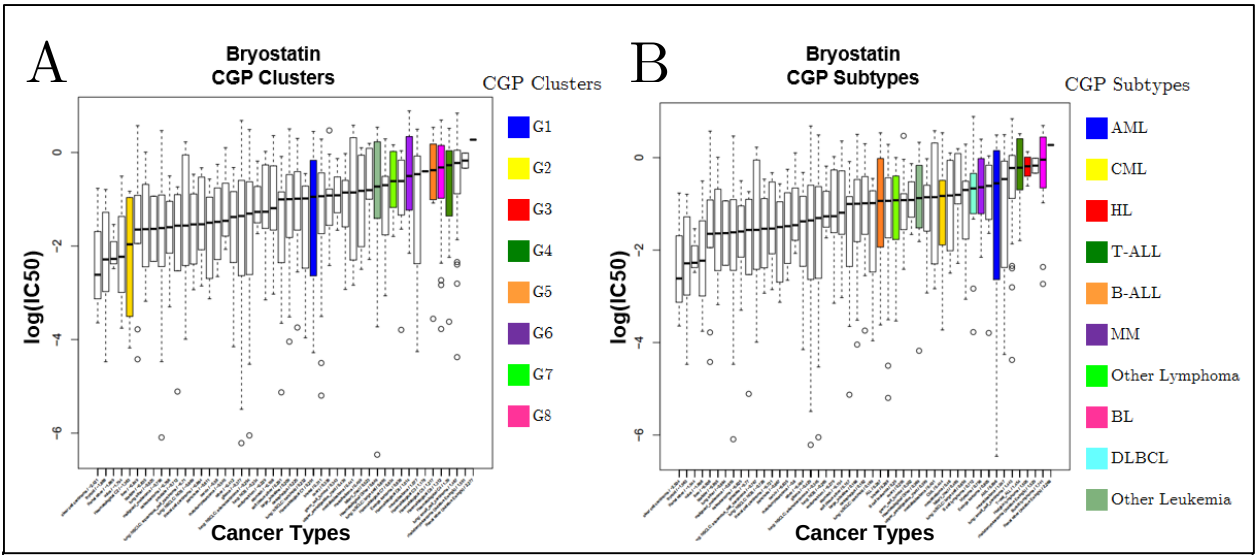

Figure 3.2: Global effectiveness of Bryostatin in CPG database.
The distribution of responses to Bryostatin of the cell lines from each origin is shown by boxplots of IC50 values in logarithmic scale. Each cluster (A) and classical subtype (B) of hematological cancers is shown by indicated colors. The cancer types are sorted based on median drug response.

cancer clusters or classical subtypes were sensitive to a drug. Therefore, we compared each clusters or subtypes with each other, as well as other cancer types. We found out that new clusters of hematological cancers could identify sensitive cell lines to some drugs, although all classical subtypes were resistant to those drugs compared to all other cancer types (Figure 3.2). Figure 3.2 shows the distributions of the concentrations of Bryostatin required to inhibit growth 50% (IC50) in logarithmic scale for each cancer types, as well as our hematological cancer clusters and classical subtypes. New clustering of hematological cancer cell lines identified cluster 2 (G2) as one of the most sensitive groups to Bryostatin compared to both other hematological cancer clusters and other cancer types, although all classical subtypes of hematological cancers were resistant to Bryostatin (Figure 3.2).

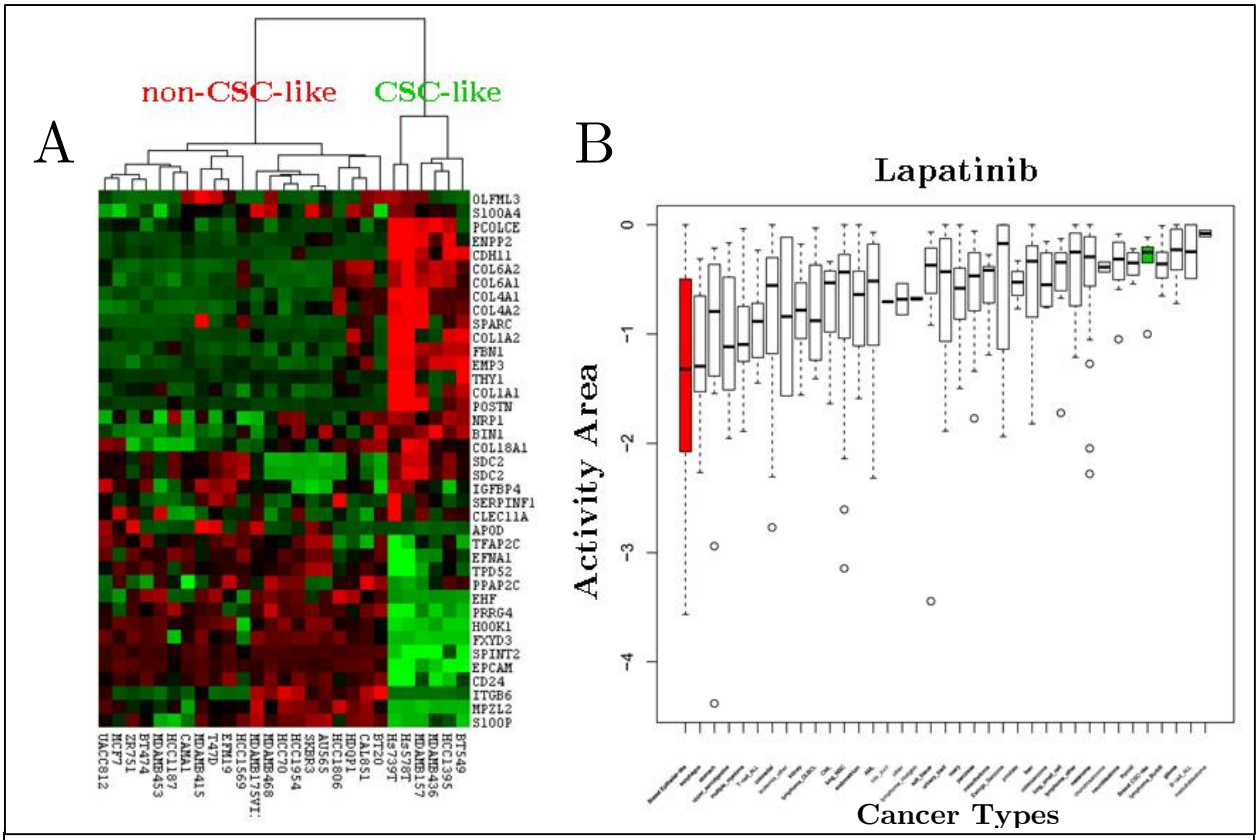

Figure 3.3: Hierarchical clustering of breast cancer cell lines in CCLE and global effectiveness distribution of Lapatinib.

(A) Hierarchical clustering of breast cancer cell lines in CCLE database identified two distinct breast cancer clusters. The cell lines (in columns) are clustered according to Gupta et al.'s gene list (in rows). Red, green and black colors in the heat-map represent high, low and intermediate gene expression. **(B)** Lapatinib response distributions of all cancer types, as well as breast cancer subtypes shown in the heat-map. The distribution of activity area (CCLE specific drug response parameter) values for each cancer type is shown by boxplots. Boxplots are sorted based on mean activity area within cancer types. Green and red boxplots represent CSC-like and non-CSC-like breast cancer cell lines. Lower the activity area is, more sensitive the cell lines are.

This way, we developed a 400-gene signature that could identify distinct subtypes with differential drug responses in hematological cancers. However, it still remains to be validated in vitro.

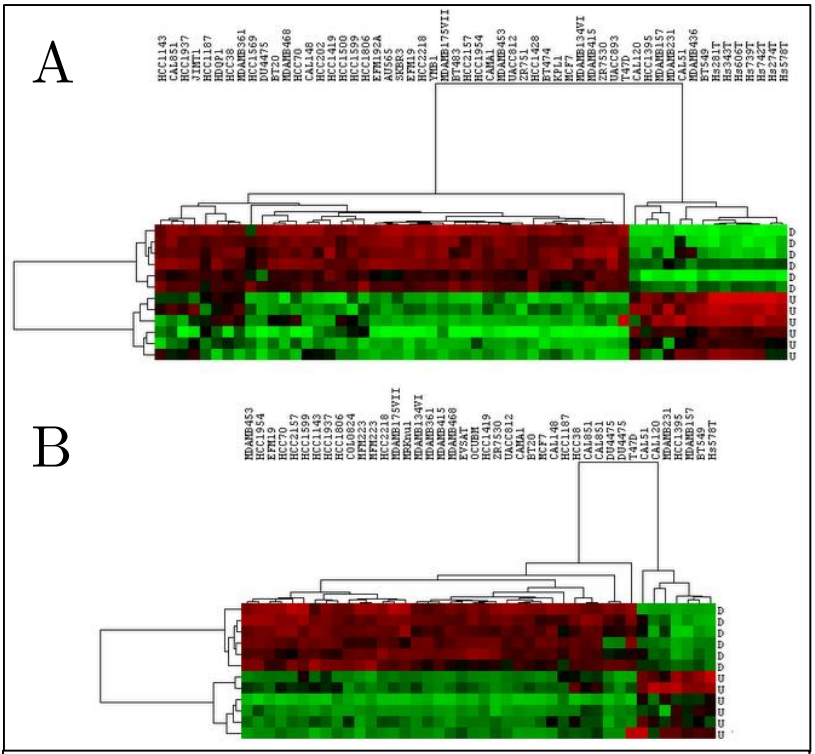

Figure 3.4: Hierarchical clustering of breast cancer cell lines in CCLE (A) and CGP (B) databases based on 12-gene signature.
Red, green and black colors define high, low and intermediate expression, respectively. The genes that are up- and down-regulated in CSC-like cells were shown by U and D, respectively.

3.1.2. Breast Cancer Cell Lines

Besides molecular and intrinsic subtypes, breast cancer can be sub-divided into two groups according to stemness properties. Gupta et al. [22] published a list of genes up- or down-regulated in cell populations enriched for stem-like cells. We showed in Isbilen et al. [23] that those genes identified two groups of breast cancer cell lines (Figure 3.3A) with differential drug responses. We have found out that CSC-like cells were very resistant to Lapatinib compared to non-CSC-like cells and other cancer types, although Lapatinib has been approved for the treatment of HER2- and hormone-positive breast cancers (Figure 3.3B).

We have also developed 12-gene signature that distinguishes breast cancer cell lines according to their cancer stemness. We found the most differentially expressed genes between CSC-like and non-CSC-like breast cancer cell lines and selected 12 of them as the best classifiers. Those 12 genes separated the common cell lines in CCLE and CGP into the same clusters sharply in both databases (Figure 3.4).

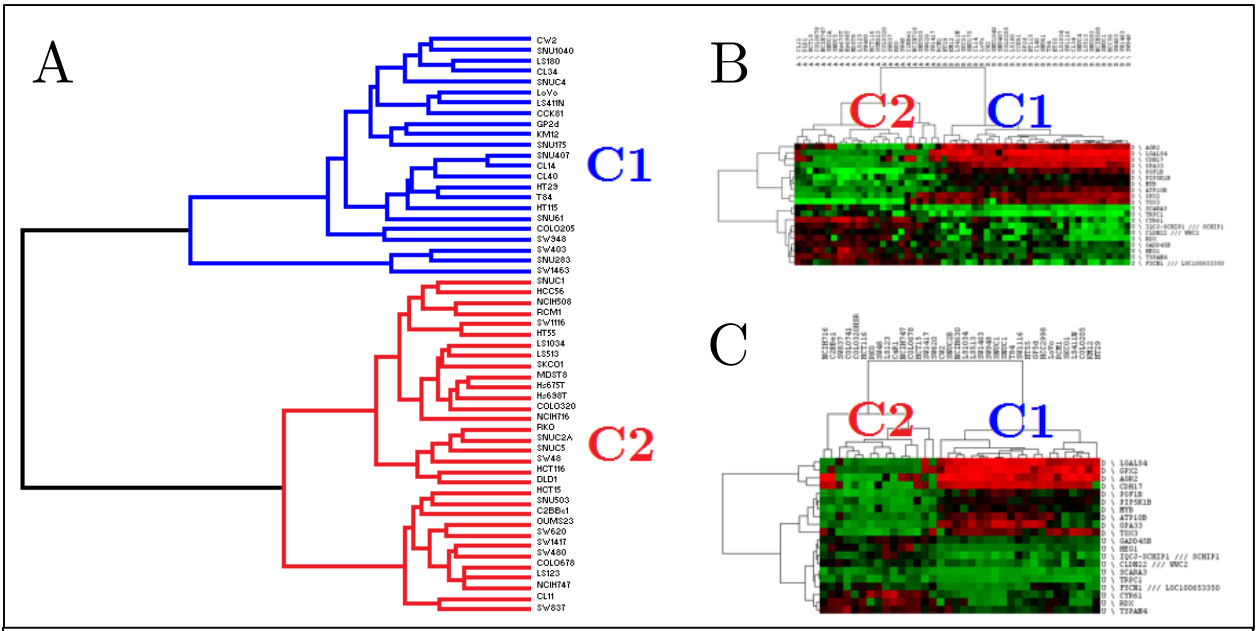

Figure 3.5: Hierarchical clustering of CRC cell lines in CCLE and CGP databases. WTHC result for CRC cell lines **(A)** in CCLE database is shown as a dendogram. Hierarchical clustering of CRC cell lines in CCLE **(B)** and CGP **(C)** databases based on the most differentially expressed genes between clusters defined by WTHC (C1 and C2) are shown as heat-maps with dendograms. Red, green and black colors in the heat-maps represent high, low and intermediate expression, respectively.

3.1.3. Colorectal Cancer Cell Lines

We also aimed to find out distinct subtypes of CRC cell lines with differential drug responses. We used gene expression and drug response data for the CRC cell lines from CCLE and CGP to identify gene signatures for subtypes and drug response. According to the hierarchical clustering with whole transcriptome, we identified 2 distinct clusters of CRC cell lines in CCLE database (Figure 3.5A). In order to find a gene signature that could distinguish those clusters, we found the differentially expressed genes between those clusters by t-test in CCLE database. We selected 20 out of 200 most differentially expressed genes as a gene signature for the clusters by Maximum Relevance Minimum Redundancy (MRMR) approach. Those 20-gene signature separated the same cell lines into same clusters in both CCLE and CGP databases (Figure 3.5B-C).

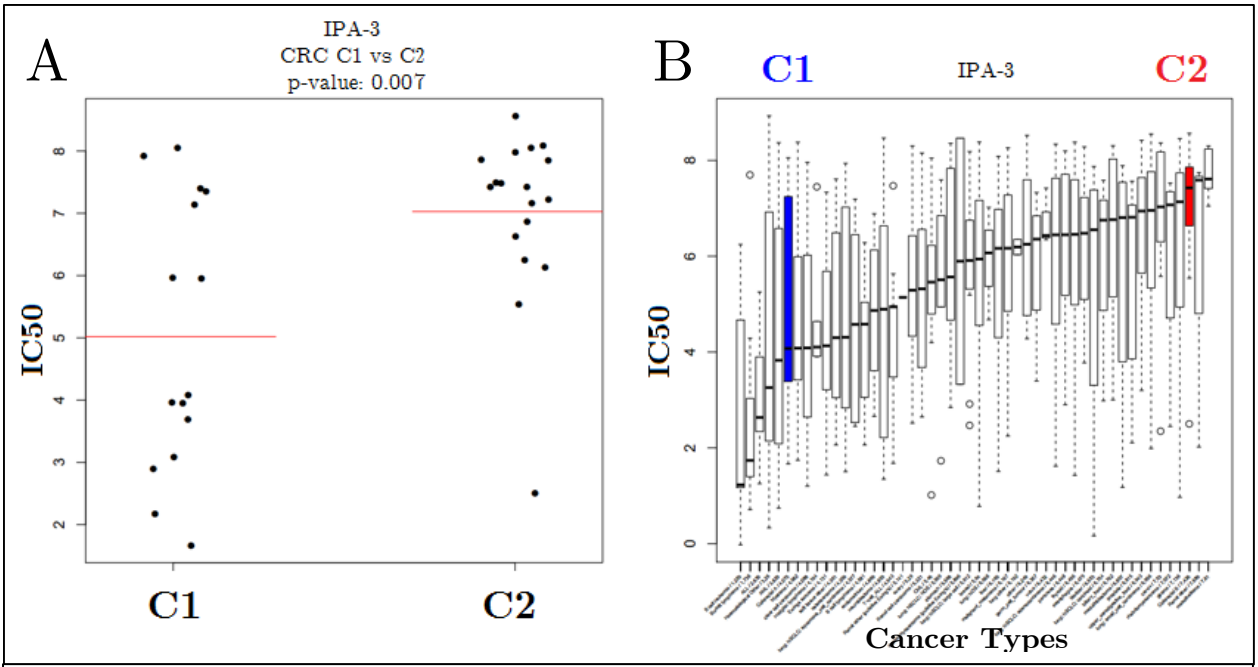

Figure 3.6: Comparison of IPA-3 response of CRC cell line sub-groups and global effectiveness of IPA-3.
IPA-3 IC50 distributions of clusters of CRC cell lines are shown in jitter plots **(A)** and all cancer types in boxplots **(B)**. Red horizontal lines in **(A)** represent average IC50 values for C1 and C2. Blue and red boxplots represent IPA-3 IC50 distributions C1 and C2 CRC clusters, respectively.

We also found some drugs for which the clusters exhibited differential drug responses. Cluster 1 cells were significantly more sensitive to IPA-3 than cluster 2 cells. Cluster 1 cells were also one of the most sensitive cell types to IPA-3 among all other cancer types, as cluster 2 cells were one of the most resistant cancer types compared to others (Figure 3.6).

This way, we developed a 20-gene signature that could identify distinct CRC cell line subtypes with differential drug response. However, it still remains to be validated in vitro.

3.2. Identification of Prognostic Subtypes in Tumors

Our next aim was to develop prognostic gene signatures in tumors using the same approach that we used for the identification of chemotherapeutic subtypes. We performed

hierarchical clustering with 1000 most variant genes using GSE17536 CRC dataset, but we could not identify distinct clusters with different prognostic outcomes according to the expression of those most variant genes (data now shown), probably due to heterogeneity of tumors.

We decided to develop new algorithms that we can use to find new prognostic gene markers. In concordance with the report published by Venet et al. [24] in which they claim that most of the random signatures with high number of genes were significantly associated with clinical outcome, we decided to find single-gene markers instead of multi-gene signatures. Therefore, we decided to develop two different R-based programs, called SSAT and USAT, which analyze normalized microarray datasets with different statistical methods and 12 different algorithms using clinical data to identify prognostic single-gene markers.

We analyzed three different microarray datasets with SSAT and USAT to identify candidate prognostic gene markers in CRC. We used GSE17536, GSE17537 with SSAT and GSE17536 and GSE41258 with USAT (Table 3.1) and identified different candidate genes with each approach. We tried to validate those results with qPCR experiments of identified genes by using a unique CRC cohort (Table 3.1).

3.2.1. SSAT

SSAT is an R-based program that can analyze microarray datasets by using Log-Rank test with different expression cut-offs of genes in order to find prognostic gene markers in cancers. SSAT uses only one probeset, which has the highest coefficient of variation value, for each gene. A gene is considered as significant if its corresponding probeset is significant ($p<0.05$) in Log-Rank test for given expression cut-off values.

Table 3.1: Clinical Characteristics and univariate CoxPH analysis results for GSE17536, GSE17537, GSE41258 and Ankara cohort.

Characteristics	GSE17536 (n = 177)				GSE17537 (n = 55)				GSE41258 (n = 182)				Ankara Cohort (n = 79)			
	No. of Patients	%	HR	p-value	No. of Patients	%	HR	p-value	No. of Patients	%	HR	p-value	No. of Patients	%	HR	p-value
Age																
<=60	59	33.3%			30	54.5%			62	34.1%			38	48.1%		
>60	118	66.7%	0.76	0.330	25	45.5%	1.03	0.968	120	65.9%	0.63	0.065	41	51.9%	1.60	0.191
Gender																
Female	81	45.8%			29	52.7%			86	47.3%			36	45.6%		
Male	96	54.2%	1.19	0.530	26	47.3%	0.57	0.375	96	52.7%	1.49	0.110	43	54.4%	0.93	0.840
TNM Stage																
1	24	13.6%			4	7.3%			28	15.4%			7	8.9%		
2	57	32.2%			15	27.3%			48	26.4%			-	0.0%		
3	57	32.2%			19	34.5%			49	26.9%			64	81.0%		
4	39	22.0%	3.62	1.4E-11	17	30.9%	13.00	0.001	57	31.3%	7.25	0.000	8	10.1%	3.42	0.001
Recurrence																
No	109	61.6%			36	65.5%			103	56.6%			NA	NA		
Yes	36	20.3%	42.90	3.1E-07	19	34.5%	3.5E+09	1.9E-07	36	19.8%	132.00	1.7E-06	NA	NA		
Other/Unknown	32	18.1%			-	0.0%			43	23.6%			NA	NA		
Grade																
Well Differentiated	16	9.0%			1	1.8%			NA	NA			31	39.2%		
Moderately Differentiated	134	75.7%			32	58.2%			NA	NA			43	54.4%		
Poorly Differentiated	27	15.3%	2.13	0.006	3	5.5%	3.26	0.230	NA	NA			1	1.3%	1.47	0.240
Other/Unknown	-	0.0%			19	34.5%			NA	NA			4	5.1%		
Tumor Localization																
Cecum	NA	NA			NA	NA			29	15.9%			4	5.1%		
Ascending Colon	NA	NA			NA	NA			31	17.0%	1.54	0.160	12	15.2%	1.03	0.950
Transverse Colon	NA	NA			NA	NA			15	8.2%	1.49	0.320	7	8.9%	0.75	0.690
Descending Colon	NA	NA			NA	NA			25	13.7%	0.92	0.820	2	2.5%	2.55	0.360
Sigmoid Colon	NA	NA			NA	NA			47	25.8%	0.83	0.530	18	22.8%	1.67	0.200
Rectum	NA	NA			NA	NA			14	7.7%	0.62	0.360	36	45.6%	0.77	0.470
Other/Unknown	NA	NA			NA	NA			21	11.5%			-	0.0%		

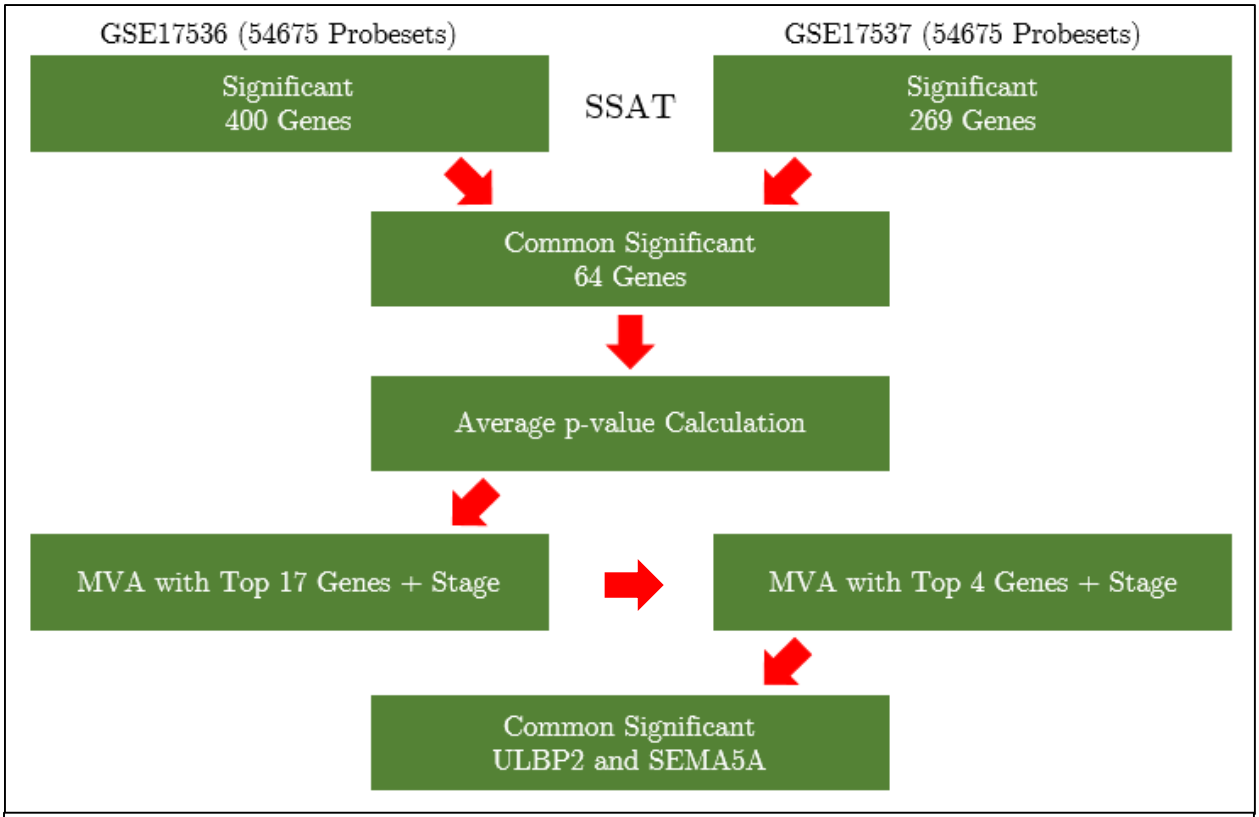

Figure 3.7: Schematic representation of SSAT approach.
Each green box shows a step for the selection of genes for further analysis. The number of significant genes in each step are indicated.

SSAT discretized the expression values for each gene as explained in Discretization of Expression Values section. We determined 8 different expression intervals for each gene. We used sequential combinations of those intervals for each gene to separate patients into high and low expression groups for Log-Rank test, e.g. ULBP2 (123) vs. ULBP2 (45678). We performed Log-Rank tests for 7 different interval cut-offs in 8 categories for each gene.

3.2.1.1. Gene Selection for Validation

We analyzed two colon cancer datasets, GSE17536 and GSE17537, for all the genes according to those interval cut-off values and identified the ones significant in both datasets. We identified 400 and 269 and 64 genes with different expression cut-offs in GSE17536, GSE17537 and common, respectively (Figure 3.7). We averaged Log-Rank p-

Table 3.2: Multivariate Cox proportional Hazard regression results and corresponding statistics of selected genes in GSE17536 and GSE17537 datasets.

	Coef.	SE	Wald	df	p-value	OR	95.0% CI for OR Lower	Upper
GSE17536 Forward Wald								
AJCC Stage	1.650	0.234	49.564	1	0.000	5.205	3.288	8.239
ULBP2 12vs38	1.102	0.326	11.389	1	0.001	3.009	1.587	5.705
SEMA5A 13vs48	-0.744	0.337	4.880	1	0.027	0.475	0.246	0.920
PCDH7 13vs48	1.137	0.412	7.608	1	0.006	3.117	1.390	6.990
GSE17536 Backward Wald								
AJCC Stage	1.731	0.247	48.962	1	0.000	5.648	3.478	9.173
ULBP2 12vs38	0.987	0.327	9.109	1	0.003	2.682	1.413	5.090
SEMA5A 13vs48	-0.778	0.333	5.463	1	0.019	0.459	0.239	0.882
PCDH7 13vs48	1.017	0.424	5.741	1	0.017	2.765	1.203	6.354
EBF1 12vs38	0.496	0.297	2.784	1	0.095	1.642	0.917	2.940
GSE17537 Forward Wald								
AJCC Stage	2.627	0.819	10.284	1	0.001	13.829	2.777	68.870
ULBP2 12vs38	1.759	0.704	6.243	1	0.012	5.808	1.461	23.090
SEMA5 13vs48	-1.358	0.667	4.149	1	0.042	0.257	0.070	0.950
GSE17537 Backward Wald								
AJCC Stage	3.105	1.147	7.330	1	0.007	22.312	2.357	211.224
ULBP2 12vs38	2.307	0.937	6.060	1	0.014	10.041	1.600	63.003
SEMA5 13vs48	-1.882	0.856	4.833	1	0.028	0.152	0.028	0.815

values of the genes with the same interval cut-off in both datasets and selected top 17 and 4 genes in GSE17536 and GSE17537, respectively, for multivariate logistic regression with forward and backward Wald selection method. ULBP2 and SEMA5A were the only significant independent prognostic genes in both datasets (Table 3.2). We decided to measure qPCR expression values of ULBP2 and SEMA5A in Ankara cohort for the validation of its association with prognosis.

3.2.1.2. Validation of ULBP2 and SEMA5A as Prognostic Gene Markers

Dr. Emre Dayanç with his intern students and Seçil Demirkol performed qPCR experiments for ULBP2 and SEMA5A in Ankara cohort and we analyzed the qPCR expression values with Log-Rank test and Kaplan-Meier curves in order to validate the association of ULBP2 with prognosis. We could not use SSAT to validate qPCR expression values of ULBP2 and SEMA5A, because SSAT was developed to analyze microarray expression, for which the values were normalized by GC-RMA method to be between 0 and 16, by categorizing them into 8 categories as explained in previous section.

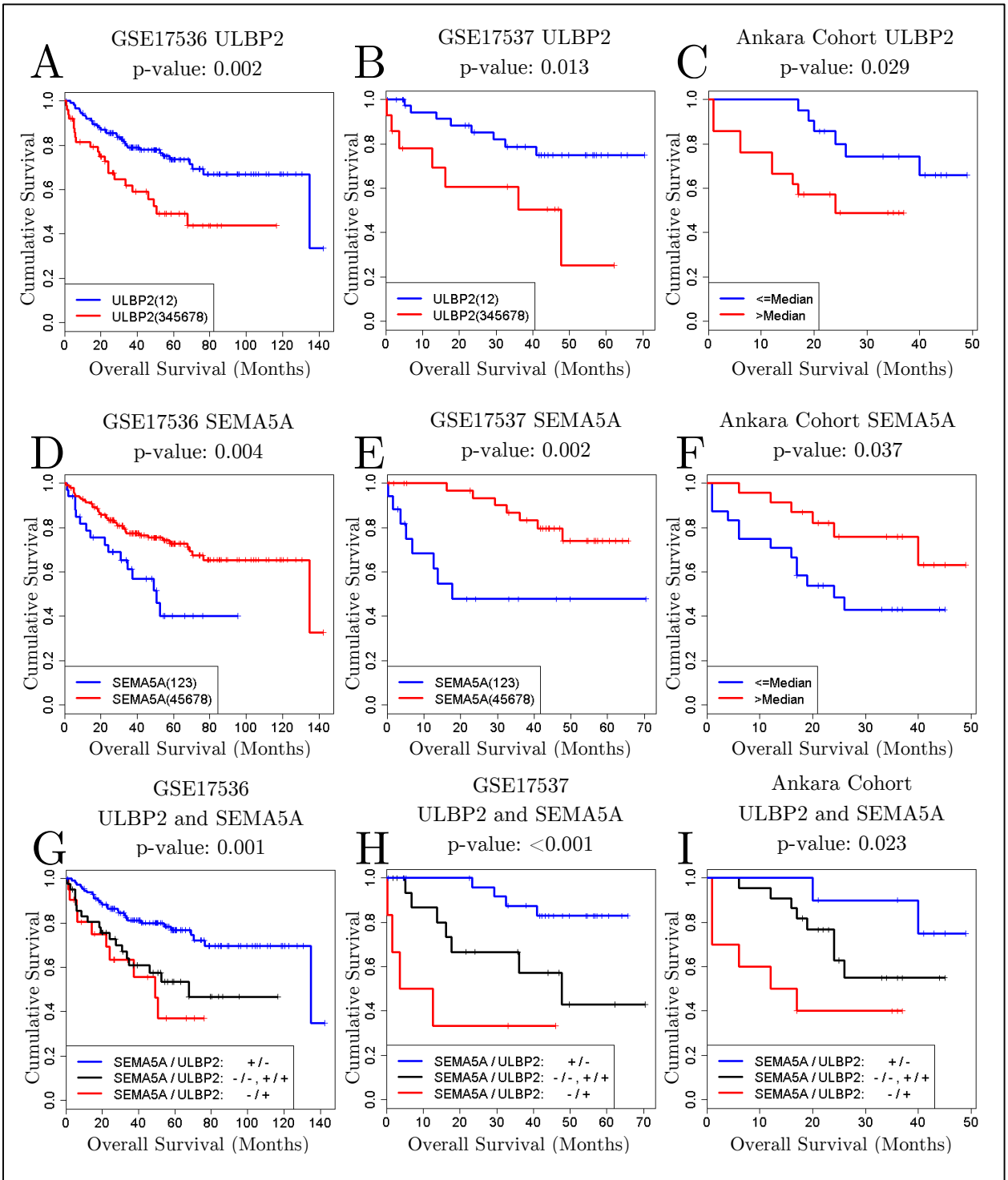

Figure 3.8: Kaplan-Meier Curves for ULBP2, SEMA5A and combination of both. Kaplan-Meier Curves for ULBP2 (A, B, C), SEMA5A (D, E, F) and combination of both (G, H, I) in GSE17536, GSE17537 and Ankara cohorts, respectively. **(A, B, C, D, E, F)** Blue and red colors represent low and high expression. **(G, H, I)** Blue, black and red colors represent good, intermediate and poor prognosis patient groups. The expression cut-off values are 4 and 6 for ULBP2 and SEMA5A, respectively, in GSE17536 and GSE17536 and median values in Ankara cohort for both genes.

Table 3.3: Stepwise multivariate Cox proportional hazard regression results and corresponding statistics of prognosis separation by ULBP2 and SEMA5A in Ankara cohort.

						95.0% CI for OR	
Parameters	**Coef.**	**SE**	**Z**	**p-value**	**OR**	**Lower**	**Upper**
AJCC Stage	1.358	0.565	5.7806629	0.016	3.890	1.285	11.7737488
Good/Int/Poor	0.936	0.425	4.8453446	0.028	2.549	1.108	5.86227734

Because qPCR values were not exposed to such a normalization, we could not apply SSAT directly to qPCR values. Instead, we used median expressions as expression thresholds to separate patients into high and low expression groups and performed Log-Rank test for those separations in Ankara cohort (Figure 3.8). We identified high expression of ULBP2 and low expression of SEMA5A as significantly associated with poor prognosis in all three cohorts. We also showed that combined analysis of ULBP2 and SEMA5A further identified better poor and good prognosis groups, as well as an intermediary prognosis group, in all three cohorts (Figure 3.8).

We also performed stepwise multivariate Cox Proportional Hazard Regression analysis with backward Wald method to see whether ULBP2 and SEMA5A were prognostic classifiers independent of other clinical parameters, like age, gender, stage and grade, in Ankara cohort. The analysis results showed that ULBP2 and SEMA5A together could define good, intermediate and poor prognostic subgroups independent of other clinical parameters (Table 3.3).

3.2.1.3. Prediction of Chemotherapy Benefit with ULBP2 and SEMA5A

We also analyzed predictive capability of ULBP2 and SEMA5A in CRC cell lines from CGP database. We identified many drugs effective to poor prognosis sub-group of CRC cell lines determined by the combination of ULBP2 and SEMA5A gene expression. One of the drugs with the greatest differential effectiveness CRC cell lines was AZ628 (Figure 3.9). According to global effectiveness of AZ628, poor prognosis sub-group of CRC cell

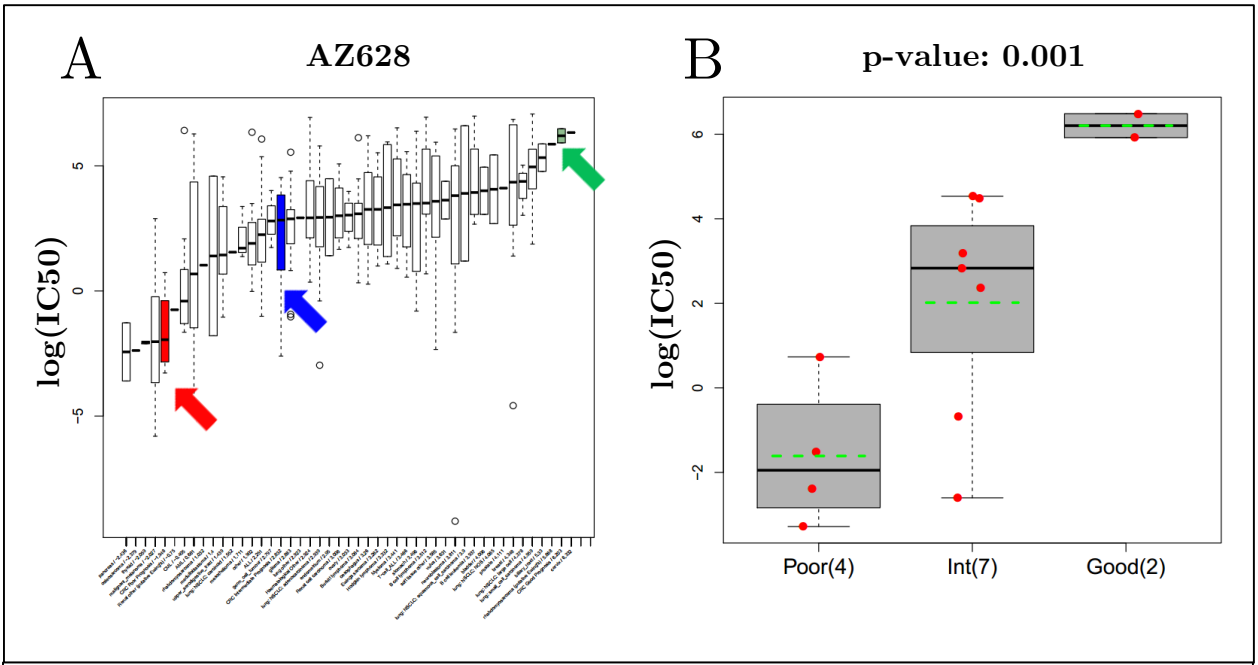

Figure 3.9: Global effect distribution of AZ628 (A) and differential response of prognostic CRC sub-groups to AZ628 (B).

(A)The distributions of AZ628 log(IC50) values are shown by boxplots for each cancer type. CRC cell lines were divided into three prognostic sub-groups: poor (red), intermediate (blue) and good (green) prognosis groups. (B) The distributions of AZ628 log(IC50) values of prognostic CRC sub-groups were compared with ANOVA. Each red dot represent a cell line with response to AZ628 shown in the y-axis. The distribution of the drug responses for each sub-group are shown in boxplots. Green dashed lines represent the average drug responses of sub-groups.

lines was one of the most sensitive cancer types to AZ628, although good prognosis subtype was almost the most resistant one among all cancer types and intermediary sub-group was somewhere in between with an average response (Figure 3.9A).

This way, we showed that ULBP2 and SEMA5A might have chemotherapy benefit prediction power, besides prognostic power. However, these findings still remain to be validated in vitro.

3.2.1.4. Conclusion

We identified ULBP2 and SEMA5A as independent prognostic factors in colon cancer with SSAT approach and validated their association with prognosis in an independent cohort. Moreover, we validated the independence of their association with prognosis when they were used together to assess the prognostic subgroups in colon cancer.

3.2.2. USAT

After SSAT approach, we decided to generate a more robust survival analysis tool, because SSAT was using only Log-Rank test with 7 expression cut-off values that we determined. Therefore we decided to develop an unsupervised approach with more advanced statistical methods and we call that method Unsupervised Survival Analysis Tool (USAT).

USAT uses Cox Proportional Hazard Regression (CoxPH), Maximally Selected Rank Statistics (Maxstat) and Log-Rank tests with different types of expression data and stage stratification to determine stage independent prognostic gene markers. CoxPH is used to determine hazard ratio for one unit increase in expression or between categories, Maxstat is used to determine the best expression cut-off value that can separate patients into low and high expression groups that exhibit the best prognosis difference and Log-Rank test is used to see whether the prognosis difference between high and low expression groups is significant or not. Any gene is considered as significant when any of their probesets are significant in all those three statistical tests.

Before analyzing any datasets with USAT, we performed secondary analyses to calculate true and false positivity rates in order to determine how reliable USAT's results are. Finally, we developed different versions of USAT by generating different models to

improve true and false positivity rates and started analyzing different CRC datasets with the latest version.

3.2.2.1. True Positivity of USAT

The first version of USAT analyzed stage-stratified categorical expression values with CoxPH, continuous expression values with Maxstat and Log-Rank tests (CGS-MC-LC, see Model Generation in Methods section). To determine the true positivity of this approach, we used 48-gene CRC prognostic gene list published by O'Connell et al. [12] as reference. In other words, we checked how many of those 48 genes were present as significant in USAT results of the test set (GSE17536). We were able to analyze 44 of those 48 genes due to lack of probesets for 4 of the gene symbols in the HG-U133 Plus 2.0 platform. USAT validated 21 genes among those 44 genes (Table 3.4), which makes the true positivity of the first version of USAT ~48%. 48% true positivity suggested that we would not be validating half of the genes, which we found by USAT in one dataset, with another dataset. Therefore, we decided to generate different models with different combinations of statistical tests, types of expression and stage stratification to see whether we can improve USAT's true positivity.

3.2.2.1.1. *USAT with Different Models*

In the first version of USAT, we were using categorical expression values and stage stratification with CoxPH, continuous expression values with Maxstat and Log-Rank tests (CGS-MC-LC). Latter, we generated 11 new models to find the best one that could find the maximum number of 48 genes published by O'Connell et al. as prognostic genes in CRC. When we analyzed the test set with 12 different models of USAT, we realized that all the models could identify similar numbers of genes. However, each model had the ability to identify different genes than the common ones in all models, because USAT was able to identify 33 of 44 genes when we considered the union of the results acquired by

Table 3.4: Number of significant genes among 48 genes in USAT models.

Models	Number of Genes (Total 44)
CGS-MC-LC	21
CG-MC-LC	20
CGS-MG-LG	24
CG-MG-LG	23
CCS-MG-LG	23
CC-MG-LG	23
CCS-MC-LC	22
CC-MC-LC	21
MC-C12S-LC	23
MC-C12-LC	23
MG-C12S-LG	23
MG-C12-LG	24
Union	**33**

all 12 models (Table 3.4). In other words, if we consider a gene as significant when it is significant in at least one of the models, then we are able to identify 33 of 44 genes with USAT. Therefore, we decided to use all 12 models in USAT, thus increasing the true positivity to from 48% to 75%.

3.2.2.2. False Positivity of USAT

We calculated false positivity rate of USAT by calculating the proportion of the probesets that exhibit opposite association with prognosis in common results of GSE17536 and GSE41258 CRC microarray datasets. USAT identified 3985 probesets corresponding to 3270 genes as CRC prognostic gene markers in GSE17536 dataset and 2408 probesets corresponding to 2025 genes in GSE41258 dataset (Figure 3.10). We found that 433 probesets corresponding to 366 genes were common in both datasets. 395 out of those 433 probesets have the same significant association with prognosis in both datasets, although the remaining 38 probesets have opposite associations with prognosis in each dataset. Thus, the proportion of the probesets that signify opposite associations with prognosis in the datasets is nearly 9%. In other words, the false positivity rate of USAT is 9% and we expect to invalidate almost 1 out of 10 genes that we identify by USAT.

3.2.2.3. Gene Selection for Validation

USAT identified 395 probesets corresponding to 329 genes as candidate CRC prognosis gene markers that exhibit the same association with prognosis in common results of

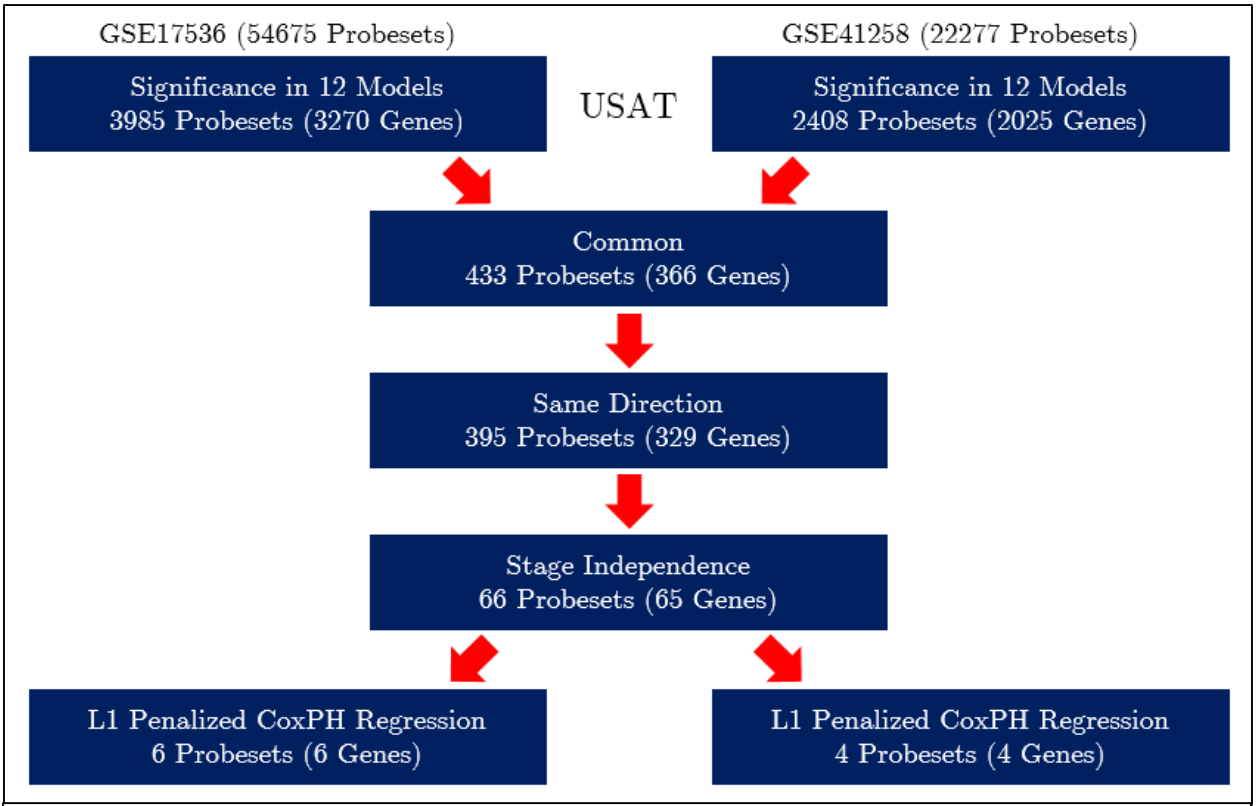

Figure 3.10: Schematic representation of USAT approach.
Each blue box shows a step for the selection of genes for further analysis. The number of significant probesets and corresponding genes in each step are also indicated.

GSE17536 and GSE41258 (Figure 3.10). In order to determine the genes for validation, we narrowed down the gene list according to independence of cancer stage and applied L1 penalized CoxPH regression to each dataset to find the best candidates. After selection through stage independence, we got 66 probesets corresponding to 65 genes. We applied L1 penalized CoxPH regression to those 66 probesets and identified 6 probesets corresponding to 6 genes in GSE17536 and 4 probesets corresponding to 4 genes in GSE41258 (Figure 3.10).

We selected 5 of 6 genes (DUSP10, PCSK5, TGFB1I1, PLS3 and CLCN7) from GSE17536 and 3 of 4 genes (PTRF, CLDN3 and KLF9) from GSE41258 for validation. Table 3.5 and Table 3.6 show the results of the statistical analyses used in USAT with continuous expression values for GSE17536 and GSE41258 and the results of 12 models of USAT in

Table 3.5: Results of USAT analyses with continuous microarray expression values for GSE17536 and GSE41258.

GSE17536		CoxPH		Maxstat		Log-Rank			Chi-Sq Test	
Probeset ID	Gene Name	p	HR	p	Threshold	p	MS-L	MS-H	p	Percentage (L - H)
208789_at	PTRF	6.45E-05	1.649	3.80E-05	9.984	7.88E-07	NA	45.92	0.096	65% - 35%
213652_at	PCSK5	2.83E-04	1.476	5.54E-05	6.077	2.32E-06	NA	45.92	0.620	69% - 31%
209651_at	TGFB1I1	5.19E-04	1.648	4.23E-04	8.710	1.19E-05	NA	57.73	0.095	58% - 42%
215501_s_at	DUSP10	3.19E-03	1.526	1.05E-02	7.190	6.53E-04	NA	76.6	0.085	39% - 61%
203543_s_at	KLF9	4.83E-03	1.425	1.78E-02	9.837	7.14E-04	NA	45.92	0.191	87% - 13%
209235_at	CLCN7	4.29E-02	0.369	4.71E-02	5.843	1.42E-03	52.5	134.9	0.066	19% - 81%
203953_s_at	CLDN3	4.51E-02	0.866	4.32E-02	10.334	2.12E-03	67.82	134.9	0.445	32% - 68%
201215_at	PLS3	6.19E-03	2.426	3.08E-02	11.992	2.46E-03	NA	134.9	0.556	29% - 71%

GSE41258		CoxPH		Maxstat		Log-Rank			Chi-Sq Test	
Probeset ID	Gene Name	p	HR	p	Threshold	p	MS-L	MS-H	p	Percentage (L - H)
208789_at	PTRF	1.33E-02	1.301	2.66E-02	10.177	4.61E-04	169	38	0.065	78% - 22%
213652_at	PCSK5	4.16E-01	1.141	1.63E-01	4.028	1.35E-02	NA	169	0.106	49% - 51%
209651_at	TGFB1I1	1.18E-02	1.291	3.42E-02	8.279	9.81E-04	169	47	0.181	76% - 24%
215501_s_at	DUSP10	1.22E-03	1.489	1.57E-02	6.646	6.31E-04	169	76	0.063	58% - 42%
203543_s_at	KLF9	1.92E-02	1.274	2.74E-02	5.796	1.54E-03	NA	99	0.293	52% - 48%
209235_at	CLCN7	2.08E-01	0.793	8.40E-03	4.599	8.90E-05	28	169	0.192	15% - 85%
203953_s_at	CLDN3	1.21E-01	0.889	8.24E-02	11.949	7.61E-03	156	NA	0.380	73% - 27%
201215_at	PLS3	2.09E-01	1.218	1.14E-01	12.073	5.43E-03	169	72	0.215	84% - 16%

both datasets, respectively. USAT identified PTRF, PCSK5, TGFB1I1, DUSP10, KLF9 and PLS3 as poor prognosis markers and CLCN7 and CLDN3 as good prognosis markers in both datasets (Table 3.5 and Table 3.6). Although PTRF, PCSK5, TGFB1I1, DUSP10 and PLS3 were significant for all 12 models in only GSE17536, KLF9 is the only gene that was significant for all 12 models in both datasets and GSE41258 itself (Table 3.6). Chi-square test results also showed that there was no significant association between cancer stage and expression of those genes in both datasets (Table 3.5).

3.2.2.4. Validation of Microarray Expression with qPCR Expression

We identified candidate genes for validation in two different microarray datasets with USAT and we used qPCR to measure the expression of those candidate genes in Ankara cohort for validation. The validation of microarray expression by qPCR expression is still a controversial issue. There are many reports that emphasize the general problems on validation of microarray expression with qPCR expression, like variability in hybridization efficiency of sequences, differences in design, differences in probes and target labeling etc.

Table 3.6: Results of USAT's 12 Models for GSE17536 and GSE41258.

GSE17536			
Gene Name	**Result Number**	**HR>1**	**HR<1**
PTRF	12	12	0
PCSK5	12	12	0
TGFB1I1	12	12	0
DUSP10	12	12	0
KLF9	12	12	0
CLCN7	5	0	5
CLDN3	4	0	4
PLS3	12	12	0

GSE41258			
Gene Name	**Result Number**	**HR>1**	**HR<1**
PTRF	4	4	0
PCSK5	2	2	0
TGFB1I1	7	7	0
DUSP10	7	7	0
KLF9	12	12	0
CLCN7	2	0	2
CLDN3	4	0	4
PLS3	4	4	0

even in the same cohorts [25, 26, 27, 28]. In this step, we try to validate microarray results of two cohorts with qPCR expression of another cohort. In other words, we have both different gene expression measurement techniques and different cohorts for validation. Therefore, we wanted to make sure that microarray chips and qPCR primers detect and measure the expression of the same transcripts, so that we measure the same thing in different cohorts. For that reason, we performed qPCR experiments for the selected genes using 7 different CRC cell lines whose microarray expression data was available in CGP database. We compared qPCR and microarray expression of the selected genes in those 7 cell lines and found out that the qPCR primers of 6 of 8 genes (PTRF, TGFB1I1, DUSP10, KLF9, CLCN7 and CLDN3) could validate the microarray expression of the cell lines (Figure 3.11). So, we decided to perform qPCR experiments in our cohorts with the genes whose qPCR expression validated the microarray counterparts.

3.2.2.5. Validation of USAT Results

We performed qPCR experiments in Ankara cohort for the genes whose microarray expression were validated with qPCR expression. We performed USAT analysis in Ankara cohort using qPCR expression values. Table 3.7 shows the results of USAT analysis performed with continuous qPCR expression values in Ankara cohort for the selected

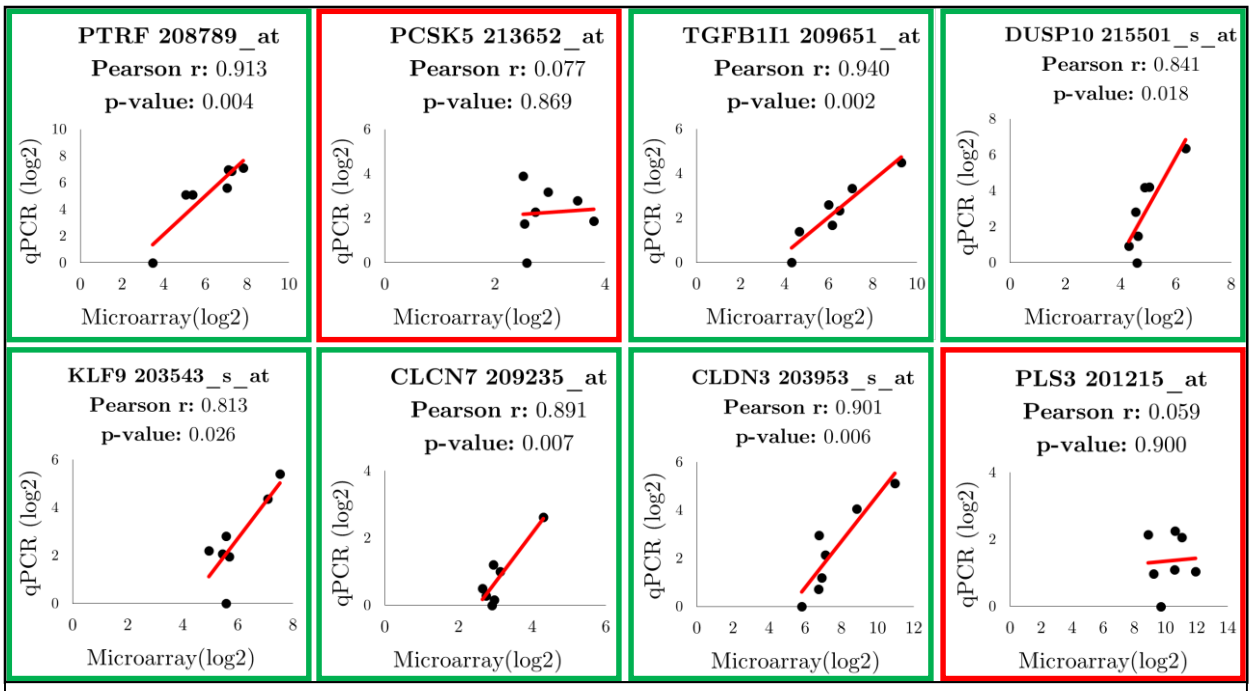

Figure 3.11: Concordance between qPCR and microarray expression of the genes selected for validation.

Each dot represents the microarray and qPCR expression of the indicated gene in the x-axis and y-axis, respectively. Red line is the best linear fit. Pearson r and corresponding p-values are indicated on top of the graphs. Green and red squares indicate concordance and discordance, respectively, between microarray and qPCR expression.

genes for validation and corresponding 12-model results. As shown in Table 3.7, the only gene whose association with prognosis could be validated by qPCR experiments was CLCN7. Surprisingly, all the other genes identified in microarrays were identified as either oppositely associated with prognosis (PTRF, TGFB1I1, DUSP10) or unrelated to prognosis at all (KLF9, CLDN3) in Ankara cohort.

3.2.2.5.1. *Log-Rank Analysis with Multiple Cut-off Values (LRMC)*

After we identified opposite associations with prognosis in Ankara cohort for most of the selected genes, we thought that the problem could be in the analysis of each gene with a single expression threshold value determined by Maxstat analysis. Maxstat analysis determines an expression cut-off that gives the biggest absolute log-rank statistics.

Table 3.7: Results of USAT analyses with qPCR expression for Ankara cohort.

	CoxPH		Maxstat		Log-Rank				12 Models		
Gene Symbol	p	HR	p	Threshold	p	MS-L	MS-H	Percentage (L - H)	Result Number	HR>1	HR<1
PTRF	0.065	0.806	0.009	5.365	0.001	32	NA	74% - 26%	8	0	8
TGFB1I1	0.039	0.771	0.049	4.778	0.005	32	NA	62% - 38%	6	0	6
DUSP10	0.023	0.543	0.010	1.012	<0.001	20	NA	19% - 81%	4	0	4
KLF9	0.050	0.753	0.056	3.254	0.006	40	NA	67% - 33%	0	0	0
CLCN7	0.018	0.744	0.077	2.216	0.003	21	NA	14% - 86%	6	0	6
CLDN3	0.032	0.758	0.120	5.455	0.017	40	NA	77% - 23%	0	0	0

Analysis of each gene with a single expression cut-off may have led us to ignore other possible cut-off values that show different associations with prognosis for the same gene. Therefore, we decided to analyze all possible expression cut-off values with Log-Rank test for each gene in terms of the direction of their association with prognosis. Figure 3.12 shows a representative LRMC analysis result for PTRF in Ankara cohort. The lower graph in Figure 3.12 shows the Log-Rank p-values in the y-axis calculated for expression threshold values in the x-axis, which separate patients into high and low expression groups for Log-Rank test. Although most of the threshold values could not pass 0.05 significance level, high expression of PTRF is associated with good prognosis in almost all the threshold values, including those were significant as well. That graph suggests that whichever PTRF expression cut-off value is selected for grouping patients into high and low expression groups, high PTRF expression suggests an association with good prognosis in Ankara cohort. So, that graph gives insight into the general behavior of a gene regarding the direction of the overall association with prognosis.

We used LRMC graphs to understand general behavior of our genes in both microarray datasets and Ankara cohort (Figure 3.13). The only gene for which we identified significant expression cut-offs that exhibited associations with prognosis in both microarray datasets and Ankara cohort in the same direction was CLCN7 (green box in Figure 3.13). High CLCN7 expression was associated with good prognosis in all three cohorts, especially for the expression cut-offs just below the 1st quartile of the expression values.

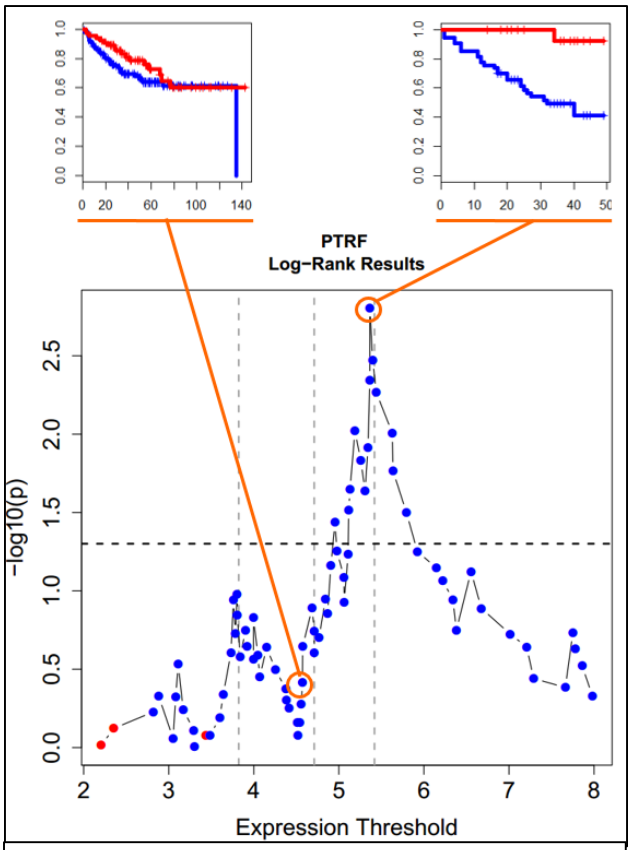

Figure 3.12: LRMC analysis of PTRF in Ankara cohort.

In the scatter plot, each dot represents a Log-Rank p-value in y-axis for a given expression cut-off value in x-axis, by which the samples are separated into high and low expression groups for Log-Rank analysis. Blue and red dots represent association with good and poor prognosis, respectively. Horizontal dashed line shows 0.05 p-value threshold. Vertical gray dashed lines show 1st, 2nd and 3rd quartiles from left to right, respectively. Kaplan-Meier curves for a non-significant (left) and a significant (right) Log-Rank results are shown above. Blue and red lines represent the cumulative survival as a function of overall survival for the patients who have low and high expression of the indicated gene, respectively, based on the expression cut-offs indicated in the scatter

High expressions of PTRF, TGFB1I1, DUSP10 and KLF9 (the genes in red boxes in Figure 3.13) were associated with poor prognosis in GSE17536 and GSE41258 for all significant expression cut-offs, although the high expressions of those genes were associated with good prognosis in Ankara cohort for all significant expression cut-offs. CLDN3 (in yellow box in Figure 3.13) was the only gene for which we could not find any significant expression cut-off values. Actually, we got a very similar LRMC graph in Ankara cohort with GSE41258 for CLDN3. However, neither graphs looked like the CLDN3 LRMC graph of GSE17536, for which there were many significant expression cut-offs associated with good prognosis. Nevertheless, all the expression cut-offs for CLDN3 in all three cohorts suggested an association with good prognosis despite lack of significance.

One of the most interesting points was that the LRMC graphs of microarray datasets and Ankara cohort were very similar to each other, especially for PTRF and

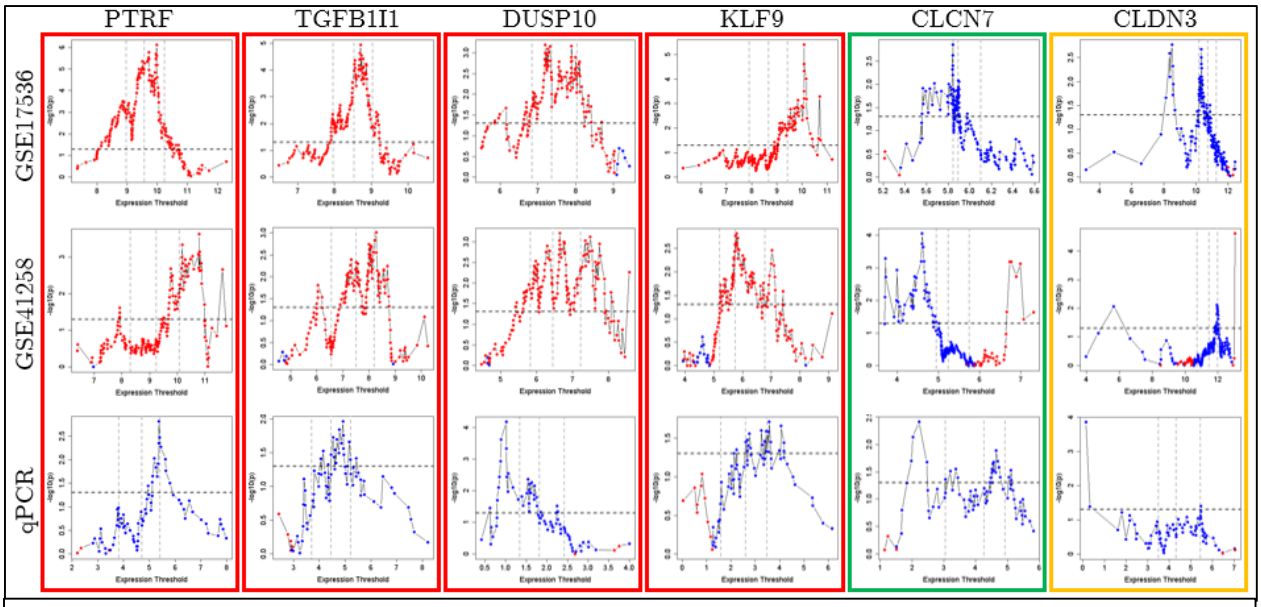

Figure 3.13: LRMC analysis of the selected genes in GSE17536, GSE41258 and Ankara cohort.

Genes and cohorts are represented in columns and rows, respectively. Each dot represents a Log-Rank p-value in y-axis for a given expression cut-off value in x-axis, by which the samples are separated into high and low expression groups for Log-Rank analysis. Blue and red dots represent association with good and poor prognosis, respectively. Horizontal black dashed lines show the 0.05 p-value threshold. Vertical gray dashed lines show 1st, 2nd (median) and 3rd quartiles from left to right, respectively. Red, green and yellow rectangles show the opposite, similar and insignificant associations with survival, respectively.

TGFB1I1, although they exhibited opposite associations with prognosis. PTRF LRMC graphs contained two different peaks, which were around the 3rd quartiles and 1st quartiles of the expression values, in LRMC graphs of all three cohorts. Moreover, the most significant PTRF expression cut-off values were around the 3rd quartile of the expression values in all cohorts. In TGFB1I1 LRMC graphs, the most significant expression cut-offs were between the median and the 3rd quartile of the expression values in all the cohorts. KLF9 LRMC graph also contained similar patterns in GSE41258 and Ankara cohort. The smaller KLF9 expression cut-off values suggested opposite associations (despite being insignificant) compared to the greater KLF9 expression cut-offs in both GSE41258 and Ankara cohort.

3.2.2.6. Further Analyses to Explain Opposite Results

Such opposite results led us to further analyze the data for an explanation. We thought that theoretically there could be explanations of the opposite results: (1) there could be two different CRC subtypes for which the genes exhibiting different associations with prognosis. (2) Ankara cohort was quite different colorectal cancer cohort than microarray datasets.

We performed hierarchical clustering analyses with different gene lists to identify subtypes like Ankara cohort. We also compared the expressions of the selected genes in GSE17536, GSE41258 and Ankara cohort to see whether Ankara cohort was a different cohort than the microarray datasets.

3.2.2.6.1. *Identification of CRC Subtypes like Ankara Cohort*

For the first idea, we tried to identify different CRC subtypes, for which the genes exhibit associations with prognosis opposite to each other or opposite to overall association, by performing whole transcriptome hierarchical clustering (WTHC) analysis for unknown subtypes and unsupervised hierarchical clustering analysis with published gene lists for known subtypes. We performed WTHC analysis with GSE17536 gene expression data and LRMC analyses for the subtypes defined by WTHC (Figure 3.14). WTHC analysis identified 4 different clusters (Figure 3.14A). We performed LRMC analyses for the selected genes using the samples in each cluster separately (Figure 3.14B). Those results showed that the significance of the association of the genes with prognosis differed according to the clusters. Interestingly, LRMC results showed that only cluster 3 gave the results that we got by using whole dataset, although it had a very low sample size compared to other clusters. Most of the genes were unrelated to prognosis in cluster 1, although it includes almost half of the samples. Nevertheless, there was no clusters giving similar results to those we had got from Ankara cohort. In other words, we could not

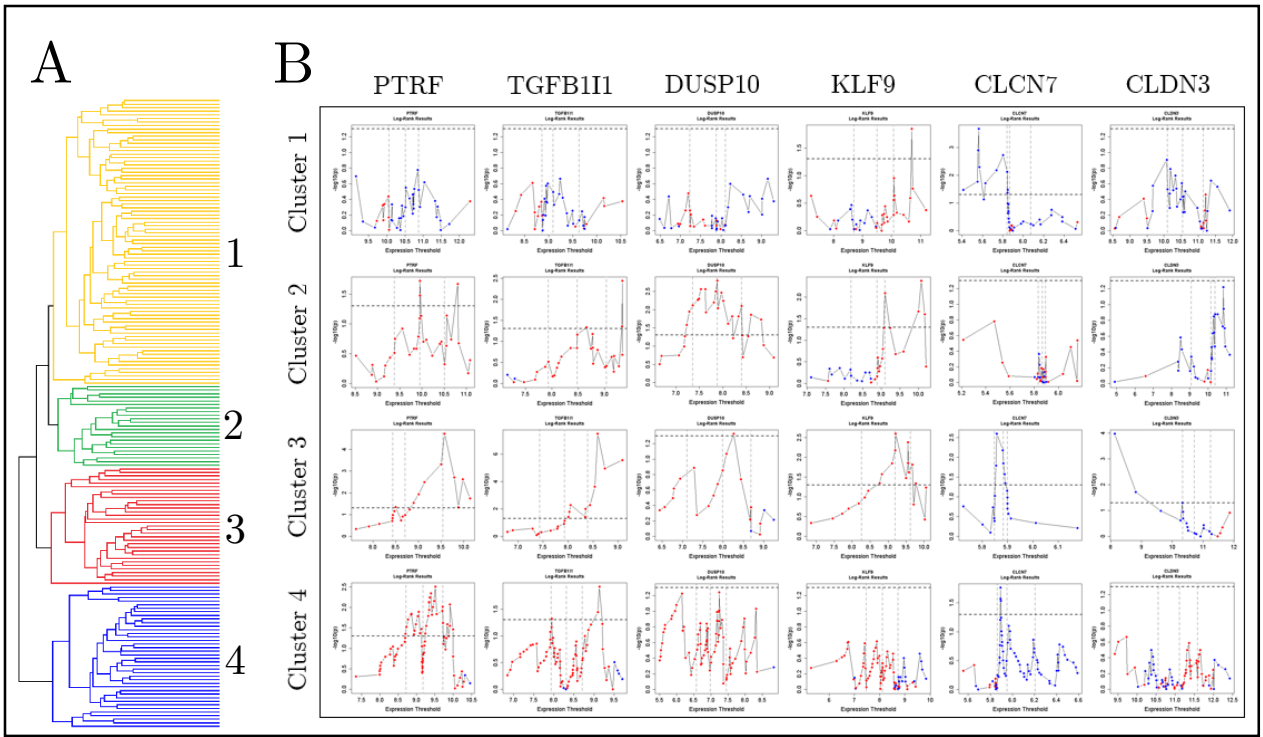

Figure 3.14: WTHC analysis of GSE17536 and LRMC analyses of selected genes for the clusters.

(A) WTHC analysis of GSE17536 identified 4 distinct clusters. (B) LRMC analysis of the selected genes (in columns) shows the association of each gene in distinct clusters of GSE17536 (in rows). Each dot represents a Log-Rank p-value in y-axis for a given expression cut-off value in x-axis, by which the samples are separated into high and low expression groups for Log-Rank analysis. Blue and red dots represent association with good and poor prognosis, respectively. Horizontal black dashed lines show the 0.05 p-value threshold. Vertical gray dashed lines show 1st, 2nd (median) and 3rd quartiles from left to right, respectively.

identify CRC subtypes that exhibited opposite association with prognosis by WTHC analysis.

We also used previously published prognostic gene lists by other groups [12, 29, 30, 31, 32, 33] to define different prognostic subtypes in CRC. We performed unsupervised hierarchical clustering in GSE17536 using O'Connell et al.'s [12] prognostic 48-gene list and identified 2 distinct clusters (Figure 3.15A). Unfortunately, LRMC analyses of both clusters showed that neither clusters resembled Ankara cohort in terms of the association of the genes with survival (Figure 3.15B). We performed similar analyses for many gene

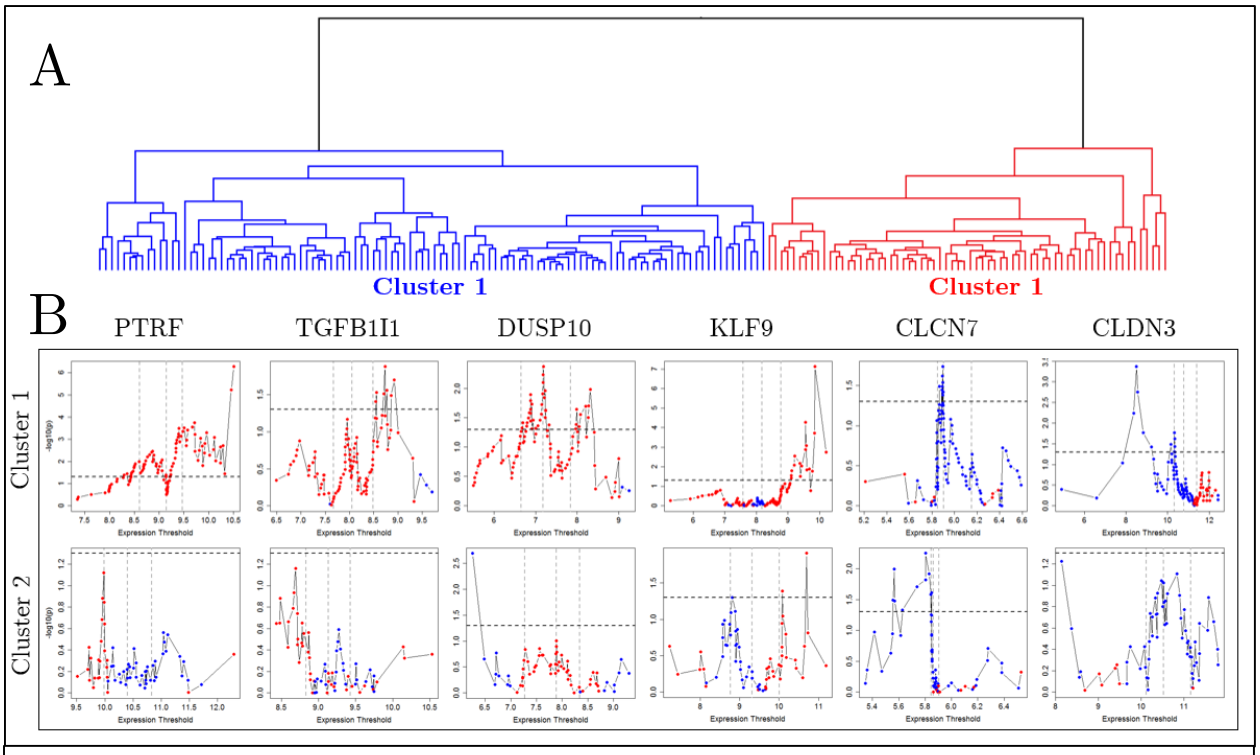

Figure 3.15: Hierarchical clustering of GSE17536 based on 48-gene list and LRMC analyses of selected genes based on the clusters.

(A) Hierarchical clustering of GSE17536 samples with the expression of O'Connell et al.'s 48 prognostic genes identified two distinct clusters. **(B)** LRMC analyses of the selected genes (in columns) shows the association of each gene in both clusters (in rows) of GSE17536. Blue and red dots represent association with good and poor prognosis, respectively. Horizontal black dashed lines show the 0.05 p-value threshold. Vertical gray dashed lines show 1st, 2nd (median) and 3rd quartiles from left to right, respectively.

lists, which identify different CRC subtypes with different biological backgrounds, published by Loboda et al. [29] (EMT), Marisa et al. [30] (prognosis), Sadanandam et al. [31] (response to chemotherapy and stemness), Sveen et al. [32] (prognostic) and Budinska et al. [33] (Morphology) but, like O'Connell et al.'s gene list, those gene lists could not identify subtypes resembling Ankara cohort in terms of association of the genes with prognosis (data not shown).

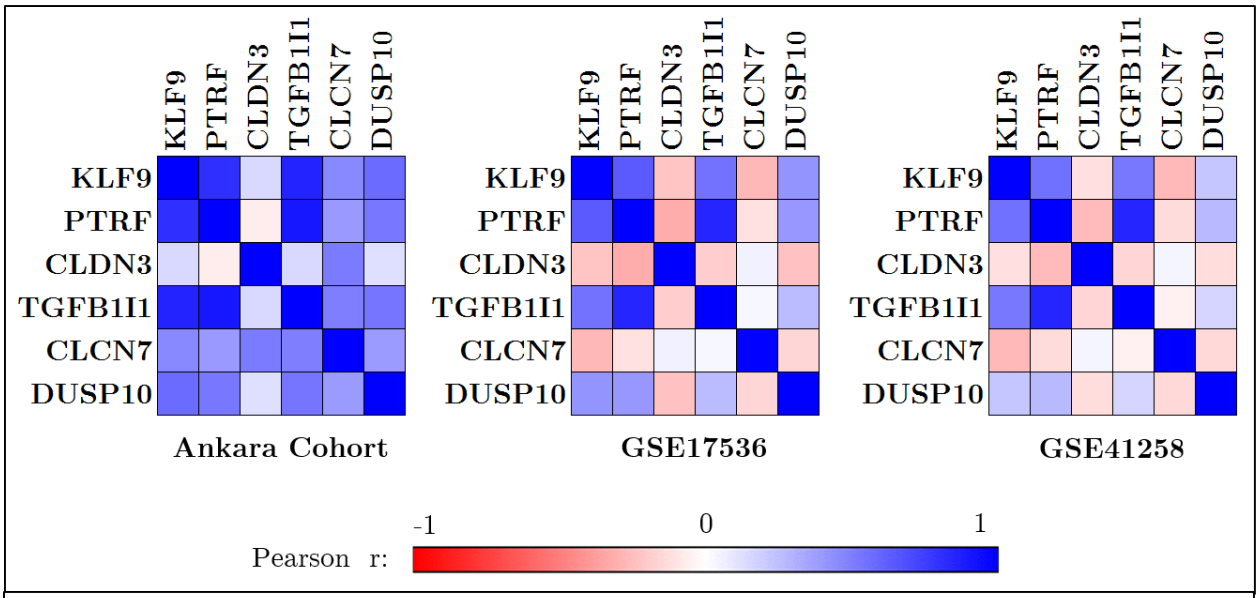

Figure 3.16: Pearson r correlation heat-maps of the expression of the selected genes for Ankara cohort, GSE17536 and GSE41258.
The qPCR expressions in Ankara cohort and the microarray expressions in GSE17536 and GSE41258 were used for correlation calculations. The correlation values were calculated for the expressions within the same cohort. Blue, red and white colors show positive, negative and no correlation, respectively.

3.2.2.6.2. *Comparison of Cohorts by Correlation between Genes*

Our second idea was that Ankara cohort and microarray cohorts might be quite different CRC cohorts. To check whether this was the case, we examined the relationship between the genes within each cohort using Pearson's r correlation coefficients. The correlation heat-maps in Figure 3.16 showed that almost all six genes in Ankara cohort correlated positively with each other. On the other hand, CLCN7 and CLDN3 correlated negatively or did not correlate at all with other genes in GSE17536 and GSE41258, unlike Ankara cohort.

This data suggested us that the biological backgrounds of Ankara cohort and microarray cohorts might be different. Thus, this data supported our second idea that Ankara cohort and microarray cohorts might be quite different CRC cohorts.

3.3. Conclusion

First of all, we developed SSAT and identified ULBP2 and SEMA5A as prognostic gene markers in two different datasets with SSAT approach. We validated both ULBP2 and SEMA5A as independent prognostic gene markers in CRC.

Latter, we developed USAT, to improve SSAT, with 75% true positivity and 9% false positivity rates. We identified 65 different candidate genes associated with prognosis independent of stage in two different microarray datasets. We narrowed down the number of the genes to 10 by L1 penalized Cox Proportional Hazard Regression and decided to validate 8 of those 10 genes.

In order to validate the association of those genes with prognosis, we should have made sure that qPCR primers measured the same expression with microarray probes, because we were going to validate the association of microarray expression with prognosis using qPCR expression instead of microarray technology. Therefore, we used 7 different CRC cell lines, whose microarray expression was present in CGP database, for validation of the expression of the genes. We were able to validate the expression of 6 of those 8 genes and went on further analyses with those 6 genes.

After we measured the qPCR expression of those 6 genes in Ankara cohort, we validated only CLCN7 with USAT analyses of qPCR expression values. PTRF, KLF9, TGFB1I1 and DUSP10 had opposite associations with prognosis in Ankara cohort compared to microarray results. Therefore, we analyzed all possible threshold values of the genes with LRMC analyses to see if there could be other expression cut-offs than Maxstat cut-off might exhibit expected associations with prognosis. However, all significant cut-off values of the genes exhibit the same associations with Maxstat cut-offs. Then, we thought that there could be some confounding factors, like different subtypes with opposite associations

with prognosis, and some bias in qPCR expression. Although we could identify different subtypes whose associations with prognosis were different, we could not identify any subtypes that had similar associations with Ankara cohort. We also realized that the associations of those genes with prognosis were specific to subtypes.

We also realized that the expression pattern of the genes in Ankara cohort was not as expected as present in microarray cohorts. All the genes were positively correlated in Ankara cohort although CLCN7 and CLDN3 were negatively correlated with other genes in microarray cohorts. So, we concluded that Ankara cohort and microarray cohorts may have different biological backgrounds and this might explain the different pattern in association with prognosis in the cohorts.

As a conclusion, we validated ULBP2 and SEMA5A as prognostic factors in CRC with SSAT approach but we could validate only CLCN7 as a prognostic factor in CRC among 6 genes with USAT approach.

4. Discussion

In this study, we identified chemosensitivity gene expression markers in different cancer types with commonly used techniques and prognostic gene expression markers with chemotherapy benefit prediction power in CRC by developing new algorithms, SSAT and USAT. We validated ULBP2 and SEMA5A as prognostic factors by SSAT approach, and CLCN7, among 5 other genes, by USAT approach. We further showed in silico that combination of ULBP2 and SEMA5A could be a marker for the benefit of chemotherapy in CRC, although it still remains to be validated in vitro.

4.1. Chemotherapeutic Signatures

4.1.1. Gene Signatures in Hematological Cancer Cell Lines

We started this study by the identification of subgroups with differential drug responses in hematological, breast and colorectal cancer cell lines using CCLE and CGP drug and expression databases. We first tried to identify specific gene expression signatures that can identify distinct sub-groups of cancer cell lines. Then, we compared drug response distributions of those sub-groups by global effectiveness graphs of each drug and t-test. We showed that our new classification identified a new sub-group which is very sensitive to Bryostatin, although all of classical classes of hematological cancers were resistant to that drug Figure 3.2.

The approach, global effectiveness of drugs, which we used for the identification of the drugs differentially effective between subtypes of interest, has some advantages and disadvantages. This type of graphs give insights into how effective drugs are against wide variety of cancer types. Besides showing the most sensitive and resistant cancer types to drugs, it reveals the overall activity of drugs against cancer types biologically relevant to each other based on the position in the effectiveness range. This way, it could be very

easy to determine the drugs that may be the most appropriate chemotherapy for different cancer types. On the other hand, it does not tell whether the drugs kill the cancer cells or not, because the effectiveness of the drugs were relative to each other. For instance, there may be some drugs which are ineffective against all cancer types, although the range of effectiveness may be wide. Global effectiveness graphs of such drugs may misdirect us to suggest them as the best drug for cancer types. Moreover, such graphs does not have any statistical meaning. In other words, the effectiveness difference of a drug against the most sensitive and resistant cancer types may not be significant, although the cancer types were the most sensitive and resistant ones. Nonetheless, it is important to determine the drugs to which a specific cancer type is the most sensitive for a better chemotherapy benefit and global effectiveness graphs are useful to foresee such associations.

Bryostatin was one of the drugs that we identified with global effectiveness graphs. It is a modulator of protein kinase C and it was previously shown that it mediates apoptosis in imatinib mesylate-dependent quiescent (G0/G1) CML cell lines [34]. We investigated the genes up-regulated in G2 group of hematological cancer cell lines using GSEA C2 curated gene sets and realized that one-fourth of the genes were included in Graham et al.'s gene list which include genes that are up-regulated in quiescent CML cells compared to quiescent normal, dividing CML and dividing normal cells [35]. We also showed in a previous study [21] that G2 group of cells mostly comprised of CML and AML cell lines. Therefore, we thought that G2 group of hematological cancer cell lines may be prone to apoptosis activation by Bryostatin, as they up-regulated quiescence-related genes. In fact, we know from the literature that PKC phosphorylates BCL-2 family proteins [36] and BCL-2 family proteins may have pro-apoptotic [37] and anti-apoptotic [38] functions upon phosphorylation. Therefore, Bryostatin should have a role in activation of pro-apoptotic proteins, like Bax, BAD and Bak etc. through activation of PKC in G2 group of hematological cell lines.

The signature we identified may contribute to acquisition of information about the biology of the tumors besides the classification of cell lines. For example, G1 and G6 cell lines up-regulated the expression of mesenchymal and cancer-testis antigen genes, respectively. Thus, G1 cells may be targeted through response to in vivo differentiation and G6 cells can be treated through immunotherapy against cancer-testis antigens.

Because these findings were revealed through in silico analyses, they have to be validated in vitro. Unfortunately, apart from the drugs differentially effective between the clusters identified, we also found some drugs for which our classification were not promising. For instance, we found that TAE684, which is one of the common drugs in CCLE and CGP, was mostly effective against G6 in CCLE but G7 in CGP. One of the main reason for such inconsistencies was previously described as the difference in the cytotoxicity measurement methods between different studies [39]. Nevertheless, which cytotoxicity measurement method at best remains unclear and can be determined by in vivo analyses. Therefore, we avoided validating drug response analyses between databases.

4.1.2. Gene Signatures in Breast Cancer Cell Lines

Our recent results in Isbilen et al. [23], in Muhammad Waqas Akbar's Master's thesis [40] and in this study showed that breast cancer cell lines can be subdivided into categories based on stemness properties, besides intrinsic and molecular classifications. We believe that such a stemness classification, combined with global effectiveness distributions of drugs, may contribute to the determination of chemotherapy benefit. Regarding this, we already revealed in Isbilen et al. that drugs like LBW242 and Raf265 were more effective against CSC-like breast cancer cell lines but non-CSC-like breast cancer cell lines were resistant. Interestingly, we found that non-CSC-like breast cancer cell lines were the most sensitive cells to Lapatinib compared to other cancer types and CSC-like breast cancer cell lines were the most resistant cells, although Lapatinib has been approved for

treatment of hormone- and HER2-positive breast cancer. According to our results, Lapatinib treatment may kill only non-CSC-like cells. However, another drug, which kills CSC-like cells, may be used in treatment in combination with Lapatinib to kill both subtypes of breast cancer cells. Therefore, these findings suggests that combination therapy may be promising for breast cancer treatment.

We also developed a 12-gene signature for the stemness of breast cancer cell lines in CCLE with a supervised approach by finding the most differentially expressed genes between two clusters of breast cancer cell lines defined by Gupta et al.'s gene list. Although we already validated that gene list in CGP database, it is still required to show stemness characteristics of CSC-like cells to prove their functional similarities to actual CSCs.

4.1.3. Gene Signatures in Colorectal Cancer Cell Lines

We tried to identify chemosensitivity gene markers in CRC by a semi-supervised approach. We first use whole gene expression data to cluster the cell lines through WTHC analysis and found the most differentially expressed genes, among which we identified the best classifiers by MRMR method. Finally, we got a 20-gene signature which could cluster the cells best in both CCLE and CGP. The next step included the analysis of differential drug response between two clusters of CRC cell lines. Among 138 drugs in CGP, we identified 5 drugs with differential effectiveness between the clusters of CRC cell lines with 99% significance level ($p<0.01$). Among those drugs, IPA-3 was one of the interesting ones, because CRC cluster 1 was one of the most sensitive cell lines although cluster 2 was one for most resistant group compared to all other cancer types. We compared clusters 1 and 2 according to their EMT properties with VIM and CDH1 expression in CCLE database and we realized that almost all the cells in cluster 2 up-regulated CDH1 expression and down-regulated VIM expression, unlike the cells in the cluster 1. Moreover, all the cells with up-regulated VIM and down-regulated CDH1 were in the cluster 1.

Thus, our signature can be a signature for EMT in CRC cell lines, as well as EMT can be signature for drug response, as Pitts et al. [41] claimed.

IPA-3 is an inhibitor of p21-activated kinases (PAK), which are responsible for many cellular activities, like cell proliferation, migration etc. and expressed in many cancers. Pitts et al. showed that another PAK inhibitor, PF-3758309, was more effective to mesenchymal CRC cells rather than epithelial CRC cells. Therefore, it might be better to focus on EMT mechanism in CRC for the prediction of chemotherapy benefit. However, combination therapy would not be a good option as a treatment, like we suggested for breast cancer treatment, because EMT is much dynamic [42] process and there could be cells in between epithelial and mesenchymal states. Such cells can escape from toxic effects of the drugs by processing through either status or oscillating in between. Therefore, it could be worthy to target rather those cells and find signatures for targeting them.

4.2. Theranostic Gene Signatures in CRC

After identification of chemotherapeutic subtypes in cancer cell lines, identification of prognostic subtypes was critical, because it is important to predict whether patients with poor prognosis can benefit from chemotherapy or not. Thus, we tried to identify prognostic signatures in CRC and determine the drugs might be effective against patients with poor prognosis rather than the ones with better prognosis. However, aforementioned supervised and unsupervised gene signature identification techniques used for cell lines could not identify distinct clusters in CRC tumors. We thought that it could be due to heterogeneity of tumors compared to cell lines or the presence of different types of cells inside tumors, like immune-related cells. Such a heterogeneity and impurity in fresh frozen tumor samples might cause fluctuations in measurement of gene expression profiles of tumors. Therefore, we decided to identify single gene markers for prognosis in CRC rather than multi-gene signatures.

A meta-analysis published by Venet el al. [24] revealed that most of multi-gene signatures published in the literature cannot predict clinical outcomes better than randomly selected gene-signatures. The authors in this report also suggested that most of the random signatures with high number of genes were significant in predicting clinical outcomes in tumors, although there is no biological association between the functions of the genes and the clinical outcome. Therefore, we designed our algorithms so that we find individual genes associated with prognosis and combine them for prediction of both prognosis and chemotherapy benefit.

4.2.1. SSAT and USAT

We developed two approaches, SSAT and USAT, which were similar in purpose but different in methodology for the identification of theranostic signatures or gene markers in cancers using gene expression data. The ultimate goal in both approaches was to identify prognostic gene markers that can define two distinct prognostic subgroups based on the best expression threshold with some statistical calculations. We thought that SSAT was less powerful in terms of statistical approaches compared to USAT, because SSAT used only Log-Rank test for the comparison of prognostic groups. On the other hand, USAT used three different statistical methods, which are CoxPH, Maxstat and Log-Rank tests, to identify the best candidate prognostic gene markers.

Both approaches has advantages and disadvantages. SSAT used only categorized expression values into 8 different expression levels and the probesets whose expression squeezed into an expression category were eliminated from the analysis. Although the expression of such a gene was different in decimal points, their expression value were almost the same for all the samples. Therefore, we thought that such genes could not explain prognosis difference between patients. Otherwise, it would not be possible to differentiate between expression values of those genes with very low variance with qPCR

technique. In other words, the probesets with very low variance would have expression values very close to each other throughout the cohort, which could make us unable to differentiate the expression difference between samples with qPCR. Therefore, such a categorization provided us with the elimination of statistically meaningful but biologically meaningless probesets from SSAT analyses.

On the other hand, SSAT did the survival analyses based on single expression threshold values, which separated patients into high and low expression groups, and there were not any methods in it to check the linear association between expression and prognosis. Therefore, it was likely to get genes as a result with Simpson's paradox. Simpson's paradox is a phenomenon in statistics, in which the association of the parameters in subgroups conflicts with the overall association. To give an example, if we could find genes that were associated with good prognosis in each subtypes of CRC but associated with poor prognosis when the subtypes were considered all together, then the analysis results would be biased by Simpson's paradox. So, it is possible to identify genes with wrong association with prognosis in SSAT approach.

SSAT also selected the probeset with the highest coefficient of variation among the ones that hit the same gene and eliminated the others from the analysis, in contrast to USAT. Therefore, SSAT could select the probesets with the highest variance. Nevertheless, we could not have any idea about the association of the expression of eliminated probesets with prognosis.

USAT was a more robust approach based on statistical power compared to SSAT. USAT uses three different statistical tests to test the association of the genes with prognosis in terms of three different aspects. CoxPH was used to check the linear association of expression with prognosis and to get hazard ratio values to measure the risk rates of the prognostic groups. CoxPH also eliminated the possibility that we could get results biased

with Simpson's paradox, because we would not get significant linear associations in CoxPH if there were conflicts between associations in subtypes and overall association. The second test was Maxstat, which was used to identify the expression threshold value that could separate patients best according to prognosis by performing a test of independence between prognosis and gene expression. Maxstat calculates standardized Log-Rank statistics for each expression threshold values in between 10%-90% quantiles and identify the threshold with the highest standard statistics as the best threshold to predict prognosis of the patients. It also calculates a p-value for the test of independence between prognosis and gene expression. We eliminated the gnes that were independent of prognosis with respect to Maxstat (with $p>0.05$), because we were looking for the genes to which prognosis was dependent. USAT also analyzed many expression threshold values to determine the best one. Nevertheless, we could not have any information about the rest of the threshold values which were ignored by Maxstat in terms of prediction of prognosis. The third test was Log-Rank test, which determines the significance of the prognosis differences between high and low expression groups defined by the Maxstat threshold. As a summary, these three methods examine the association between gene expression and prognosis in three different aspects to eliminate biases and get the optimum results.

Like SSAT, the main disadvantage of USAT was the analyses with a single expression threshold. We thought that the expression threshold values other than Maxstat threshold values might give us critical information about the association of genes with prognosis. It was also important to identify the threshold values where the association with prognosis changed direction. This way we might have identified different subtypes with different associations with prognosis. In other words, we could have revealed Simpson's effect for such genes. Therefore, we developed another approach, called LRMC, alternative to Maxstat. We discussed LRMC in the subsequent sections.

4.2.1.1. Identification and Validation of Candidate Genes by SSAT

SSAT was performed by GSE17536 and GSE17537 colon cancer microarray datasets and ULBP2 and SEMA5A were identified as the only independent prognostic genes as a result of step-wise multivariate analysis of common significant genes. We did not perform SSAT with GSE41258, because the dataset was not published when we developed SSAT and the probeset significant for ULBP2 was not present in HGU-133A platform.

SSAT approach contains 5 basic steps as explained in Figure 3.7. In the first and the second steps, we perform SSAT analysis with two cohorts and find the common results between them, respectively. We got 400 and 269 genes significant in GSE17536 and GSE17537, respectively. These numbers corresponded to only 1.9% and 1.3% of all 21325 genes in the datasets, which were much smaller than 5% of all probesets. Theoretically, we expected to get significant results more than 5% of all genes, because besides real associations, 5% of all genes were expected to be significant by chance with 95% significance level. The reason why we got such small number of significant genes was that there were gene elimination steps in SSAT, as well. SSAT eliminated the genes for which more than 90% of all categorized expression values in the same expression category, from the analysis. The number of significant genes, 64 genes, in common shows that many genes were significant by chance in SSAT analyses. In the third step, we crudely calculate average of Log-Rank p-values of each gene across datasets in order to sort the genes based on their significance in both datasets and select the best genes for multivariate analysis. Although calculation of p-value averages does not have any statistical meaning, we assumed that it may give an insight into overall significance order of the genes. For the fourth step, we selected the first 17 and 4 genes based on average p-values, apart from stage information, for step-wise multivariate analyses in GSE17536 and GSE17537, respectively. We determined those numbers so that the number of genes was at most one

tenth of the sample size to prevent over-fitting in multivariate regression. Otherwise, multivariate regression model might describe random error (noise) instead of actual relationships between parameters. Finally in the fifth step, ULBP2 and SEMA5A were the only common genes that can predict prognosis independent of stage. Therefore, we decided to validate prognostic value of ULBP2 and SEMA5A in Ankara cohort through qPCR experiments.

We identified ULBP2 and SEMA5A as prognostic factors with different expression cut-offs. ULBP2 was the most significant when the patients in the expression intervals 1 and 2 were compared to patients in the expression intervals 3, 4, 5, 6, 7 and 8, as SEMA5A was the most significant when the patients in the expression intervals 1, 2 and 3 were compared to patients in the expression intervals 4, 5, 6, 7 and 8. Therefore, we determined expression thresholds 4 and 6 for ULBP2 and SEMA5A, respectively in GC-RMA normalized microarray datasets. We could not use such threshold values in qPCR expressions of Ankara cohort, because those threshold values were adjusted for GC-RMA normalized expression values, which are generally in between 0 to 16. Instead, we used median expression as threshold for qPCR values in Ankara cohort.

To analyze prognostic powers of ULBP2 and SEMA5A further, we performed multivariate CoxPH regression with clinical parameters and Log-Rank test with the combination of the genes. Multivariate analysis results showed that ULBP2 and SEMA5A were independent prognostic factors in microarray datasets, although ULBP2 and SEMA5A were independent factors when they were used together to assess the prognosis in Ankara cohort. We performed the experiments with colon cancer samples and removed rectal cancer samples, as the discovery datasets were colon cancer datasets and we found out that ULBP2 and SEMA5A were not significant prognostic factors in rectal cancer.

It may be also very critical in clinical use that we can further separate colon cancer patients into high, intermediate and low risk groups through combination of ULBP2 and SEMA5A. The patients with high expression of SEMA5A and low expression of ULBP2, the patients with low expression of SEMA5A and high expression of ULBP2 and the patients expressing both genes high or low could be separated into high, low and intermediary risk groups significantly in both microarray datasets and we validated this association in Ankara cohort (Figure 3.8). Therefore, we believe that ULBP2 and SEMA5A can be used for prediction of prognosis of colon cancer patients in clinical trial.

We also showed in silico that the combination of ULBP2 and SEMA5A could be prognostic factors with chemotherapy benefit, as well. We identified many drug effective to high risk group according to ULBP2 and SEMA5A expression in CCLE and CGP cell line databases. Global effectiveness of AZ628 showed that high risk colon cancer cell lines were one the most sensitive cell lines to AZ628, as low risk colon cancer cell lines were almost the most resistant (Figure 3.9). It was also important that the effectiveness of AZ628 increases gradually from low risk group to high risk group, including intermediary risk group. On the other hand, some other ERK pathway inhibitors FTI-277, Tipifarnib, CEP-701 and PF-02341066 had the same effect with AZ628 on prognostic groups of CRC cell lines, although other BRAF inhibitors, SB590885 and PLX4720, did not (data not shown). Sensitivity to AZ628 also were also positively correlated with number of mutated genes in CRC cell lines in CGP (data not shown). Nevertheless, multivariate linear regression analyses showed that SEMA5A could predict AZ628 response independent of number of mutated genes, although ULBP2 could not (data not shown). The number of mutated genes was also correlated with prognostic classification of CRC cell lines (data not shown). Mutation accumulation increased from good prognosis cell lines to poor prognosis cell lines. In spite of the fact that those results have to be validated in vitro and

in vivo with tumors, we believe that they may pioneer the research based on identification of prognostic factors with chemotherapy benefit power.

ULBP2 and SEMA5A encode for transmembrane proteins, which may lead to activation of many signaling pathways with oncogenic or tumor suppressor effects. ULBP2 is a ligand for natural killer cell activating receptor NKG2D and generally expressed in transformed and stressed cells [43] to mediate immunosurveillance of such cells by activating NK cells. It has also been shown that it can be present in cancer patients' sera as a soluble protein [44, 45]. SEMA5A is a ligand for Plexin-3B transmembrane receptor [46], which can take place in many signaling pathways to regulate cell proliferation, invasion [47] and adhesion [48] etc. It has been shown that SEMA5A can have both oncogenic and tumor suppressor effects in various cancers. There is also sufficient evidence that SEMA5A can be in secreted form in sera and may enhance tumor aggressiveness, proliferation and invasion in pancreatic and gastric cancer [49, 50, 51]. On the other hand, the expression of SEMA5A was found to be suppressed in colon tumors [52]. Moreover, it has been shown that soluble SEMA5A strongly increases proliferation of T-cells and NK cells and secretion of pro-inflammatory cytokines in vitro [53]. Combining all these information, we thought that soluble SEMA5A may contribute to immunosurveillance of tumor, as soluble ULBP2 may contribute to tumor escape from immune system through blocking NKG2D receptors to prevent activation of NK cells. Therefore, we believe that SEMA5A-positive/ULBP2-negative low risk colon cancer patients can induce immune response against tumors, although SEMA5A-negative/ULBP2-positive high-risk colon cancer patients may not be able to develop immune systems to recognize tumor cells due to blockage of the receptors of NK cells by soluble ULBP2. Intermediary risk colon cancer patients, who express both genes high or low in their tumors, may cancel out the effect of both genes by producing more NK cells through soluble SEMA5A but blocking them through soluble ULBP2, or vice versa.

AZ628 is a RAF inhibitor, which acts on wild type and V600E BRAF as well as C-RAF activities [54] effecting ERK pathway. It has been previously shown that CRC, melanoma and thyroid cancer cell lines with mutated BRAF are sensitive to AZ628 along with the suppression of ERK pathway [55]. Although ULBP2 and SEMA5A have not been associated with BRAF up to now, there are studies that shows the association between ERK signaling and either ULBP2 or SEMA5A in different cancer types. Xiaosong et al. claimed that ULBP2 expression was increased in myeloma cell lines as a result of increased ERK signaling activity [56]. They also asserted that PD98059, an ERK signaling pathway inhibitor, led to decrease in ULBP2 expression in myeloma cells [56]. Sadanandam at al. also revealed a mice model that secreted form of SEMA5A from pancreatic cancer cell lines enhanced ERK phosphorylation in NK cells [57]. Based on these information, we may hypothesize that ULBP2-positive/SEMA5A-negative CRC cell lines were more sensitive to AZ628 due to increased ERK signaling activity. We believe that ERK signaling pathway could be promising to start to understand the association between AZ628 and ULBP2/SEMA5A in colon cancer.

4.2.1.2. Identification and Validation of Candidate Genes by USAT

We performed USAT analyses with GSE17536 and GSE41258. The reason why we did not use GSE17537 was the low sample size. We realized that the statistical tests that we used in USAT analysis were sensitive to sample size so that the number of significant results increased with increasing sample sizes (data not shown). We could only identify 7 genes in common in three datasets, among which 2 of them had the association with prognosis in opposite directions between datasets. We thought that the low number of common genes was due to low sample size of GSE17537, because GSE17537 had very low number of significant genes in common with both GSE17536 and GSE41258. Therefore, we decided to remove GSE17537 from USAT analyses.

Like SSAT approach, USAT approach contains 5 steps as explained in Figure 3.10. In the first and the second steps, we performed USAT with GSE17536 and GSE41258 datasets and find the common results between them, respectively. In those steps, we identified 3985 and 2408 probesets, which were 7.3% and 10.1% of all probesets, in GSE17536 and GSE41258, respectively, although around 20,000 probesets in GSE17536 and 10,000 probesets in GSE41258 could not be analyzed due to model fitting problems. Those percentages shows that USAT could identify real associations besides random genes with associations by chance. Nevertheless, we performed some extra analyses to calculate true and false positivity of USAT in order to make sure that USAT could identify true associations. Those analyses are discussed in this and the following paragraphs under this section, as well. In the third step, we determined the genes with common directionality based on association with survival. We also identified 38 probesets that were associated with prognosis in opposite directions between the datasets. We calculated the ratio of opposite associations with respect to common results, 9%, as false-positivity of USAT. Then in the fourth step, we performed Chi-square test to make sure that the expression of the genes were not dependent to stage, although we already performed multivariate analyses in some of the models used in USAT using stage information of the patients. Finally in the fifth step, we performed L1 penalized Cox proportional hazard regression to determine the genes with the best associations for validation. We identified 6 and 4 genes and selected 5 and 3 of those genes in GSE17536 and GSE41258, respectively, for validation.

Before we determined these steps of USAT approach, we calculated true and false positivity of USAT using GSE17536 dataset. We used the directionality of the associations with prognosis as reference for false positivity, as we explained in the previous paragraph. Whereas, we used a list of 48 genes, which were already published by O'Connell et al. [12] as CRC prognosis predictors, as a reference for true positivity of USAT. USAT was first

built with a single model (CGS-MC-LC), in which it performed CoxPH with categorical expression data and stage stratification and Maxstat and Log-Rank tests with continuous expression data. This model could identify 21 of 44 genes, which were present in GSE17536 dataset among 48 genes. In order to increase the number of significant genes, we built 11 other models with combinations of expression types and statistical tests. Although all the models could identify similar numbers of genes, we realized that each model had a power to identify the genes that the others cannot. So, considering the genes significant in at least one model, we could identify 33 of those genes as the union, thus making the true positivity of USAT 75%.

Before performing qPCR experiments for the resulting genes, we decided to check whether qPCR primers could detect the same expression that microarray probes detected. In other words, we checked whether qPCR expression of a gene represents microarray expression of the corresponding probeset. We selected 7 CRC cell lines from CGP database, for which we had microarray gene expression data, and performed qPCR experiment on these cell lines for the selected genes to compare qPCR and microarray expression patterns of these cell lines. We performed this step to show that the difference in the expression of the genes between microarray and qPCR due to detection with different probe sequences may be negligible. We showed that 6 of 8 genes exhibited significant concordance between qPCR and microarray in 7 CRC cell lines (Figure 3.11) and we decided to perform qPCR experiment with these significant genes on Ankara cohort, because the qPCR expression of the discordant genes did not represent the microarray expression of the corresponding probesets.

We performed qPCR on Ankara cohort for these 6 genes and analyzed the results with USAT for validation. Interestingly and unexpectedly, all 6 genes except for CLCN7 had the opposite significant associations with prognosis, compared to microarray results. Such

a high false positivity rate was much bigger than what we calculated for USAT. Truthfully, we expected some disassociations with prognosis in validation cohort but getting opposite significant associations for most of the genes was very surprising for us. Therefore, we decided to analyze the results further to understand the biological reasons behind such results.

We also performed analyses to predict the prognosis of patients in GSE17536 and Ankara cohorts by the combination of ULBP2, SEMA5A and CLCN7. We could show that this combination can predict prognosis in colon cancer patients independent of tumor stage, when the patients were divided into 4 prognostic groups based on the number of the genes with which they were predicted as good prognosis patients (data not shown).

4.2.1.3. Further Analyses to Understand the Reasons Behind Opposite Results

As we explained in previous paragraphs, one of the disadvantages of USAT was that it performs survival analysis based on a single Maxstat threshold. We thought that there might be different significant expression threshold values for a gene, even with opposite associations with prognosis. We decided to analyze all possible expression thresholds for all the genes and developed a graph, called Log-Rank with Multiple Cut-offs, by which we could visualize both the significance and the direction of the association of each possible expression threshold values.

LRMC graphs of the selected genes in microarray datasets showed that the genes were significantly associated with prognosis for many expression threshold values in the same direction with Maxstat cut-offs, although we expected thresholds with opposite associations. However, there were some fluctuations in the graphs, where the significance of thresholds peaked. Such patterns in those graphs led us to think that there might be subtypes of CRC that were associated with prognosis in opposite directions. To give an example, the significance of TGFB1I1 in GSE41258 peaked at the threshold values just

below the 1st quartile and went down peaking again at the thresholds around median. Such a pattern suggested that the CRC samples between the 1st quartile and the median might represent a subtype of CRC, because the significance of the thresholds in that range decreased dramatically, indicating that those samples might have an opposite association with survival compared to others. We identified the samples corresponding to the thresholds decreasing the significance of from the peaks and performed gene set enrichment analysis to compare these samples with the others. However, there were no significant gene sets differentially expressed in these samples, suggesting that the difference between those groups did not have a biological meaning. Therefore, we tried to cluster the microarray samples with the gene lists that were published or that we identified in order to see whether different subtypes had opposite associations with survival. We clustered GSE171536 samples based on the gene lists with different biological backgrounds like EMT, stemness, cell morphology and prognosis etc. However, we could not identify any subtypes with opposite associations with prognosis compared to overall association. We also generated new gene lists to identify different subtypes of CRC samples, but again the resulting clusters did not exhibit opposite associations with overall associations. Moreover, we realized that the genes we identified might be prognostic factors for only some of the subtypes. For instance, none of the genes, except for CLCN7, were significant in LRMC graphs of the first cluster we identified by WTHC analysis in GSE17436 (Figure 3.14), although it contained almost half of the samples. The only cluster concordant with the overall association for all the genes was the cluster 3. Nonetheless, we could not identify any clusters exhibiting associations with prognosis opposite to overall association.

Our next step was to understand whether the biological background of Ankara cohort and microarray cohort was different or not. We compared the relationships between the candidate genes in Ankara cohort, GSE17536 and GSE41258 by Pearson's r correlation and realized that the overall relationships within Ankara cohort were opposite to the

relationships in GSE17536 and GSE41258. For instance, CLDN3 and CLCN7 correlated positively with the other 4 genes in Ankara cohort, although they correlated negatively or did not correlate at all in GSE17536 and GSE41258 (Figure 3.16). We know that an analysis with only six genes will not be enough to make a conclusion about the biological difference between cohorts. Nevertheless, such difference in the relationship between the genes within the cohorts gave us a clue about that the cohorts might be composed of biologically distinct CRC samples.

We thought that one of the reasons why we got such biological differences between Ankara cohort and microarray datasets in terms of gene expression might be partial degradation of RNA molecules in Ankara cohort before or after RNA extraction with Trizol reagent. qPCR experiments showed that there were two groups of tumors in terms of GAPDH CT values, one of which included those with CT values around 18 and the other included those with CT values around 25. Nevertheless, LRMC analyses within each of two GAPDH CT groups exhibited similar results to those that we got by whole Ankara cohort (data not shown). In fact, considering the fact that RNA degradation occurs from 5' to 3', we did not expect to find an association between RNA degradation and direction of the association with prognosis, because we already used the best coverage TaqMan primers that were already designed to detect 3' ends of the RNAs.

We were able to validate the prognostic power of ULBP2 and SEMA5A that we identified with SSAT approach, although their chemotherapeutic power still remains to be validated in vitro. It was interesting to validate ULBP2 and SEMA5A in Ankara cohort, although we had some clues about that Ankara cohort might be a distinct CRC cohort compared to microarray datasets. Nevertheless, this validation supported the independent prognostic power of ULBP2 and SEMA5A, because they were still validated in spite of some differences between cohorts. On the other hand, we identified exact opposite

associations with prognosis in Ankara cohort with USAT approach. One of the underlying reasons of opposite associations could be lack of biological bases while we selected the genes for validation in USAT approach. In SSAT approach, we had comprehensive arguments about how biological backgrounds of ULBP2 and SEMA5A were related to prognosis in CRC, besides statistical significance. However, we selected the genes in USAT approach based only on statistical power, regardless of any biological associations. In fact, the genes selected with USAT approach have previously been implicated in oncogenic and tumor suppressor mechanisms in many cancers. PTRF and TGFB1I1 have been shown to have tumor suppressor characteristics in prostate cancer and NSCLC [58, 59, 60, 61, 62], as KLF9 have been shown to act as a tumor suppressor in many cancer types [63, 64, 65, 66, 67, 68], as well as CRC [69]. Whereas, DUSP10 have been suggested to promote prostate cancer cell proliferation [70] and as up-regulated in colon cancer [71]. CLDN3 has been indicated in many cancers as a prognostic marker, as it is a MET marker. However, there are not many studies about the relationship between CLCN7 and clinical characteristics of cancers. CLCN7 has been patented as one of the diagnostic markers for breast cancer [72]. Moreover, it has been shown that mutations in CLCN7 is associated with cervical cancer [73].

4.3. Conclusion

In this study, we tried to identify prognostic and chemotherapeutic (or theranostic) gene markers in different cancer types. Although we were able to validate theranostic powers of some genes, there still remains some problems about the identification process. First of all, we used overall survival information of the patients as representative of prognosis. Overall survival is a complex parameter, which is affected by many factors unrelated to cancer, like eating habits, living conditions etc. When we try to associate the expression of a single gene with overall survival, we theoretically ignore such parameters. Overall

survival is also two sided censored data, which means that we do not know when cancer occurred in patients before diagnosis and when they died due to cancer (for the patients which were not followed). Therefore, there are many uninformative or misleading data in cohorts. For instance, the patients who were not followed or died due to causes other than cancer 2 months after diagnosis do not give any information about the severity of the tumors they have, because we do not know whether those patients will survive too many years or not. Including such patients into analyses may decrease the significance of the analyses in terms of actual associations (not statistical significance).

The other problem is that overall survival also includes disease-free survival and disease-specific survival together. The patients with early recurrent but non-aggressive tumors will be considered similar to the ones with late recurrent but aggressive tumors. Thus, overall survival combines recurrence and tumor aggressiveness with equal weights and may make a bias in prognosis. Therefore, it could be a good idea to follow-up patients for many clinical parameters and find biomarkers specific to those specific clinical parameters as representatives of prognosis.

We also assume that gene expression represents functional activity in tumors, while associating gene expression with prognosis. However, there are strong evidence that high transcription rate does not imply high translation or functional activity rate. In other words, although biological background of genes may imply association with prognosis, gene expression does not necessarily exhibit such associations. Therefore, the associations in expression level should be validated in protein activity level, as well. Otherwise, it could be highly likely to get false positive gene markers that are not even better than any random genes, as Venet et al. suggested [24]. Before model development for prognostic and chemotherapeutic prediction, the genes should also be validated in vivo with mice

models. Latter, the prediction power of the model should be validated in independent patient cohorts.

To summarize, SSAT and USAT approaches are comprised of conventional statistical tests and analyze conventional clinical parameters. This study showed us that new methods that can analyze many parameters with higher true positivity rate are still required for the identification of theranostic markers. Concordantly, we were able to validate SSAT results with a higher success rate compared to USAT results. We believe that one of the main reason behind this consequence is that we identified genes associated with clinical outcome independent of other clinical parameters using multivariate statistical analyses in SSAT approach, but not in USAT approach. Therefore, it is more likely to validate the genes identified as associated with the outcome independent of other parameters. Nevertheless, prospective stepwise multivariate analyses of the genes that we identified by USAT approach showed us that all the genes were still prognostic factors independent of age, gender, stage and grade (data not shown). However, we still could not validate these genes in Ankara cohort. We believe that many parameters irrespective of relation to cancer biology may affect prognosis and such parameters should be considered in survival analyses to achieve higher rates of validation.

Despite having such problems we were able to validate both genes we identified with SSAT approach and one gene with USAT approach in Ankara cohort, which is an independent cohort from microarray cohorts. These analyses showed us that gene expression can be used to assess the prognosis of cancers, when they are adjusted with other clinical parameters. After optimization studies, accurate and robust statistical models can be built for the prediction of cancer prognosis using gene expression patterns along with clinical parameters.

Bibliography

[1] Ferlay J et al., GLOBOCAN 2008 v1.2, Cancer Incidance and Mortality Worldwide: IARC CancerBase No. 10, Lyon, France: International Agency for Research on Cancer, 2010.

[2] "Food, nutrition, physical activity, and prevention of cancer: a global perspective," AICR, Washington, DC, 2007.

[3] Etzioni R et al., "The case for early detection," *Nat Rev Cancer,* vol. 3, pp. 243-252, 2003.

[4] American Joint Committee on Cancer, AJCC cancer staging manual, 5th edition, Philadelphia: Lippincott-Raven, 1997.

[5] Locker et al., "ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer.," *Journal of Clinical Oncology,* vol. 24, pp. 5313-27, 2006.

[6] Benson AB et al., "American Society of Clinical Oncology Recommendations on Adjuvant Chemotherapy for Stage II Colon Cancer," *Journal of Clinical Oncology,* vol. 22, no. 16, 2004.

[7] Schaeybroeck SV et al., "Implementing prognostic and predictive biomarkers in CRC clinical trials," *Nat Rev Clinical Oncology,* vol. 8, pp. 222-232, 2011.

[8] Wang Y et al., "Gene Expression Profiles and Molecular Markers To Predict Recurrence of Dukes' B Colon Cancer," *Journal of Clinical Oncology,* vol. 22, no. 9, pp. 1564-1571, 2004.

[9] Barrier A et al., "Colon cancer prognosis prediction by gene expression profiling," *Oncogene,* vol. 24, pp. 6155-6164, 2005.

[10] Barrier A et al., "Stage II Colon Cancer Prognosis Prediction by Tumor Gene Expression Profiling," *Journal of Clinical Oncology,* vol. 24, no. 29, pp. 4685-4691, 2006.

[11] Kerr D et al., "A quantitative multigene RT-PCR assay for prediction of recurrence in stage II colon cancer: Selection of the genes in four large studies and results of the independent, prospectively designed QUASAR validation study," *Journal of Clinical Oncology,* 27:15s, 2009 (suppl; abstr 4000).

[12] O'Connell MJ et al., "Relationship Between Tumor Gene Expression and Recurrence in Four Independent Studies of Patients With Stage II/III Colon Cancer Treated With Surgery Alone or Surgery Plus Adjuvant Fluorouracil Plus Leucovorin," *Journal of Clinical Oncology,* vol. 28, no. 25, pp. 3937-3944, 2010.

[13] Gray GR, "Validation Study of a Quantitative Multigene Reverse Transcriptase-Polymerase Chain Reaction Assay for Assessment of Recurrence Risck in Patients with Stage II Colon Cancer," *Journal of Clinical Oncology,* vol. 29, no. 35, pp. 4611-19, 2011.

[14] Salazar R, "Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer," *Journal of Clinical Oncology,* vol. 29, no. 1, pp. 17-24, 2011.

[15] Kennedy et al. , "Development and Independent Validation of a Prognostic Assay for Stage II Colon Cancer Using Formalin-Fixed Paraffin-Embedded Tissue," *Journal of Clinical Oncology,* vol. 29, no. 35, pp. 4620-26, 2011.

[16] O'Conner et al., "Adjuvant chemotherapy for stage II colon cancer with poor prognostic features.," *Journal of Clinical Oncology,* vol. 29, no. 25, p. 3381–88, 2011.

[17] Barretina et al., "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature,* no. 483, pp. 603-607, 2012.

[18] Garnett et al., "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature,* no. 483, pp. 570-575, 2012.

[19] R Core Team, *R: A language and environment for statistical computing.,* Vienna, Austria: R Foundation for Statistical Computing, 2013.

[20] Terry M. et al (2000), "Modelling Survival Data: Extending the Cox Model," Springer, New York, ISBN 0-387-98784-3.

[21] Isbilen et al., "Identifying Effective Molecularly Targeted Drugs for Hematological Cancers," *Türkiye Klinikleri Jourlan of Hematology-Special Topics,* vol. 7, no. 1, pp. 1-7, 2014.

[22] Gupta et al., "Identification of selective inhibitors of cancer stem cells by high-throughput screening," *Cell,* vol. 138, no. 4, pp. 645-59, 2009.

[23] Isbilen et al., "Predicting Chemotherapy Sensitivity Profiles for Breast Cancer Cell Lines with and Without Stem Cell-Like Features," *Current Signal Transduction Therapy,* vol. 8, pp. 268-73, 2013.

[24] Venet et al., "Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome," *PloS Computational Biology,* vol. 7, no. 10, 2011.

[25] Bustin SA, "Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems," *Society for Endocrinology,* vol. 29, p. 23:39, 2002.

[26] Chuaqui RF et al., "Post-analysis follow-up and validation of microarray experiments," *Nature Genetics,* vol. 32, pp. 509-514, 2002.

[27] van der Spek PJ et al., "Are gene expression microarray analyses reliable? A review of studies of retinoic acid responsive genes," *Genomic Proteomics Bioinformatics,* vol. 1, no. 1, pp. 9-14, 2003.

[28] Rajeevan MS et al., "Validation of array-based gene expression profiles by real-time (kinetic) RT-PCR," *J Mol Diagn,* vol. 3, no. 1, pp. 26-31, 2001.

[29] Loboda et al., "EMT is the dominant program in human colon cancer," *BMC Medical Genomics,* pp. 4-9, 2011.

[30] Marisa et al., "Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value," *PLoS Med,* vol. 10, no. 5, 2013.

[31] Sadanandam et al., "A colorectal cancer classification system that associates cellular phenotype and responses to therapy," *Nature Medicine,* vol. 19, no. 5, pp. 619-26, 2013.

[32] Sveen et al., "ColoGuidePro: A Prognostic 7-Gene Expression Signature for Stage III Colorectal Cancer Patients," *Clin Cancer Res,* vol. 18, no. 21, pp. 6001-10, 2012.

[33] Budinska et al., "Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer," *J Pathol,* vol. 231, pp. 63-76, 2013.

[34] Jorgensen et al., "Enhanced CML stem cell elimination in vitro by bryostatin priming with imatinib mesylate," *Experimental Hematology,* vol. 33, no. 10, pp. 1140-46, 2005.

[35] Graham et al., "Transcriptional Analysis of Quiescent and Proliferating CD34+ Human Hemopoietic Cells from Normal and Chronic Myeloid Leukemia Sources," *Stem Cells,* no. 25, pp. 3111-3120, 2007.

[36] May et al., "Interleukin-3 and bryostatin-1 mediate hyperphosphorylation of Bcl2a in association with suppression of apoptosis," *Journal of Biological Chemistry,* no. 269, pp. 26865-70, 1994.

[37] Yang et al., "Bad, a heterodimeric partner for Bcl-XL and Bcl2, displaces Bax and promotes cell death," *Cell,* no. 80, pp. 285-291, 1995.

[38] Ito et al., "Bcl2 phosphorylation required for anti-apoptosis function," *Journal of Biological Chemistry,* no. 272, pp. 11671-73, 1997.

[39] Haibe-Kains et al., "Inconsistencies in large pharmacogenomic studies," *Nature,* vol. 504, no. 7480, pp. 389-393, 2013.

[40] Akbar M.W., "Characterization of chemosensitivity profiles of breast cancer cell lınes, with and without stem cell like features," (Master Thesis), Department of Molecular Biology and Genetics, Graduate School of Engineering and Science of Bilkent University, Ankara, 2014. Retrived from Bilkent University Library Reserve Theses. Call Number: XX(896295.1).

[41] Pitts et al., "Association of the epithelial-to-mesenchymal transition phenotype with responsiveness to the p21-activated kinase inhibitor, PF-3758309, in colon cancer models," *Front Parmacol,* vol. 4, no. 35, 2013.

[42] Yılmaz-Ozcan et al., "Epigenetic Mechanisms Underlying the Dynamic Expression of Cancer-Testis Genes, PAGE2, -2B and SPANX-B, during Mesenchymal-to-Epithelial Transition.," *PloS One,* vol. 9, no. 9, 2014.

[43] Gonzales et al., "Immunobiology of human NKG2D and its ligands," *Curr. Top. Microbiol. Immunol.,* no. 298, pp. 121-138, 2006.

[44] Champsaur et al., "Effect of NKG2D ligand expression on host immune responses," *Immunol. Rev,* no. 235, pp. 267-285, 2010.

[45] Raulet et al., "Regulation of ligands for the NKG2D activating receptor," *Annu. Rev. Immunol.,* no. 31, pp. 413-441, 2013.

[46] Artigiani et al., "Plexin-B3 is a functional receptor for semaphorin 5A," *EMBO Rep.,* no. 5, pp. 710-714, 2004.

[47] Sadanandam et al., "Semaphorin 5A promotes angiogenesis by increasing endothelial cell proliferation, migration, and decreasing apoptosis," *Microvasc. Res.,* no. 79, pp. 1-9, 2009.

[48] Casazza et al., "Semaphorin signals in cell adhesion and cell migration: functional role and molecular mechanisms," *Adv. Exp. Med. Biol.,* no. 600, pp. 90-108, 2007.

[49] Sadanandam et al. , "High gene expression of semaphorin 5A in pancreatic cancer is associated with tumor growth, invasion and metastasis," *Int. J. Cancer,* no. 127, pp. 1373-83, 2010.

[50] Pan et al., "Elevated expression of semaphorin 5A in human gastric cancer and its implication in carcinogenesis," *Life Sci.,* no. 86, pp. 139-144, 2010.

[51] Pan et al., "Expression of semaphorin 5A and its receptor plexin B3 contributes to invasion and metastasis of gastric carcinoma," *World J. Gastroenterol.,* no. 15, pp. 2800-04, 2009.

[52] Liu et al., "Genome-wide association and fine mapping of genetic loci predisposing to colon carcinogenesis in mice," *Mol. Cancer Res.,* no. 10, pp. 66-74, 2012.

[53] Gras et al., "Secreted Semaphorin 5A Activates Immune Effector Cells and Is a Biomarker for Rheumatoid Arthritis," *Arthritis & Rheumatology,* vol. 66, no. 6, pp. 1461-71, 2014.

[54] Shen et al., "Linking molecular characteristics to the pharmacological response of a panel of cancer cell lines to the BRAF inhibitor, AZ628," in *AACR 98th Meeting*, Los Angeles, 2007.

[55] McDermott et al., "Identification of genotype-correlated sensitivity to selective kinase inhibitors using high-throughput tumor cell line profiling," *Proc Natl Acad Sci U S A,* vol. 104, no. 50, pp. 19936-41, 2007.

[56] Xiaosong et al., "Valproic Acid Upregulates NKG2D Ligand Expression through an ERK-dependent Mechanism and Potentially Enhances NK Cell-mediated Lysis of Myeloma," *Neoplasia,* vol. 14, no. 12, pp. 1178-89, 2012.

[57] Sadanandam et al., "Secreted semaphorin 5A suppressed pancreatic tumour burden but increased metastasis and endothelial cell proliferation," *Br J Cancer,* vol. 107, no. 3, pp. 501-507, 2012.

[58] Gomez-Pozo et al., "PTRF/Cavin-1 and MIF Proteins Are Identified as Non-Small Cell Lung Cancer Biomarkers by Label-Free Proteomics," *PLoS One,* vol. 7, no. 3, 2012.

[59] Nassar et al. , "PTRF/Cavin-1 decreases prostate cancer angiogenesis and lymphangiogenesis," *Oncotarget,* vol. 4, no. 10, pp. 1844-55, 2013.

[60] Nassar et al., "Caveola-forming proteins caveolin-1 and PTRF in prostate cance," *Nature Reviews Urology,* vol. 10, pp. 529-536, 2013.

[61] Rahman et al., "Inactivation of androgen receptor coregulator ARA55 inhibits androgen receptor activity and agonist effect of antiandrogens in prostate cancer cells.," *Proc Natl Acad Sci U S A,* vol. 100, no. 9, pp. 5124-9, 2003.

[62] Fujimoto et al., "Cloning and Characterization of Androgen Receptor Coactivator, ARA55, in Human Prostate," *The Journal of Biological Chemistry,* vol. 274, no. 12, pp. 8316-21, 1999.

[63] Huang et al., "Krüppel-like factor 9 inhibits glioma cell proliferation and tumorigenicity via downregulation of miR-21.," *Cancer Lett,* vol. 3835, no. 14, p. In Press Accepted Manuscript, 2014.

[64] Sun et al., "Transcription factor KLF9 suppresses the growth of hepatocellular carcinoma cells in vivo and positively regulates p53 expression.," *Cancer Lett,* vol. 3835, no. 14, p. In Press Accepted Manuscript, 2014.

[65] Zhang et al., "Lentivirus-mediated knockdown of Krüppel-like factor 9 inhibits the growth of ovarian cancer.," *Arch Gynecol Obstet,* p. Epub ahead of print, 2014.

[66] Shen et al., "KLF9, a transcription factor induced in flutamide-caused cell apoptosis, inhibits AKT activation and suppresses tumor growth of prostate cancer cells.," *Prostate,* vol. 74, no. 9, pp. 946-58, 2014.

[67] Mannava et al., "KLF9 is a novel transcriptional regulator of bortezomib- and LBH589-induced apoptosis in multiple myeloma cells.," *Blood,* vol. 119, no. 6, pp. 1450-58, 2012.

[68] Simmen et al., "The Krüppel-like factor 9 (KLF9) network in HEC-1-A endometrial carcinoma cells suggests the carcinogenic potential of dys-regulated KLF9 expression.," *Reprod Biol Endocrinol,* vol. 6, no. 41, 2008.

[69] Kang et al., "Downregulation of Krüppel-like factor 9 in human colorectal cancer.," *Pathol Int,* vol. 58, no. 6, pp. 334-338, 2008.

[70] He et al., "miR-92a/DUSP10/JNK signalling axis promotes human pancreatic cancer cells proliferation.," *Biomed Pharmacother,* vol. 68, no. 1, pp. 25-30, 2014.

[71] Nomura et al., "Novel function of MKP-5/DUSP10, a phosphatase of stress-activated kinases, on ERK-dependent gene expression, and upregulation of its gene expression in colon carcinomas.," *Oncol Rep,* vol. 28, no. 3, pp. 931-6, 2012.

[72] P. B. H. D. Xiaojun Li, "Breast Cancer Specific Markers and Methods of Use". United States of America Patent US 12/538,045, 29 July 2010.

[73] Johannesson et al., "A systematic validation of hypothesis-driven candidate genes for cervical cancer in a genome-wide association study," *Carcinogenesis,* vol. 35, no. 9, pp. 2084-88, 2014.