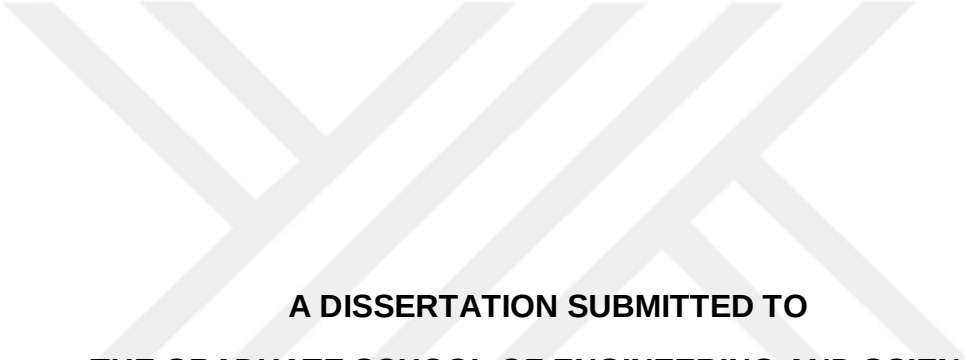


PREDICTION OF PROGNOSIS AND CHEMOSENSITIVITY IN GASTROINTESTINAL CANCERS



**A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
MOLECULAR BIOLOGY AND GENETICS**

BY

SEÇİL DEMİRKOL

DECEMBER 2017

PREDICTION OF PROGNOSIS AND CHEMOSENSITIVITY IN GASTROINTESTINAL
CANCERS

By Seil Demirkol

December 2017

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

Ali Osmay Gre (Advisor)

zlen Konu Karakayalı

Bala Gr Dedeoėlu

zgr Ŗahin

AyŖegl ner

Approved for the Graduate School of Engineering and Science:

Ezhan KaraŖan

Director of the Graduate School

ABSTRACT

PREDICTION OF PROGNOSIS AND CHEMOSENSITIVITY IN GASTROINTESTINAL CANCERS

Seçil Demirkol

Ph.D. in Molecular Biology and Genetics

Supervisor: Ali Osmay Güre

December 2017

Colon and gastric cancers are the third and fifth cancer types with the poorest survival. Surgery is considered the primary treatment option, which can be curative. However the decision as to whether chemotherapy administration after surgery is needed, is critical for especially stage 2 colon cancer; since a group of patients who receive chemotherapy do not have significantly improved clinical outcome. For this purpose, clinical risk factors are currently evaluated to determine patients with a high risk of progression. However this is not standardized by guidelines yet. Therefore, in this thesis my first aim was to *ex vivo* validate two novel independent mRNA based biomarkers, ULBP2 and SEMA5A, which have been previously identified in our lab, and to generate a prognostic signature which could stratify colon cancer patients with differential prognostic profiles using the best stratification method. I showed that a 3-group prognostic signature, SU-GIB, based on expression of these two genes is associated with cancer-specific, disease-free, and overall survival independent of clinical confounding factors in colon cancer. I performed *in silico* analysis in order to understand the biological details of the prognostic distinction, and revealed that patients with poorer prognosis show higher expression of pro-inflammatory cytokines and a more mesenchymal profile. Patients with better clinical outcome exhibit a more epithelial profile with higher levels of phosphorylated EGFR and Shc proteins. Analysis of high-throughput drug cytotoxicity databases also showed that colon cancer cell lines with 'Bad' SU-GIB signature are more sensitive to a dual PI3K-MTOR inhibitor, BEZ235.

In this thesis, my second goal was to define prognostic and molecular sub-groups for gastric cancer and characterize the specific biology of each sub-group. Utilizing an unsupervised approach on publicly available microarray data, I identified 3 biological groups, which harbor differential characteristics related to ECM involvement, EMT, proliferation and cell cycle. Moreover I identified distinct prognostic groups which are related to this molecular

classification which can further stratify patients with known pathologic subtypes (diffuse and intestinal).

I found that EMT was an important parameter for molecular and prognostic classifications for both types of cancers. I therefore, studied high-throughput drug cytotoxicity databases in order to identify selective compounds with selective growth inhibition on epithelial or mesenchymal cancer cells. I identified EGFR inhibitors as being significantly more effective on epithelial cancer cells regardless of the tissue type. My future plans include the large scale validation of SU-GIB, *ex vivo* validation of the gastric prognostic signature, and *in vitro* studies that would demonstrate effectivity of EGFR inhibitors on epithelial cancer cells and combination of EGFR inhibitors with MET inducers.

Keywords: EMT, colon cancer, gastric cancer, prognosis, biomarker

ÖZET

GASTROİNTESTİNAL KANSERLERDE PROGNOZ VE TEDAVİ CEVABI TAHMİNİ

Seçil Demirkol

Moleküler Biyoloji ve Genetik, Doktora

Tez Danışmanı: Ali Osmay Güre

Aralık 2017

Kolon ve mide kanserleri en kötü sağkalım görülen üçüncü ve beşinci kanser türleridir. Her iki kanserde de cerrahi rezeksiyon temel tedavi seçeneği olarak sunulmakta ve kuratif olabilmektedir. Ancak operasyon sonrası adjuvan kemoterapi ihtiyacı olup olmadığı kararı, özellikle evre 2 hastalar için önem arz etmektedir. Zira kemoterapi alan hastalardan bir kısmının prognozunda anlamlı bir iyileşme gözlenememektedir. Rekürans riski yüksek hastaları tahmin etmek amacıyla klinik risk faktörleri değerlendirilmektedir, ancak bu yaklaşım henüz standart bir uygulama haline getirilememiştir. Bu sebepten, bu tezde ilk amacım daha önce laboratuvarımızda keşfi yapılmış olan iki özgün mRNA temelli biyobelirteci, ULBP2 ve SEMA5A, *ex vivo* doğrulamak ve en iyi stratifikasyon metodunu kullanarak kolon kanseri hastalarını farklı prognoz gösteren gruplara ayırabilecek bir genetik test oluşturmaktır. Bu genlerin ifadeleri temelli oluşturduğum 3-gruplu bir prognostik imzanın (SU-GIB) kanser spesifik, hastalıksız ve genel sağkalım ile klinik karıştırıcı faktörlerden bağımsız olarak ilişkili olduğunu gösterdim. Bu prognostik ayrımın biyolojik temellerini anlayabilmek için yaptığımız *in silico* çalışmalar ile kötü prognozlu hastaların tümörlerinde proinflatuar sitokinlerin ifadesinin arttığını ve daha mezenkimal bir profil sergilediklerini gösterdim. İyi prognozlu hastaların tümörlerinin ise daha epitel bir profil gösterdiğini, ve EGFR ve Shc proteinlerinde de artmış fosforilasyon bulundurduklarını gözlemledim. Geniş ölçekli ilaç hassasiyet veri tabanlarının kullanarak yaptığım analizler ise 'Kötü' SU-GIB imzasına sahip kolon kanserli hücrelerin ikili bir PI3K-MTOR inhibitörü olan BEZ235'e daha hassas olduğunu gösterdi.

Bu tezdeki ikinci amacım mide kanserinde prognostik ve moleküler grupları tanımlamak ve her alt-grubun kendine özel biyolojisini karakterize etmektir. Kullanıma açık mikroarray verisetlerine uyguladığımız yönlendirilmemiş yaklaşım ile ECM, EMT, proliferasyon ve hücre döngüsü ile alakalı farklılıklar içeren 3 biyolojik grup tanımladım. Buna ek olarak, bu moleküler

gruplar ile alakalı ve bilinen patolojik alt-grupları (difüz ve intestinal) da ayırabilen prognozistik gruplar tanımladım.

Çalışmalarımda EMT'nin iki gastrointestinal kanser türünde de moleküler ve prognozistik tiplenmeyle alakalı olan önemli bir parametre olduğu sonucuna vardım. Burdan yola çıkarak, epitel ve mezenkimal kanser hücreleri üzerinde seçici etkisi olabilecek bileşikler bulmak amacıyla büyük ölçekli ilaç sitotoksikite verisetleri ile çalıştım. EGFR inhibitörlerinin epitel kanser hücreleri üzerinde, doku tipinden bağımsız ve istatistiksel olarak anlamlı şekilde daha etkili olduğunu gösterdim. Gelecekteki çalışma planlarımız dahilinde; SU-GIB'nin büyük ölçekli doğrulanması, mide kanseri prognozistik imzasının *ex vivo* doğrulanması ve EGFR inhibitörlerinin epitel kanser hücreleri üzerinde MET indükleyen bileşiklerle kombine edilerek çalışılması yer almaktadır.

Anahtar sözcükler: EMT, kolon kanseri, mide kanseri, prognoz, biyobelirteç



I dedicate my thesis to my family for their endless support and love...

ACKNOWLEDGEMENTS

I would like express my sincere gratitude to my advisor Dr. Ali O. Güre for his endless patience, support and guidance during times of both success and failure. He always lead me to ways of finding the answer rather than giving me the answer directly which is the only way I believe to develop scientific thinking in the way to become an independent researcher. I also would like to thank my advisor for being the politest person I have ever known, and being understanding whenever I needed. I am grateful to his trust in me, which was very valuable to me.

I would like to thank my thesis progression committee members Dr. Özlen Konu and Dr. Bala Gür Dedeoğlu for their contributions and valuable insights in terms of science, and also their positivity and warm smiles which gave me relief in every committee presentation.

I thank Murat İşbilen who had irreplaceable contributions to my PhD study and for being always there when I needed help about R even after his studies finished in Bilkent.

I owe special thanks to my dear friends Şükrü, Mehdi, Seyhan, Barış, Alper, Gülçin and Waqas for the joy they brought to my daily lab routine. I also thank Pınar, Şahika, Damla and Bircan for their precious friendship; Zeynep for being more than a friend and all the wonderful memories together. I thank Özlem and Erdem for being a huge support even from overseas distances.

The biggest thanks should go to my mother for letting me ask all the questions I had during my childhood, and her endless patience for trying to answer all my 'Why's, which routed me to ask more questions in science. Without her, I definitely would not be here. I thank my father for his absolute faith in me and for being besides me whatever the decision I make during all these years. I am grateful to my dear uncle for always supporting me in my academic and non-professional life. I also would like to thank my sister for being a part of me during good and bad times. I owe many thanks to my fiancée for all his priceless love, patience and mental support.

I thank TUBITAK BİDEB 2211 scholarship for their financial support during my PhD study.

CONTENTS

1. INTRODUCTION	1
1.1. Colon cancer survival - treatment	1
1.2. Colon cancer clinical prognostic markers	4
1.2.1. Major clinical factors.....	4
1.2.2. Microsatellite stability	5
1.2.3. Mutations	5
1.3. Colon cancer prognostic gene lists	6
1.4. <i>In silico</i> discovery of two novel prognostic genes for colon cancer.....	8
1.5. ULBP2 and SEMA5A.....	12
1.7. Gastric cancer statistics.....	14
1.8. Gastric cancer heterogeneity	14
1.9. Gastric cancer biomarkers.....	15
1.10. EMT and cancer	17
1.10.1. E-cadherin and Vimentin as epithelial and mesenchymal markers	18
2. AIM.....	19
3. MATERIALS AND METHODS	20
3.1. MATERIALS	20
3.1.1. General laboratory reagents.....	20
3.1.2. General Lab Equipment	21
3.1.3. Cell Culture Reagents	21
3.1.4. General lab solutions	21
3.1.5. Cell Culture Solutions.....	22
3.1.6. Primers and probes.....	22
3.1.7. <i>In silico</i> cohorts and datasets	23
3.1.8. <i>Ex vivo</i> validation Cohort.....	24
3.2. METHODS	24

3.2.1. Cell Culture Methods.....	24
3.2.2. General Lab Methods.....	26
3.2.3. Bioinformatic Methods.....	29
4. RESULTS.....	33
4.1. NOVEL MRNA BASED BIOMARKERS OF PROGNOSIS AND TREATMENT IN COLON CANCER	33
4.1.1. Analysis of ULBP2 and SEMA5A at multiple cut-offs and in stage 2&3 disease	33
4.1.2. SSAT true positivity analysis	37
4.1.3. <i>Ex vivo</i> validation of ULBP2 and SEMA5A gene expression in Turkish cohort	39
4.1.4. Generation of 'SU-GIB' prognostic signature.....	40
4.1.5. SU-GIB based prognostic classification of MSI and MSS tumors	45
4.1.6. Association SU-GIB with disease-free survival (DFS).....	46
4.1.7. Comparison of SU-GIB with other molecular and prognostic classifiers	51
4.1.8. Biological characteristics of SU-GIB groups	54
4.1.9. Response to chemotherapeutics in SU-GIB groups	58
4.2. NOVEL mRNA BASED BIOMARKERS OF PROGNOSIS IN GASTRIC CANCER	60
4.2.1. Development of a methodology for the identification of prognostic genes in gastric cancer.....	60
4.2.2. APPROACH 1 – Identification of prognostic groups in gastric cancer.....	62
4.2.3. APPROACH 2 – Identification of molecular group-specific prognostic markers in gastric cancer	76
4.2.4. EMT based classification of gastric cancer cell lines	94
4.3. <i>IN SILICO</i> IDENTIFICATION OF COMPOUNDS WITH SELECTIVE ACTION ON EPITHELIAL AND MESENCHYMAL CELLS	96
4.3.1. Generation of a in-house EMT scoring method	96
4.3.2. Identification of novel compounds that are preferentially effective on epithelial or mesenchymal cells, independent of tissue-of-origin in NCI60	99
4.3.3. Identification of novel compounds that are preferentially effective on epithelial or mesenchymal cells, independent of tissue-of-origin in CCLE and CGP.....	104

4.3.4. Identification of novel and tissue-independent parameters in addition to EMT, which will improve the prediction of sensitivity to EGFR	109
5. DISCUSSION AND CONCLUSION.....	121
5.1. Prognostic SU-GIB signature identified in colon cancer	121
5.2. Molecular and prognostic sub-groups of gastric cancer	125
5.3. Selective targeting of epithelial and mesenchymal cells in cancers.....	128
6. BIBLIOGRAPHY	131
7. APPENDIX	138
7.1. APPENDIX A.....	138
7.2. APPENDIX B.....	139
7.3. APPENDIX C.....	141
7.4. APPENDIX D.....	146
7.5. APPENDIX E.....	148
7.6. APPENDIX F.....	151
7.7. APPENDIX G.....	156
7.8. APPENDIX H.....	160
7.9. APPENDIX I.....	162
7.10. APPENDIX J.....	164
7.11. APPENDIX K.....	166
7.12. APPENDIX L	167
7.13. APPENDIX M	169
7.14. APPENDIX N.....	170
7.15. APPENDIX O.....	171
8. PUBLICATION	173

List of Figures

Figure 1.1: Representation of current clinical decision in treatment of colorectal cancer based on conventional risk factors.....	3
Figure 1.2: Survival graphs of colon cancer patients stratified based on either ULBP2 or SEMA5A gene expression I (GSE17536).....	11
Figure 1.3: Survival graphs of colon cancer patients stratified based on either ULBP2 or SEMA5A gene expression II (GSE17537).....	11
Figure 4.1: Log rank multiple cut-off (LRMC) graphs of colon cancer patients stratified based on either ULBP2 or SEMA5A gene expression (GSE17536-GSE17537).	35
Figure 4.2: Kaplan-Meier graphs for CSS of GSE17536 stratified based on either ULBP2 or SEMA5A gene expression for stage 2&3 patients	36
Figure 4.3: Survival graphs of colon cancer patients stratified based on either ULBP2 or SEMA5A gene expression III (ex vivo analysis).	39
Figure 4.4: SEMA5A/ULBP2 (SU) gene signature for colon cancer prognostication.....	41
Figure 4.5: Kaplan-Meier graphs based on the SU-GIB signature for GSE17536 (A), GSE17537 (B) and the Ankara cohort (C); restricted to stage 2&3 patients for <i>in silico</i> (A&B) and stage 3 patients for <i>ex vivo</i> (C) analysis	43
Figure 4.6: Kaplan-Meier graphs for CSS of GSE17536 microsatellite-unstable (MSI) and stable (MSS) patients as stratified by SU-GIB.	45
Figure 4.7: Kaplan-Meier graphs for DFS of GSE17536 and GSE17537 patients as stratified by SU-GIB. G, I and B stratification of patients in both cohorts reveals that this can predict disease-free survival in both cohorts. ULBP2 and SEMA5A cut-offs were determined by SSAT (4 and 6, respectively). Log-rank test p values are shown.	47
Figure 4.8: SU-GIB can predict RFS especially for microsatellite stable patients in GSE39582.. ..	47
Figure 4.9: SU-GIB can predict recurrence-free survival for microsatellite stable patients with stage 2 or 3 disease in GSE39582.....	48
Figure 4.10: Kaplan-Meier graphs for SU-GIB of MSS patients in GSE39582 as stratified by KRAS&BRAF mutational status	50
Figure 4.11: Kaplan-Meier graphs for SU-GIB of patients in GSE39582 as stratified by tumor location. Cut-off values for both genes were the most significant within the 25 th and 75 th interquartile range. Log rank p values were shown.	50
Figure 4.12: TCGA based proteome analysis of colon cancer tumor tissue in SU-GIB prognostic groups.....	55

Figure 4.13: TCGA based RNA seq. analysis of colon cancer tumor tissue reveals increased inflammatory cytokine gene expression in the bad prognosis group.....	56
Figure 4.14: SU-GIB stratification is associated with EMT. Scatter plot for tumors in the GSE39582 dataset based on E-cadherin (201131_S_AT) and Vimentin (201426_S_AT) gene expression	57
Figure 4.15: Chemosensitivity profiles of colon cancer cell lines corresponding to SU G, I and B phenotypes	58
Figure 4.16: <i>In silico</i> selection of prognostic genes for gastric cancer.	62
Figure 4.17: Hierarchical clustering analysis of gastric cancer tumors with prognostic genes....	63
Figure 4.18: Kaplan Meier plots for prognostic groups of gastric cancer tumors.....	64
Figure 4.19: Hierarchical clustering and kaplan meier analysis of gastric cancer tumors with prognostic genes in validation cohort	65
Figure 4.20: Kaplan Meier plot for prognostic groups of gastric cancer tumors with disease free survival as an end point measure	68
Figure 4.21: Kaplan Meier plot for prognostic groups of tumors in diffuse and intestinal type gastric cancers.....	70
Figure 4.22: Prognostic classification of TCGA gastric tumors based on RNAseq data.....	71
Figure 4.23: Enrichment plots of highly enriched gene sets listed in Table 4.14 in GSE62254 dataset indicating an enrichment of genes related to the extracellular matrix, cytoskeleton and a reduction of metabolic processes and cell growth	74
Figure 4.24: Enrichment plots of highly enriched gene sets listed in Table 4.14 in GSE15459 dataset show a similar pattern to that observed for GSE62254	75
Figure 4.25: Summary of high variance based stratification in gastric cancer.....	76
Figure 4.26: Determination of molecular classification via unsupervised hierarchical clustering	77
Figure 4.27: Enrichment plots of top 4 genesets enriched in GSE29272 HIGH group show an enrichment of extracellular matrix related genes	79
Figure 4.28: Enrichment plots of top 4 genesets enriched in GSE29272 LOW group indicate increased cell cycle, proliferation and mitosis.....	80
Figure 4.29: Molecular classification of gastric tumors in GSE62254 dataset.....	82
Figure 4.30: Molecular classification of gastric tumors in GSE35809 dataset.....	85
Figure 4.31: Hierarchical clustering of GSE29272 dataset with generic EMT marker genes.....	88
Figure 4.32: Molecular and EMT based classification of gastric tumors in GSE22377 dataset. .	90
Figure 4.33: Molecular classification of gastric tumors in GSE15459 dataset.....	91
Figure 4.34: Kaplan Meier plots for molecular sub-groups of gastric cancer tumors.....	93
Figure 4.35: Venn diagram of prognostic probesets in each molecular sub-group.....	94

Figure 4.36: Hierarchical clustering analysis of gastric cancer cell lines with Tan et al. [1] generic EMT signature for cell lines	95
Figure 4.37: Scatter plot of EMT scoring of NCI60 cancer cell lines	97
Figure 4.38: Correlation of a multi-gene EMT score and the CDH1-VIM score	98
Figure 4.39: Correlation plot of the multi-gene EMT score and CDH1-VIM score	98
Figure 4.40: Stem cell marker expression in EMT groups determined by CDH1-VIM score.	99
Figure 4.41: Pubchem 2D structure of compounds selectively effective on mesenchymal cells.	102
Figure 4.42: Correlation matrix of drugs which are effective on epithelial cancer cells.....	103
Figure 4.43: Sensitivity of epithelial and mesenchymal NCI60 cell lines to berberine derivatives	104
Figure 4.44: Equalization of distribution of epithelial and mesenchymal cells in CGP.....	105
Figure 4.45: Equalization of distribution of epithelial and mesenchymal cells in CCLE.	105
Figure 4.46: Sensitivity of CDH1-VIM based epithelial (EPI) and mesenchymal (MES) CGP and CCLE cell lines to EGFR inhibitors.....	106
Figure 4.47: Tissue specific sensitivity of epithelial and mesenchymal CGP cell lines to BIBW2992, EGFR inhibitor.....	107
Figure 4.48: Tissue-specific sensitivity of epithelial and mesenchymal CCLE cell lines to Lapatinib and Erlotinib, EGFR inhibitors.....	108
Figure 4.49: Analysis of mutational profile for associations with sensitivity to EGFR inhibitors	112
Figure 4.50: Microarray based expression level of EGFR in NCI60 cancer cell lines.....	113
Figure 4.51: Microarray based expression level of EGFR ligands in NCI60 cancer cell lines...	114
Figure 4.52: Venn diagram of probesets which significantly linearly correlated with cytotoxicity of BIBW2992.....	115
Figure 4.53: Venn diagram of non-coding transcripts which significantly linearly correlated with cytotoxicity of BIBW2992.	115
Figure 4.54: Correlation of cRAF protein level expression and EGFR inhibitors in CCLE and CGP databases.....	117
Figure 4.55: Linear regression analysis of cRAF and p38 MAPK for prediction of sensitivity to BIBW2992.....	118
Figure 4.56: Sensitivity to EGFR inhibitors in 4 groups of endometrial cancer cells divided by EMT and cRAF-p38.	119
Figure 4.57: Sensitivity to EGFR inhibitors in 4 groups of breast cancer cells divided by EMT and cRAF-p38.....	120

List of Tables

Table 1.1: Prognostic and predictive roles of clinicopathological parameters in colorectal cancer	4
Table 1.2: Genes with strongest survival associations by SSAT	9
Table 3.1: General laboratory reagents	20
Table 3.2: Cell Culture Reagents	21
Table 3.3: Primers	22
Table 3.4: GEO Datasets used	23
Table 4.1: Clinicopathological characteristics of patient cohorts II (DFS)	34
Table 4.2: Univariate cox regression analysis of ULBP2 and SEMA5A gene expression - CSS36	
Table 4.3: O`Connell genes validated via SSAT	38
Table 4.4: Multivariate analysis of clinicopathological parameters and SU-GIB	42
Table 4.5: Stage 2&3 restricted MVA of clinicopathological parameters and SU-GIB in GSE17536 - CSS	44
Table 4.6: Cox regression analysis with GSE39582 MSS samples - DFS	49
Table 4.7: Molecular sub-groups in Marisa et al. and SU-GIB	51
Table 4.8: MVA analyses for stage 2 & 3 patients	52
Table 4.9: MVA analyses for stage 2 & 3 patients including both Oncotype and SU-GIB	53
Table 4.10: Drugs that are significantly correlated with both SEMA5A and ULBP2 expression in colon cancer cell lines according to CGP	59
Table 4.11: Methodology of biomarker identification in colon and gastric cancers	61
Table 4.12: Univariate Cox regression analysis of clinical parameters in discovery and validation cohorts – OS	66
Table 4.13: MVA cox regression analysis of prognostic classification and stage in discovery cohorts - OS	67
Table 4.14: Gene set enrichment analysis comparing UP-DOWN prognostic groups	73
Table 4.15: Univariate Cox regression analysis of 134 probesets (significant associations reported) in GSE29272 dataset – OS	78
Table 4.16: Genesets enriched common in HIGH group	83
Table 4.17: Genesets enriched common in LOW group	84
Table 4.18: Genesets enriched common in LOW group	86
Table 4.19: Molecular and EMT based classification of gastric cancer tumors –GSE29272	89
Table 4.20: Molecular and EMT based classification of gastric cancer tumors –GSE22377	89
Table 4.21: Molecular and pathologic classification of gastric cancer tumors –GSE15479	92

Table 4.22: Molecular and prognostic classification of gastric cancer tumors –GSE15459 & GSE62254	92
Table 4.23: Cell and gene distribution in hierarchical clustering analysis in terms of epithelial-mesenchymalness	95
Table 4.24: Selection of compounds in NCI60.....	100
Table 4.25: NSC codes and Pubchem IDs of compounds selectively effecting mesenchymal cells	101
Table 4.26: List of <i>in silico</i> analyses performed	109
Table 4.27: GSEA analysis between cells lines which are sensitive and resistant to EGFR inhibitor, BIBW2992	110
Table 5.1: Characteristics associated with 'GOOD' and 'BAD' prognostic groups of SU-GIB...	124

CHAPTER 1

1.INTRODUCTION

The gastrointestinal system (GIS) is comprised of gastrointestinal tract (GI tract) and organs which are involved in digestion; pancreas, liver, gallbladder and salivary glands. Food ingestion and digestion, absorption of nutrients and disposal of waste are performed by gastrointestinal system. The GI tract includes oral cavity, pharynx, esophagus, stomach, small and large intestine, rectum, anal canal and anus. The layers of gut are mucosa, sub mucosa, muscularis externa and either adventitia or serosa. These layers are critical in evaluating tumor stage [4].

Although the parts of gastrointestinal system are composed of similar layers, the driver mutations and other risk factors, incidence and prognosis are variable among cancers of the GIS [5].

We have worked on gastrointestinal cancers originating from colon and stomach and identified biomarkers for prognosis and response to therapy. In addition to that we consistently observed a relationship of EMT with prognosis, molecular heterogeneity and drug sensitivity of the tumors/cancer cells. From this viewpoint, we studied other cancer types originating from various tissues in the body, and found that EMT is a major biological process which has a role in determining chemosensitivity in cancer regardless of the tissue of origin.

1.1. Colon cancer survival - treatment

Colorectal cancer is the third leading cancer type and the third leading cause of cancer death in both men and women in USA [6]. Almost 130,000 new cases and 50,000 deaths were recorded for colorectal cancer in 2015 [6]. According to the statistics of the Health ministry of the Turkish Republic, colon cancer is the third most common cancer among women with an incidence of 7.8%, and the fourth among men with an incidence of 7.5 [7]. 5 year survival of patients with colorectal cancer is 58% according to Cancer Research UK's statistics [8]. Although the primary treatment option for stage 2 and 3 colon cancer is surgery followed by adjuvant chemotherapy, patients receiving the same treatment regimen are known to have very different outcomes. Therefore it is of utmost importance to especially identify stage 2 patients with bad, and stage 3 patients with good prognosis, who need, and don't need therapy,

respectively. Surgery is considered curative for earlier stages of colon cancer, but still 30-50% of patients undergoing curative resection have recurrence after surgery [9].

Clinical decision making based on risk stratification in early-mid stage colon cancer is summarized in **Figure 1.1**. As can be seen, while treatment decision for stage 1 and 3 patients are standardized (stage 1 patients are kept under observation without therapy and stage 3 patients receive standard chemotherapy after surgery) the treatment decision for stage 2 patients is dependent on the assessment of high risk factors (high grade, lymphovascular invasion, perforation, obstruction, <12 lymph node examination, perineural invasion and indeterminate or positive margin) [10]. However The National Comprehensive Cancer Network (NCCN) Guidelines for colon cancer do not state clear decision routes for patients in these situations, in other words how those patients who would benefit from adjuvant chemotherapy can be determined is still unclear [11]. Therefore direct methods of assessing risk groups are essential for critical decision making in colon cancer.

FU- based chemotherapy has been a standard therapy since 1990s in order to reduce risk of recurrence [12]. Postoperative chemotherapy is recommended for stage 3 patients, however it is still under debate for stage 2 patients [13]. Fluoropyrimidine based therapies reduce risk of death 10-16 % in stage 3, which was increased 4-6% more with oxaliplatin addition, whereas FU-based treatment results in only 2-5% absolute survival benefit in stage 2 disease [14]. In other words, a group of stage 2 and 3 patients with good prognosis do not benefit from treatment but still receive chemotherapy and are exposed to toxic effects of the treatment unnecessarily. In contrast there are stage 2 patients who develop post-operative recurrence but do not receive chemotherapy because no risk of recurrence was predicted with available clinical data of the patient. In large studies, it has been shown that patients who receive post-operative chemotherapy at stage 2 or later, have recurrence of more than 20% [15], showing the need for new and most likely molecular methods for assessment of recurrence risk groups, and in line with this; potential patient subgroups with appreciable treatment benefit. Additionally, given the 5% survival benefit of standard chemotherapy as described above, novel targeted therapy options should be also utilized for specified patient groups.

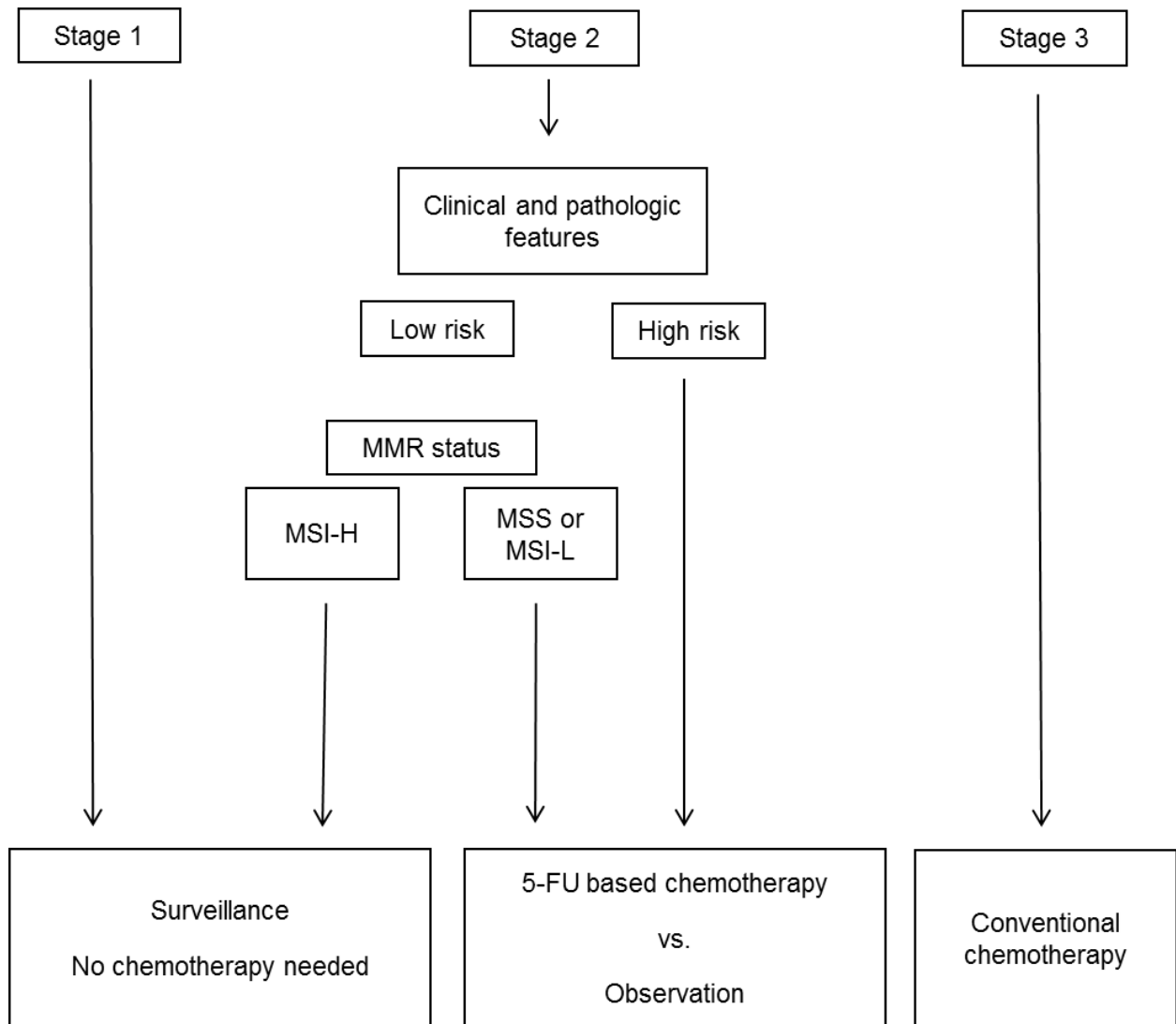


Figure 1.1: Representation of current clinical decision in treatment of colorectal cancer based on conventional risk factors. Adopted from [10]

1.2. Colon cancer clinical prognostic markers

1.2.1. Major clinical factors

Many studies on more than 400 patients in total, have evaluated the prognostic value of clinical parameters as, WBC (white blood cell), serum albumin, liver transaminases, hemoglobin, platelets, age, body fitness, performance status, elevated CEA, Ca19.9, ALP levels and pathological grading, location of the primary tumor. However it is still unstandardized which of these factors should be considered and taken into account for the assessment of prognosis [16]. Divitiis et al. summarized the major clinical prognostic factors for colon cancer (**Table 1.1**), and although prognostic value exists for these factors, prediction of response to therapy can be only poorly addressed or cannot be assessed. This non-consensus created the need for molecular tests for prediction of patient outcome that could be used with a standardized methodology. This was the driving force for the studies presented in this thesis which resulted in the validation of two novel mRNA based prognostic biomarkers and generated a combined prognostic signature which is independent of all available clinical confounding factors.

Table 1.1: Prognostic and predictive roles of clinicopathological parameters in colorectal cancer

	Positive	Negative	Methodology	Predictive of response to therapy
Young vs old*	No	Yes	CE	No
Fit vs unfit*	Yes	No		Limitations to choose the T
PS-WHO: ≤ 1	Yes	No	CE	No
PS-WHO: ≥ 2	No	Yes	CE	No
HIV positive vs negative*	No	Yes	LT, CE	Poor
Advanced clinical state	No	Yes	TNM, CE	No
Elevated CEA levels	No	Yes	LT	No
Elevated Ca19.9 levels	No	Yes	LT	No
Elevated ALP levels	No	Yes	LT	Poor

PS-WHO: Performance status- World Health Organization. CE: Clinical evaluation; LT: Laboratory test; T: Treatment (including surgery, chemotherapy, radiation therapy); TNM: Tumor node metastasis. Negative: association with poor survival. Positive: association with good survival. *For categorical analysis, the latter group was considered as reference. Adopted from [16].

Accumulation of transcriptomic and genomic data in recent years indicates that tumor location should be also considered as a confounding factor for colon cancer. It has been found that proximal tumors harbor higher levels of genome-wide hypermethylation, microsatellite instability and BRAF mutation [17]. In terms of clinical outcome, 5-year survival of patients with proximal tumors were 13 % less when compared to distal tumors (OR 0.87; 95% CI, 0.82–0.91) [18] and mortality of proximal disease is significantly higher than distal tumors [19]. Our prognostic signature was a predictor of prognosis primarily in distal cancers.

1.2.2. Microsatellite stability

Colon cancer patients with microsatellite instability (MSI) have been defined as a patient group with favorable prognosis, and with less aggressive clinical behavior when compared to patients with stable microsatellites (MSS). This prognostic pattern of MSI-high patients has been shown in recent large trials and a meta-analysis and many smaller scale retrospective studies [20]. Microsatellite instability status is also known to be associated with chemotherapy response. Patients with dMMR (deficient mismatch repair status) receiving FU showed no improvement in DFS (disease free survival) whereas patients with pMMR (proficient mismatch repair status) significantly improved DFS [21].

The underlying biology behind the better prognostic profile of colon cancer patients with MSI tumors is not clear yet [22]. However it has been shown that coding microsatellites and mutations in TGFBR2 and ACVR2A cause loss of their protein expression in MSI patients [23]. The allele loss or mutation in DCC, TP53 and KRAS which are associated with worse survival are not common in MSI tumors [24,25]. Researchers were also interested in likelihood of a cellular immune response caused by accumulation of frameshift proteins as an outcome of MSI. In this line, a higher lymphocyte infiltration rate was observed in tumors with MSI [26,27]. Furthermore, absence of MSI phenotype has been associated with expression of EMT markers [28], which could also have a role in prognostic profile of MSI tumors. Me and my colleagues work resulted in the definition of prognostic sub-groups of especially MSS tumors in colon cancer.

1.2.3. Mutations

As evaluated in 2721 patients from 13 countries, around 40% of colon cancers patients harbor a particular KRAS mutation [24] which is associated with both worse prognosis and resistance to treatment. Codon 12 glycine to valine mutation of KRAS is significantly associated with failure-free survival ($P = 0.004$, HR 1.3) and overall survival ($P = 0.008$, HR 1.29) [29]. More

importantly, KRAS mutation status is now considered as the only standard clinical biomarker for EGFR-targeted therapies in colon cancer, since many reports supported that patients with KRAS mutation do not respond to cetuximab or panitumumab [30].

BRAF mutations that occur in 7.9% of tumors have not been associated with relapse-free survival but have been found prognostic for overall survival in the PETACC-3 study which included 3278 colorectal cancer patients with stage 2 and 3 disease [31]. In case of stage 4 colorectal cancer patients, another study showed that BRAF V600E mutation, which was observed in 8.7% of the patients, was an indicator of shorter progression-free and overall survival [32]. BRAF mutant metastatic colorectal cancer patients were reported as unresponsive to EGFR-targeted therapies [33].

Furthermore, while NRAS mutations have been also found as poor prognostic indicators, studies about prognostic associations of PTEN and PIK3CA were contradictory [34-36].

Although these achievements pave the way for the prediction of response to EGFR-targeted therapy for metastatic colon cancer patients to a certain extent, other molecular tests are still required in order to make clinical decisions for the other patients with wild type tumors.

Our prognostic signature was independent of KRAS&BRAF mutational status, while patients with poor prognosis as defined by our method harbor more KRAS&BRAF mutations when compared to patients with good prognosis.

1.3. Colon cancer prognostic gene lists

Gene expression profiling assays are being investigated as an alternative approach for prediction of prognosis. Among many gene sets studied, Oncotype Dx® (Genomic Health, Inc., CA, USA) and ColoPrint (Agendia, BV., Holland) are the most promising, but still are not used in routine patient care [37]. These prognostic gene expression based tests were designed to assess the risk of recurrence after surgical removal of the primary tumor via generating risk groups.

ColoPrint is an 18-gene signature which was used for 188 fresh frozen tumor tissues from stage 1 to 4 using Agilent microarray technology and validated in an independent cohort of 206 patients from stage I to III; concluding that ColoPrint classified low risk stage 2 patients had significantly longer RFS (HR:2.5, $p=0.005$) [38]. A revalidation by Salazar et al. in 416 stage 2 colon cancer patients showed that the ColoPrint defined risk groups were significantly

associated with risk of relapse, independent of microsatellite instability and number of lymph nodes assessed in stage 2 colon cancer [39].

Studies based on fresh frozen tumor samples like Coloprint have limited clinical utility as they do not utilize the paraffin embedded sections of the pathology archives [40]. In contrast, Oncotype Dx Colon, a 12 gene qPCR based assay, was applied to formalin-fixed paraffin embedded (FFPE) tissues, which are widely used for many clinical tests and archived for each patient, and therefore have been tested in more than 1800 patients from 4 large trials [41]. The QUASAR trial [42], include 1436 patients among which 711 were treated with surgery alone and 725 patients with surgery plus chemotherapy. Patients who were treated with surgery only were divided into low, intermediate and high risk groups depending on the recurrence score of 0 to 100 (<30:low risk, 30-40:intermediate risk, >40: high risk) with Oncotype Dx. 12%, 18% and 22% risk recurrence was observed for each risk group respectively, with a HR of 1.47 (high vs low risk) and $p=0.046$. However the 11 gene based treatment score of Oncotype Dx could not be validated for prediction of fluorouracil and leucovorin based treatment response of the 725 patients. For instance the absolute chemotherapy benefit in Oncotype Dx Colon Cancer low risk and high risk groups are 3% and 6%, respectively [43].

The prognostic classification provided by these signatures is valuable, however the contribution of these signatures to clinical decision making is limited [14]. Di Narzo et al. tested 4 different prognostic gene signatures including Coloprint and Oncotype Dx in an independent cohort of 688 stage 2-3 patients, and although the results indicated a statistically significant association with OS or RFS for all 4 tests in univariate models, they contributed little to risk assessment done with already used parameters as T-stage, N-stage, and MSI [44].

A comprehensive evaluation of gene expression based tests by AHRQ (Agency for Healthcare Research and Quality) concludes that, most of the prognostic/diagnostic tests available in the literature lack multivariate analysis (MVA) and thus analysis of independence from confounding factors [45]. Therefore the value these tests add on top of evaluations done with classical clinical parameters/methods is unclear. In addition to this, most tests consider only one type of clinical outcome, however AHRQ strongly recommends that prognostic tests should be tested with multiple survival parameters including overall survival, disease-free survival, cancer-specific survival etc. [45]. In order to overcome identification of cohort-specific biomarkers, AHRQ states that evaluation of prognostic tests in multiple study cohorts is crucial [45]. Another important deficiency emphasized in this report was the assessment of treatment response prediction resulting in significant and valid changes in treatment decision [45].

As a consequence of these, there is an urgent need for studies regarding prognostic biomarkers in cancer with the requirements of AHRQ fulfilled. Therefore we included MVA, evaluated multiple survival measures in multiple independent study cohorts and also worked on possible predictive value of our prognostic indicator using publically available drug cytotoxicity databases. Our novel prognostic classification was associated with cancer-specific, overall and disease free survival significantly in multiple microarray based cohorts and was independent of all available confounding factors, but most importantly from stage. We validated our *in silico* findings *ex vivo* in fresh frozen tumors of an independent cohort. We also demonstrated that our prognostic classification has a predictive role for sensitivity to a dual PI3K/mTOR inhibitor.

1.4. *In silico* discovery of two novel prognostic genes for colon cancer

Risk assessment in especially stage 2 colon cancer is highly important for decision making to either offer chemotherapy after surgery or not, however this process has not been standardized in clinics [11]. Therefore there is an urgent need for a standardized method of risk assessment in colon cancer. Two novel mRNA based biomarkers were previously discovered in our lab for this purpose. To identify novel mRNA based biomarkers that can prognosticate colon cancer patients based on the variance in their gene expression levels; we used two microarray datasets, GSE17536 and GSE17537, with available survival data as discovery cohorts. Median survival times for patients within the GSE17536 and GSE17537 cohorts were 42.2 (0.9-142.6) and 41.7 (0.4-70.5) months respectively. Clinicopathologic characteristics of these patients are summarized in (APPENDIX M and **Table 4.1**). We used SSAT (Semi-supervised survival analysis) tool which is developed by Murat İşbilen (see Methods). SSAT algorithm was run for microarray dataset GSE17536 which resulted in 400 genes with significant association with cancer specific survival. Similarly, 269 genes were significantly associated with survival in GSE17537 dataset, and 64 of these genes were common in both datasets with hazard ratios in the same direction and shared the same threshold value. Therefore these 64 genes were analyzed further.

According to the sample sizes of GSE17536 and GSE17537 datasets, weighted averages of the two p values were calculated with a weight of 78% for GSE17536 and 22% for GSE17537 for each gene. The most significant top 20 genes (**Table 1.2**) were further used in Multivariate Cox Regression analysis (MVA) including stage as the confounding factor with the GSE17536 dataset (APPENDIX N).

Table 1.2: Genes with strongest survival associations by SSAT

Gene		NCBI gene ID	Rank
ADAM12	(ADAM metallopeptidase domain 12)	8038	1
SFRP2	(secreted frizzled-related protein 2)	6423	2
ULBP2	(UL16 binding protein 2)	80328	3
KAL1	(Kallmann syndrome 1 sequence)	3730	4
SEMA5A	(sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A)	9037	5
PXDN	(peroxidasin homolog (Drosophila))	7837	6
PCDH7	(protocadherin 7)	5099	7
CTGF	(connective tissue growth factor)	1490	8
PRKD1	(protein kinase D1)	5587	9
COL11A1	(collagen, type XI, alpha 1)	1301	10
EBF1	(early B-cell factor 1)	1879	11
SCFD1	(sec1 family domain containing 1)	23256	12
SPP1	(secreted phosphoprotein 1)	6696	13
PIM1	(Pim-1 proto-oncogene, serine/threonine kinase)	5292	14
HTRA1	(HtrA serine peptidase 1)	5654	15
SPON1	(spondin 1, extracellular matrix protein)	10418	16
CRYAB	(crystallin, alpha B)	1410	17
FABP4	(fatty acid binding protein 4, adipocyte)	2167	18
CLDN11	(claudin 11)	5010	19
LAMC2	(laminin, gamma 2)	3918	20

Out of 20 genes, ULBP2, SEMA5A and PCDH7 genes were found as stage independent prognostic factors (APPENDIX N). Next these genes were used in MVA using the GSE17537 dataset which showed that ULBP2 and SEMA5A genes were independent prognostic factors of cancer specific survival for both datasets analyzed (APPENDIX N). Kaplan Meier plots indicated that ULBP2 and SEMA5A were associated significantly with shorter and longer cancer specific survival, respectively when all stages were included in the analysis (**Figure 1.2** and **Figure 1.3**).

These data generated a basis for my studies and I *ex vivo* validated ULBP2 and SEMA5A in primary colon tumor tissues and performed further *in silico* analysis for the identification of best stratification method with these two genes and understanding biological differences among the prognostic groups.



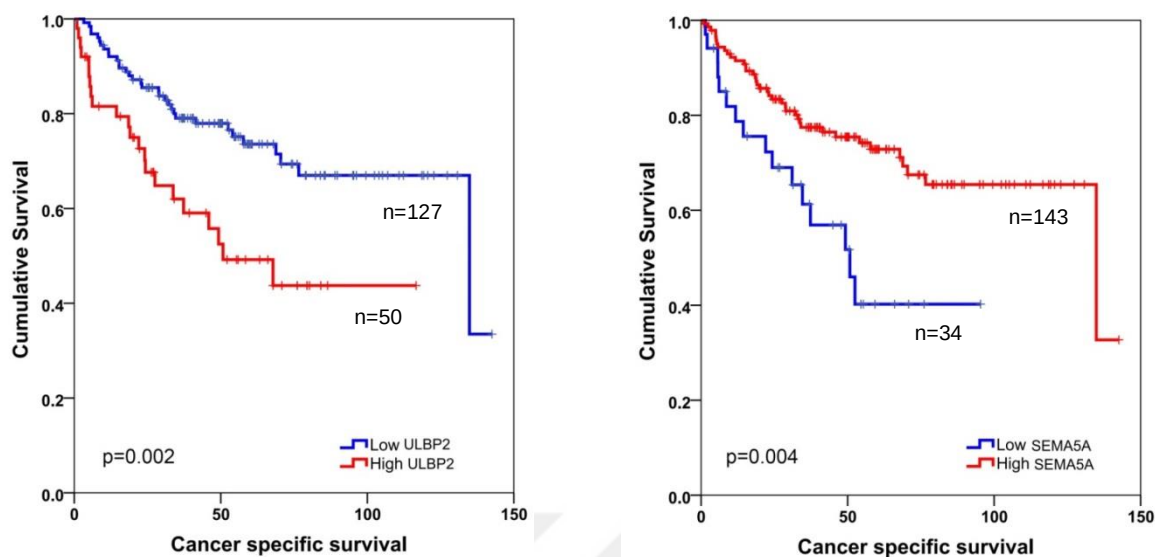


Figure 1.2: Survival graphs of colon cancer patients stratified based on either ULBP2 or SEMA5A gene expression I (GSE17536). Kaplan-Meier analyses (*in silico*) comparing “high” and “low” expression groups as defined by SSAT cut-off values for ULBP2 (A) and SEMA5A (B), which were 4 and 6, respectively. Log-rank p values are indicated. Cancer specific survival is in months.

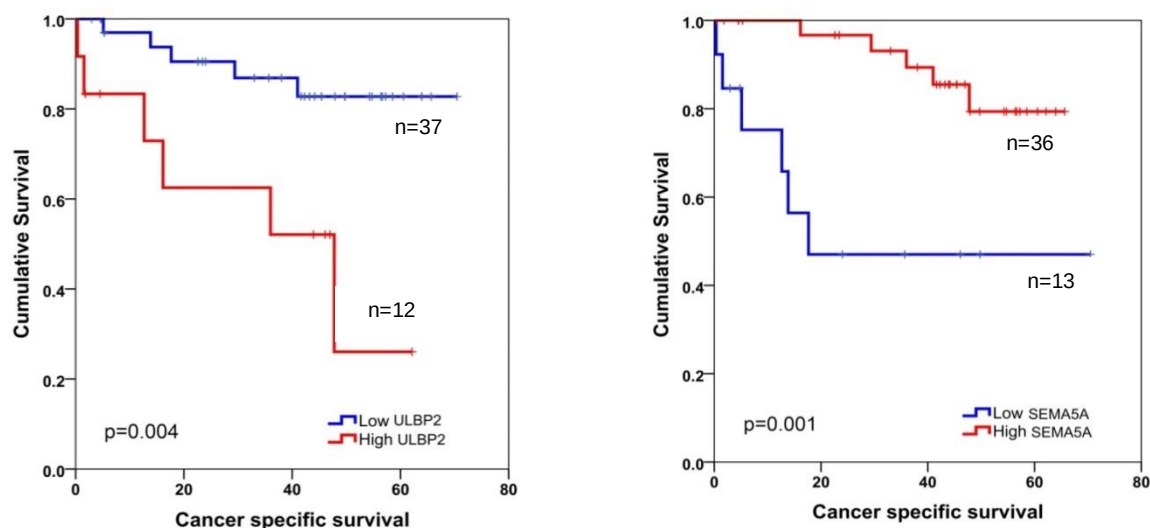


Figure 1.3: Survival graphs of colon cancer patients stratified based on either ULBP2 or SEMA5A gene expression II (GSE17537). Kaplan-Meier analyses (*in silico*) comparing “high” and “low” expression groups as defined by SSAT cut-off values for ULBP2 (A) and SEMA5A (B), which were 4 and 6, respectively. Log-rank p values are indicated. Cancer specific survival is in months.

1.5. ULBP2 and SEMA5A

Semaphorins were identified as axon guidance molecules contributing to regulation of neural connections [46]. Type 5 semaphorins including SEMA5A, are transmembrane proteins and are present in invertebrates and vertebrates [47]. They have unique thrombospondin type-1 (TSP-1) repeats in addition to the common sema domain in the extracellular region [48]. Plexin B3 is a known receptor for SEMA5A [49]. Semaphorins are involved in various cellular processes like tumor suppression, apoptosis and migration [50]. SEMA5A has been linked with invasion and migration of cancerous cells previously. SEMA5A transfection to pancreatic cancer cell lines resulted in higher cell invasion *in vitro* and injection of those cells to nude mice enhanced tumorigenicity and tumor growth *in vivo* [51]. Promotion of *in vitro* migration and invasion by SEMA5A has been also shown in gastric cancer cell lines [52]. In contrast, SEMA5A impedes motility of human glioma cells through its receptor plexin 3B [53], suggesting its functions differ based on the tissue that is analyzed. In terms of clinical outcome, up-regulation of SEMA5A has been associated with good prognosis in NSCLC in non-smoking women [54]. Other proteins in the semaphorin family have also been associated with good or bad clinical outcome [55]. SEMA3A up-regulation has been found to be significantly linked to good prognosis in NSCLC [56], tongue cancer [57], but with shorter survival in pancreatic cancer [58]. High expression of SEMA4D has been associated with poor clinical outcome in cervical cancer [59], and SEMA3G up-regulation was associated with better clinical outcome in glial tumors [60]; cumulatively suggesting that these molecules affect clinical outcome in a somewhat tissue-dependent fashion, similar to their molecular effects. In our studies, we found that higher expression of SEMA5A was associated with longer cancer specific, overall and disease free survival.

Accumulating studies showed that ULBP2 protein expression and secretion has been associated with immune response, via blocking natural killer cell mediated cytotoxicity. Natural killer (NK) cells target tumor cells and are thought to be involved in host defense against malignancy [61]. NKG2D receptors found on NK cells and cytotoxic T cells interact with their ligands like MICA and MICB which are members of MHC class 1 family and also UL16 binding proteins ULBP 1-6 molecules. MICA and MICB have restricted expression in normal tissue whereas they are highly expressed on tumor cells of lung, breast, kidney, ovary, prostate, and colon [62]. ULBP glycoproteins are highly expressed in transformed or stressed cells. They are shown to be linked to cellular membrane with or without a GPI (glycosylphosphatidylinositol) anchor and provide interaction with NK cells inducing cytokine and chemokine release [63,64].

This interaction has been shown to promote perforin bridge mediated delivery of cytotoxic proteases to tumor cell which in turn causes cell death [65].

ULBP proteins are highly expressed in colon and stomach tumors when compared to corresponding normal tissue [63]. Such cell surface expression of ULBPs was also correlated with the susceptibility of cells to NK cell mediated cell lysis [66]. Therefore ULBPs has been thought to be involved in tumor immunity in various cancer types which is supported also by defective tumor immunosurveillance in NKG2D deficient mice in a prostate cancer model [67]. GPI-anchored ULBP2 has been found to be removed from cell surface via metalloproteinases which reduced NK mediated cell lysis [68]. Therefore ULBP2 was considered a major player in tumor immunity and was studied as a prognostic marker in ovarian cancer where high expression of ULBP2 is found to be an indicator of poor prognosis [69,70].

ULBP2 was proposed as a serum biomarker for determining the prognosis and diagnosis of several other cancer types. A comprehensive analysis of 200 patients' sera showed that soluble ULBP2 is related to poor prognosis in melanoma [71]. ULBP2 has been proposed as an early screening serum biomarker for pancreatic cancer (PC) which was even superior to CA 19-9, a clinically used serum biomarker for PC, in terms of discriminating early stage PC patients from healthy controls [72]. Multivariate analysis showed that ULBP2 serum level is a prognostic marker for non-small cell lung cancer (NSCLC) and it is a diagnostic marker for squamous cell carcinoma patients [73]. In leukemia, serum ULBP2 was found to be an independent indicator of treatment free survival [74]. In order to utilize these data clinically; a recent study showed that a fusion protein combining human ULBP2 and an antibody derived chain against tumor carcino-embryonic antigen (CEA) has been found to bring NK cells and tumor cells together and induce NK mediated cell death *in vitro* and show anti-tumor activity *in vivo* [75]. Not in tumor cells but in colorectal cancer associated fibroblasts, ULBP2 has been included in a 5 –gene classifier that can predict relapse as determined by RT-PCR [76]. Our studies revealed that high expression of ULBP2 in primary colon tumors is associated with shorter cancer specific, overall and disease free survival.

1.7. Gastric cancer statistics

Gastric cancer is the fifth cancer type with the worst prognosis [6]. 5 year survival expectancy of gastric cancer patients in developed countries is 29%. The incidence in Turkey is 16.3 for men, 8.8 for women in 100.000, thus the 4th and 3rd most common cancer type, respectively [77]. Although surgical resection is considered curative, the majority of the patients are diagnosed at the later stages of the disease when the cancer has invaded other tissues, which hampers the opportunity of curative resection resulting in poor prognosis [78]. The overall survival could be only increased to 35 months with the conventional chemoradiotherapy when compared to the placebo arm with 27 months of overall survival in 559 patients with primaries $\geq$ T3 and/or node-positive gastric cancer [79]. Both predictive and prognostic biomarkers may contribute to individualized treatment of gastric cancer patients. In addition new targeted treatment strategies are necessary to improve the clinical outcome in gastric cancer [80].

1.8. Gastric cancer heterogeneity

Gastric cancer is a highly heterogeneous disease with molecular aberrations and changes at DNA, RNA and protein levels [81]. Intestinal type and diffuse type gastric cancers constitute the two major histological classes, the so called Lauren classification. In 2010; WHO (World Health Organization) defined four major histological classes which are papillary, tubular, mucinous and poorly cohesive (including signet ring cell carcinoma), and uncommon histologic variants [82]. While the diffuse type gastric cancer is known to be driven mostly by intrinsic factors (genetic abnormalities), the intestinal type is more related to extrinsic factors such as life style, diet and *Helicobacter Pylori* infection [82] which is an important extrinsic promoter of gastric cancer due to long-term inflammation and induction of change in gene expression [83].

The familial diffuse type cancers are also distinguished by EMT-like features of cells since they harbor germ-line mutations of CDH1 (E-cadherin) gene which codes for an adherence junction protein. Thus the resulting downregulation of this protein leads to dissemination of tumor cells leading to a more mesenchymal phenotype [84]. CDH1 gene alterations include large deletions, point mutations and frameshift mutations, in 46% of high-risk families [85]. Reduction in the protein level of CDH1 can be also caused by CpG hypermethylation or transcriptional repression by SNAIL [84]. Comparison of gene expression between diffuse and intestinal type cancers revealed that the ratio of only two genes, CDH1 and CDH2, is sufficient to distinguish these two sub-types- high CDH1/CDH2 ratio demonstrates a

diffuse phenotype, and diffuse type tumors have highly dominant mesenchymal features and have a worse prognostic profile [86].

Although gastric tumor biology is better understood in recent decades, there is still no treatment protocol for groups of patients with molecularly distinct tumors. Therefore understanding the molecular characteristics of gastric cancer and identification of prognostic and predictive biomarkers are critical for improved clinical outcome [87]. In this thesis, I identified molecular and prognostic sub-groups among gastric tumors with apparent differences in the biology revealing involvement of ECM, EMT and proliferation/cell cycle. In addition to that, I found that the histologically classified diffuse and intestinal type tumors can be further sub-classified into two groups with highly distinct prognostic profiles.

1.9. Gastric cancer biomarkers

There are a number of prognostic and predictive biomarkers identified for gastric cancer. A receptor tyrosine kinase, MET, which is activated by scatter factor (SF) or hepatocyte growth factor [88] that induces activation of PI3K-AKT and RAS-MAPK pathways, is overexpressed and amplified in 50% and around 20% of gastric cancer tumors, respectively [89,90]. The overexpression and amplification of MET is shown to be associated with worse survival and liver metastasis in gastric cancer [89,91].

Overexpression of HER family members HER2 and EGFR are observed in 23-26% and 27% of gastric cancers, respectively [92-94]. The transformation from low-grade to high-grade carcinoma was correlated with increasing HER2 expression [95] and was also a biomarker of treatment benefit for trastuzumab, the only targeted agent approved for gastric cancer [96,97]. Amplification of EGFR has been recorded in 5% of patients and its overexpression is associated with higher grade, higher stage and unfavorable prognosis [94,98].

Mutations and amplification in PI3K have been recorded in 7% and 67% of gastric cancer patients [99] and cause frequent activation of PI3K/mTOR pathway [100]. Deregulation in this pathway by activated mTOR has been detected in 51% of gastric cancer patients and was associated with lymph node metastasis, shorter disease-free and overall survival [101,102]. However, a phase 3 study of the mTOR inhibitor, everolimus, including 656 patients revealed that treatment did not result in a significant overall survival benefit when compared to placebo, therefore suggesting no predictive role for mTOR [103].

Several studies have aimed the identification of prognostic biomarkers in gastric cancer, In the majority of these, however, the biomarkers were not investigated in individual histological subtypes of gastric cancer (intestinal and diffuse) [104-107] or have been found as prognostic in only one subtype [108,109]. Therefore in this thesis, I define a set of prognostic biomarkers which can divide prognostic subgroups both in intestinal and diffuse type gastric cancer.

My studies in colon and gastric cancers resulted in novel biomarkers which could be used for molecular sub-grouping, prognosis and prediction of drug sensitivity. *In silico* analysis with tumors and cancer cell lines of both cancer types revealed EMT as a major player in prognosis, drug sensitivity and molecular heterogeneity; therefore I concentrated on studying the prediction of drug sensitivity with regard to EMT status of cancer cells.



1.10. EMT and cancer

Epithelial-mesenchymal transition (EMT) describes a biological process by which epithelial cells lose their attachments to the neighboring cells and the basement membrane causing loss of their cell polarity by transcriptional, proteomic, epigenetic and biochemical changes that leads to enhanced cell motility, invasiveness and resistance to apoptosis. Upon degradation of basement membrane, newly mesenchymal cells gain the ability to migrate away from the epithelial layer [110]. This biological process is well studied for its role in embryonic development and is also involved in organ fibrosis in adult tissues [111]. Lately, accumulating evidence support the reinitiation of EMT during malignant transformation and metastasis. EMT, supported by the tumor microenvironment, allows cancer cells to metastasize [112].

The process of invasion and metastasis via EMT, involves reduction of proliferative activity, activation of anti-apoptotic processes, changes in cellular features from epithelial to mesenchymal, decreased cell-cell attachment, upregulation of adhesion molecules required for cell motility and degradation of ECM [113].

Accumulating evidence reveals a relationship of EMT also with stem like properties and with drug resistance [114]. Snail, one of the transcription factors that drives EMT process, has been shown to attenuate cell cycle, and cause drug resistance via withdrawal of survival factors [115]. Another study reported that Snail and Slug have been involved in direct transcriptional repression of factors which have a role in apoptosis thus providing resistance to DNA damage inducers [116]. Overexpression of these EMT transcription factors also upregulate ABC transporters enhancing multi-drug resistance causing 10 fold higher resistance to doxorubicin in breast cancer cells [117]. Another data showed that induction of EMT via downregulation E-cadherin resulted in resistance to paclitaxel and doxorubicin in breast cancer cells [118]. In my data, cancer cells exhibiting a mesenchymal gene expression profile showed a higher resistance to EGFR inhibitors when analyzed as a pool of all cancer types and cancer specifically in multiple independent drug cytotoxicity databases.

1.10.1. E-cadherin and Vimentin as epithelial and mesenchymal markers

The protein coded by the E-cadherin (CDH1) gene, located on chromosome 16, attaches to the cadherins of adjacent cells generating a link to the intercellular cytoskeleton of epithelial cells. α -catenin links actin filaments within the cells to β -catenin which interacts with the intracellular domain of CDH1 supporting cell polarity. In addition E-cadherin is involved in differentiation, migration and behaves as a tumor suppressor gene [119,120]. In carcinoma cells, the downregulation, complete shut-down, or inactivation of E-cadherin has been observed resulting from mechanisms like mutations, promoter methylation, or transcriptional repressors [121,122]. Loss of the cadherin-cadherin cytoskeleton is associated with transformation of benign lesions to more aggressive metastatic tumors [121].

Vimentin (VIM) is a major marker of EMT. It is a type III intermediate filament which is expressed in various tissues in the human body including endothelial cells, hematopoietic cells, glial cells and fibroblasts [123,124]. Vimentin forms homopolymers and heteropolymers with other intermediate filaments forming a stable structure via presence of coiled-coiled alpha helical domains [125], playing a role in cytoarchitecture and tissue integrity [126]. Vimentin connects the nucleus to the cellular membrane and plays a major role in vimentin-associated matrix adhesions that are dynamically involved in migration [127]. Vimentin deficiency causes impaired mechanical stability, migration and contractile capacity [128]. Increased expression of VIM is strongly correlated with invasive and metastatic behaviour of tumors in many cancer types and therefore has role in tumor progression [129-132].

Lately, the EMT status of cancer cells has been evaluated in terms of their expression of several epithelial and mesenchymal markers or generating complex scoring algorithms based on EMT gene expression [1]. I developed a novel and simpler scoring method, using E-cadherin and Vimentin mRNA expression levels, for assessment of EMT status in various cancer types that is almost perfectly correlated with more complex scoring algorithms. I also identified novel compounds from the NCI60 database which selectively target epithelial and mesenchymal cells in all cancer types analyzed. I found out that EMT is a biomarker of sensitivity to EGFR targeted agents *in vitro*. This could be useful especially for metastatic colon cancer patients with wild type KRAS who are candidates for anti-EGFR therapy, since a large clinical trial indicated no benefit of cetuximab addition to oxaliplatin plus 5-FU based therapy for those patients in terms of overall and progression-free survival [133]. My data showed that a sub-group of those patients with epithelial type tumor may benefit from EGFR inhibitor, and I could show this to be the case regardless of the cancer type *in vitro*.

CHAPTER 2

2.AIM

The prediction of prognosis and risk assessment in colon cancer has not yet been standardized in clinical practice especially for stage 2 disease where this is very critical. Considering the hazardous side effects of chemoradiotherapy and very little survival benefit of FU based treatment, predicting a sub-group of patients with high risk of progression as candidates for adjuvant chemotherapy after surgery is of utmost importance. For this purpose, in order to stratify colon cancer patients according to prognosis independent of clinical confounding factors as stage, grade, lymphovascular invasion, age, gender etc., two novel mRNA biomarkers, ULBP2 and SEMA5A were discovered previously in our lab. My first aim was to *ex vivo* validate these biomarkers and to identify the best stratification method for the prognostic testing. Together with my colleagues we identified and validated novel prognostic markers and generated a combined 3-class prognostic signature “SU-GIB”. My second goal was to resolve the underlying biology of SU-GIB signature; and the data revealed an inflammatory and more mesenchymal gene expression in patients with poor prognosis, and a higher EGFR phosphorylation with higher epithelial gene expression in patients with good prognosis.

Although gastric cancer is well known for its heterogeneity, no standardized treatment strategy has been developed for molecularly defined patient groups. Therefore I next aimed to define prognostic and molecularly distinct biological sub-groups for gastric cancer and to reveal the biological differences between molecular and prognostic distinctions among defined groups. I found a major involvement of EMT, ECM related cellular components; and cell cycle and proliferation related processes to two distinct gastric cancer subgroups.

Since I observed a major role for EMT in prognosis and molecular heterogeneity of two gastrointestinal cancer types; we aimed to identify compounds which have selective cytotoxic/cytostatic effect on either epithelial and mesenchymal cells. For this purpose we used publically available drug cytotoxicity databases and found that EGFR inhibitors are cancer-independent blockers of cell growth for cancer cells with an epithelial phenotype. Hence we then aimed to identify other biomarkers of sensitivity to EGFR inhibitors based on proteomic, mutational and high-throughput gene expression data and found two protein biomarkers.

CHAPTER 3

3. MATERIALS AND METHODS

3.1. MATERIALS

3.1.1. General laboratory reagents

Chemicals such as ethanol, isopropanol and chloroform were purchased from SIGMA-ALDRICH (St. Louis, MO, USA). Nuclease Free Water and RNAase Zap were bought from Ambion (Carlsbad, CA, USA) and DMSO (Dimethyl Sulfoxide) was purchased from AppliChem (Darmstadt, Germany).

Table 3.1: General laboratory reagents

Reagent	Company
DNA-free™ Kit	Invitrogen (Carlsbad, CA, USA)
Qubit™ RNA BR Assay Kit	Invitrogen (Carlsbad, CA, USA)
Agilent RNA 600 Nano Kit	Agilent technologies (Santa Clara, CA, USA)
RevertAid First Strand cDNA Synthesis Kit	Thermo Scientific (Waltham, MA, USA)
OneTaq Quick-Load 2X Master Mix with Standard Buffer	New England Biolabs (Ipswich, MA, USA)
Taqman Universal PCR Master Mix	Applied Biosystems (Carlsbad, CA, USA)
Taqman Gene Expression Assays	Applied Biosystems (Carlsbad, CA, USA)
Tri-reagent®	Ambion (Carlsbad, CA, USA)
SYBR® Green PCR Master Mix	Applied Biosystems (Carlsbad, CA, USA)

3.1.2. General Lab Equipment

Power supply and Gel Electrophoresis tank were purchased from Hoefer Inc (Holliston, MA, USA). In order to quantify and measure the purity of RNA samples, Nanodrop (Thermo Scientific, MA, USA) was used. Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA) was used to double-check the RNA quantity. Microplate Spectrophotometer μ Quant (Biotek, VT, USA) was used for luminescence reading while performing mycoplasma test and measuring cell growth inhibition.

3.1.3. Cell Culture Reagents

Table 3.2: Cell Culture Reagents

Reagent	Company
Fetal Bovine Serum	Biowest (Nuaille, France)
Rosswell Park Memorial Institute Medium (RPMI)	GIBCO (Carlsbad, CA, USA)
Dulbecco`s Modified Eagle Medium (DMEM)	Lonza (Basel, Switzerland)
Eagle`s Minimum Essential Medium (EMEM)	Lonza (Basel, Switzerland)
Trypsin-EDTA	SIGMA-ALDRICH (St. Louis, MO, USA)
L-glutamine	Hyclone (Rockford, USA)
Penicillin - Streptomycin	Hyclone (Rockford, USA)
Cell Titer Glo® Luminescent Cell Viability Assay	Promega (Fitchburg, Wisconsin, USA)
MycoAlert™ Mycoplasma Detection Kit	Lonza (Basel, Switzerland)

3.1.4. General lab solutions

50X Tris-Acetate-EDTA (TAE)

For 1 L of 50X TAE, 242 g of Tris base, 37.2 g of Tritiplex III (EDTA) and 57.1 ml of glacial acetic acid were dissolved in ddH₂O.

Ethidium Bromide

5 μ l of 10 mg/ml concentrated Ethidium Bromide was added to 100 ml Agarose Gel.

3.1.5. Cell Culture Solutions

10x Phosphate Buffered Saline

14.4 g of Na₂HPO₄·7H₂O, 80 g NaCl, 2 g KCl, 2.4 g KH₂PO₄ were dissolved in 800 ml of ddH₂O. After adjusting the pH to 7.4, the volume is filled to 1 L. 1X solution was prepared via adding 450 ml of ddH₂O to 50 ml of 10 X PBS solution. 1X PBS was autoclaved and filtered before use in cell culture.

Complete Media

500 ml of RPMI, DMEM and EMEM media all were supplemented with 50 ml of Fetal Bovine Serum, 5 ml of L-Glutamine and 5 ml of Penicillin-Streptomycin constituting approximately 10%, 1% and 1% of the final media, respectively.

Freezing Media

90% DMSO and 10% FBS solution was prepared as freezing media.

3.1.6. Primers and probes

Custom designed Taqman primer probes were used for quantification of ULBP2 in frozen tissues, (Hs00607609_mH). GAPDH, ACTB were used as reference genes (4352934E, 4352935E). Primers used for SYBR green experiments are listed below. GAPDH, ACTB and 18S were used as reference genes for SYBR green based qPCR experiments.

Table 3.3: Primers

SEMA5A_F	TCTCTCTCCTTGGCACTTTCC
SEMA5A_R	ACTGGATGCTCGGTTCTCTG
ULBP2_F	GTGCAGGAGCACCACTCG
ULBP2_R	CATACACCGTAGGTCGTGGG
GAPDH_F*	GGAGCGAGATCCCTCCAAAAT
GAPDH_R*	GGCTGTTGTCATACTTCTCATGG
ACTB_F	AGAGCTACGAGCTGCCTGAC
ACTB_R	AGCACTGTGTTGGCGTACAG
18S_F	CGTGCAATTATCAGATCAAAACCAACC
18S_R	ATGGTAGGCACGGCGACTAC

*Primer sequences were obtained from [134].

3.1.7. *In silico* cohorts and datasets

Microarray datasets that were downloaded from GEO (<http://www.ncbi.nlm.nih.gov/gds/>) of NCBI were listed in **Table 3.4**. Clinical data of each dataset were obtained from Array express (<http://www.ebi.ac.uk/arrayexpress/>) of EBI or supplementary data of the related research paper shown in **Table 3.4**. RNAseq and reverse phase protein array (RPPA) outputs of colon adenocarcinoma samples are downloaded from <https://tcga-data.nci.nih.gov>. Cancer genome project expression data (E-MTAB-783) was downloaded from Array express and corresponding drug sensitivity data was obtained from <http://www.cancerrxgene.org>. CCLE (Cancer cell line encyclopedia) gene expression data were downloaded from GEO (GSE36133) and cytotoxicity data were obtained from supplementary files of Barretina et al [135]. Raw microarray data of NCI60 database were downloaded from GEO database with accession code of GSE5720 in U133A platform. Cytotoxicity data of NCI60 were downloaded from <https://discover.nci.nih.gov/cellminer/> website of Genomics and Bioinformatics group of National Cancer Institute [136]. All clinical analyses were performed with highest number of available patient data at the time of the analysis.

Table 3.4: GEO Datasets used

Primary tissue dataset	Sample size	Survival data type	Platform	Ref.
GSE17536	177	OS*, DFS**	HG-U133 Plus 2	[137]
GSE17537	55	OS*, DFS**	HG-U133 Plus 2	[137]
GSE39582	585	RFS***	HG-U133 Plus 2	[138]
GSE62254	300	OS*	HG-U133 Plus 2	[139]
GSE15459	192	OS*	HG-U133 Plus 2	[140]
GSE29272	126	OS*	HG-U133 A	[141]
GSE35809	70	-	HG-U133 Plus 2	[142]
GSE22377	43	-	HG-U133 Plus 2	[143]

* Overall Survival

** Disease Free Survival

*** Recurrence Free Survival

3.1.8. Ex vivo validation Cohort

46 Patients over the age of 20 years with histologically confirmed primary adenocarcinoma of the colon were offered participation in the study. Patients who received prior radiation or chemotherapy were ineligible for the study. Colon carcinoma specimens were provided by Ankara Yuksek Ihtisas Hospital. All patients or their legal guardians provided signed informed consent for all human tissue samples prior to sample acquisition. All specimens underwent routine macroscopic and microscopic analysis by the pathologist. The observation period ranged from 1 to 49 months.

3.2. METHODS

3.2.1. Cell Culture Methods

3.2.1.1. Cell Lines and Maintenance of Cells in Culture

Colon cancer cell lines COLO205, HCT15, LOVO, SW837 cells were grown in RPMI; WIDR cells in EMEM, and KM12, HCT116, HT29, HCT8, SW620, SW48, SW480, LS513 cells in DMEM. All media were supplemented with 10% FBS, 1% Penicillin/Streptomycin, 1% L-glutamine. All cells were tested for Mycoplasma contamination prior to experiments. Cells were cultured in a humidified atmosphere with 5% CO₂ in air (20% O₂) at 37°C. The culture medium was changed every second day.

3.2.1.2. Thawing Cryopreserved Cells

Each cryovial of cell was thawed by hand using body temperature and the cell content is transferred to a 15 ml falcon. 5 ml of medium was added on the cell content drop by drop while shaking the falcon gently by hand in order to provide slow mixing of cell with media preventing the burst of cells. The cells were centrifuged at 1000 rpm for 3 min. After the removal of supernatant, the pellet was dissolved in 5-6 ml of complete medium and transferred to 25 cm² cell culture flasks. Flasks are kept at 37°C incubator. When cell attachment was accomplished, usually after a day, cell culture medium was changed.

3.2.1.3. Passaging Cultures

After a culture was reached to 70-80 % of confluency, cells were passaged to a new flask. First medium was aspirated, and cells were washed with 5 ml of 1X autoclaved and filtered PBS. After aspirating PBS, 1.5 ml of Trypsin-EDTA was added on the cells to detach the cells from the flask. The flask is then placed in 37°C incubator for 2-3 min. Upon the observation

of cell detachment roughly, 5 ml of cell culture medium was used for resuspending cells. The cells were then transferred to a 15 ml falcon and centrifuged at 1000 rpm for 3 minutes. After the removal of the supernatant, the pellet was dissolved in 5 ml of medium. 1-2 ml of the cell content, depending on the growth rate of the cell line, was transferred to a new flask with 10 ml of fresh medium.

3.2.1.4. Assessment of Mycoplasma Contamination

MycoAlert™ Mycoplasma Detection Kit (Lonza, Basel, Switzerland) was used for the assessment of mycoplasma contamination before freezing cells and before experiments. The kit reagent and substrate were prepared via reconstituting the lyophilized pellets as recommended. 2-3 ml of at least a day old cell supernatant was transferred into a 15 ml falcon and centrifuged at 1000 rpm for 3 min in order to get rid of the cellular content. 20 µl of the supernatant was put in a well of opaque 96-well plate. 20 µl of MycoAlert™ Reagent was added onto the supernatant. Plate was then incubated at room temperature for 5 min. Luminescence was measured by Microplate Spectrophotometer µQuant (Biotek, Winooski, USA) and the reading was saved as 'Reading A'. Then 20 µl of MycoAlert™ Substrate was added to the supernatant and reagent mix. After 10 min incubation at room temperature, luminescence was measured and recorded as 'Reading B'. The same procedure is repeated in a separate well with fresh media as negative control. Reading B divided by Reading A above 1.2 indicates mycoplasma contamination, below 0.9 shows no signs of mycoplasma, and a score between 0.9 - 1.2 requires a retest on 24 hours.

3.2.1.5. Freezing cells

After confirmation of the absence of mycoplasma contamination, cells were frozen to be stored in the nitrogen tank. Cell medium was aspirated and cells were washed with 5 ml of 1X PBS. After trypsinization of cells with 1.5 ml Trypsin-EDTA and centrifugation at 1000 rpm for 3 minutes, the supernatant is sucked and the pellet was dissolved in freezing media. The amount of the cellular content was the determinant of the volume of the freezing media used. 1 ml of cells in freezing media is transferred to each cryovial and kept at -20°C for 2-3 hours. Cryovials were next placed at -80°C overnight and then stocked in nitrogen tank.

3.2.1.6. Cell viability Assays

BEZ235 drug was kindly provided by Rengül Çetin Atalay laboratory in METU. 500 cells were seeded into 96-well plates in 90 µl aliquots. 10 different concentrations of BEZ235 (1000, 300, 100, 30, 10, 3, 1, 0.3, 0.1, 0.03 µM) in 1% DMSO were added onto cells in 10 µl at day 1 so that DMSO concentration reaches to 0.1% in the wells which is not toxic to cells. After 72

hours of incubation at 37 °C, the viability of cells was measured using Cell Titer Glo Reagent (G7572, Promega, Fitchburg, Wisconsin, USA). 20 µl of CellTiter-Glo® Reagent was added to each well, and pipetted. Plate was orbitally shaken for 2 min to induce cell lysis and then incubated at room temperature to stabilize luminescent signal. Luminescence was then recorded using The Reporter Microplate Luminometer (Turner Designs, CA, USA). 'No treatment' and '0.1 % DMSO' controls were included in each plate in order to use in calculations of percent cell viability and confirm the lack of DMSO toxicity. All treatment and control experiments were performed in 5 replicates. Percent cell viability was calculated by the following formula; % Cell Viability = ((OD of treated well)/(Average OD of DMSO control wells))*100. The percent viability values were then used for growth inhibition curves and calculation of IC50 values in Graphpad Prism Version 5 for Windows (Graphpad Software Inc., San Diego, CA, USA), www.graphpad.com.

3.2.2. General Lab Methods

3.2.2.1. RNA extraction

Cell pellet was prepared at 70-80 % confluency prior to RNA extraction in an eppendorf tube. Pellet was dissolved with 1 ml of Tri-reagent® (Ambion, Carlsbad, CA, USA) and incubated for 5 min at room temperature. 200 µl of chloroform was added pipetted vigorously. After incubation at room temperature for 10 min, the samples were centrifuged at 13000 rpm for 15 min at 4°C. Upper phase was then transferred to another eppendorf tube and 500 µl of isopropanol was added on it. Tubes were inverted gently and incubated at room temperature for 10 minutes. After centrifugation at 13000 rpm for 15 min at 4°C, supernatant was removed. 1 ml of ethanol was added onto each pellet and centrifuged at 13000 rpm for 15 min at 4°C after vortexing. Supernatant was removed carefully under the light and pellets were air dried. The pellets were then resuspended in a certain amount of nuclease free water determined by the quantity of the pellet. RNA samples were stored at -80 °C.

3.2.2.2. DNAase Treatment

DNA-free™ Kit (Invitrogen, Carlsbad, CA, USA) was used for the removal of any DNA in RNA samples. First, RNA concentration was measured by Nanodrop (Thermo Scientific, Wilmington, USA) and/or Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA). If the RNA concentration is above 200 ng/µl, it should be diluted to 200 ng/µl or less. 50 µl of reaction was set up for each sample with 44 µl of RNA sample, 1 µl of DNAase I enzyme and 5 µl of DNAase buffer. Reaction mixes were placed on the thermolyne heater and incubated at 37°C for 30 min.

3.2.2.3. RNA Quantification and Purity Assessment

Total RNA concentration and purity were measured using Nanodrop spectrophotometer (ND-1000). Samples with A260/A230 ratio between 2.0-2.2 and A260/A280 ratio between 1.8 - 2.1 were assumed pure and their concentrations were measured using Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA) which is a fluorescence based assay for more precise quantification of RNA.

Qubit™ RNA BR Assay Kit (Invitrogen, Carlsbad, CA, USA) was used for measurements. 200 µl of working solution was prepared using Quant-it reagent at 1/199 ratio in Quant-it buffer for each sample and standards. Calibration of the Qubit instrument was performed before measurement as follows; 10 µl of standards were added to 190 µl of working solution, vortexed, incubated at RT for 2 min and measured. 2 µl of each RNA sample is added on 198 µl of working solution, vortexed, incubated at RT and measured by Qubit.

3.2.2.4. cDNA Synthesis

200-500 ng of RNA was reverse transcribed into cDNA using Revert-aid first strand cDNA synthesis kit (Thermo Fisher Scientific, Boston, MA, USA). 1 µl of random hexamer primer was added on 11 µl of 500 ng RNA sample and nuclease free water. RNA and primer were incubated at 65 °C for 5 min. Samples are then chilled on ice and spined down and placed back on ice. Reaction master mix was prepared containing 1 µl of 5X reaction buffer, 1 µl of Ribolock RNase inhibitor, 2 µl 10 mM dNTP mix and 1 µl of RevertAid M-MuLV RT for each sample. 8 µl of reaction master mix was added on RNA+primer mix completing the reaction volume to 20 µl. Each sample was mixed gently and centrifuged briefly, then incubated at 25 °C for 5 min and 42 °C for 60 min. Termination of the reaction was performed by incubating samples at 70 °C for 5 min. –RT controls, reactions run with nuclease free water instead of reverse transcriptase enzyme, were run for each sample to assess genomic DNA contamination. No template control, a reaction with no sample but nuclease free water instead, was run to control reagent contamination. cDNAs were kept at -20 °C for short-term and -80 °C for long-term storage.

3.2.2.5. End Point Polymerase Chain Reaction

All samples in cDNA synthesis, including negative controls, were amplified with primers of 18S RNA in order to check for genomic DNA contamination. 18S RNA forward primer was 5'-CGTGCAATTTATCAGATCAAAACCAACC-3' and reverse primer was 5'-ATGGTAGGCACGGCGACTAC-3'. cDNA and genomic DNA were expected to be amplified since 18S has only one exon, and primers were designed within this exon. Reaction master mix

was prepared in a way that each reaction included 5 µl of 2X OneTaq™ QuickLoad® Master Mix (New England Biolabs, MA, USA), 0.2 µl of 10 µM forward and reverse primers each, 3.4 µl of nuclease free water. 1.2 µl of samples were added on 8.8 µl of reaction master mix. Initial denaturation was performed at 94 °C for 30 seconds. 35 cycles of 94 °C incubation for 30 seconds, 60 °C for 30 seconds and 72 °C for 30 seconds were performed. Reaction was terminated with 72 °C incubation for 5 min.

3.2.2.6. Agarose Gel Electrophoresis

In order to detect reagent contamination and genomic DNA contamination, end point PCR products were run on an agarose gel and visualized. 1 gr of Agarose was dissolved in 100 ml of 1X TAE solution via heating in microwave for 3-4 min. 5 µl of Ethidium Bromide was added on top, after immediate evaporation of agarose gel solution. An appropriate size of gel tank was selected and the gel solution was poured into the tank. The comb is placed. After the gel solidified, end point PCR products were loaded into the gel. 5 µl of Gene Ruler 50 bp DNA ladder (#SM0373 Thermo Scientific Waltham, MA, USA) was loaded into the last well. The gel was run at 120 V for 30 min and visualized.

3.2.2.7. Real Time PCR Experiments and Analysis

Real time PCR experiments were run using Applied Biosystems 7500 Instrument (Applied Biosystems, Carlsbad, CA, USA), using either Taqman universal master mix (Applied Biosystems, Carlsbad, CA, USA) or Power Sybr Green master mix (Applied Biosystems, Carlsbad, CA, USA). Each quantification was run in triplicates and no template control was added to each reaction plate in order to check reagent contamination. ULBP2 gene quantification was performed using Taqman technology and each reaction was prepared with 10 µl of Taqman master mix (Applied Biosystems, Carlsbad, CA, USA), 7 µl of nuclease free water, 1 µl of Taqman primer-probe mix ((Hs00607609_mH), and 2 µl of cDNA. GAPDH and ACTB genes were used as reference genes and quantifications of these genes were included for each sample in each plate using 4352934E and 4352935E primer-probes, respectively. SEMA5A gene quantification was performed using Sybr green technology and each reaction was prepared with 10 µl of Sybr Green master mix (Applied Biosystems, Carlsbad, CA, USA), 6.8 µl of nuclease free water, 0.6 µl of forward and reverse primers each and 2 µl of cDNA. GAPDH, ACTB and 18S genes were used as reference genes and quantifications of these genes were included for each sample in each plate using primers listed in primers-probes section. The assay's thermal conditions were 50°C for 2 min for UNG enzyme activity, an initial denaturation stage of 95°C 10 min, followed by 45 cycles of 95°C 15 sec denaturation, 60°C 1 min of annealing and extension. The data was analyzed using 7500 Software v2.0.4 (Applied

Biosystems, Carlsbad, CA, USA), and gene expression is normalized to two control genes GAPDH, ACTB for Taqman experiments and three control genes GAPDH, ACTB, 18S for Sybr Green experiments. The data was analyzed by using 2^{-DDCT} method [144].

3.2.3. Bioinformatic Methods

3.2.3.1. Normalization of Microarray Data

Raw microarray data were downloaded as explained in '*In silico* cohorts and datasets' and normalized with RMA or GC-RMA methods using GeneSpring 12.0 (Agilent Technologies, Santa Clara, CA, USA) or BRB array-tools [145].

3.2.3.2. SSAT – Semi supervised survival analysis tool

SSAT analysis was developed and performed using R (version 3.1.2) utilizing a script that was written by Murat İşbilen. Among multiple probesets of a gene, the one with the highest coefficient of variance was selected. 7 thresholds of gene expression were determined and named as 1,2,3,4,5,6 and 7 at microarray expression values 2, 4, 6, 8, 10, 12 and 14 in order to be used in assigning high and low expression groups. The high and low expression groups determined by each one of the 7 thresholds were used in Log-Rank analysis and the threshold with the highest significance was reported for each gene. p values below 0.05 were considered significant. Each gene has 2 thresholds for GSE17536 and GSE17537 datasets with the highest significance separately, therefore the genes which have different cut-offs from two datasets were eliminated to rule out the possibility of inconsistent cut-off based predictions. The SSAT determined interval based cut-offs for ULBP2 and SEMA5A genes were 2 and 3 which correspond to 4 and 6 as expression values. These cut-offs were used for survival analysis performed with microarray data.

3.2.3.3. mUSAT – MVA Unsupervised survival analysis tool

mUSAT code was written by Murat İşbilen using R (version 3.1.2). Cox proportional hazards regression (Cox PH), Maximally selected rank statistics (Max Stat) and Log-rank tests were used for prognostic associations with gene expression.

For each probeset, univariate Cox PH was performed with continuous expression values in order to assess the significance of the linear relationship between gene expression and survival. The probesets with p value below 0.05 were selected as 'prognostic probesets'. Then MVA was performed with the prognostic clinical parameters and each 'prognostic probeset'. As an output, the script reported the probesets which are significantly associated with survival at

the first step 'prognostic probesets', and which are also found as independent of confounding factors in the MVA 'independent prognostic probesets' (second step). In addition to regression analysis, the script performed Max stat tests for each probeset and that indicated the best cut-off between 10-90 percentile range. This was the cut-off with which the highest survival difference between the two groups was observed. Then Log rank test was used to test the significance of the survival difference between the defined groups by Max stat.

3.2.3.4. Log Rank with multiple cut-offs (LRMC) analysis

Gene expression values of each patient were taken as a threshold separating patients into low and high expression groups. Log-Rank tests were performed between these two groups at each threshold. Log-Rank p-values were plotted versus corresponding thresholds. Blue and red colors were used to indicate the association of the gene expression with good or poor prognosis, respectively. Horizontal dotted line corresponds to the significance threshold of 0.05. Vertical lines correspond to the 25th, 50th and 75th percentile expression values.

3.2.3.5. Hierarchical Clustering Analysis

Cluster 3.0 and Treeview programs were downloaded from <http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm>. Hierarchical clustering was performed by clustering both genes and arrays using Euclidian Distance as similarity metric and Complete Linkage as clustering method. Clustered data in .cdt file was then standardized with the following formula, `(Value-Average)/Standard Deviation` for each gene or probeset in order to observe within gene variations. Cdt file was then visualized with Treeview.

3.2.3.6. Microarray based Oncotype Dx risk group assessment

The recurrence risk score of Oncotype Dx is composed of 7 genes related to recurrence and 5 genes as reference genes. Each sample in GSE17536 was assigned to low or high risk groups of Oncotype Dx as described by Marisa et al [138].

3.2.3.7. Microarray based microsatellite instability assessment

This analysis was performed by Murat İşbilen. Affymetrix probesets corresponding to 46 of 64 genes published by Tian et al. [146] available in HGU 133A platform were identified and median of the expression values of the probesets were calculated in GSE41258 human colon cancer dataset for MSI-High + MSI-Low and MSS subgroups. These median values were used as reference centroids for the classification of the samples in GSE26682 HGU133 Plus 2.0 and HGU133A datasets (validation) and GSE17536 (prediction). Cosine correlation was performed between reference median values (MSI and MSS median values of 46 probesets) and the

expression of the samples that were used for prediction. The difference between MSI and MSS cosine correlation values was calculated as the MSI index for each sample. A threshold index value was determined by the ROC curves such that the maximum accuracy was obtained in both platforms of GSE26682. The threshold value used was 0.45, above which the samples were classified as MSI, but MSS otherwise.

3.2.3.8. TCGA mRNA and protein level expression based analysis

Level 3 RNA sequencing (RNAseq) and RPPA (reverse phase protein array) based proteomic data of 132 colon cancer primary tumor tissues were downloaded from cancergenome.nih.gov via the TCGA data portal. Median expression of ULBP2 and SEMA5A genes were used as cut-off for the determination of SU-GIB groups, generating 37, 55 and 40 tumors in 'Good', 'Intermediate' and 'Bad' groups, respectively. Ttest was performed between 'Good' and 'Bad' groups for all proteins in the RPPA data. ANOVA test was used to evaluate the proteins in proteomic data in terms differential expression between SU-GIB groups. ANOVA was performed for expression of proinflammatory genes in RNA seq data within SU-GIB groups. P values below 0.05 were considered significant.

3.2.3.9. GSEA – Gene set enrichment analysis

GSEA desktop application was downloaded from Broad Institute. (<http://software.broadinstitute.org/gsea/downloads.jsp>) . GSEA analysis was performed according to GSEA guidelines (<http://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html>). Gene sets were collapsed to gene symbols and maximum probe is selected for collapsing mode. Gene sets from Gene Ontology 'c5_all' was selected as gene sets database.

3.2.3.10. Generation of correlation matrix of compounds in NCI60

-logGI50 values of 58 compounds with selective action on epithelial cells in NCI60 panel were used as an input for hierarchical clustering analysis. The clustered order of compounds were extracted from hierarchical clustering output (CDT file). Pearson correlation coefficients were calculated using the sensitivity profile of each compound with all others. Correlation matrix was generated with pearson r values in the clustered order of compounds. Red: high positive correlation, green: high negative correlation. The diagonal red line with pearson r values of 1, shows 100% correlation of each compound's sensitivity profile with itself.

3.2.3.11. CDH1-VIM based EMT scoring of cancer cell lines

E-cadherin (CDH1/201131_s_at) expression in cancer cell lines were standardized as the minimum CDH1 expression will be 0 and the maximum CDH1 expression will be -1. Vimentin (VIM/201426_s_at) expression in cancer cell lines were standardized as the minimum VIM expression will be -1 and the maximum VIM expression will be 0. Therefore the sum of the standardized expression values of CDH1 and VIM genes so called the 'CDH1-VIM score' ranges between -2 to 0 (from the most epithelial to the most mesenchymal). CDH1-VIM score of -1 was used as a cut-off for separating cells into epithelial and mesenchymal groups.

3.2.3.12. Statistical Analysis

Statistical analyses were performed using Graphpad Prism Version 5 for Windows (Graphpad Software Inc., San Diego, CA, USA), www.graphpad.com. Univariate and Multivariate Cox Regression Analysis were performed and Kaplan Meier plots were drawn using SPSS Statistics 19 (IBM, Chicago, IL, USA). P values below 0.05 were considered statistically significant. Cox PH, max stat and log rank test included in SSAT and mUSAT were performed using "survival" package in R [147].



CHAPTER 4

4.RESULTS

4.1. NOVEL MRNA BASED BIOMARKERS OF PROGNOSIS AND TREATMENT IN COLON CANCER

4.1.1. Analysis of ULBP2 and SEMA5A at multiple cut-offs and in stage 2&3 disease

To test the robustness of the survival associations of ULBP2 and SEMA5A, here I analyzed them further at various cut-off points and within stage 2&3 patients to identify the best stratification method for a future prognostic test.

Clinicopathologic characteristics of the patients that have been used for the discovery and further analyses of ULBP2 and SEMA5A genes are summarized in APPENDIX M and **Table 4.1**. Log rank multiple cut-off plots indicated that ULBP2 and SEMA5A were associated significantly with shorter and longer cancer specific survival (CSS) at almost all possible cut-offs, respectively when all stages were included in the analysis (**Figure 4.1**). Kaplan meiers showed that the association of high ULBP2 and SEMA5A expression with shorter and longer CSS were also evident when the analysis was restricted to stage 2&3 patients (**Figure 4.2**). In addition ULBP2 and SEMA5A gene expression were significantly and linearly related to cancer-specific survival (**Table 4.2**).

Table 4.1: Clinicopathological characteristics of patient cohorts II (DFS)

Patient characteristics	GSE17536 (n = 177)			GSE17537 (n = 55)			GSE39582 (n=566)		
	Nr. ^{\$}	HR*	p	Nr.	HR*	p	Nr.	HR*	p
Age									
≤65 (ref.)	83	0.585	0.114	33	0.677	0.429	150	0.859	0.317
>65	94			22			415		
Unknown	-			-			1		
Gender									
Male (ref.)	96	1.000	0.999	26	2.037	0.150	310	0.790	0.124
Female	81			29			256		
TNM Stage**									
0	-	2.047	0.001	-	13.988	<0.001	4	2.605	<0.001
1	24			4			33		
2	57			15			264		
3	57			19			205		
4	39			17			60		
Grade									
Well Differentiated	16	1.848	0.076	1	1.347	0.721	NA		
Moderately	134			32			NA		
Poorly Differentiated	27			3			NA		
Other/Unknown	-			19			NA		
Microsatellite									
Low or stable (ref.)	146	1.152	0.285	NA			NA		
High	31			NA			NA		
Mismatch repair									
Proficient (ref.)	NA			NA			444	0.358	0.002
Deficient	NA			NA			75		
Unknown	NA			NA			47		
KRAS or BRAF									
Wild Type (ref.)	NA			NA			255	1.210	0.019
Mutant	NA			NA			268		
Unknown	NA			NA			43		

*Cox proportional hazards regression

**Stage: For Ankara cohort - treated as a continuous variable (1: stage 1, 2: stage 3A and 3B, 3: stage 3C and 4). For GSE17536 and GSE17537 - treated as a continuous variable (1, 2, 3, 4)

***MSI status: Determined *in silico* (see Methods)

^{\$}Nr.: case numbers

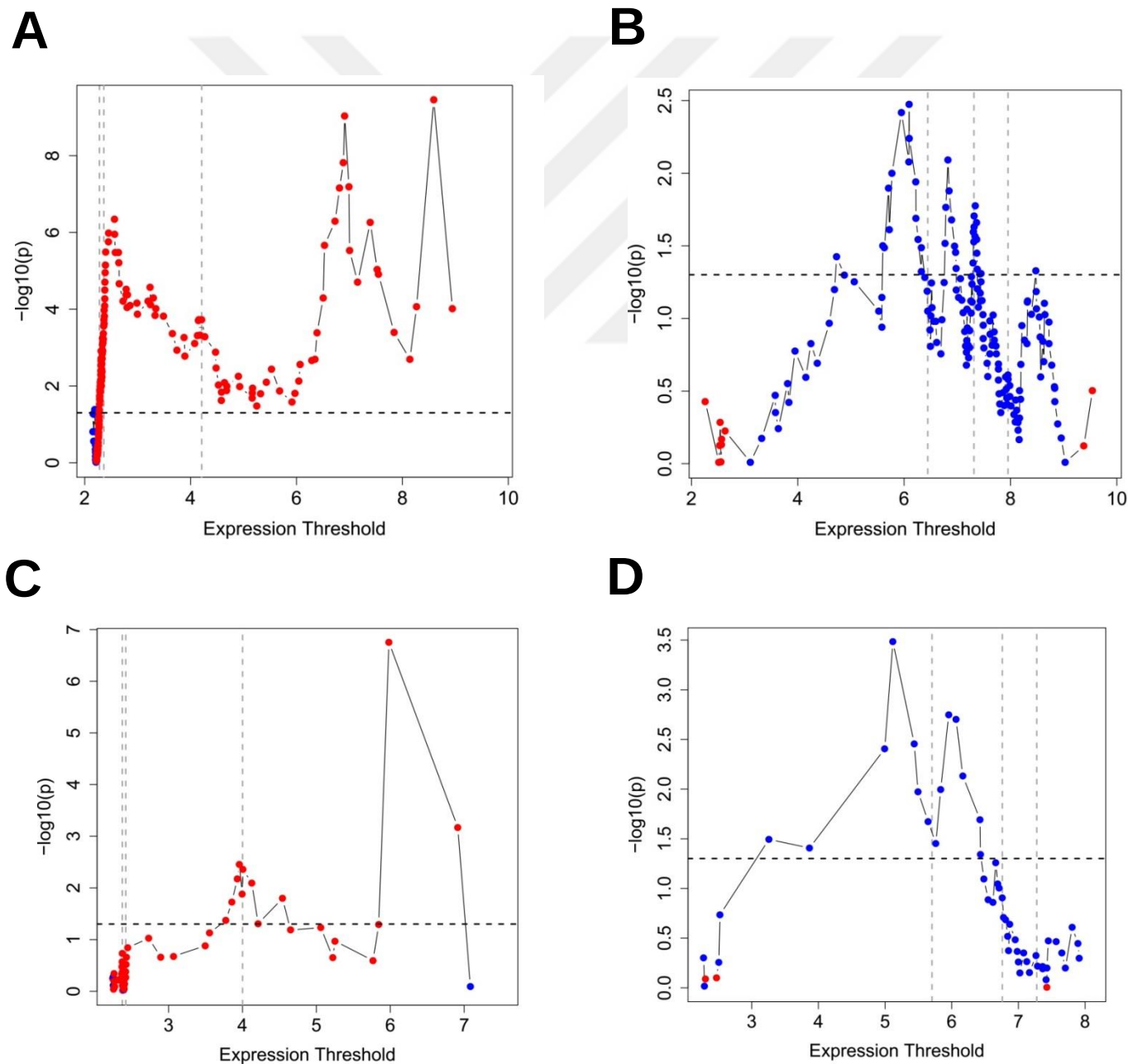


Figure 4.1: Log rank multiple cut-off (LRMC) graphs of colon cancer patients stratified based on either ULBP2 or SEMA5A gene expression (GSE17536-GSE17537). Log-rank test results plotted at all cut-off values: the graphic indicates log-rank values (shown as dots) obtained at every possible cut-off for ULBP2 (A,C) and SEMA5A (B,D) in GSE17536 (A,B) and GSE17537 (C,D). A red dot indicates the p value is associated with a HR larger than 1, when the low expression group is the reference, while blue indicates the reverse. Horizontal dotted line $p=0.05$. Vertical dotted lines: (from left to right) first 25th %, median, and 75th percentile.

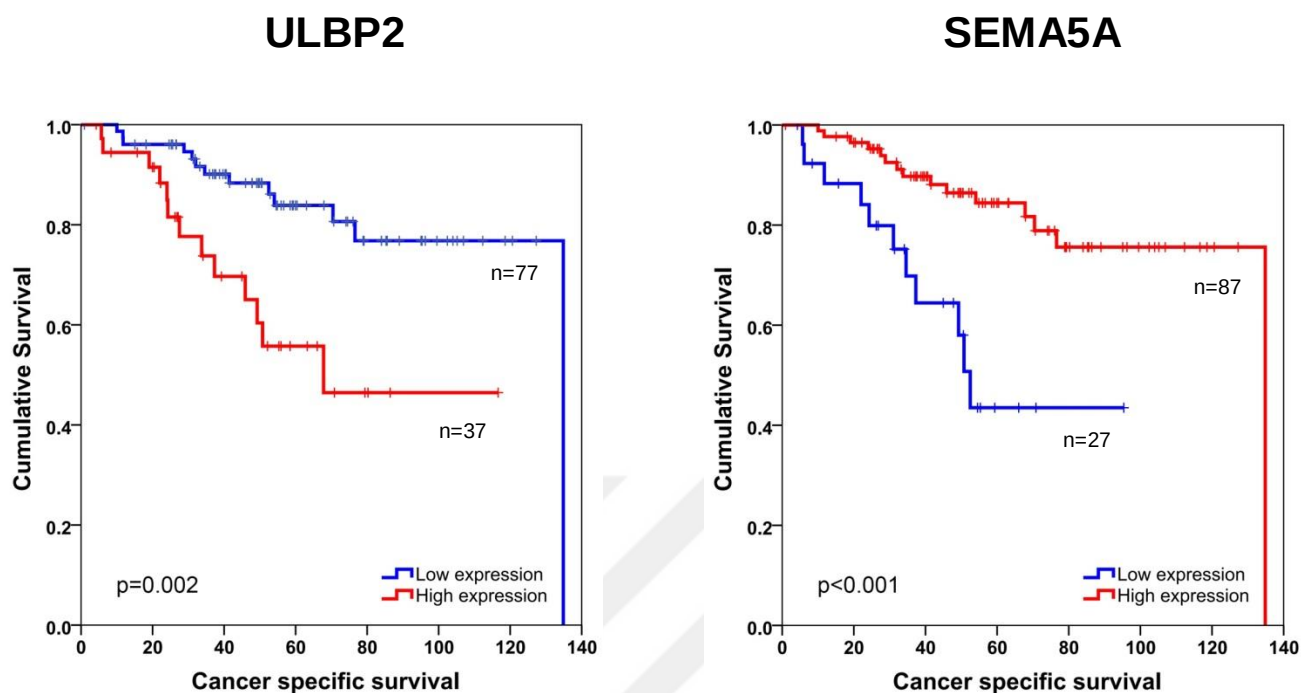


Figure 4.2: Kaplan-Meier graphs for CSS of GSE17536 stratified based on either ULBP2 or SEMA5A gene expression for stage 2&3 patients. Respective log-rank p values are shown. Survival times are in months. ULBP2 and SEMA5A expression can predict cancer specific survival when restricted to stage 2&3 patients. ULBP2 and SEMA5A cut-off values were 4 and 6, respectively.

Table 4.2: Univariate cox regression analysis of ULBP2 and SEMA5A gene expression - CSS

Dataset	Parameters	Hazard Ratio	95%CI	P*
GSE17536	ULBP2**	1.335	1.173 - 1.519	<0.001
	SEMA5A**	0.862	0.747 - 0.996	0.044
GSE17537	ULBP2**	1.670	1.119 - 2.492	0.012
	SEMA5A**	0.719	0.541 - 0.956	0.023

*Cox proportional hazards regression

**Treated as a continuous variable

4.1.2. SSAT true positivity analysis

A previously published gene list of O'Connell includes 48 genes with significant association with recurrence risk identified via quantitative polymerase chain reaction of a panel of 761 candidate genes in formalin fixed, paraffin-embedded (FFPE) tumor blocks [41]. In order to check the reliability of our method, I analyzed the survival associations of these 48 genes in our discovery cohort (GSE17536), and observed that 24 out of 45 genes (among 48) available in HGU 133 +2 platform could be associated significantly with cancer specific survival (CSS) (**Table 4.3**). When direction of survival associations were compared with O'Connell results, 23 out of 24 genes that were significant indicators of prognosis in both datasets had hazard ratios in the same direction (**Table 4.3**), which gives a true positivity rate of 53%. Although there are differences between our approach and theirs; in terms of the technology used (qPCR and microarray for O'Connell et al's and GSE17536, respectively), tissue material (FFPE and frozen tissue for O'Connell et al's and GSE17536, respectively) and the clinical endpoint measure (recurrence-free survival and cancer specific survival for O'Connell et al's and GSE17536, respectively), the fact that I can validate 53% of the genes identified in O'Connell et al supports the reliability of our approach.

Table 4.3: O'Connell genes validated via SSAT

Gene Symbol	Description	Log Rank p value	O'Connell		SSAT	
			HR> 1	HR<1	HR> 1	HR<1
AKAP12	A-kinase anchoring protein 12	<0.0001	yes		yes	
SPARC	secreted protein acidic and cysteine rich	0.0001	yes		yes	
INHBA	inhibin beta A subunit	0.0002	yes		yes	
KLK6	kallikrein related peptidase 6	0.0013	yes		yes	
SFRP2	secreted frizzled-related protein 2	0.0014	yes		yes	
MMP11	matrix metalloproteinase 11	0.0016	yes		yes	
IGFBP3	insulin like growth factor binding protein 3	0.0021	yes		yes	
FAP	fibroblast activation protein alpha	0.0024	yes		yes	
PDGFC	platelet derived growth factor C	0.0029	yes		yes	
BGN	biglycan	0.0031	yes		yes	
TIMP3	TIMP metalloproteinase inhibitor 3	0.0072	yes		yes	
DLC1	DLC1 Rho GTPase activating protein	0.0078	yes		yes	
CTHRC1	collagen triple helix repeat containing 1	0.0102	yes		yes	
LAMC2	laminin subunit gamma 2	0.0118	yes		yes	
EGR1	early growth response 1	0.0128	yes		yes	
CALD1	caldesmon 1	0.0144	yes		yes	
SERPINE1	serpin family E member 1	0.0162	yes		yes	
MYC	v-myc avian myelocytomatosis viral oncogene homolog	0.0175		yes		yes
SFRP4	secreted frizzled related protein 4	0.0188	yes		yes	
TIMP2	TIMP metalloproteinase inhibitor 2	0.0269	yes		yes	
GADD45B	growth arrest and DNA-damage-inducible 45 beta	0.032	yes		yes	
MKI67	marker of proliferation Ki-67	0.0322		yes	yes	
AKT3	AKT serine/threonine kinase 3	0.0336	yes		yes	
COL1A1	collagen type I alpha 1	0.0483	yes		yes	

4.1.3. Ex vivo validation of ULBP2 and SEMA5A gene expression in Turkish cohort

We had a validation cohort of 46 Turkish colon cancer patients enrolled in Ankara Yuksek Ihtisas Hospital with frozen tumor tissue samples and the overall survival as a measure of clinical outcome. Clinical data was summarized in (APPENDIX M). Median survival was 23 months for this cohort, as most all patients had stage 3 disease. When median of quantitative RT-PCR based expression of ULBP2 and SEMA5A in primary tumors were used as a cut-off, I confirmed that high expression of ULBP2 and SEMA5A genes were associated with shorter and longer overall survival *ex vivo*, confirming our *in silico* findings (**Figure 4.3**)

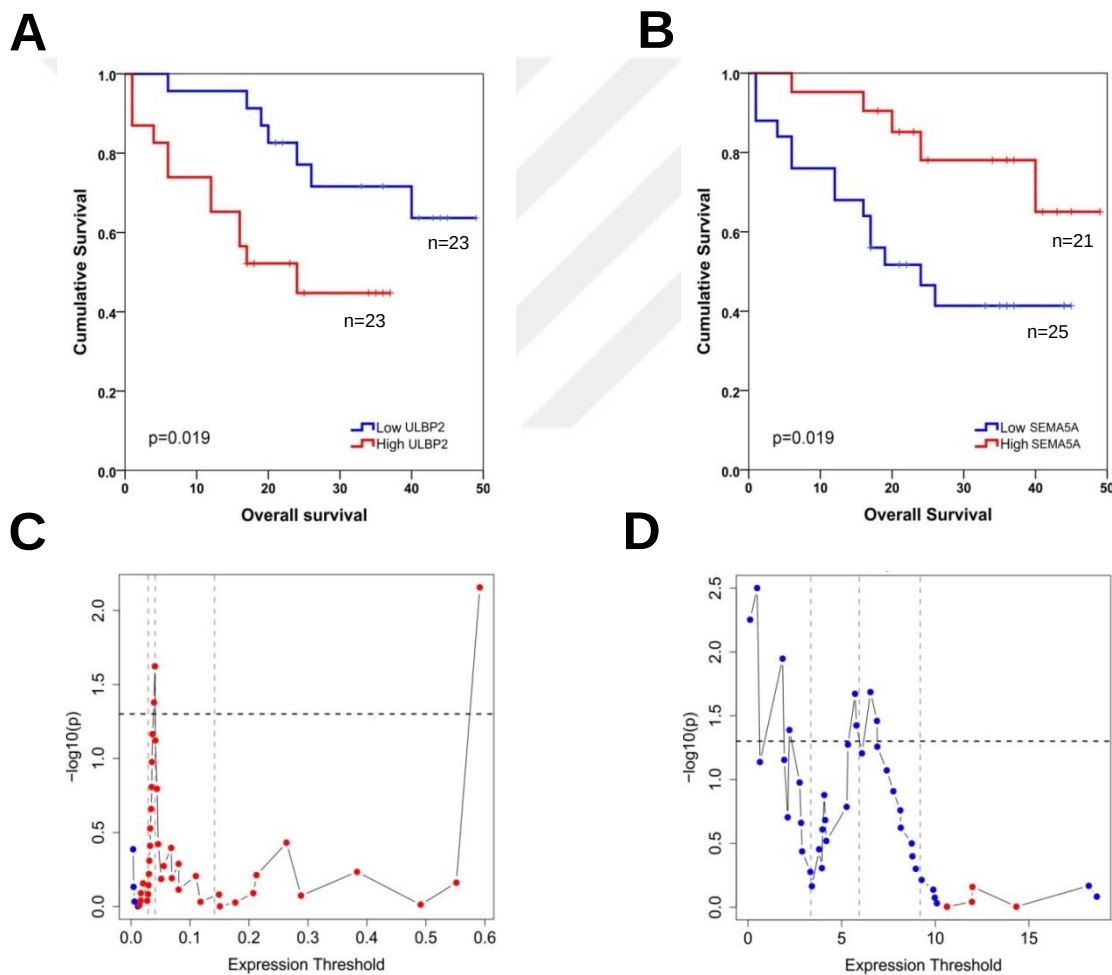


Figure 4.3: Survival graphs of colon cancer patients stratified based on either ULBP2 or SEMA5A gene expression III (ex vivo analysis). Kaplan-Meier analyses (*ex vivo*) comparing “high” and “low” expression groups based on the cut-off with the smallest log rank p value within the 25-75th % interquartile ranges for ULBP2 (A) and SEMA5A (B) for the Ankara cohort. Log-rank p values are indicated. Overall survival is given in months. C & D: Log-rank test results plotted at all cut-off values: the graphic indicates log-rank values (shown as dots) obtained at every possible cut-off for ULBP2 (C) and SEMA5A (D). A red dot indicates the p value is associated with a HR larger than 1, when the low expression group is the reference, while blue indicates the reverse. Horizontal dotted line $p=0.05$. Vertical dotted lines: (from left to right) first 25th %, median, and 75th percentile. Figure obtained with permission from Demirkol et al. [2]

4.1.4. Generation of 'SU-GIB' prognostic signature

Since the prognostic associations of ULBP2 and SEMA5A genes were independent in MVA (**Table 4.4 and 4.5**), theoretically a combination of these genes should improve the prediction of prognosis. Therefore we divided colon cancer patients into 3 groups such that patients with ULBP2 high-SEMA5A low expression were in “Bad” prognosis group, ULBP2 low-SEMA5A high expression were in “Good” prognosis group, and the rest were in “Intermediate” prognosis group. We called this prognostic classification SU-GIB (SEMA5A-ULP2 Good-Intermediate-Bad) signature. As expected, a combined score based on the evaluation of both ULBP2 and SEMA5A expression, i.e. SU-GIB, identified highly significant prognostic groups *in silico* and *ex vivo* (**Figure 4.4**) independent of confounding factors with hazard ratios ranging from 2.4 to 4.5 (**Table 4.4**), better than when patients were stratified using only one gene. SU-GIB stratification generated stage independent prognostic classification of tumors when the analysis was restricted to stage 2&3 patients (**Figure 4.5, Table 4.5**).

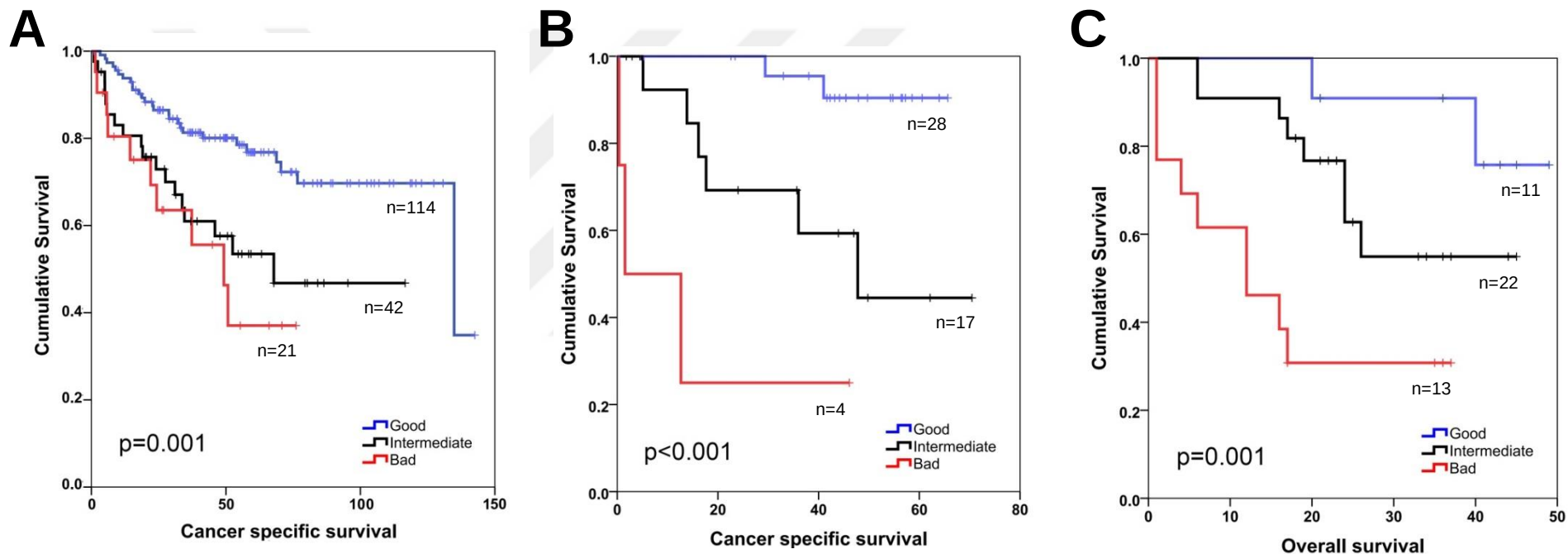


Figure 4.4: SU-GIB gene signature for colon cancer prognostication. Kaplan-Meier graphs based on the SU-GIB signature for GSE17536 (A), GSE17537 (B) and the Ankara cohort (C), and their respective log-rank p values are shown. Survival times are in months. A combined score based on ULBP2 and SEMA5A expression that separates patients into good (SEMA5A high, ULBP low), bad (SEMA5A low, ULBP high) and intermediate (patients with both SEMA5A and ULBP2 high or both low) groups results in better stratification of colon cancer patients. ULBP2 and SEMA5A cut-off values were 4 and 6, respectively for *in silico* analyses. The cut-offs with the smallest log rank p value within the 25-75th % interquartile ranges for both genes were used in the Ankara cohort. Figure obtained with permission from Demirkol et al. [2] A and B are rerepresented from Murat İşbilen's thesis [3], C is generated from my qPCR based study of *ex vivo* cohort.

Table 4.4: Multivariate analysis of clinicopathological parameters and SU-GIB

GSE17536 - CSS	Hazard Ratio	95%CI	P*
Grade**	1.165	0.685 - 1.979	0.573
Stage†	5.028	3.133 - 8.068	<0.001
Age (above 65 vs equal to or below 65)	0.818	0.455 - 1.467	0.500
Gender (female vs. male)	1.018	0.567 - 1.831	0.951
MSI_transcription based (stable vs.	1.319	0.558 - 3.119	0.529
SU-GIB [#]	2.452	1.616 - 3.720	<0.001
GSE17537 - CSS	Hazard ratio	95% CI	P*
Stage†	13.662	2.769 - 67.400	0.001
Age (equal to or below 65 vs. above 65)	1.370	0.292 - 6.434	0.690
Gender (female vs. male)	0.924	0.183 - 4.668	0.924
SU-GIB [#]	4.502	1.557 - 13.017	0.005
Ankara Cohort - OS	Hazard ratio	95% CI	P*
Grade**	1.171	0.409 - 3.358	0.769
Stage‡	4.481	1.597 - 12.574	0.004
Age (above 65 vs equal to or below 65)	2.300	0.857 - 6.174	0.098
Gender (female vs. Male)	1.062	0.407 - 2.775	0.902
SU-GIB [#]	3.481	1.514 - 8.004	0.003

*Cox proportional hazards regression

**Grade: Treated as a continuous variable; Poorly differentiated (1), moderately differentiated (2), well differentiated (3)

†Stage: Treated as a continuous variable (1, 2, 3, 4)

‡Stage: Treated as a continuous variable (1: stage 1, 2: stage 3A and 3B, 3: stage 3C and 4: stage 4)

[#] SU-GIB: Treated as a continuous variable (1:Good, 2:Intermediate, 3:Bad)

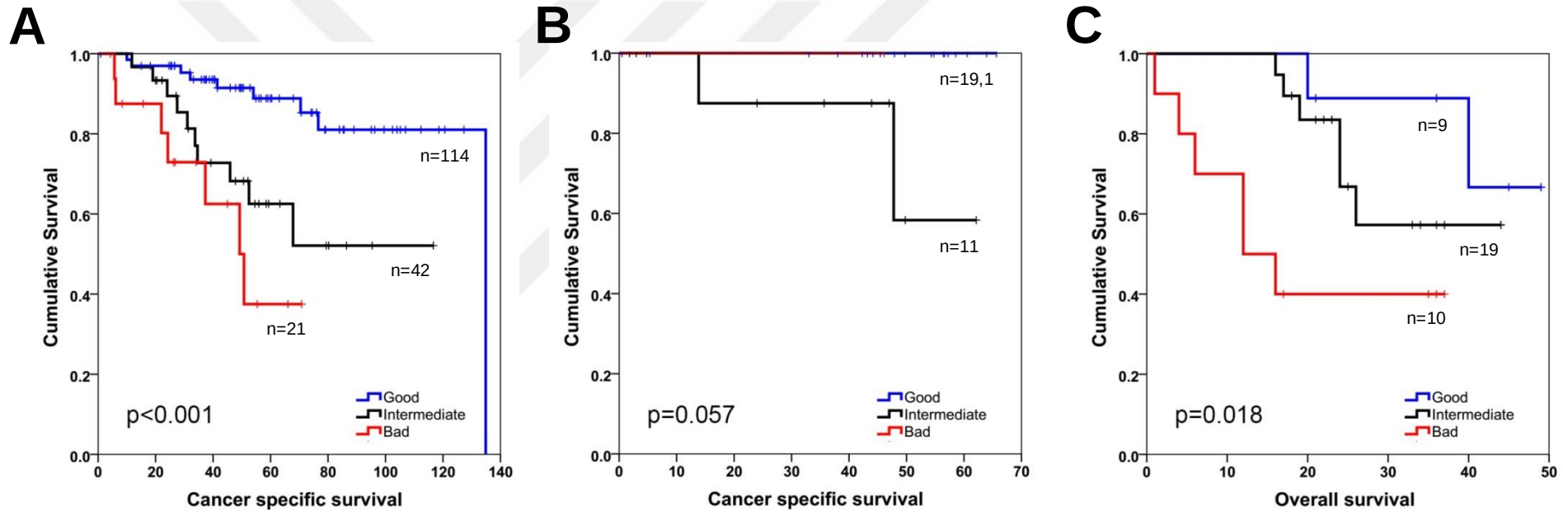


Figure 4.5: Kaplan-Meier graphs based on the SU-GIB signature for GSE17536 (A), GSE17537 (B) and the Ankara cohort (C); restricted to stage 2&3 patients for *in silico* (A&B) and stage 3 patients for *ex vivo* (C) analysis. Respective log-rank p values are shown. Survival times are in months. SU-GIB signature cut-off values for ULBP2 and SEMA5A were 4 and 6, respectively for *in silico* analyses. For the Ankara cohort, cut-off values correspond to the median values which were identical to the cut-off with the smallest log rank p value within the 25-75th % interquartile range for both genes. Figure obtained with permission from Demirkol et al. [2]

Table 4.5: Stage 2&3 restricted MVA of clinicopathological parameters and SU-GIB in GSE17536 - CSS

GSE17536		Hazard ratio	95% CI	P*
Stage (stage 3 vs 2)		2.176	0.936 - 5.058	0.071
SU-GIB	Baseline	1		0.001
	Intermediate vs good	3.501	1.374 - 8.922	0.009
	Bad vs good	7.176	2.546 - 20.225	<0.001

*Cox proportional hazards regression

4.1.5. SU-GIB based prognostic classification of MSI and MSS tumors

Favorable prognosis of colon cancer patients with deficient mismatch repair status has been supported by large-scale studies, especially in stage 2 and 3 tumors [20]. Since I have identified a prognostic classification, we wondered if our prognostic grouping can further stratify patients with and without microsatellite instability. For our discovery cohorts, mismatch repair status information of patients was not available, therefore I used a previously published gene list of Tian et al [146] which can define microsatellite instability status of patients according to the expression pattern of 64 genes (see Methods). Thus, patients with microsatellite stable (MSS) tumors could successfully be divided into SU-GIB groups with significantly different clinical outcome. The prognostic difference of SU-GIB within microsatellite instable (MSI) patients was not significant probably due to the small sample size, although the trend remains the same in both groups (**Figure 4.6**).

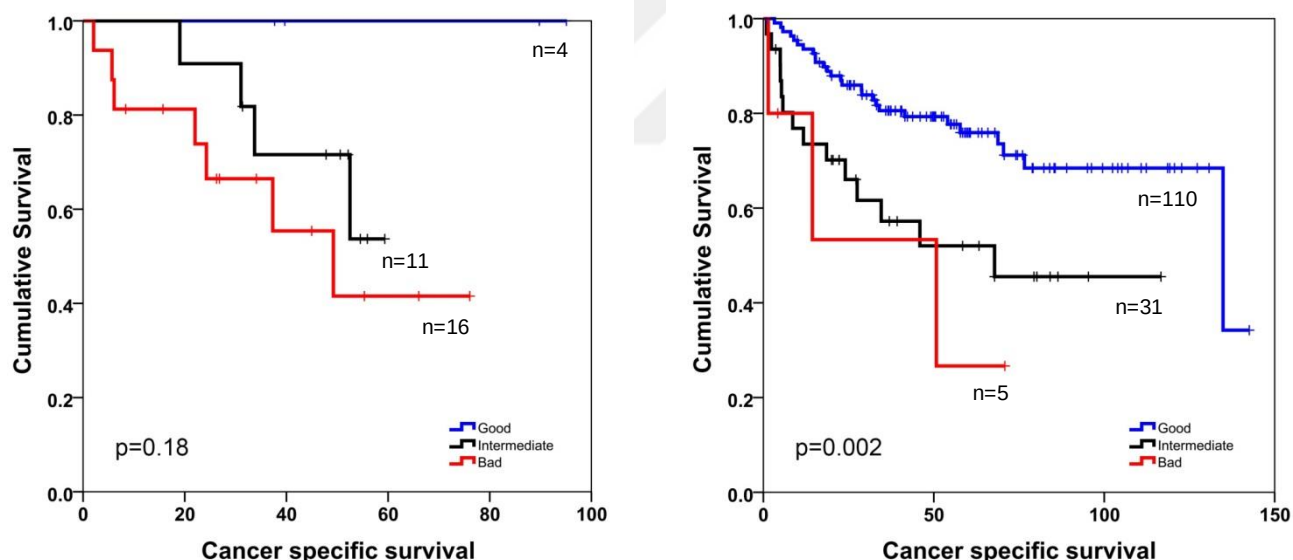


Figure 4.6: Kaplan-Meier graphs for CSS of GSE17536 microsatellite-Instable (MSI) and stable (MSS) patients as stratified by SU-GIB. ULBP2 and SEMA5A cut-offs were determined previously SSAT (4 and 6, respectively). * MSI and MSS subgroups were predicted *in silico* (see Methods). The GSE17536 cohort, separated into MSI and MSS subgroups shows that the G, I and B stratification results in the expected trend in either MSI or MSS patient groups, although only the latter shows significance by the log-rank test.

4.1.6. Association SU-GIB with disease-free survival (DFS)

SU-GIB generated 3 groups of patients with significantly different disease-free survival (**Figure 4.7**) in GSE17536 and GSE17537 datasets. When the analysis was performed separately for MSI and MSS patients in the GSE39582 cohort, a significant prognostic stratification was observed especially in the MSS group (**Figure 4.8**). In addition to that the significant prognostic association of SU-GIB was also observed in stage 2 patients with MSS tumors, and an insignificant but apparent trend in stage 3 patients with MSS tumors (**Figure 4.9**). Clinical data of GSE39582 dataset included KRAS and BRAF mutation information which have an incidence of 45% and 34% shown in a cohort of 1110 colon cancer patients, respectively [148]. Mutations in both genes are widely studied prognostic predictors of poor clinical outcome [34]. A MVA including SU-GIB, stage and MSI status demonstrated that KRAS-BRAF mutational status was not an independent prognostic factor among the patients with MSS tumors when stage and SU-GIB were included in the equation (**Table 4.6**). This revealed that SU-GIB was an independent prognostic factor for DFS when even the strongest known prognostic confounding factors, MSI, stage and KRAS-BRAF mutational status, were included in the MVA.

In order to further understand the optimum group of patients which can be stratified prognostically with high significance I analyzed MSS tumors stratified by KRAS or BRAF mutation and found that SU-GIB signature was a better prognostic indicator in terms of RFS (recurrence free survival) in the mutant group than those with wild type genotype (**Figure 4.10**)

In recent years; the location of the tumor in colorectal tumors has been reported as another parameter which is associated with patient survival. Proximal tumor location is associated with higher mortality and reduced chance of 5-year survival when compared to distal tumors [18,19]. Thus I wondered if the SU-GIB signature could stratify prognostic groups further in proximal and distal cancers. This revealed that SU-GIB could generate highly significantly distinct prognostic groups in terms of RFS within distal tumors, but not in proximal tumors (**Figure 4.11**).

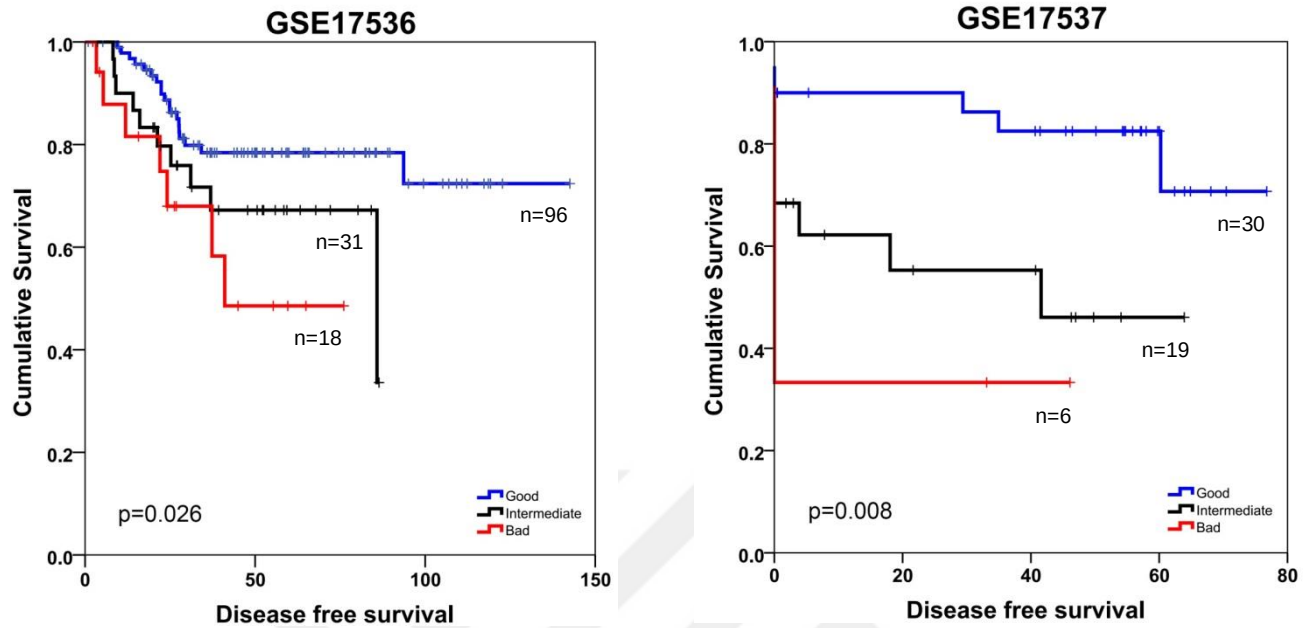


Figure 4.7: Kaplan-Meier graphs for DFS of GSE17536 and GSE17537 patients as stratified by SU-GIB. G, I and B stratification of patients in both cohorts reveals that this can predict disease-free survival in both cohorts. ULBP2 and SEMA5A cut-offs were determined by SSAT (4 and 6, respectively). Log-rank test p values are shown.

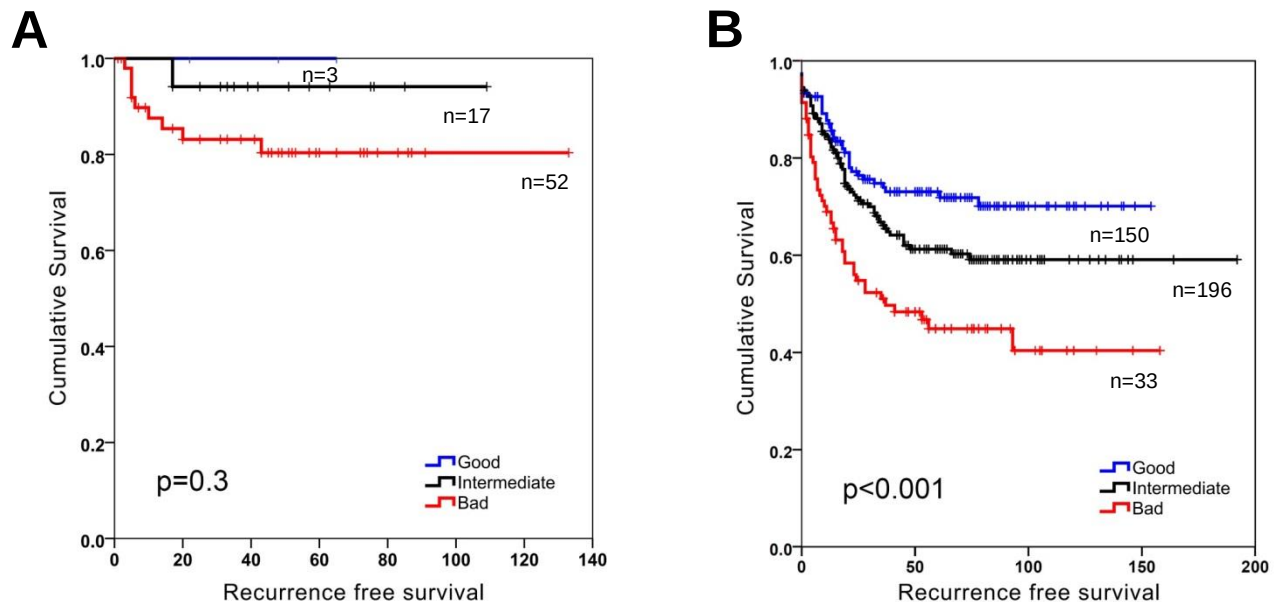


Figure 4.8: SU-GIB can predict RFS especially for microsatellite stable patients in GSE39582. Kaplan-Meier graphs of microsatellite instable (MSI) and stable (MSS) patients stratified independently using SU-GIB, and log-rank p values are shown. Cut-off values for both genes were the most significant within the 25th and 75th interquartile range.

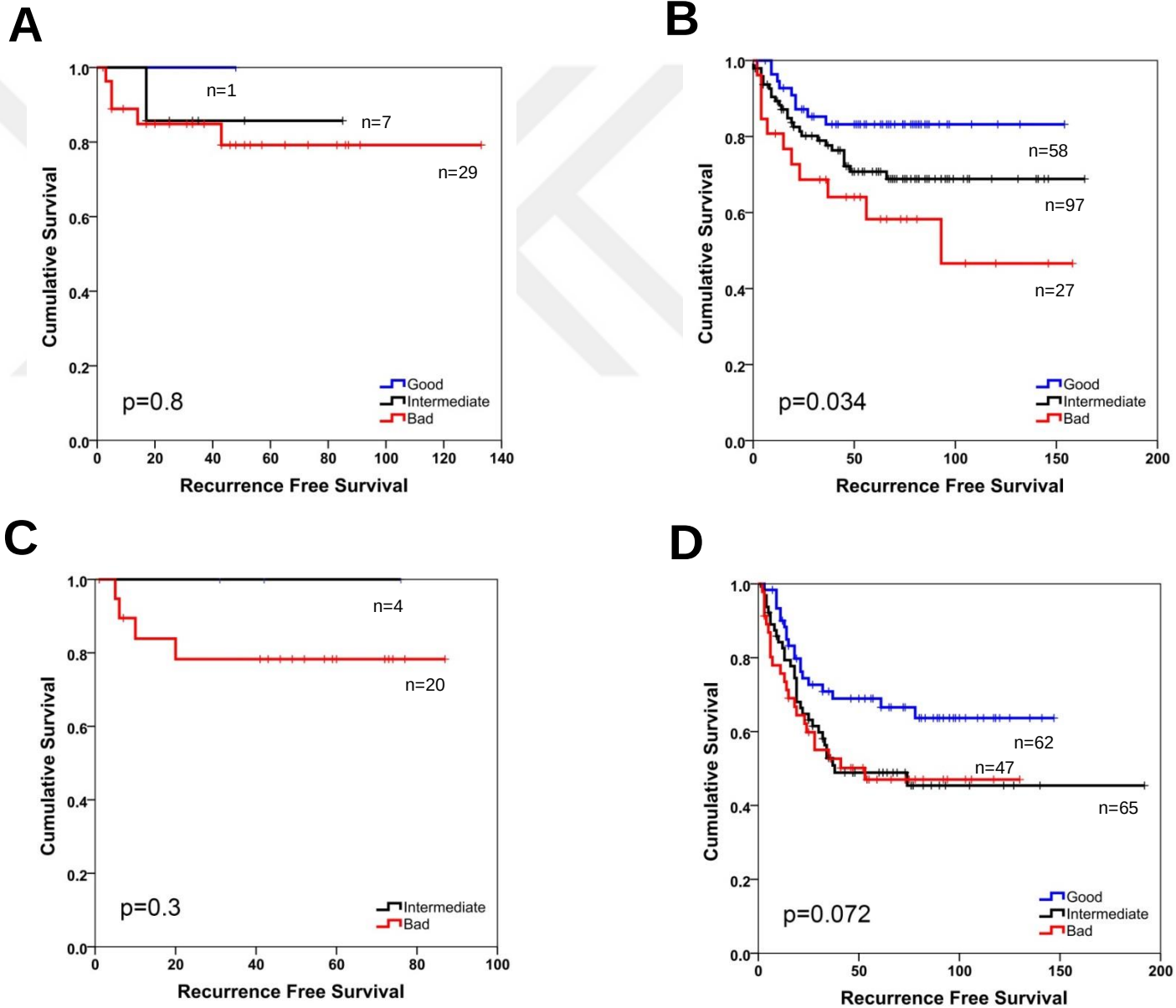


Figure 4.9: SU-GIB can predict recurrence-free survival for microsatellite stable patients with stage 2 or 3 disease in GSE39582. Kaplan-Meier graphs and log-rank p values are shown for stage 2 (A, B) and 3 (C, D) patients with microsatellite unstable (A, C) and stable (B, D) tumors. Cut-off values for both genes were the most significant within the 25th and 75th interquartile range.

Table 4.6: Cox regression analysis with GSE39582 MSS samples - DFS

KRAS or BRAF mutation (Mutant vs. Wild type)**		Hazard Ratio	95%CI	P*
		1.135	0.797 - 1.616	0.482
Stage	Baseline			<0.001
	Stage 2 vs 1	6.139	0.844 - 44.637	0.073
	Stage 3 vs 1	10.297	1.423 - 74.501	0.021
	Stage 4 vs 1	36.488	4.973 - 267.726	<0.001
SU-GIB	Baseline			0.016
	Intermediate vs good	1.472	0.959 - 2.259	0.077
	Bad vs good	1.978	1.241 - 3.154	0.004

*Cox proportional hazards regression

**KRAS or BRAF mutation: Samples are considered as wild type if only both genes are wild type, other situations are considered mutated.

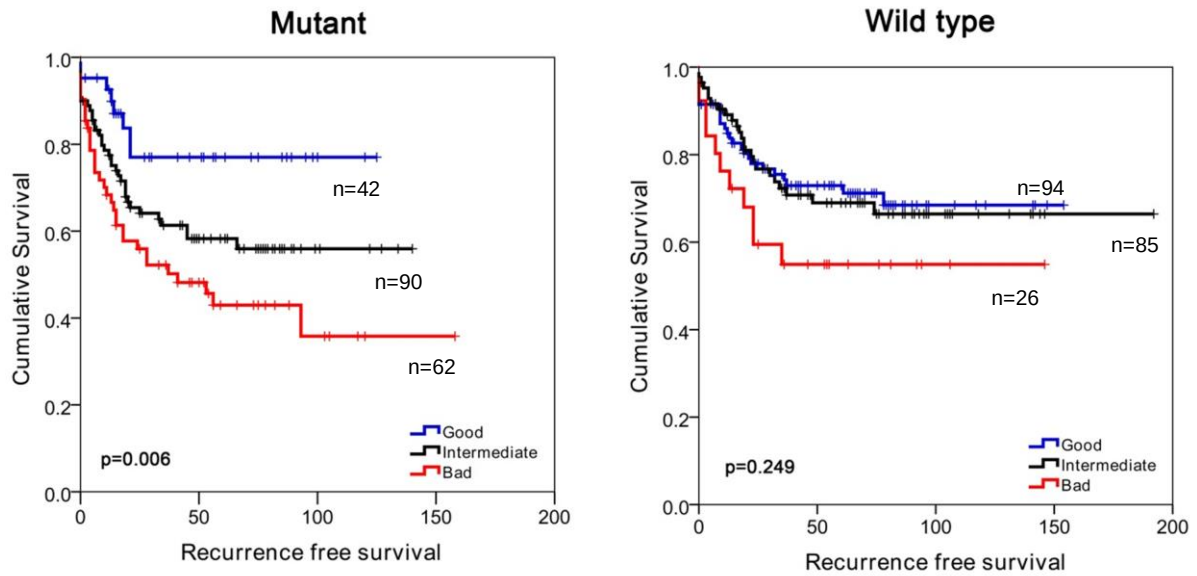


Figure 4.10: Kaplan-Meier graphs for SU-GIB of MSS patients in GSE39582 as stratified by KRAS&BRAF mutational status. Samples are considered as wild type if only both genes are wild type, other situations are considered mutated. Cut-off values for both genes were the most significant within the 25th and 75th interquartile range. Log rank p values were shown.

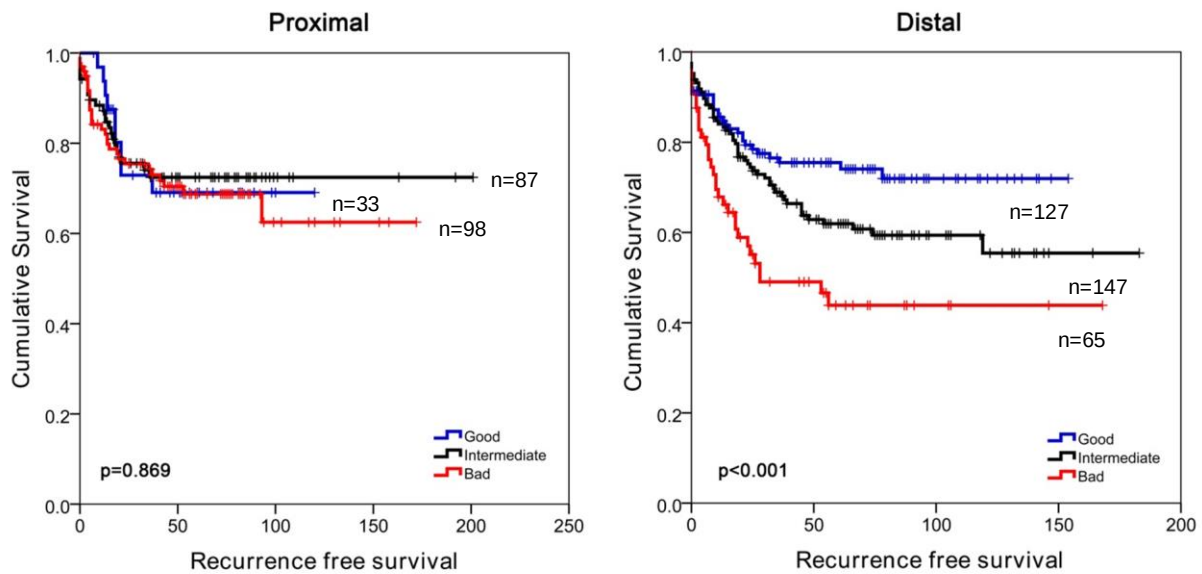


Figure 4.11: Kaplan-Meier graphs for SU-GIB of patients in GSE39582 as stratified by tumor location. Cut-off values for both genes were the most significant within the 25th and 75th interquartile range. Log rank p values were shown.

4.1.7. Comparison of SU-GIB with other molecular and prognostic classifiers

Among various prognostic gene signatures that have been published over the years, Oncotype Dx, a 12 gene qPCR based assay applied to formalin-fixed paraffin embedded (FFPE) tissues, has been tested in more than 1800 patients from 4 large trials [41] which is also applicable to microarray (see methods).

Oncotype Dx assigns patients with stage 2&3 disease to groups of low, intermediate and high risk of recurrence. Since Oncotype Dx usage is restricted to stage 2&3 patients, I compared our SU-GIB signature with Oncotype dx in patients with stage 2&3 disease. Therefore I used a published method of *in silico* application of Oncotype Dx (see Methods). MVA analysis with confounding factors and the prognostic tests included separately in the models showed that both prognostic tests to be independent prognostic classifiers in terms of predicting DFS in GSE17536 and GSE39582 datasets (**Table 4.8**). A MVA that included both prognostic tests revealed significance for SU-GIB, but not for Oncotype Dx showing that these two prognostic tests identify overlapping groups, but SU-GIB results in a statistically better stratification. (**Table 4.9**). Furthermore, I asked if our prognostic group also overlaps with a molecular classification as defined by Marisa et al. (C1-C6) which is generated based on BRAF, KRAS and p53 mutational status, DNA alterations, methylator phenotype, MSI, chromosomal stability status, and whole genome and transcriptome data of 566 colon cancer patients [138]. Among these 6 molecular groups, 62% and 53% of all patients with a good and intermediate prognostic SU-GIB signature was in DMMR(C1) and CINwntup (C5) groups, whereas the majority of patients with bad prognosis, 43%, were in CSC group (**Table 4.7**). Since mismatch repair deficiency (DMMR) is associated with favorable prognosis [20], and a CSC-like phenotype is associated widely with more aggressive and metastatic behaviour; the distribution of these groups within our SU-GIB 'Good' and 'Bad' groups was a reasonable outcome.

Table 4.7: Molecular sub-groups in Marisa et al. and SU-GIB

SU-GIB	Good	Intermediate	Bad
DMMR (C1)	50	47	17
CSC (C2)	6	24	71
KRASm (C3)	17	39	15
CINnorm (C4)	6	23	30
CINwntup (C5)	50	78	24
CINimmunedown (C6)	31	23	6

Table 4.8: MVA analyses for stage 2 & 3 patients

GSE17536 - DFS	Hazard ratio	95% CI	P*
Age (>65 vs ≤65)	0.422	0.187 - 0.956	0.039
Gender (male vs female)	1.005	0.688 - 1.467	0.98
Stage (stage 3 vs. 2)	1.732	0.802 - 3.742	0.162
Grade***	0.91	0.424 - 1.954	0.809
MSI status (Instable vs. Stable)**	2.157	0.856 - 5.437	0.13
Oncotype (High vs. Low risk)****	2.471	1.179 - 5.181	0.017
GSE17536 - DFS	Hazard ratio	95% CI	P*
Age (>65 vs ≤65)	0.459	0.209 - 1.008	0.052
Gender (male vs female)	0.988	0.681 - 1.435	0.952
Stage (stage 3 vs 2)	1.872	0.864 - 4.057	0.112
Grade***	0.925	0.468 - 1.830	0.824
MSI status (Instable vs stable)**	1.135	0.415 - 3.107	0.805
SU-GIB*****	2.004	1.158 - 3.469	0.013

*Cox proportional hazards regression

**MSI status: Determined *in silico* (see Methods)

***Grade: Treated as a continuous variable; poorly differentiated (1), moderately differentiated (2), well differentiated (3); not available for GSE39582.

****Oncotype: Determined *in silico* (see Methods)

*****SU-GIB: Treated as a continuous variable (1: Good, 2: Intermediate, 3: Bad)

GSE39582 - DFS		Hazard ratio	95% CI	P*
Age (>65 vs ≤65)		0.921	0.647 - 1.312	0.65
Gender (female vs male)		0.795	0.557 - 1.135	0.206
Stage (stage 3 vs 2)		1.734	1.214 - 2.475	0.002
MSI status (Instable vs Stable)		2.008	1.049 - 3.842	0.035
Oncotype (High vs low risk)****		1.207	1.012 - 1.439	0.037
GSE39582 - DFS		Hazard ratio	95% CI	P*
Age (>65 vs ≤65)		1.016	0.712 - 1.450	0.929
Gender (female vs male)		0.708	0.488 - 1.026	0.068
Stage (stage 3 vs 2)		1.806	1.263 - 2.583	0.001
MSI status (Instable vs stable)		0.367	0.186 - 0.726	0.004
SU-GIB	Baseline			0.002
	Intermediate vs good	1.731	1.104 - 2.713	0.017
	Bad vs good	2.47	1.494 - 4.083	<0.001

Table 4.9: MVA analyses for stage 2 & 3 patients including both Oncotype and SU-GIB

GSE39582 – DFS - stage 2 & 3		Hazard ratio	95% CI	P*
Age (>65 vs ≤65)		1.022	0.717 - 1.458	0.903
Gender (female vs male)		0.72	0.497 - 1.043	0.082
Stage (stage 3 vs 2)		1.73	1.203 - 2.488	0.003
MSI status (Instable vs stable)		0.37	0.187 - 0.732	0.004
Oncotype (High vs low risk)**		1.131	0.944 - 1.356	0.183
SU-GIB	Baseline			0.006
	Intermediate vs good	1.665	1.059 - 2.619	0.027
	Bad vs good	2.296	1.374 - 3.837	0.002

*Cox proportional hazards regression

**Oncotype: Determined *in silico* (see Methods)

4.1.8. Biological characteristics of SU-GIB groups

Up till now, we defined prognostic groups for colon cancer independent of confounding factors which was validated in 3 *in silico* and 1 *ex vivo* cohorts using overall survival, disease-specific survival and disease-free survival. I found that this classifier works better in MSS, distal and KRAS or BRAF mutant tumors when compared to MSI, proximal and KRAS&BRAF wild type tumors, respectively. In order to understand the biology behind differences in clinical outcome, I first analyzed The Cancer Genome Atlas database RNA sequencing (RNAseq) and proteomic data. I defined SU-GIB groups using RNAseq data (see methods). Top 3 most significant output of student's ttest performed with protein expression in 'Good' and 'Bad' groups resulted in higher level of phosphorylated EGFR (Y1068) and Shc proteins (Y312) in patients with 'Good' signature, and higher cleaved caspase-7 in patients with 'Bad' signature (Ttest p values are 0.00006, 0.00005 and 0.0001, respectively) (**Figure 4.12**). I also observed that EGFR phosphorylation at another (Y992) residue was also significantly up-regulated in 'Good' group when compared to the 'Bad' group ($p=0.029$) (**Figure 4.12**). The full list of all significant proteins and their ttest p values are given in APPENDIX A.

Linear correlation analysis showed that Caspase 7 cleavage was directly correlated with ULBP2 and inversely with SEMA5A expression ($p<0.001$ for both genes by Pearson's r). It has been found that Caspase 7 promotes NF κ B target activation [149], thus I analyzed if the expression of pro-inflammatory genes were upregulated in tumors with 'Bad' prognosis gene signature using TCGA RNAseq data. As expected, inflammatory factors, IL6, IL1 β , TGF β 1, TGF β 3, IL1R2, and TNF α , were significantly upregulated in 'Bad' group (**Figure 4.13**). Induction of organ fibrosis upon inflammation leading to EMT has been associated with the progression of many cancer types as well as colon cancer [150,151]. Therefore I asked if EMT was involved in prognostication of colon cancer. The analysis of 2 genes, E-cadherin as an epithelial marker and Vimentin as a mesenchymal marker, in GSE39582 dataset revealed that most of the 'Good' type tumors were epithelial, whereas the majority of 'Bad' type tumors were mesenchymal. As expected, 'Intermediate' type tumors showed a mixed phenotype (**Figure 4.14**). In summary, I conclude that EMT, inflammation and EGFR signalling have a role in the prognosis of colon cancer, as revealed by the SU-GIB prognostic classification.

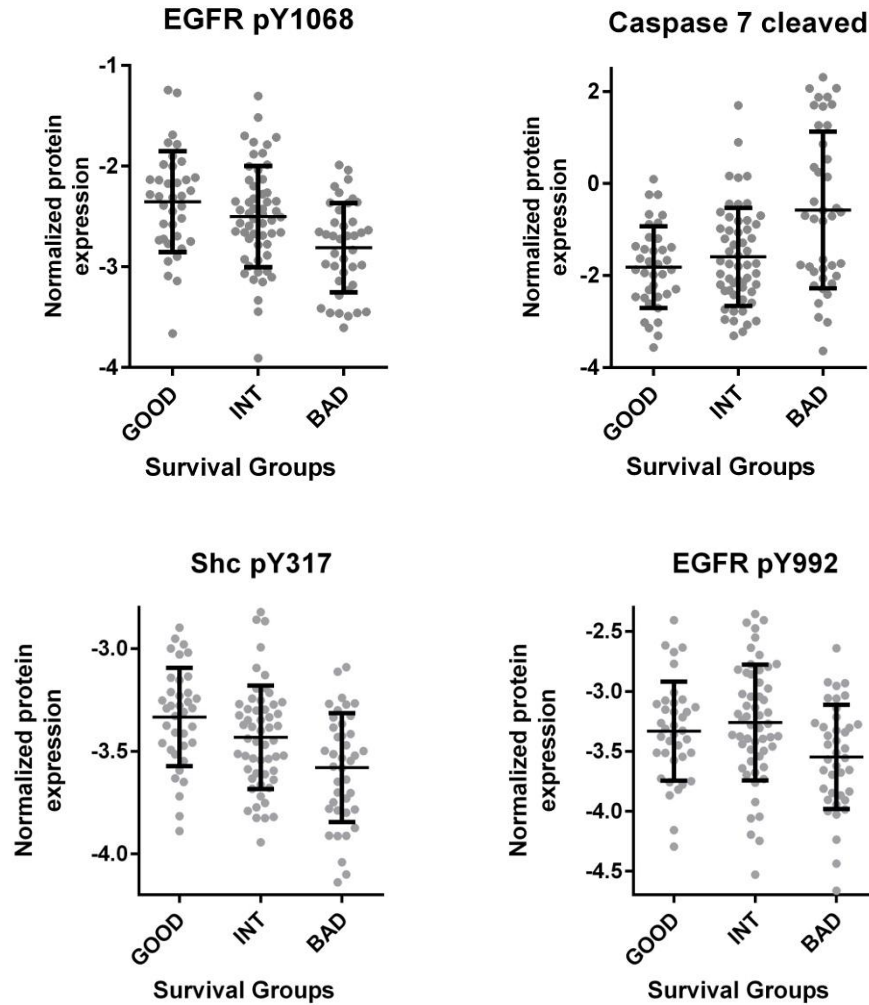


Figure 4.12: TCGA based proteome analysis of colon cancer tumor tissue in SU-GIB prognostic groups. Proteome data for 132 colon cancer primary tumor tissues downloaded from “cancergenome.nih.gov” via the TCGA data portal classified according to RNA seq. results of the same data and grouped according to the SU signature (bad survivors: 40, good survivors: 37, intermediate survivors: 55). Normalized protein expression levels are shown for EGFR pY1068, EGFR pY992 and Shc pY317 phosphorylations, and cleaved Caspase 7. The mean and standard deviation for each group are indicated.

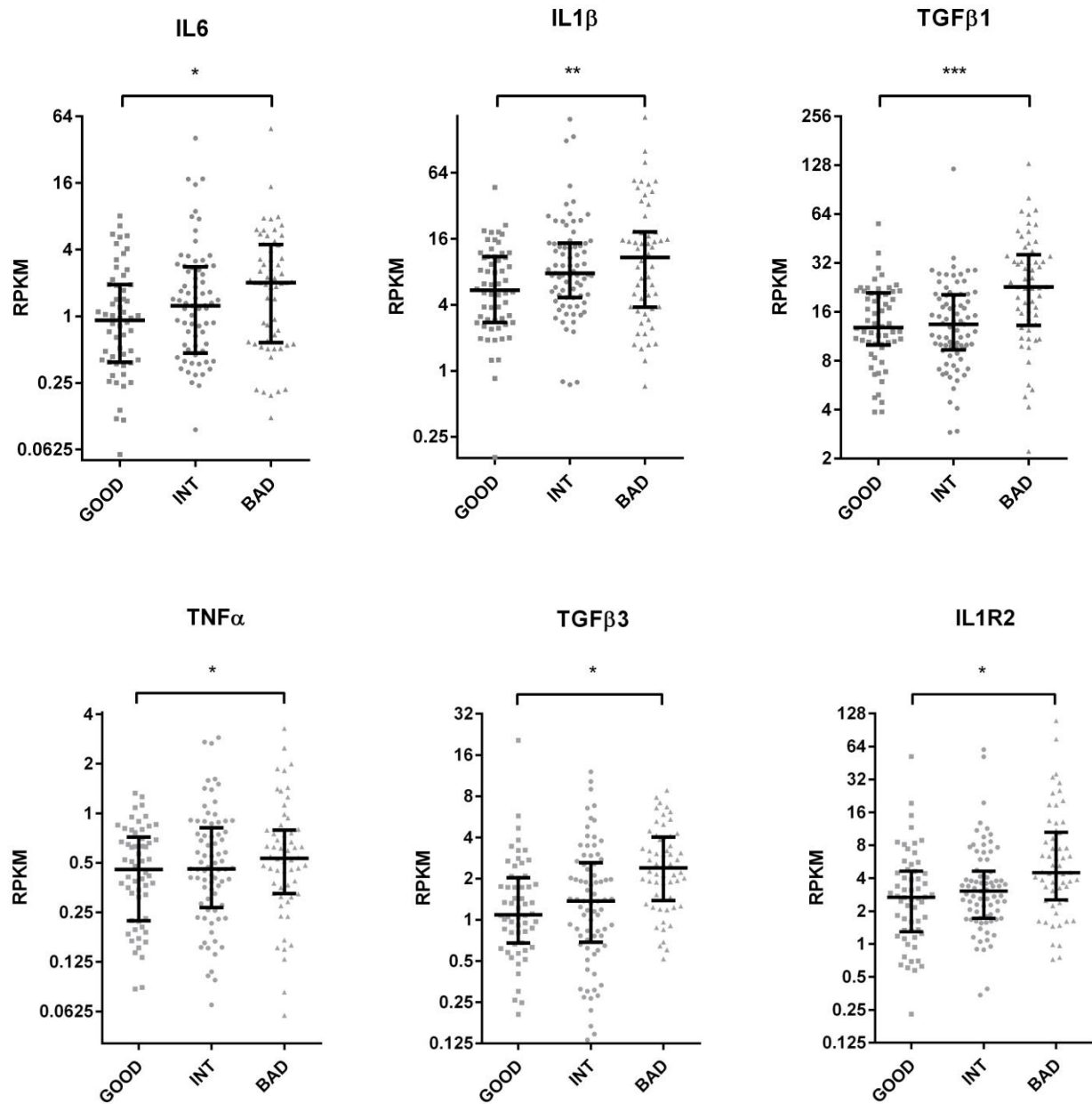


Figure 4.13: TCGA based RNA seq. analysis of colon cancer tumor tissue reveals increased inflammatory cytokine gene expression in the bad prognosis group. TCGA colon tumor samples were stratified into GIB groups using median expression values as cut-offs for ULBP2 and SEMA5A. RPKM values of IL6, IL1 β , TGF β 1, TNF α , TGF β 3 and IL1R2 are plotted for “good”, “intermediate” and “bad” groups. Given t-test p values are between “good” and “bad” groups. *p < 0.05, **p < 0.01, ***p < 0.0001. RPKM: reads per kilobase per million mapped reads. The median and inter-quartile range for each group is indicated.

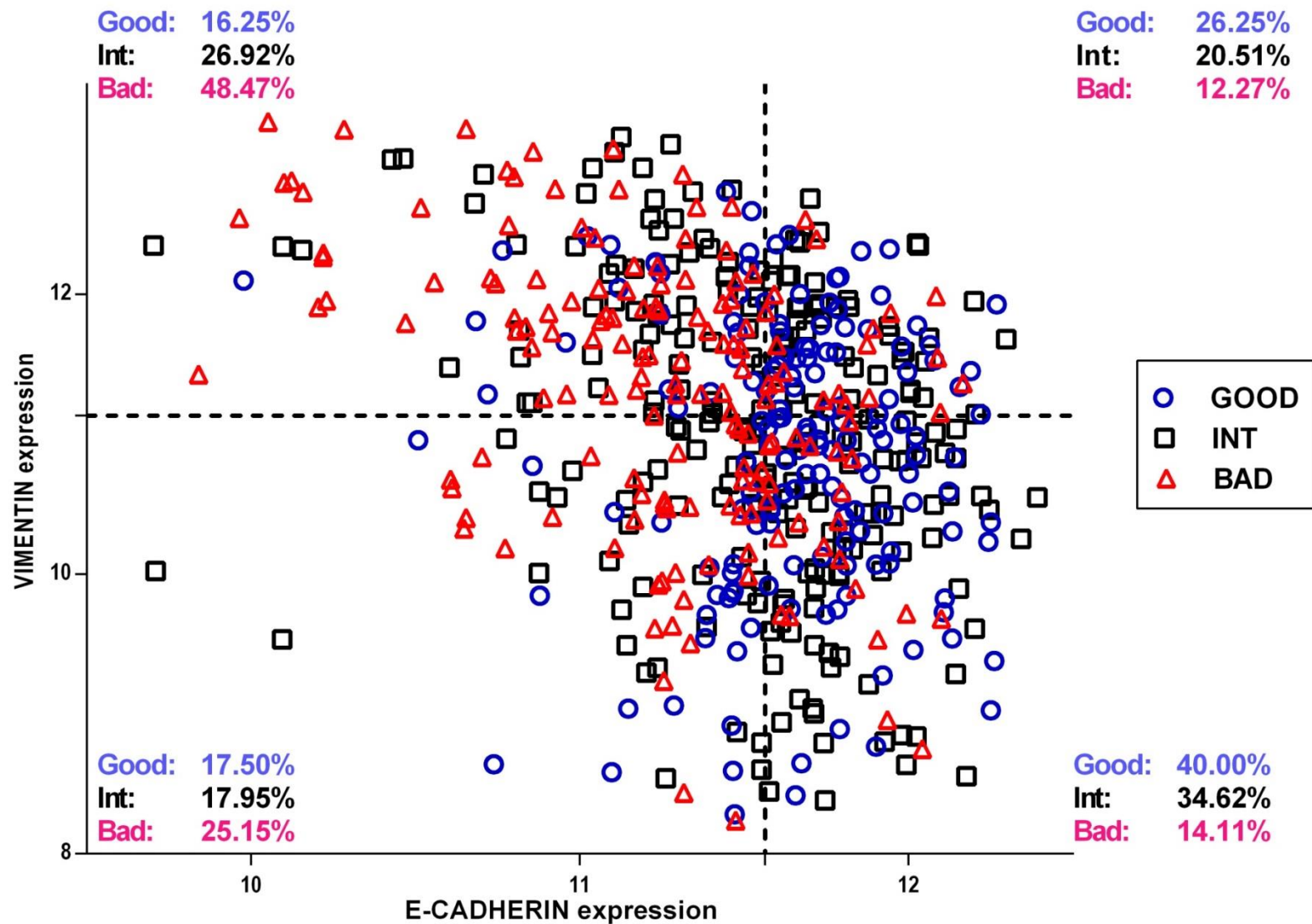


Figure 4.14: SU-GIB stratification is associated with EMT. Scatter plot for tumors in the GSE39582 dataset based on E-cadherin (201131_S_AT) and Vimentin (201426_S_AT) gene expression. While most of the patients with a good prognosis are also “E-cadherin high, Vimentin low”, patients with bad prognosis show the opposite gene expression pattern (Chi-sq. $p < 0.00001$). Median expression was used as the threshold for both genes. Dotted lines indicate median expression

4.1.9. Response to chemotherapeutics in SU-GIB groups

In the clinical setting, the prognostic tests are preferred more when they can also contribute to the treatment decision; therefore we analyzed a publically available drug cytotoxicity database to determine compounds which can especially be used for the treatment of patients with 'Bad' prognosis. The linear correlation analysis with IC₅₀ (inhibitory concentration 50) values of all drugs available in CGP database (n=142) [152], enabled identification of 13 drugs for which IC₅₀ is positively and negatively correlated with gene expression of SEMA5A and ULBP2, respectively (**Table 4.10**). Among these drugs, a dual PI3K/mTOR inhibitor NVP-BEZ235 was the most significant one with the higher growth inhibition in 'Bad' prognosis group (**Table 4.10**). I concluded that colon cancer cell lines classified as 'Bad' according to the SU-GIB signature were more sensitive to NVP-BEZ235 *in silico* (ANOVA p value<0.0001, ttest p value (BAD vs GOOD) <0.0001) (**Figure 4.15A**). The *in vitro*, NVP-BEZ235 cytotoxicity data of 13 colon cancer cell lines, albeit not statistically significant, showed the same trend as the *in silico* analysis (**Figure 4.15B**).

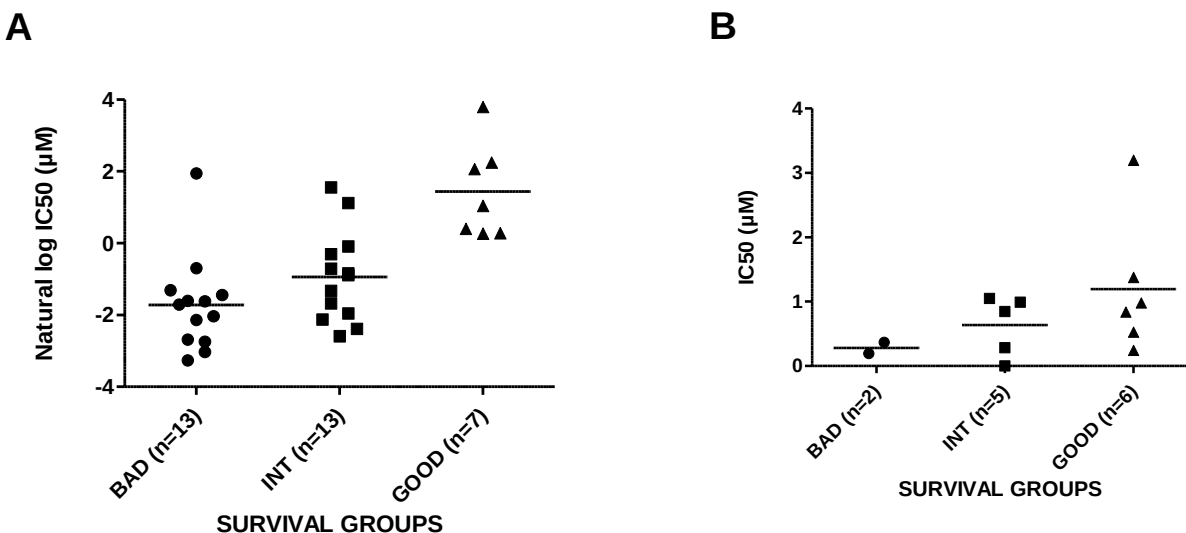


Figure 4.15: Chemosensitivity profiles of colon cancer cell lines corresponding to SU G, I and B phenotypes. *In silico* analysis of colon cancer cell lines as classified by the SU signature into G, I or B phenotypes for BEZ235 cytotoxicity based on CGP data (A) and *in vitro* validation (B). ULBP2 expression cut-off was the median for both analysis. SEMA5A cut-off was 6 for (A) and 0.2 for (B). Horizontal lines show mean values.

Table 4.10: Drugs that are significantly correlated with both SEMA5A and ULBP2 expression in colon cancer cell lines according to CGP

Drugs	Target	Rank sum	SEMA5A			ULBP2		
			Pearson's r	p	rank	Pearson's r	p	rank
NVP-BEZ235	PI3K/mTOR	5	0.60	<0.001	1	-0.47	0.006	4
Bortezomib	PARP	8	0.43	0.013	6	-0.51	0.002	2
AZ628	Farnesyltransferase	8	0.47	0.005	3	-0.47	0.006	5
Sorafenib	PDGFRA, PDGFRB, KDR, KIT, FLT3	12	0.38	0.029	9	-0.52	0.002	3
Bleomycin	DNA Damage	12	0.46	0.007	5	-0.43	0.013	7
Etoposide	Topoisomerase 2	12	0.49	0.004	2	-0.37	0.032	10
AG-014699	PARP1/2	13	0.37	0.034	12	-0.54	0.001	1
BI-D1870	HSP90	16	0.47	0.006	4	-0.35	0.048	12
Thapsigargin	Ca++ transporting ATPase	19	0.35	0.043	13	-0.47	0.006	6
AZD7762	Chk 1/2	19	0.38	0.031	11	-0.39	0.023	8
CEP-701	FLT3, JAK2, NTRK1, RET	19	0.38	0.029	10	-0.38	0.029	9
17-AAG	HSP90	19	0.38	0.028	8	-0.37	0.033	11
Tipifarnib	Farnesyl-transferase	20	0.41	0.017	7	-0.35	0.049	13

4.2. NOVEL mRNA BASED BIOMARKERS OF PROGNOSIS IN GASTRIC CANCER

4.2.1. Development of a methodology for the identification of prognostic genes in gastric cancer

Although gastric cancer is well known for its heterogeneity [81,82]; there are no specified treatment protocols for molecularly defined prognostic groups or biological subtypes. There are also no approved and clinically used tests by which gastric cancer can be prognosticated. Therefore, to determine prognostic genes in gastric cancer, we worked on the improvement of the methodology we utilized to do the same for colon cancer. **Table 4.11** describes the changes we considered. Left and right columns show the main steps of our methodology in colon cancer and gastric cancer, respectively. Prognostic gene discovery in colon cancer was performed by using 7 fixed thresholds for normalized microarray based expression data and was solely log-rank based (SSAT). For similar analysis in gastric cancer, we revised the script in a way to test all possible expression values as cut-off points. In addition to that, we combined Cox PH, Max stat and Log rank tests as described in the 'Methods' section. We named the novel method USAT for (Unsupervised Survival Analysis Tool). We believed this approach could reveal a larger number of prognostic genes as more cut-off values were considered, and our results would have even a lower false positivity as genes that could pass threshold values of three statistical tests would be studied further. In SSAT, once prognostic genes are defined, we tested the independence of the genes' prognostic prediction with clinical confounding factors using a MVA model. To improve USAT further, we included a MVA analysis that would use stage information within the code and designated this mUSAT. Another major difference between the approaches of determination of prognostic genes for colon and gastric cancer is the number of biomarkers determined as an output. During our prognostic studies for colon cancer, we were trying to develop a qPCR based prognostic kit that should also be cheaper with the involvement of as less genes as possible for clinical usage. However since the sequencing technology is getting cheaper every year [153], we considered RNA-seq based prognostic kits as a more reasonable alternative for gastric cancer project, therefore considered no limitation for the number of genes that could be utilized for the prognostic test that would be ultimately developed.

Table 4.11: Methodology of biomarker identification in colon and gastric cancers

	Colon SSAT (Semi supervised survival analysis tool)	Gastric mUSAT (MVA-Unsupervised survival analysis tool)
Statistical tests	Log rank	Cox PH, Max stat, Log rank
Thresholds	7 fixed thresholds	All expression values for a gene are possible thresholds
MVA	Performed subsequently	Included in the code
Number of biomarkers	Minimum number	No restriction

In order to prognosticate gastric cancer patients, I pursued two different approaches as described below.

4.2.2. APPROACH 1 – Identification of prognostic groups in gastric cancer

4.2.2.1. Identification of mRNA based prognostic biomarkers for gastric cancer

In order to identify prognostic genes for gastric cancer, I analyzed two discovery datasets that are GSE62254 (n=300) and GSE15459 (n=192) with the mUSAT script as shown in **Figure 4.16**. For each probeset, two Cox PH p values were obtained from the two datasets and the probesets were ranked according to their Cox PH p value from smallest to largest. Rank values were assigned to each probeset within each dataset. Then the rank values were summed and we generated a single 'rank sum' value for each probeset and performed the sorting of the probesets according to the rank sum. Top 20 genes with the lowest rank sum were selected for the further analysis.

Hierarchical clustering analysis of the 20 probesets identified from both discovery cohorts showed that the selected probesets created two tumor groups with noticeable high and low expression in both datasets, named as 'UP' and 'DOWN', respectively (**Figure 4.17**). Furthermore, I observed highly correlated expression patterns for almost all 20 prognostic probesets (**Figure 4.17**).

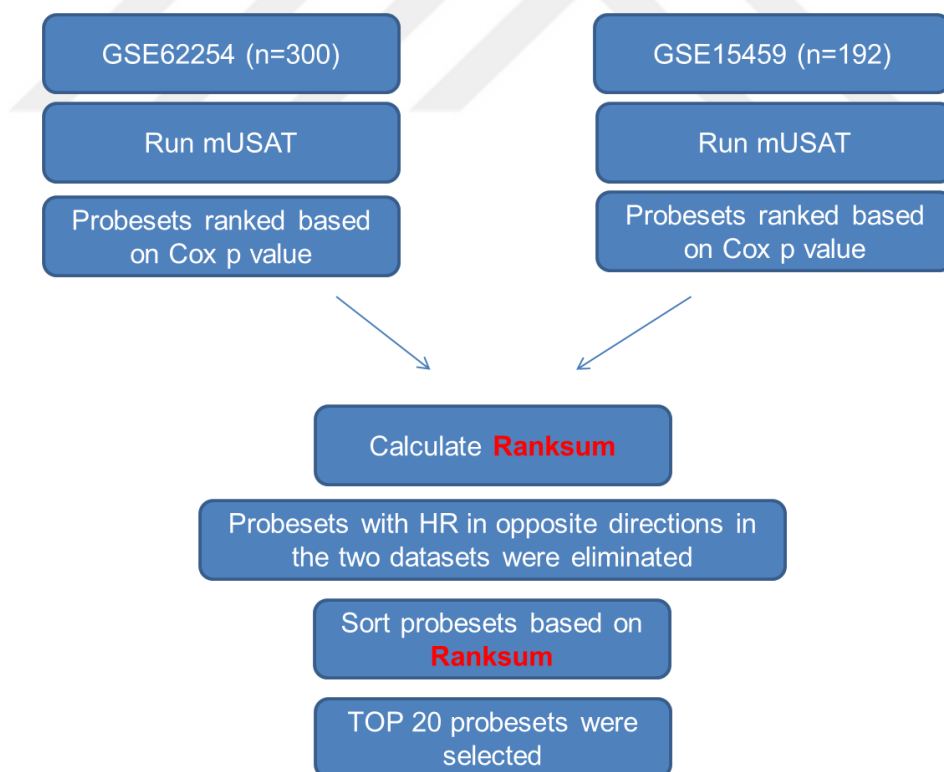


Figure 4.16: *In silico* selection of prognostic genes for gastric cancer. Probesets that fulfilled all conditions of mUSAT were ranked according to cox p values for two different gastric cancer microarray datasets, GSE62254 and GSE15459. Rank values were summed generating ranksum values. After elimination of probesets with HRs in the opposite direction, TOP 20 probesets based on ranksum values were selected for further analysis.

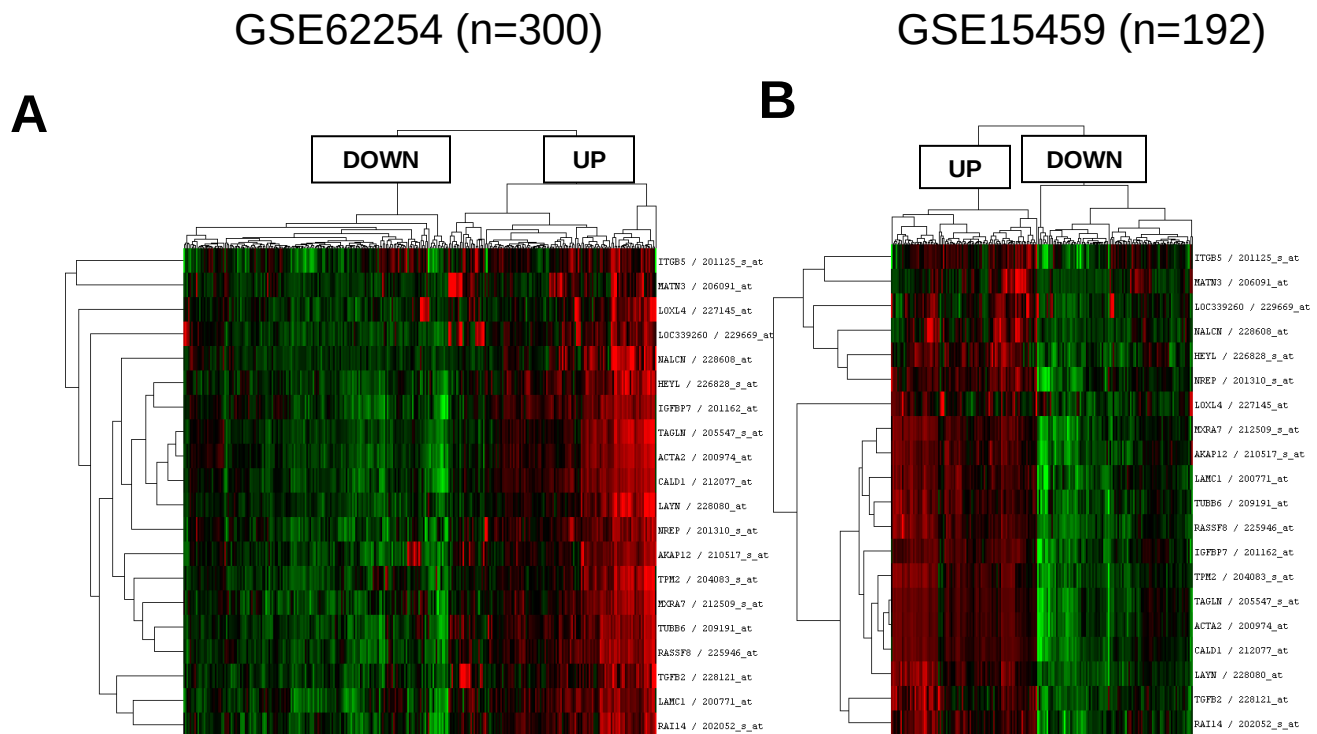


Figure 4.17: Hierarchical clustering analysis of gastric cancer tumors with prognostic genes. 20 probesets which were significantly associated with survival were used for hierarchical clustering in discovery cohorts, GSE62254 (A) and GSE15459 (B). Tumor groups with discernable high and low expression for all 20 probesets were labeled as "UP" and "DOWN" groups, respectively.

Kaplan Meier plots and log rank tests illustrated that the UP and DOWN groups are significantly different prognostically and higher expression of these genes was associated with shorter overall survival (**Figure 4.18**). The median overall survival time for UP groups were 20.2 and 26.2 months for GSE15459 and GSE62254 datasets, respectively, whereas patients in DOWN groups did not reach to median overall survival in 10 years follow-up time in both datasets indicating a clear distinction in clinical outcome (**Figure 4.18**).

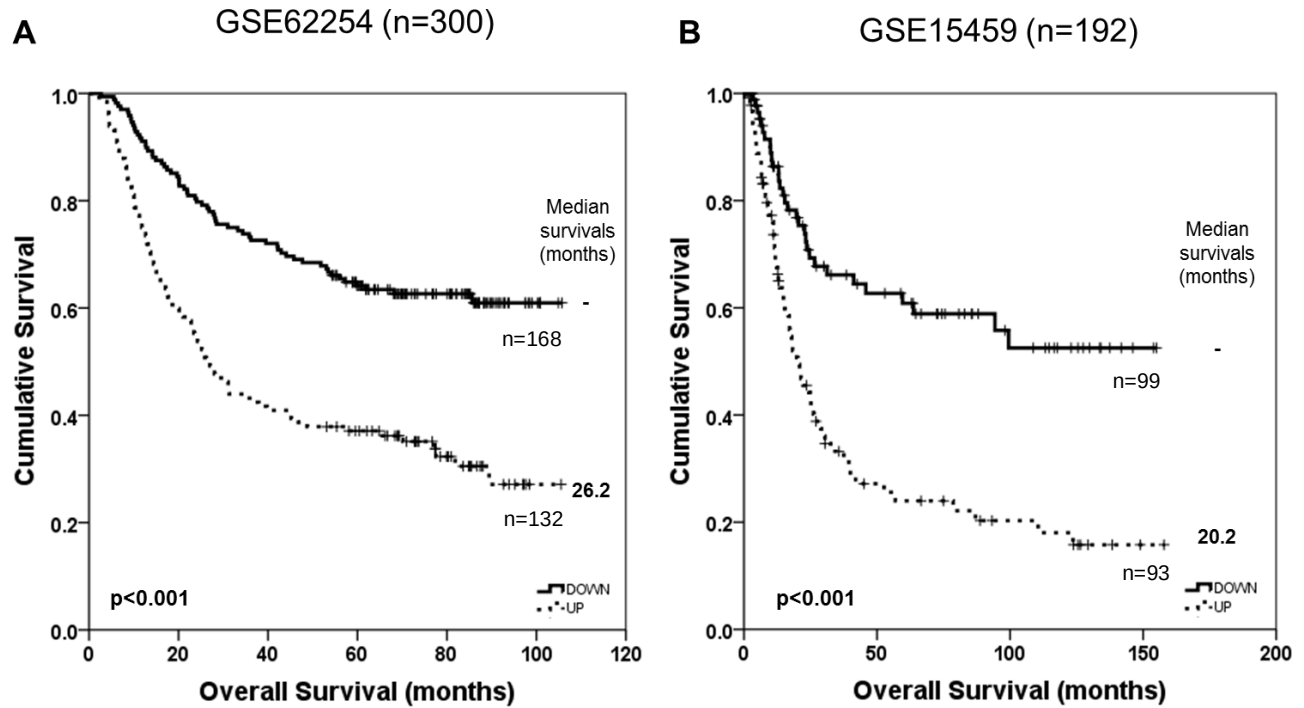


Figure 4.18: Kaplan Meier plots for prognostic groups of gastric cancer tumors. Samples stratified using UP-DOWN prognostic classification in GSE62254 (A) and GSE15459 (B). Log rank p values are shown. Median survival times for the DOWN groups were not reached at the indicated survival times, whereas for the UP groups, median overall survival times were 26.2 and 20.2 months for GSE62254 and GSE15459, respectively.

To validate these findings in an independent, third patient cohort, I used the GSE29272 dataset. However, as our discovery datasets were analyzed by the Affymetrix HGU 133 Plus 2 platform while the GSE29272 dataset was HGU133A based, we could use only 13 of the 20 probesets in our validation analysis. As seen in **Figure 4.19A**, even though the full list of probesets were not available in this dataset, hierarchical clustering analysis still could generate two major tumor groups which demonstrated a significantly different prognosis in overall survival (**Figure 4.19B**).

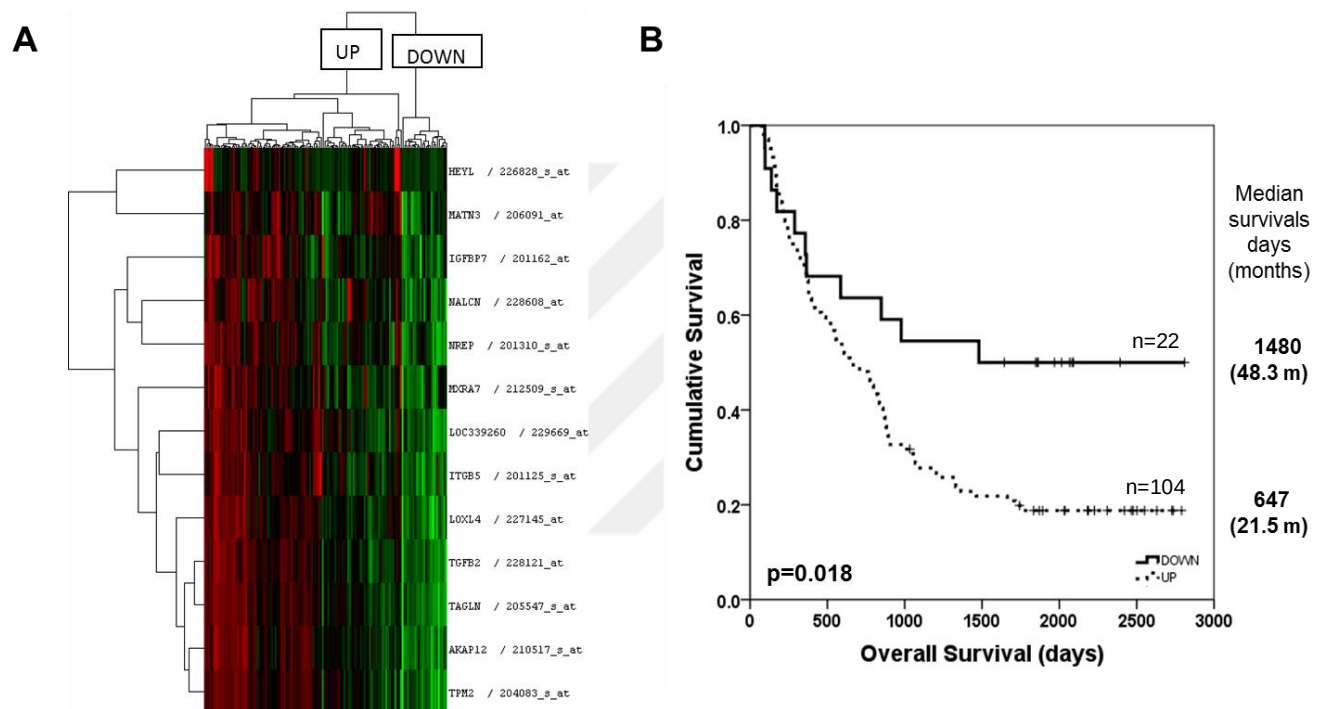


Figure 4.19: Hierarchical clustering and Kaplan Meier analysis of gastric cancer tumors with prognostic genes in validation cohort. 13 of 20 prognostic probesets which were available in the HGU133A platform were used for hierarchical clustering in the validation cohort, GSE29272 (A). Tumor groups with discernable high and low expression were labeled as "UP" and "DOWN", respectively. Kaplan Meier plot with samples stratified using UP-DOWN prognostic classification in GSE29272 (B). Log rank p value and median survival times are shown.

4.2.2.2. Stage independence of prognostic association of 20 probesets

To demonstrate that the prognostic grouping of gastric cancer patients was independent of clinical confounding factors, we first performed univariate Cox PH analysis for common confounding factors in our discovery and validation datasets to determine such prognostic clinical factors. Stage was significantly associated with survival in our discovery cohorts, whereas it was not the case for the validation cohort, which was expected since 85% of the patients had stage 3 disease (**Table 4.12**). Therefore I performed MVA analysis with our prognostic UP&DOWN signature and stage and found that UP&DOWN signature was a stage independent prognostic predictor (**Table 4.13**).

Table 4.12: Univariate Cox regression analysis of clinical parameters in discovery and validation cohorts – OS

	DISCOVERY				VALIDATION	
	GSE15459 (n=192)		GSE62254 (n=300)		GSE29272 (n=126)	
	<u>HR*</u>	<u>p-value</u>	<u>HR*</u>	<u>p-value</u>	<u>HR*</u>	<u>p-value</u>
Stage**	2.789	<0.001	2.283	<0.001	1.155	0.451
Age (< or > 60)	0.983	0.936	0.792	0.168	1.407	0.098
Gender (Female or male)	0.713	0.127	1.105	0.559	1.122	0.642
Tumor subtype***		0.578		<0.01		0.134

*Cox proportional hazards regression

**Stage: treated as a continuous variable (1, 2, 3, 4)

***Tumor subtype: treated as a categorical variable (intestinal, mixed, diffuse)

Table 4.13: MVA cox regression analysis of prognostic classification and stage in discovery cohorts - OS

GSE62254	<u>p-value</u>	<u>HR*</u>	<u>95% CI for HR</u>	
			<u>Lower</u>	<u>Upper</u>
DOWN vs UP	<0.001	0.699	0.594	0.824
Stage**	<0.001	2.146	1.754	2.625

GSE15459	<u>p-value.</u>	<u>HR*</u>	<u>95% CI for HR</u>	
			<u>Lower</u>	<u>Upper</u>
DOWN vs UP	0.002	0.493	0.316	0.770
Stage**	<0.001	2.608	1.997	3.406

*Cox proportional hazards regression

**Stage: treated as a continuous variable (1, 2, 3, 4)

4.2.2.3. Assessment of prognosis with disease-free survival

According to American Agency for Healthcare Research and Quality, studies of prognostic biomarkers should be assessed using multiple clinical outcome types [45], therefore I analyzed our prognostic signature using DFS which was available for the GSE62254 dataset. Indeed, the UP group was significantly associated with shorter disease free survival (**Figure 4.20**).

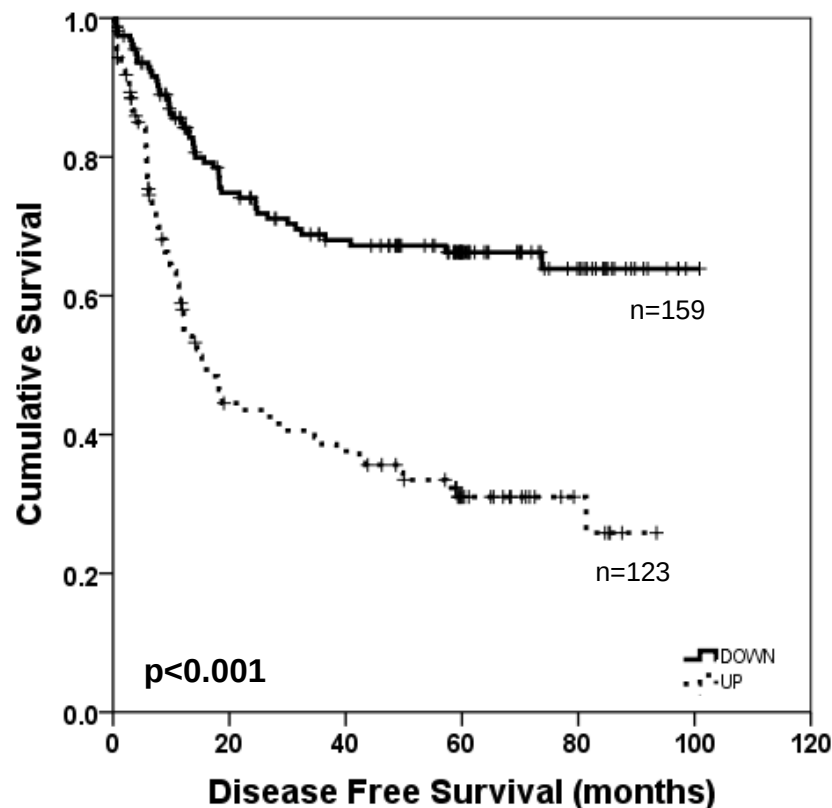


Figure 4.20: Kaplan Meier plot for prognostic groups of gastric cancer tumors with disease free survival as an end point measure. Samples are stratified using UP-DOWN prognostic classification in GSE62254 (n=282). Log rank p value is shown. Median DFS for patients in UP group was 15.3 months, whereas a median for survival time for patients in DOWN group was not reached in 10 years.

4.2.2.4. Prognostic stratification within known pathologic sub-groups

The two well-defined types of gastric cancer are intestinal type which is sporadic and mostly associated with environmental factors, and diffuse type which is highly aggressive and associated with poor prognosis according to a meta-analysis of 61,468 patients [154]. Therefore I asked if our prognostic gene signature is independent of this pathological (also known as Lauren) classification that identifies two prognostic subgroups, “intestinal” and “diffuse”. Using the clinical data available for GSE62254 and GSE15459 datasets, I stratified patients according to their pathologic subtypes, and determined the prognostic difference between UP and DOWN groups. Indeed, our prognostic signature can further divide patients with both intestinal and diffuse type tumors to subgroups with significantly different clinical outcome in two independent cohorts (**Figure 4.21**).

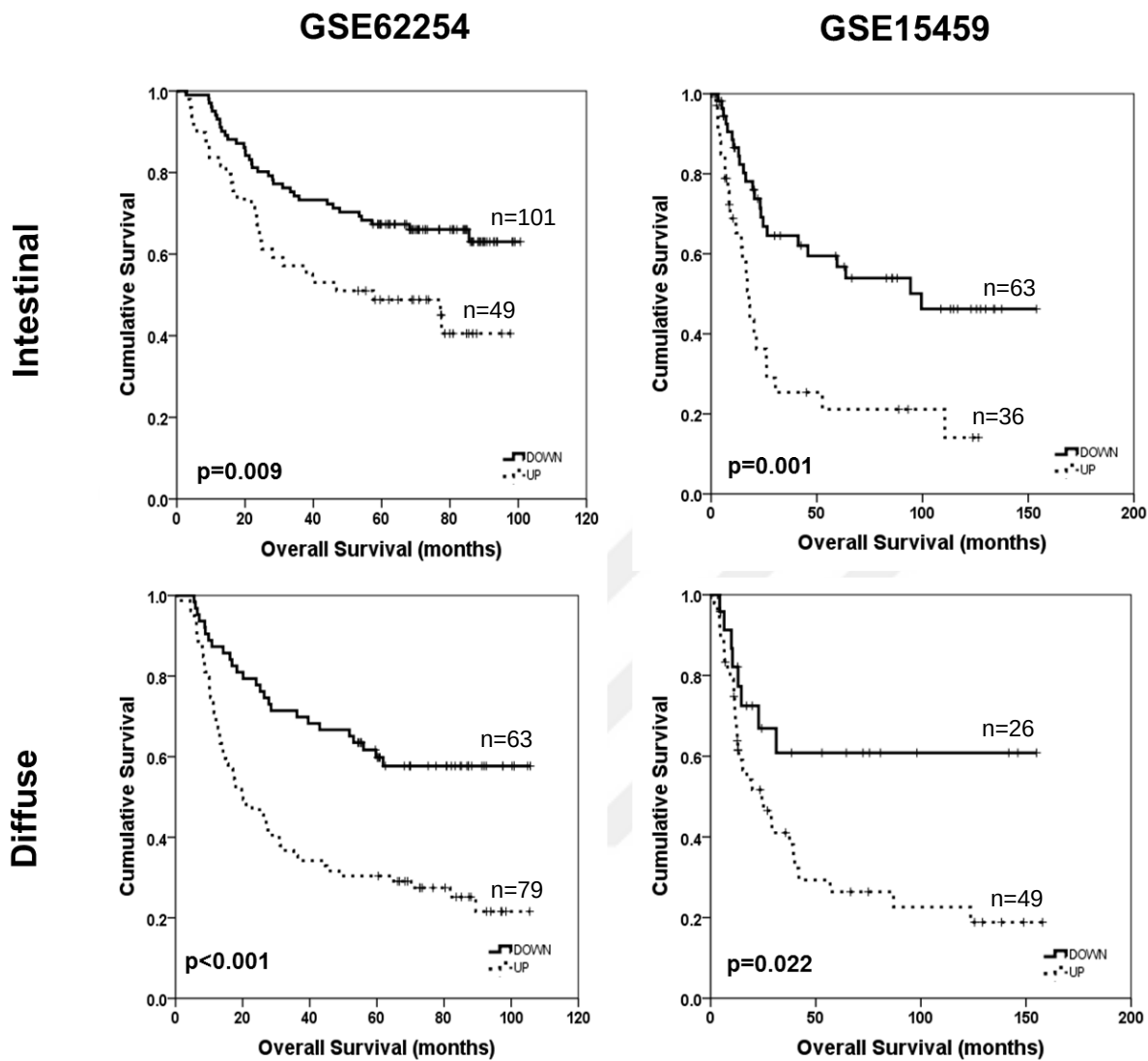


Figure 4.21: Kaplan Meier plot for prognostic groups of tumors in diffuse and intestinal type gastric cancers. Samples were stratified using UP-DOWN prognostic classification using the indicated discovery cohorts. Log rank p values are shown.

4.2.2.5. Gastric tumor sub-groups in The Cancer Genome Atlas

As yet another validation, I performed hierarchical clustering of the RNAseq data of 336 gastric cancer patients in The Cancer Genome Atlas database. We could use 18 out of 20 prognostic genes as they had counterparts in the RNA-seq data. The clustering exhibited clear-cut patient groups (**Figure 4.22**). Although the technology is quite different for quantification of the transcription, I could observe 2 patient groups with total up-regulation and down-regulation of the prognostic genes, as expected. In addition to that, I also could define a third group of patients with intermediate level of expression which could not been observed with microarray technology probably due to the restricted expression range.

The clinical data downloaded from the TCGA database included 279 of these gastric cancer patients and 263 of them were censored (94%). Therefore I could not test the differences of clinical outcome in these patient groups in TCGA data.

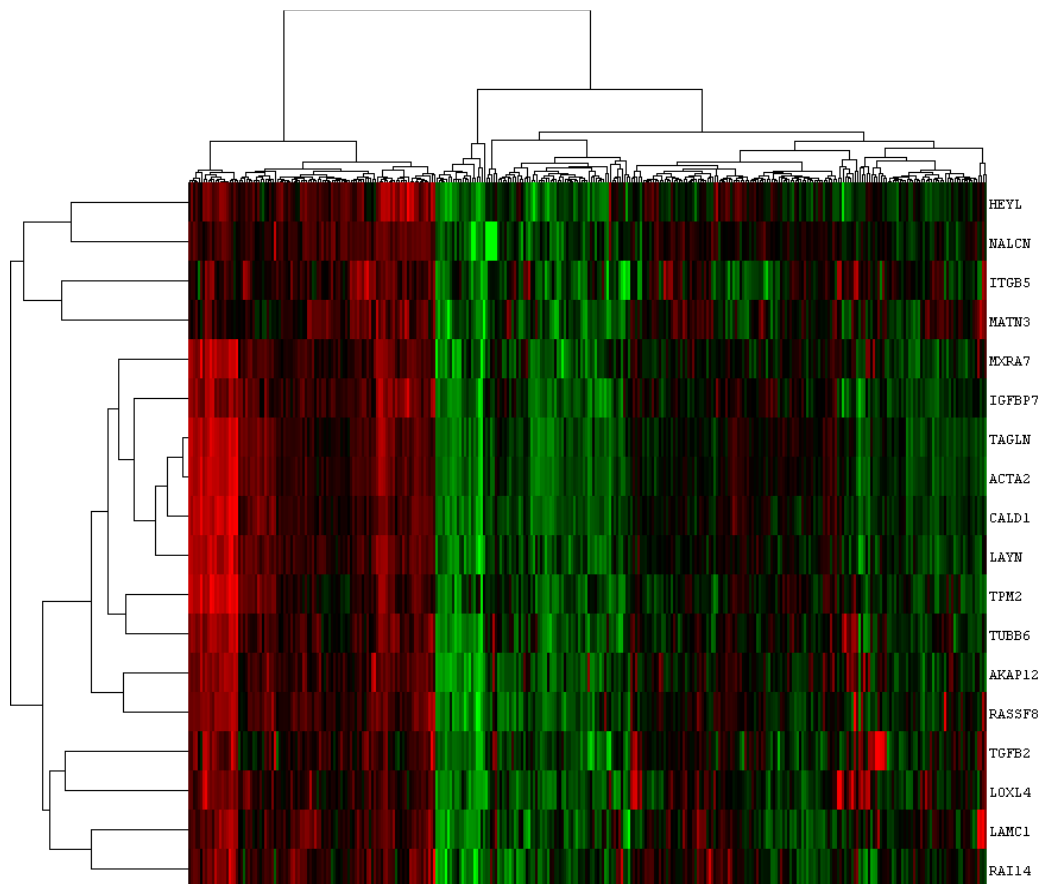


Figure 4.22: Prognostic classification of TCGA gastric tumors based on RNAseq data. Hierarchical clustering of the RNAseq level3 expression data of 336 gastric cancer patients in The Cancer Genome Atlas database with the available 18 out of 20 prognostic genes.

4.2.2.6. Biological characteristics of prognostic groups

In order to perceive the biological differences between 'UP' and 'DOWN' prognostic groups, I identified the differential gene profile between these using GSEA (Gene set enrichment analysis, see methods). An FDR q value < 0.25 was used as cut-off for all GSEA analysis. This identified 63 and 0 genesets enriched in the UP and DOWN groups, respectively, in the GSE15459 dataset (APPENDIX B). In the GSE62254 dataset, I identified 188 and 3 genesets enriched in UP and DOWN groups, respectively (APPENDIX C), indicating heterogeneity within the DOWN group. For the genesets that are

enriched in the UP group, I realized common patterns for the two datasets. Enriched genesets included those related to extracellular matrix, actin cytoskeleton and binding, and wound healing. In addition to that I also observed genesets related to the negative regulation of cell cycle and metabolism (**Table 4.14**). Since there were no genesets enriched in 'DOWN' group, I thought that there may be multiple molecular groups in 'DOWN', which I analyzed them further with our second approach described below. These findings showed that the patient group with the worse prognostic profile (UP) may have less proliferation but more extracellular matrix involvement and cell migration suggesting a more mesenchymal phenotype when compared to patients with a relatively better prognostic profile (DOWN).



Table 4.14: Gene set enrichment analysis comparing UP-DOWN prognostic groups

GSE62254*	
NAME	FDR q-val
EXTRACELLULAR_REGION	0.061
ACTIN_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.061
ACTIN_BINDING	0.061
CYTOSKELETON	0.066
EXTRACELLULAR_REGION_PART	0.067
ACTIN_CYTOSKELETON	0.068
PROTEINACEOUS_EXTRACELLULAR_MATRIX	0.083
EXTRACELLULAR_MATRIX	0.086
EXTRACELLULAR_MATRIX_PART	0.087
WOUND_HEALING	0.088
CELL_MIGRATION	0.088
NEGATIVE_REGULATION_OF_METABOLIC_PROCESS	0.060
CELL_CYCLE_ARREST_GO_0007050	0.065
NEGATIVE_REGULATION_OF_CELL_CYCLE	0.065

*Enriched gene sets with a consistent pattern of biology were listed in GSE62254 and GSE15459 datasets. FDR q values below 0.25 were considered significant.

GSE15459*	
NAME	FDR q-val
ACTIN_FILAMENT_BASED_PROCESS	0.187
CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.189
ACTIN_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.192
ACTIN_BINDING	0.197
EXTRACELLULAR_MATRIX	0.204
PROTEINACEOUS_EXTRACELLULAR_MATRIX	0.211
CYTOSKELETAL_PROTEIN_BINDING	0.213
EXTRACELLULAR_MATRIX_PART	0.224
WOUND_HEALING	0.239
NEGATIVE_REGULATION_OF_TRANSCRIPTION	0.188
NEGATIVE_REGULATION_OF_NUCLEIC_ACID_METABOLIC_PROCESS	0.188
NEGATIVE_REGULATION_OF_CELL_CYCLE	0.190
NEGATIVE_REGULATION_OF_CELLULAR_METABOLIC_PROCESS	0.191
NEGATIVE_REGULATION_OF_METABOLIC_PROCESS	0.192
CELL_CYCLE_ARREST_GO_0007050	0.235

Enrichment plots also confirm that genesets related to extracellular matrix, cytoskeletal organization and cell cycle arrest are enriched in 'UP' group in GSE62254 as well as in GSE15459 with a clearly skewed pattern (**Figure 4.23 and Figure 4.24**).

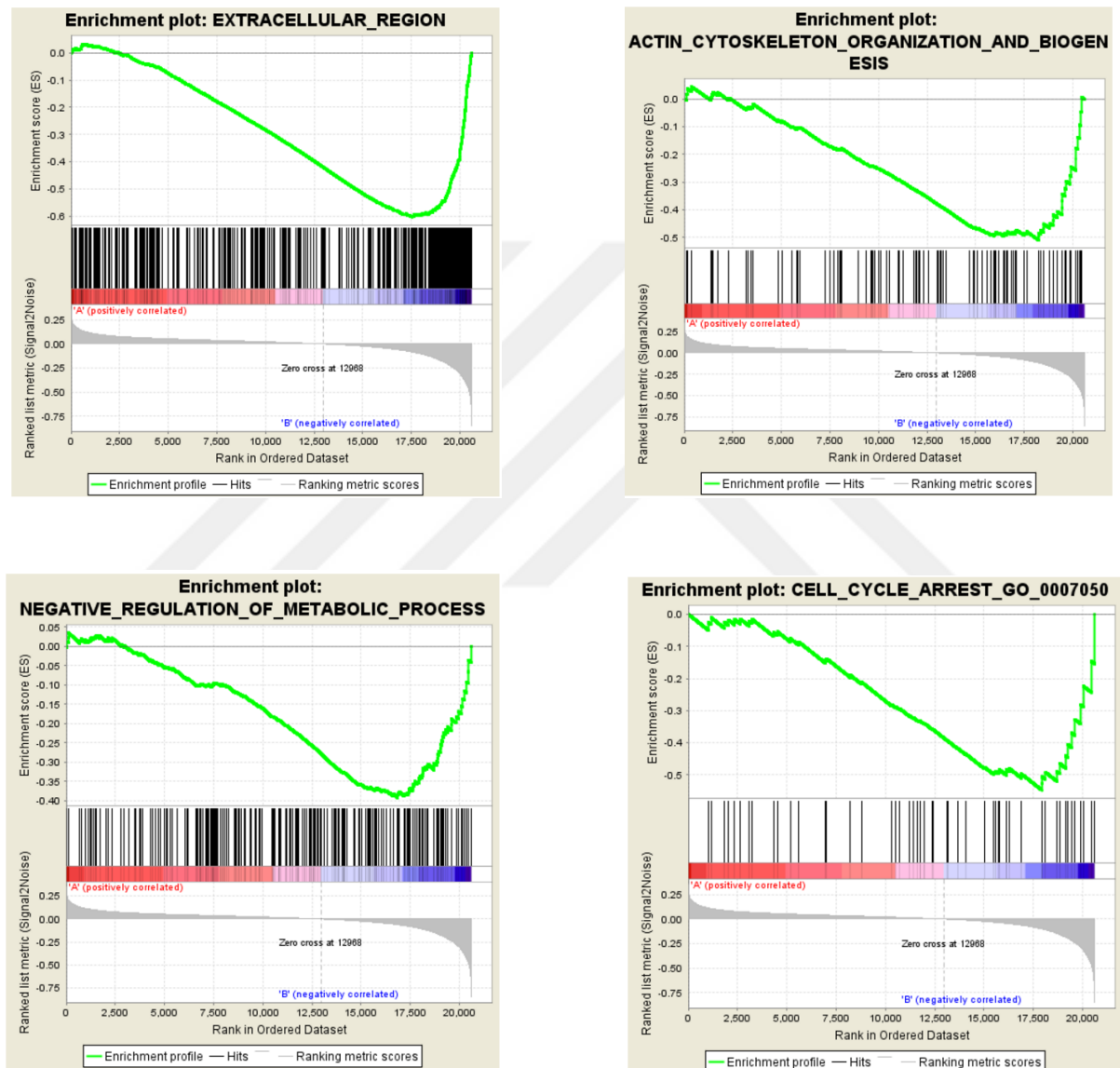


Figure 4.23: Enrichment plots of highly enriched gene sets listed in Table 4.14 in GSE62254 dataset indicating an enrichment of genes related to the extracellular matrix, cytoskeleton and a reduction of metabolic processes and cell growth

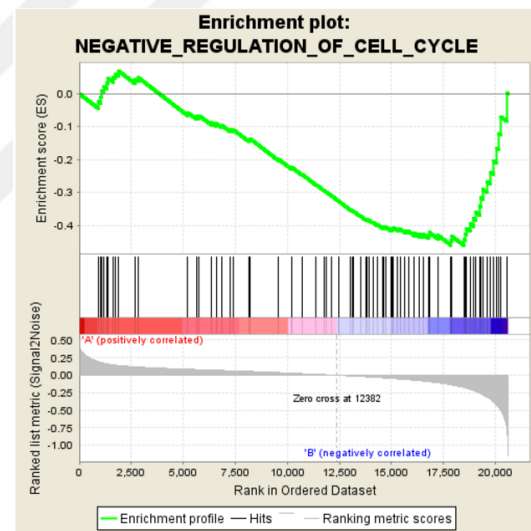
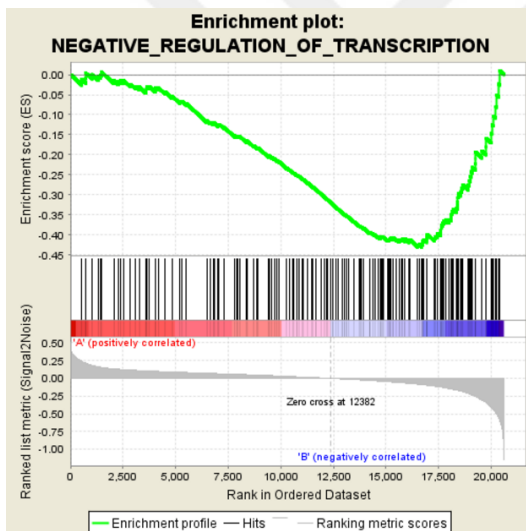
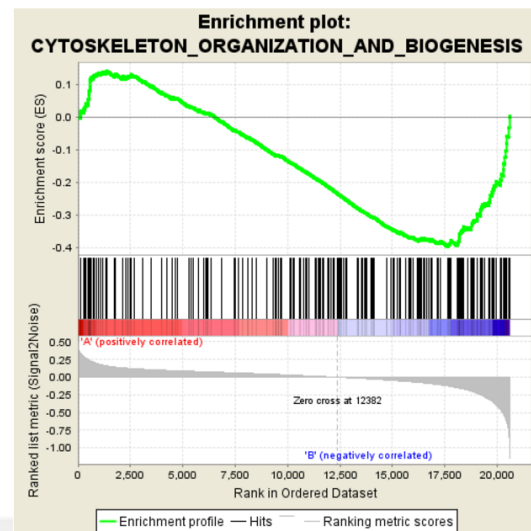
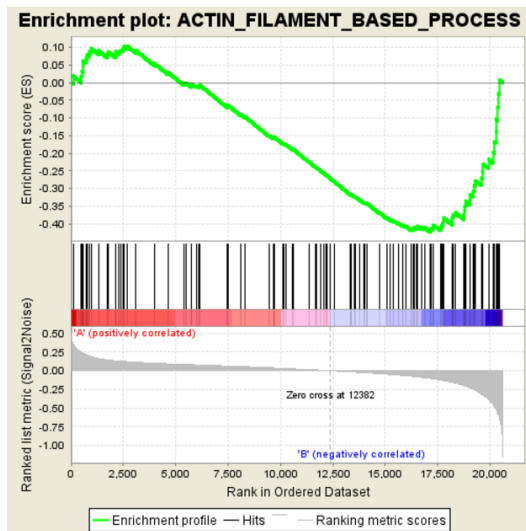


Figure 4.24: Enrichment plots of highly enriched gene sets listed in Table 4.14 in GSE15459 dataset show a similar pattern to that observed for GSE62254

4.2.3. APPROACH 2 – Identification of molecular group-specific prognostic markers in gastric cancer

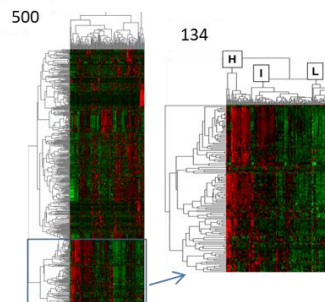
The prognostically distinct groups I identified in approach 1 successfully revealed a very clear biological pattern in ‘UP’ group, however none observed in ‘DOWN’ group. This data suggested that if such molecular sub groups are initially identified, they could be used as a basis for the more effective identification of prognostic genes for each sub group. I therefore, decided to define such groups utilizing a hierarchical clustering based approach (**Figure 4.25**). Groups thus identified were then used independently for prognostic analyses.

4.2.3.1. Identification of molecular sub-groups of gastric cancer

For an unsupervised molecular classification of gastric cancer tumors, I performed hierarchical clustering analysis with 500 probesets with the highest coefficient of variance in the GSE29272 dataset (**Figure 4.26A**), and identified 134 probesets which generated the most distinct tumor sub-groups (shown within the blue box). A dedicated hierarchical clustering analysis with these 134 probesets resulted in three distinct tumor groups with high (H), low (L) and intermediate (I) expression (**Figure 4.26B**). These 134 probesets are listed in APPENDIX D.

1. Define molecular groups of gastric cancer

Hierarchical clustering of the tumors with 500 probesets with the highest COV



134/500
probesets can
generate 3
molecular sub-
groups

2. Run mUSAT (find prognostic genes) for each molecular subtype

Figure 4.25: Summary of high variance based stratification in gastric cancer. This approach was based on the 500 probesets with highest coefficient of variation. Hierarchical clustering analysis showed that 134 of the top 500 genes were able to generate 3 major groups.

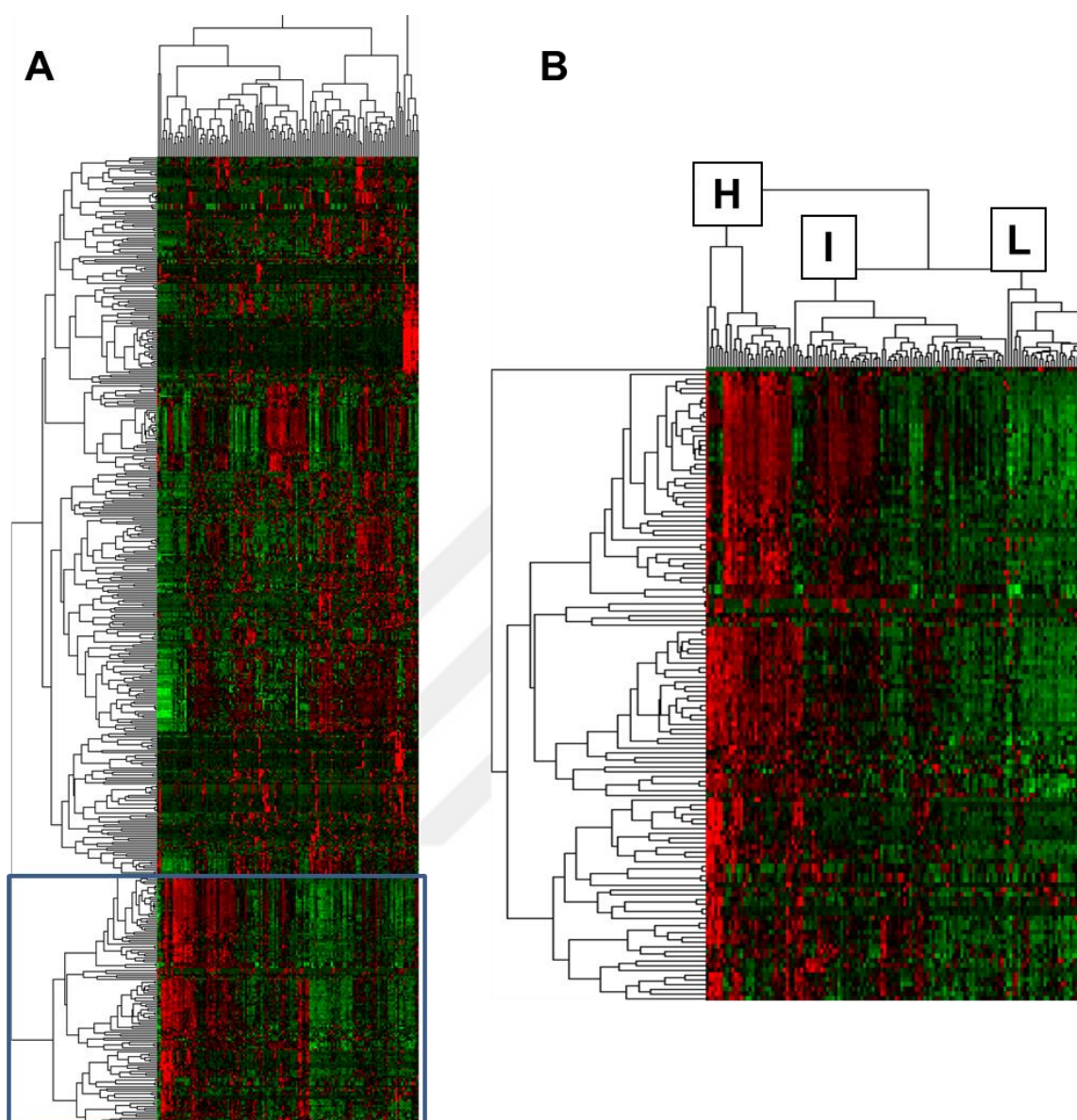


Figure 4.26: Determination of molecular classification via unsupervised hierarchical clustering. Hierarchical clustering of GSE29272 dataset (n=126) using 500 probesets with the highest variance (A). Two groups with clear discrimination of gene expression were observed (shown with blue box) which includes 134 probesets. Hierarchical clustering of GSE29272 dataset (n=126) using these 134 probesets (B). Samples were classified according to their level of gene expression as high “H”, intermediate “I” and low “L”.

When I ran mUSAT for the GSE29272 dataset, only 5 probesets among 134 probeset which are used for molecular classification of gastric tumors were significantly associated with overall survival with a Cox PH p value below 0.05. (**Table 4.15**). While these 134 probesets were classifying molecular sub-groups, they were not associated with survival individually.

Table 4.15: Univariate Cox regression analysis of 134 probesets (significant associations reported) in GSE29272 dataset – OS

Gene symbol	Probeset ID	HR	Cox p
CHI3L1	209395_at	0.810	0.014
CHI3L1	209396_s_at	0.802	0.012
MFAP5	213764_s_at	1.226	0.026
ITGBL1	205422_s_at	1.311	0.008
TNFAIP6	206025_s_at	0.803	0.029

In order to understand the biology underlying the HIL molecular classification, I performed GSEA comparing HIGH to LOW tumors. Below FDR q value cut-off of 0.25, 134 genesets were enriched in HIGH and 187 genesets were enriched in LOW group in GSE29272 dataset (APPENDIX E and F, respectively), reflecting a better biological classification then what I achieved in my first approach. When genesets were ranked in ascending order according to their FDR q values, 7 out of top 10 most enriched genesets in HIGH group were extracellular matrix (ECM) related and all top 10 gene sets in LOW group were cell cycle related indicating different biological groups; ECM involved and a highly proliferative phenotype. Enrichment plots of top 4 genesets from both lists are shown in **Figures 4.27 and 4.28**. The enrichment plots also reveal the improvement this approach has on the definition of molecular homogeneity within groups.

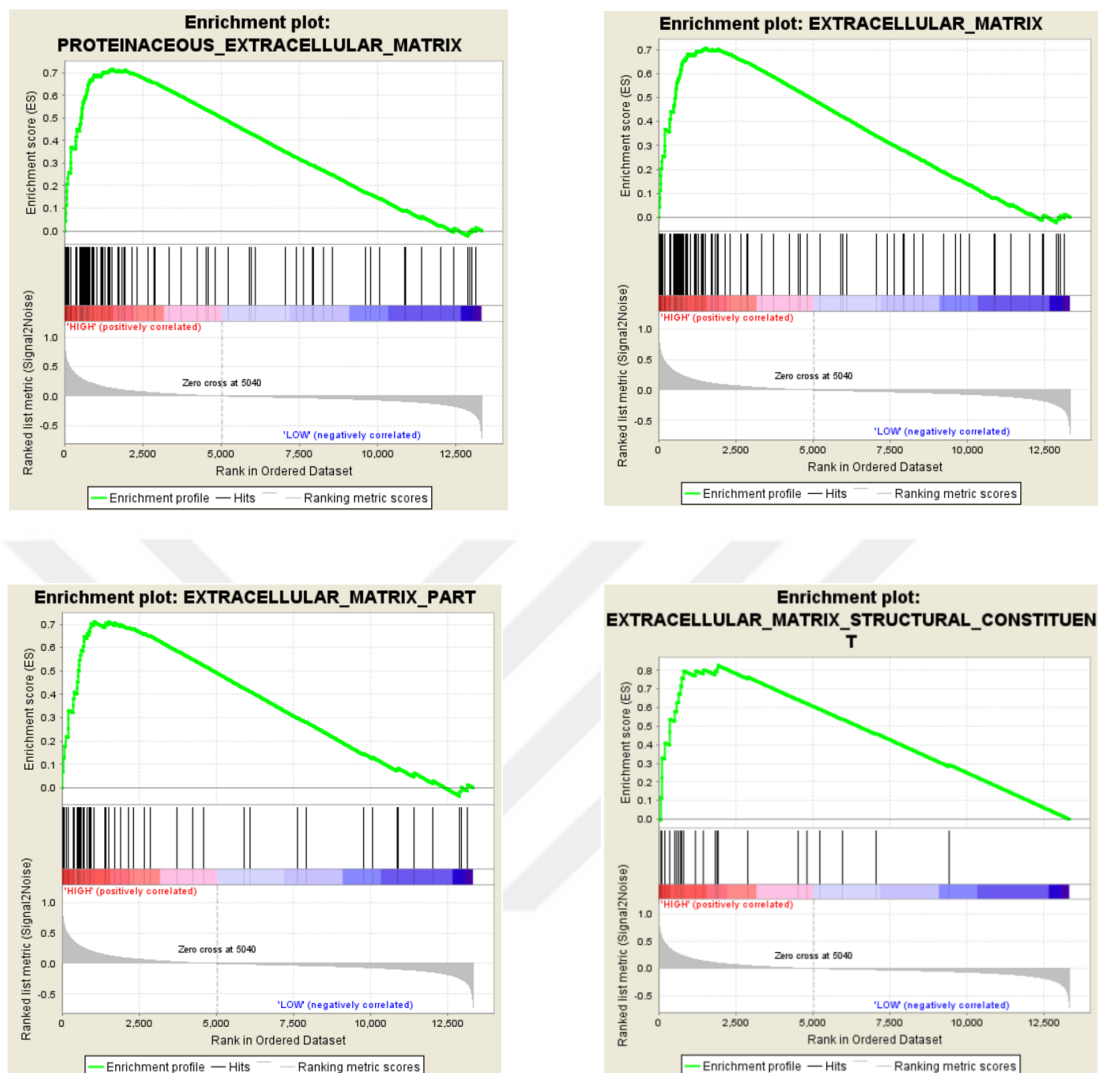


Figure 4.27: Enrichment plots of top 4 genesets enriched in GSE29272 HIGH group show an enrichment of extracellular matrix related genes

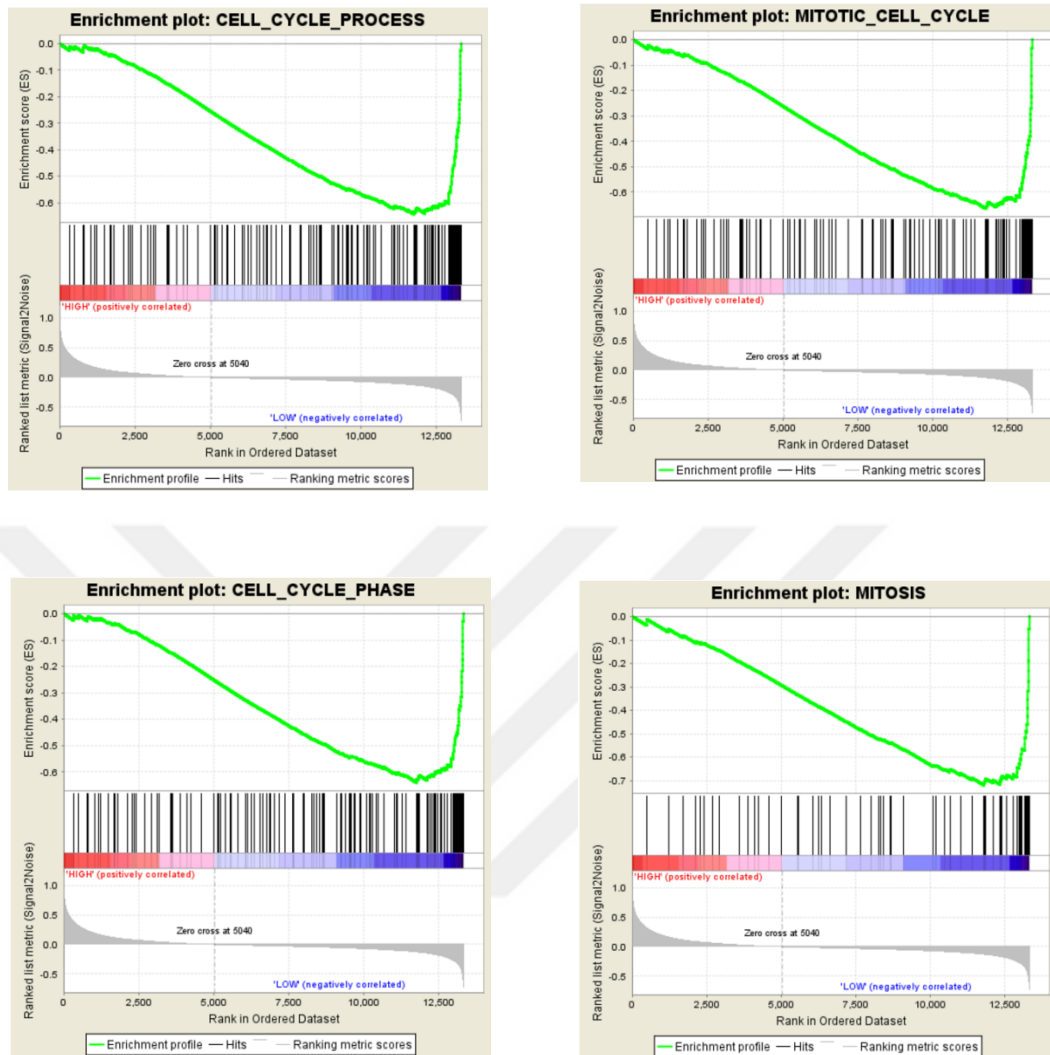


Figure 4.28: Enrichment plots of top 4 genesets enriched in GSE29272 LOW group indicate increased cell cycle, proliferation and mitosis

4.2.3.2. Validation of molecular sub-groups of gastric cancer

Hierarchical clustering using all 134 probesets of a third independent cohort (GSE62254) revealed that a small group of patients were clearly distinguished with very high expression of all genes involved (H). A second group of patients showed low expression of almost all probesets (L) with the rest of the samples showing intermediate expression (I) (**Figure 4.29**).

GSEA analysis between HIGH and LOW groups defined in GSE62254 dataset resulted in 138 and 53 genesets enriched in HIGH and LOW groups, respectively. The full list of enriched genesets are given in APPENDIXES G and H. Genesets which were enriched in HIGH in both GSE29272 and GSE62254 datasets included 'Basement membrane', 'Basolateral plasma membrane', 'Collagen', 'Adherens junction', 'Actin filament based process' and 'Extracellular structure organization and biogenesis' (**Table 4.16**) which are biologically highly similar to genesets enriched in poor prognostic 'UP' group of patients identified by my first approach. Although no consistent biological pattern was observed in good prognostic 'DOWN' group in approach 1, commonly enriched genesets included 'Kinetochore', 'Mitochondrial checkpoint', 'Response to DNA damage stimulus', 'Chromosome pericentric region', 'Cell cycle', 'Interphase', 'DNA repair' and also other genesets related to induction of cell cycle related processes were obtained for LOW group indicating the improvement in defining biological distinction via my second approach (**Table 4.17**).

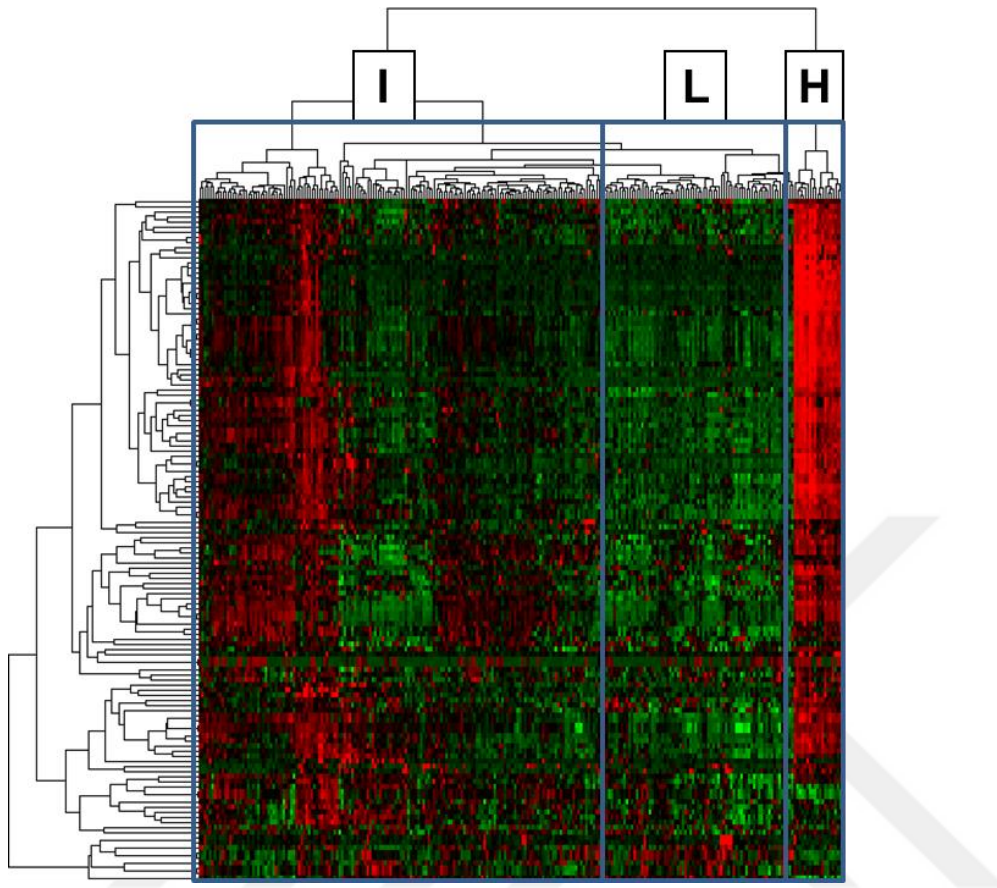


Figure 4.29: Molecular classification of gastric tumors in GSE62254 dataset. 134 probesets shown in previous figure were used for hierarchical clustering in GSE62254 dataset (n=253). Samples are classified according to their level of gene expression as high "H", intermediate "I" and low "L".

Table 4.16: Genesets enriched common in HIGH group

Genesets	FDR q-val	
	GSE29272	GSE62254
BASEMENT_MEMBRANE	0.000	0.220
BASOLATERAL_PLASMA_MEMBRANE	0.015	0.218
COLLAGEN	0.042	0.220
SKELETAL_DEVELOPMENT	0.059	0.219
PHOSPHOINOSITIDE_MEDIATED_SIGNALING	0.059	0.224
PATTERN_BINDING	0.075	0.220
CALCIUM_ION_TRANSPORT	0.085	0.219
ADHERENS_JUNCTION	0.096	0.221
GLYCOSAMINOGLYCAN_BINDING	0.096	0.223
CALCIUM_CHANNEL_ACTIVITY	0.109	0.216
POLYSACCHARIDE_BINDING	0.111	0.218
CARBOHYDRATE_BINDING	0.111	0.220
TRANSFORMING_GROWTH_FACTOR_BETA_RECEPTOR_SIGNALING_PATHWAY	0.111	0.220
HEPARIN_BINDING	0.107	0.225
TRANSMEMBRANE_RECEPTOR_PROTEIN_TYROSINE_KINASE_ACTIVITY	0.115	0.218
DI__TRI_VALENT_INORGANIC_CATION_TRANSPORT	0.141	0.218
TRANSMEMBRANE_RECEPTOR_PROTEIN_SERINE_THREONINE_KINASE_SIGNALING_PATHWAY	0.168	0.221
CELL_JUNCTION	0.180	0.222
ADENYLATE_CYCLASE_ACTIVATION	0.185	0.219
ACTIN_FILAMENT_BASED_PROCESS	0.205	0.221
CYCLIC_NUCLEOTIDE_MEDIATED_SIGNALING	0.205	0.223
TRANSCRIPTION_REPRESSOR_ACTIVITY	0.208	0.222
REGULATION_OF_SIGNAL_TRANSDUCTION	0.206	0.226
MOLECULAR_ADAPTOR_ACTIVITY	0.207	0.225
GROWTH	0.220	0.226
EXTRACELLULAR_STRUCTURE_ORGANIZATION_AND_BIOGENESIS	0.247	0.218
G_PROTEIN_SIGNALING_COUPLED_TO_CYCLIC_NUCLEOTIDE_SECOND_MESSENGER	0.249	0.223

Table 4.17: Genesets enriched common in LOW group

Genesets	FDR q-val	
	GSE29272	GSE62254
KINETOCHORE	0.000	0.181
MITOTIC_CELL_CYCLE_CHECKPOINT	0.002	0.183
RESPONSE_TO_DNA_DAMAGE_STIMULUS	0.000	0.185
CHROMOSOME PERICENTRIC REGION	0.000	0.186
CELL_CYCLE_GO_0007049	0.000	0.187
INTERPHASE	0.001	0.190
DNA_REPAIR	0.001	0.190
RIBONUCLEOPROTEIN_COMPLEX	0.000	0.192
CELL_CYCLE_CHECKPOINT_GO_0000075	0.000	0.192
DNA_METABOLIC_PROCESS	0.001	0.194
NUCLEAR_CHROMOSOME	0.002	0.195
CONDENSED_CHROMOSOME	0.002	0.198
REGULATION_OF_MITOSIS	0.000	0.200
DNA_REPLICATION	0.002	0.199
INTERPHASE_OF_MITOTIC_CELL_CYCLE	0.002	0.199
M_PHASE_OF_MITOTIC_CELL_CYCLE	0.000	0.203
SPINDLE_POLE	0.012	0.191
MITOCHONDRIAL_RIBOSOME	0.012	0.192
CHROMOSOME_SEGREGATION	0.001	0.207
MITOSIS	0.000	0.208
DNA_RECOMBINATION	0.015	0.193
CHROMOSOMAL_PART	0.000	0.216
ORGANELLAR_RIBOSOME	0.015	0.202
NUCLEAR_MEMBRANE_PART	0.003	0.217
MICROTUBULE_CYTOSKELETON	0.000	0.224
MRNA_METABOLIC_PROCESS	0.015	0.210
REGULATION_OF_MITOTIC_CELL_CYCLE	0.032	0.195
CENTROSOME	0.024	0.210
COENZYME_METABOLIC_PROCESS	0.031	0.204
NEGATIVE_REGULATION_OF_DNA_METABOLIC_PROCESS	0.049	0.188
DNA_DEPENDENT_DNA_REPLICATION	0.004	0.247
MITOCHONDRION_ORGANIZATION_AND_BIOGENESIS	0.053	0.220
MICROTUBULE_ORGANIZING_CENTER	0.032	0.249
ORGANELLE_ENVELOPE	0.032	0.250
LIGASE_ACTIVITY	0.081	0.202
REGULATION_OF_DNA_REPLICATION	0.055	0.232
COFACTOR_METABOLIC_PROCESS	0.080	0.229
ORGANELLE_INNER_MEMBRANE	0.083	0.229
MITOCHONDRIAL_PART	0.127	0.186
MITOCHONDRIAL_INNER_MEMBRANE	0.119	0.198
MITOCHONDRIAL_MATRIX	0.138	0.195
MITOCHONDRIAL_LUMEN	0.159	0.190
NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_TRANSPORT	0.158	0.225
ENDONUCLEASE_ACTIVITY	0.201	0.223

Clustering analysis of a third dataset showed two main clusters of tumors observed for GSE35809 dataset defined by hierarchical clustering with distinct expression patterns; those with clearly low expression of most of the genes (L) and those with relatively higher expression (H) (**Figure 4.30**). I did not designate any samples as intermediate (I) in this dataset due to the limited sample size.

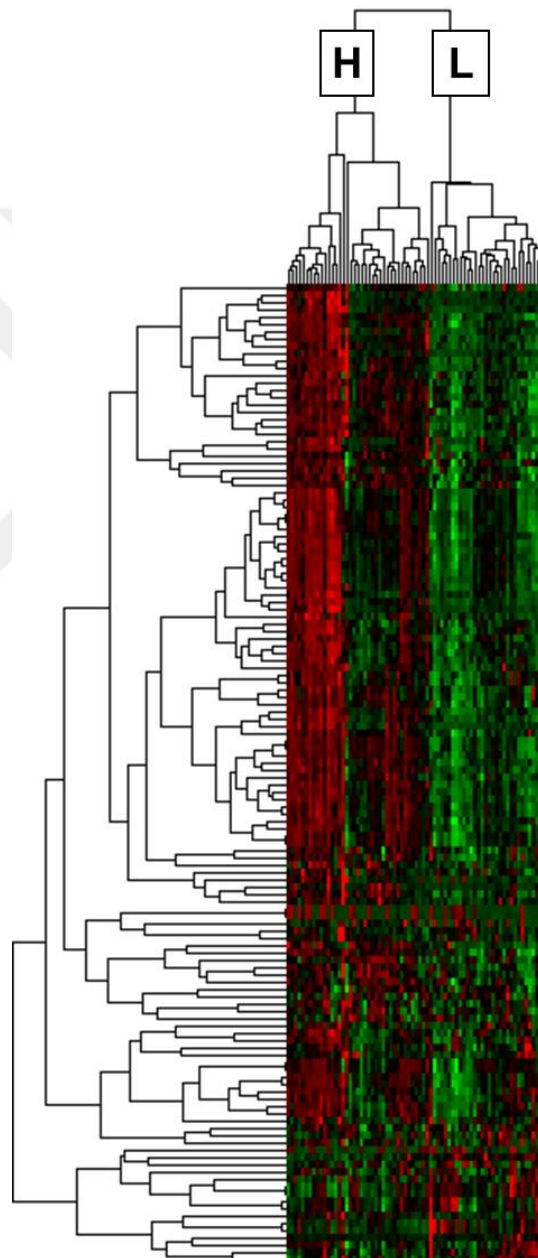


Figure 4.30: Molecular classification of gastric tumors in GSE35809 dataset. 134 probesets previously selected for molecular classification of gastric cancer tumors were used for hierarchical clustering in GSE35809 dataset (n=70). Samples are classified according to their level of gene expression as high “H”, and low “L”.

GSEA analysis of GSE35809 dataset revealed no enrichment in the HIGH group, however 62 genesets (listed in APPENDIX I) were enriched in the LOW group. 25 genesets which were enriched in all of the LOW groups of all three datasets studied, GSE29272, GSE62254 and GSE35809 are listed in **Table 4.18**. As shown, all of the genesets were related to DNA replication, cell cycle and mitosis strongly supporting a highly proliferative molecular phenotype for the samples in the LOW group.

Table 4.18: Genesets enriched common in LOW group

Genesets	FDR q-val		
	GSE29272	GSE62254	GSE35809
RESPONSE_TO_DNA_DAMAGE_STIMULUS	0.000	0.185	0.190
CELL_CYCLE_GO_0007049	0.000	0.187	0.191
INTERPHASE	0.001	0.190	0.193
CELL_CYCLE_CHECKPOINT_GO_0000075	0.000	0.192	0.196
DNA_METABOLIC_PROCESS	0.001	0.194	0.197
NUCLEAR_CHROMOSOME	0.002	0.195	0.198
CONDENSED_CHROMOSOME	0.002	0.198	0.198
REGULATION_OF_MITOSIS	0.000	0.200	0.198
INTERPHASE_OF_MITOTIC_CELL_CYCLE	0.002	0.199	0.199
M_PHASE_OF_MITOTIC_CELL_CYCLE	0.000	0.203	0.199
SPINDLE_POLE	0.012	0.191	0.200
MITOCHONDRIAL_RIBOSOME	0.012	0.192	0.200
MITOSIS	0.000	0.208	0.200
CHROMOSOMAL_PART	0.000	0.216	0.201
ORGANELLAR_RIBOSOME	0.015	0.202	0.202
NUCLEAR_MEMBRANE_PART	0.003	0.217	0.202
MICROTUBULE_CYTOSKELETON	0.000	0.224	0.203
REGULATION_OF_MITOTIC_CELL_CYCLE	0.032	0.195	0.207
CENTROSOME	0.024	0.210	0.208
NEGATIVE_REGULATION_OF_DNA_METABOLIC_PROCESS	0.049	0.188	0.208
DNA_DEPENDENT_DNA_REPLICATION	0.004	0.247	0.210
MITOCHONDRION_ORGANIZATION_AND_BIOGENESIS	0.053	0.220	0.211
MICROTUBULE_ORGANIZING_CENTER	0.032	0.249	0.214
LIGASE_ACTIVITY	0.081	0.202	0.217
REGULATION_OF_DNA_REPLICATION	0.055	0.232	0.217

4.2.3.3. Relation of molecular sub-groups of gastric cancer with EMT

Accumulating literature showed that cells undergoing EMT remodel their ECM via upregulation of collagens and fibronectin, reorganize an actin-rich membrane which facilitates cell motility [155], and also slow down proliferation [113]. Therefore GSEA results related to collagen, ECM reorganization, actin filament based processes in HIGH group; and enhanced cell cycle, interphase and DNA repair genesets in LOW group strongly suggested an involvement of EMT in molecular classification of gastric tumors.

To assess the involvement of EMT in our molecular classification I performed hierarchical clustering analysis with Tan et al's generic EMT signature which was generated for tumors and can be used independent of the tissue-of origin [1] (**Figure 4.31**). The probesets corresponding to each gene in this gene list requested from the authors are listed in APPENDIX J. Samples with easily noticeable differential expression of mesenchymal and epithelial probesets were classified as mesenchymal (M) and epithelial (E). The rest of the samples with no clear pattern are classified as intermediates (I) (**Figure 4.31**). Apparently, gastric cancer tumors can be separated into very clear groups based on an EMT signature.

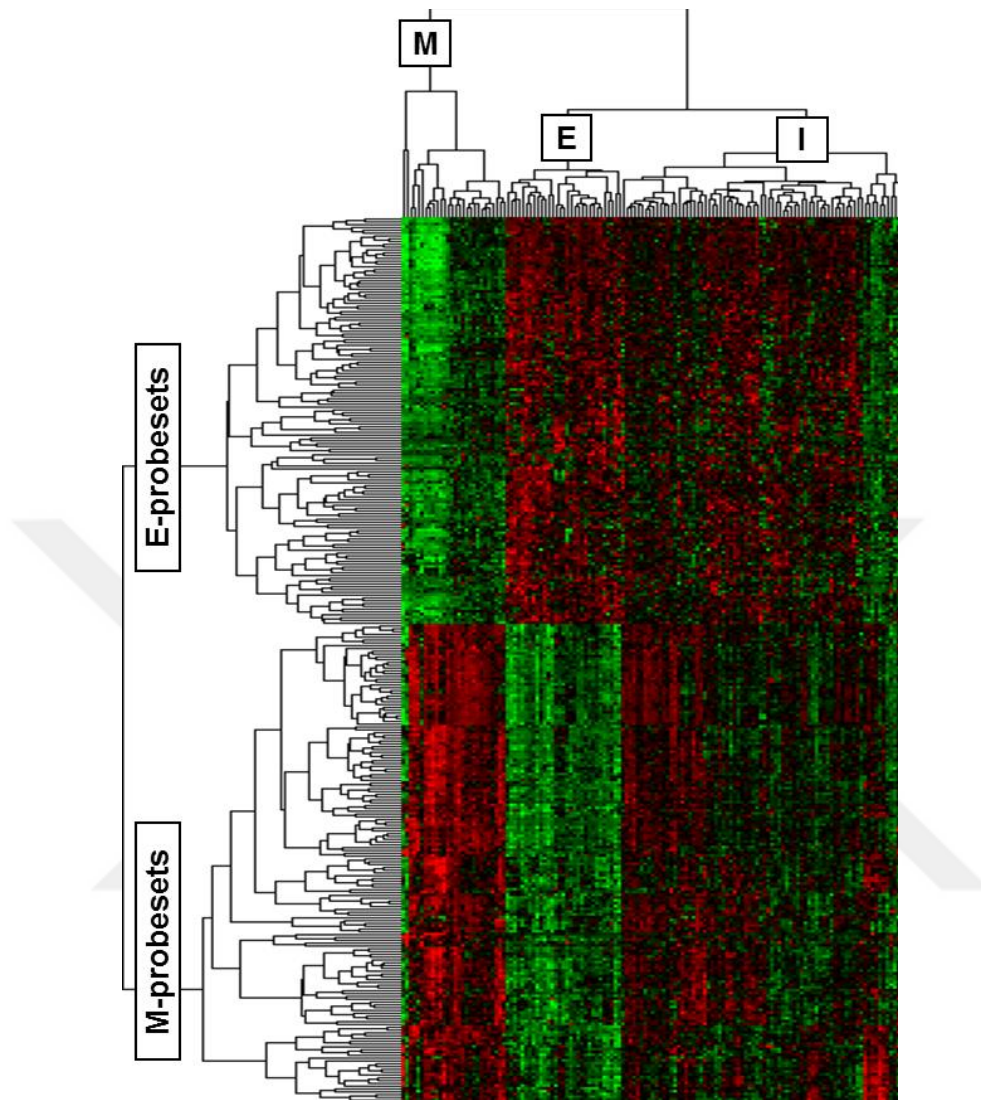


Figure 4.31: Hierarchical clustering of GSE29272 dataset with generic EMT marker genes. 315 probesets listed in APPENDIX J were used as an EMT signature for hierarchical clustering in the GSE29272 dataset (n=126). Epithelial signature (n=145) and mesenchymal signature (n=170) were perfectly clustered within each group and separated from each other as shown with “E-probesets” and “M-probesets”, respectively. Samples were classified according to their level of EMT marker expression as epithelial (E), mesenchymal (M) and intermediate (I).

When gastric tumors of GSE29272 dataset were classified with the generic EMT signature of Tan et al., we found out that our unsupervised molecular sub-groups are related to EMT (**Table 4.19**). Majority of the mesenchymal tumors were in 'High' group and most of the epithelial tumors were in 'Low' group. Intermediate samples also were found to overlap when analyzed for both classification methods. Chi square test showed that molecular sub-groups are related to EMT classification with high statistical significance ($p < 0.0001$).

Table 4.19: Molecular and EMT based classification of gastric cancer tumors –GSE29272

Molecular Sub-groups	EMT classification			Sum
	M	I	E	
HIGH	22	8	0	30
INT	4	61	10	75
LOW	2	6	21	29
Sum	28	75	31	Total: 134

Analysis of another dataset GSE22377 ($n=43$) using my molecular signature of 134 probesets and also with the EMT signature of Tan et al. showed, similar to my previous analysis, that samples generally overlapped (**Figure 4.32**). Although the sample size of this dataset was not as high as previously analyzed datasets ($n=43$), it is evident that all mesenchymal tumors were in 'High' group with a Chi square p value below 0.0001 indicating a significant relationship between molecular sub-groups and EMT classification (**Table 4.20**).

Table 4.20: Molecular and EMT based classification of gastric cancer tumors –GSE22377

Molecular Sub-groups	EMT classification			Sum
	M	I	E	
HIGH	17	0	3	20
INT	0	4	7	11
LOW	0	8	4	12
Sum	17	12	14	Total: 43

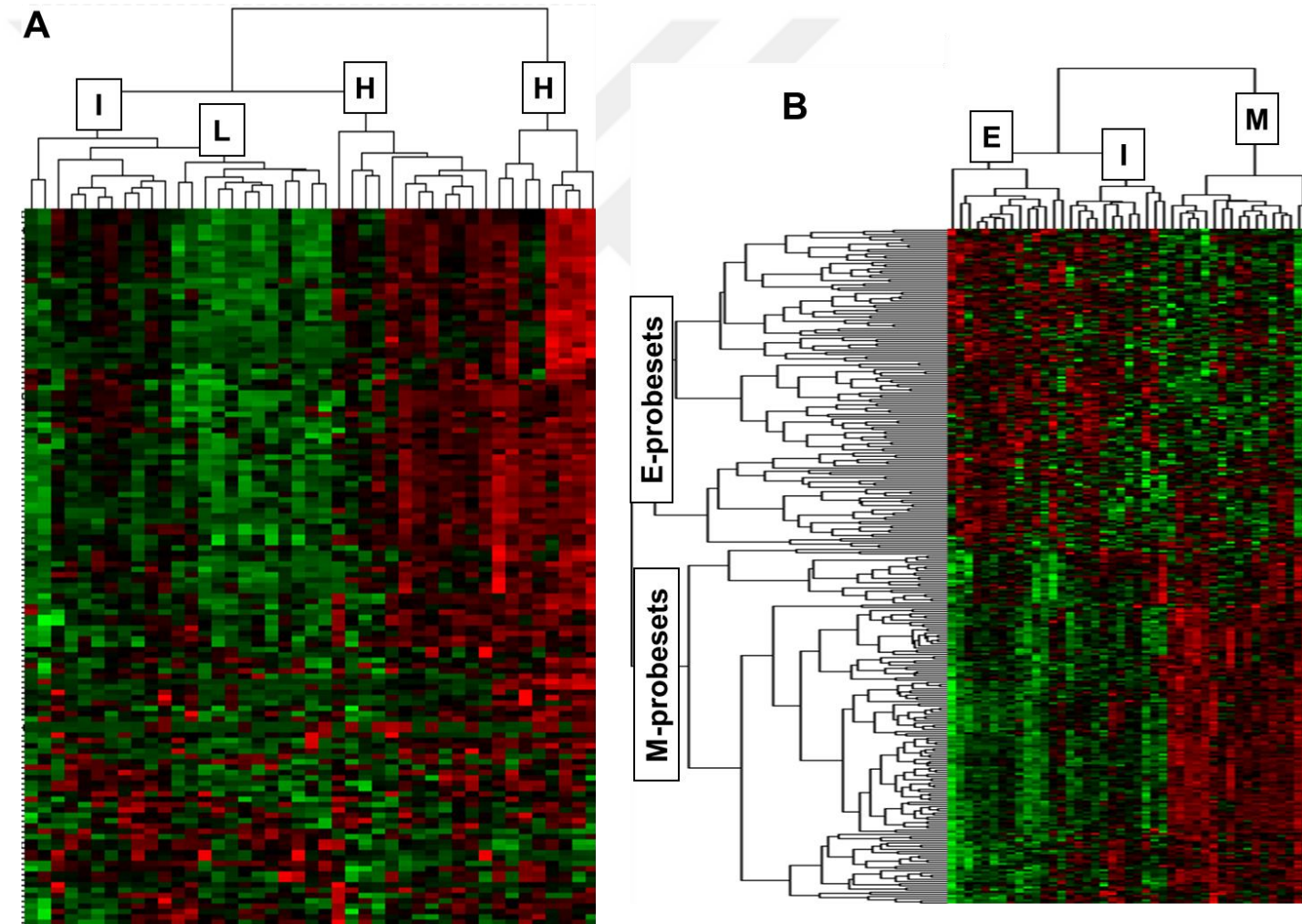


Figure 4.32: Molecular and EMT based classification of gastric tumors in GSE22377 dataset. 134 probesets were used for hierarchical clustering in GSE22377 dataset (n=43)(A). Samples were classified according to their level of gene expression as high “H”, intermediate “I” and low “L”. Tan et al. EMT signature [1] for tumors (APPENDIX J) were used for hierarchical clustering in GSE22377 (B). 165 out of 170 mesenchymal markers were clustered together (M-probesets) on the upper panel and 144 of 145 epithelial markers were clustered (E-probesets) on the lower panel. Samples were classified according to their level of EMT marker expression as epithelial (E), mesenchymal (M) and intermediate (I).

4.2.3.4. Relation of molecular sub-groups of gastric cancer with pathologic (Lauren) classification

In order to check whether my molecular sub-groups overlap with pathologic classification, we used GSE15459 (n=192) where Lauren classification of the tumors were available. Molecular sub-grouping of tumors in GSE15459 was performed with the 134 probesets. As can be observed from the heatmap (**Figure 4.33**) and **Table 4.21**, it is obvious that the molecular classification is highly related to pathologic Lauren classification (Chi square p value<0.0001). However these two classifications do not fully overlap; and my molecular classification can further sub-group the pathological subtypes defining an intermediary group between diffuse and intestinal types.

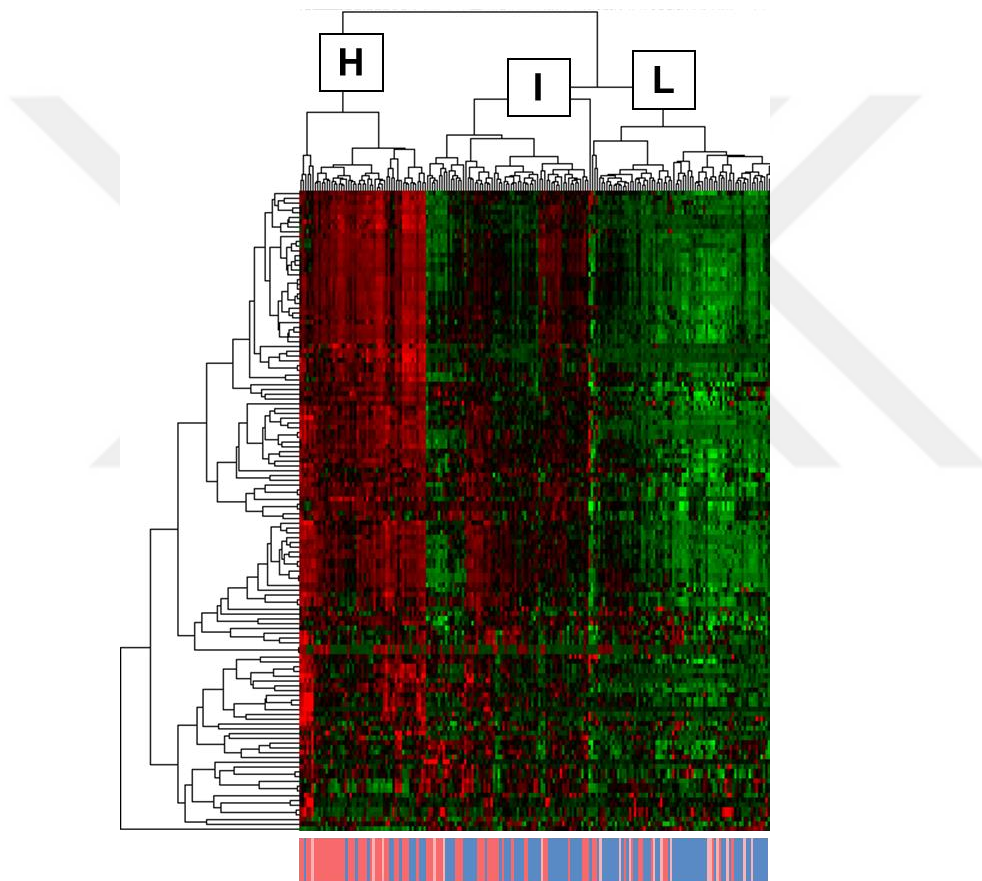


Figure 4.33: Molecular classification of gastric tumors in GSE15459 dataset. 134 probesets were used for hierarchical clustering in GSE15459 dataset (n=192). Tumor groups with discernable high and low expression were labeled with “H” and “L”, respectively. The rest was assigned to intermediate group “I”. Color panel at the bottom indicates the Lauren classification. Red: diffuse, blue: intestinal, pink: mixed.

Table 4.21: Molecular and pathologic classification of gastric cancer tumors –GSE15479

Lauren Classification	Molecular sub-groups			Sum
	HIGH	INT	LOW	
Diffuse	36	30	9	75
Mixed	3	6	9	18
Intestinal	13	48	38	99
Sum	52	84	56	Total: 192

4.2.3.5. Relation of molecular sub-groups of gastric cancer with prognostic UP-DOWN classification

I next compared our molecular groups with the prognostic groups which were identified by my first approach. As can be seen, subtypes identified by the two approaches are highly related with chi square p values below 0.0001 in both datasets (**Table 4.22**). Almost all patients in HIGH molecular group were in the UP group and a minimum of 84% of the patients in LOW group were in the DOWN group. INT group samples were almost equally distributed among DOWN and UP patients. Therefore, identification of molecular and prognostic groups in gastric cancer lead to highly similar classifications of the tumors which means that the differences in the clinical outcome are directly related to distinct biological properties of the gastric cancer tumors. With INT group defined for molecular classification, I could be able to identify the biological differences much strongly and consistently (HIGH vs LOW) than biological differences of UP-DOWN prognostic groups; however the trends were quite similar.

Table 4.22: Molecular and prognostic classification of gastric cancer tumors –GSE15459 & GSE62254

GSE15459			
Molecular Sub-groups	Prognostic groups		Sum
	DOWN	UP	
HIGH	2	50	52
INT	41	43	84
LOW	56	0	56
Sum	99	93	Total: 192

GSE62254			
Molecular Sub-groups	Prognostic groups		Sum
	DOWN	UP	
HIGH	0	19	19
INT	82	76	158
LOW	65	11	76
Sum	147	106	Total: 253

I next asked if the biological differences identified (HIL) also are associated with the clinical outcome. Log rank tests applied to HIGH, INT and LOW groups in three independent microarray datasets showed that the molecular sub-groups could not consistently classify prognostic subgroups. As shown in **Figure 4.34**, we observed significant log rank p values for GSE62254 and GSE15459 datasets, however the intermediate group can show a similar prognostic profile to HIGH and LOW depending on the cohort used. This suggests that there is still a heterogeneity within intermediate group of patients, which needs to be further investigated. In addition we could not observe a prognostic difference between HIL groups in GSE29272 dataset (**Figure 4.34**). The reasons for this result should be further studied. I therefore, pursued the identification of specific prognostic biomarkers for each molecular sub-group by mUSAT analysis. This resulted in a list of prognostic genes for each sub-group with only one probeset common to all sub-groups (**Figure 4.35**), indicating that prognostic determinants were indeed different for each subgroup and supporting the idea that the pursuit of prognostic markers which will work in all molecular groups is not a solid idea. In other words, this result demonstrated that these molecular groups are indeed biologically distinct and the prognosis in these groups cannot be explained by one set of common prognostic markers. I therefore suggest that in order to use these markers in the clinic, one would first have to determine the molecular sub-group of a specific type of tumor by using the 134 gene signature, and then use another set of genes for the assessment of prognosis for that specific molecular sub-group. A drawback of this approach, however, is that although this would lead to well-defined and more precise prediction of prognosis, it may not be practical and feasible when it comes to application. Therefore a robust and one-step prognostic association as described in approach 1 with a clear-cut distinction of clinical outcome between UP-DOWN expression groups may be more applicable to clinical practice.

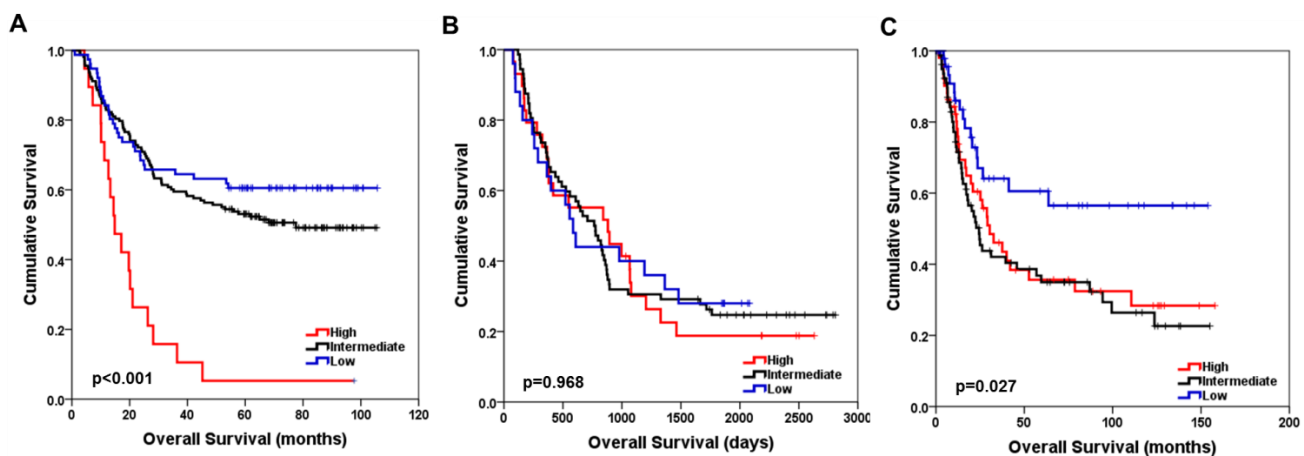


Figure 4.34: Kaplan Meier plots for molecular sub-groups of gastric cancer tumors. Samples were stratified using high-intermediate-low molecular classification in GSE62254 (A) GSE29272 (B) and GSE15459 (C). Log rank p values are shown. In H-I-L groups, there are 19, 158, 76 samples in GSE62254; 29, 72, 25 samples in GSE29272; and 52, 84, 56 samples in GSE15459, respectively.

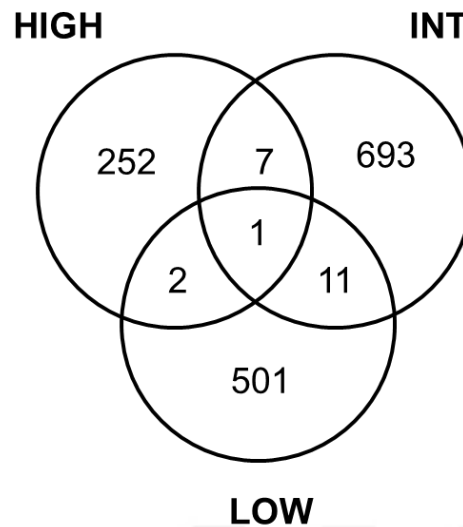


Figure 4.35: Venn diagram of prognostic probesets in each molecular sub-group. mUSAT was ran restrictedly for each molecular sub-group in GSE29272. High, intermediate and low groups included 30, 75 and 29 patients, respectively. The probesets with significant linear relationship with overall survival were shown (Univariate Cox PH p value<0.05).

4.2.4. EMT based classification of gastric cancer cell lines

An aim of my work was to associate the molecular sub-groups I identified for gastric cancer with sensitivity to a given targeted therapy agent or chemotherapeutic. This would be possible if I could identify cell lines corresponding to tumor sample sub-groups and then analyzing one of many databases where cell line cytotoxicity data is available. Therefore, I asked if I could observe similar molecular sub-groups in cell lines by applying the molecular classification to the cell lines. However the hierarchical clustering analyses with 134 probesets did not result in clear groups of cell lines with different expression patterns; most probably due to the fact that the biological differences defined in tumors by molecular classification depended highly on ECM involvement and thus microenvironment, which cells lack.

To overcome this obstacle I used a published EMT marker gene list generated specifically for cancer cell lines [1], since I have identified that molecular sub-groups of gastric cancer were highly related to EMT status. I analyzed the gastric cancer cancer cell lines in CCLE (Cancer cell line encyclopedia) and CGP (Cancer genome project) databases and cell lines with all cancer types in NCI60 (National cancer institute 60) database, in terms of EMT related gene expression and found that cancer cells are constructed of mainly two types of cells, epithelial or mesenchymal, when analyzed gastric cancer specifically (CCLE & CGP) and cancer-independently (NCI60) (**Figure 4.36, Table 4.23**).

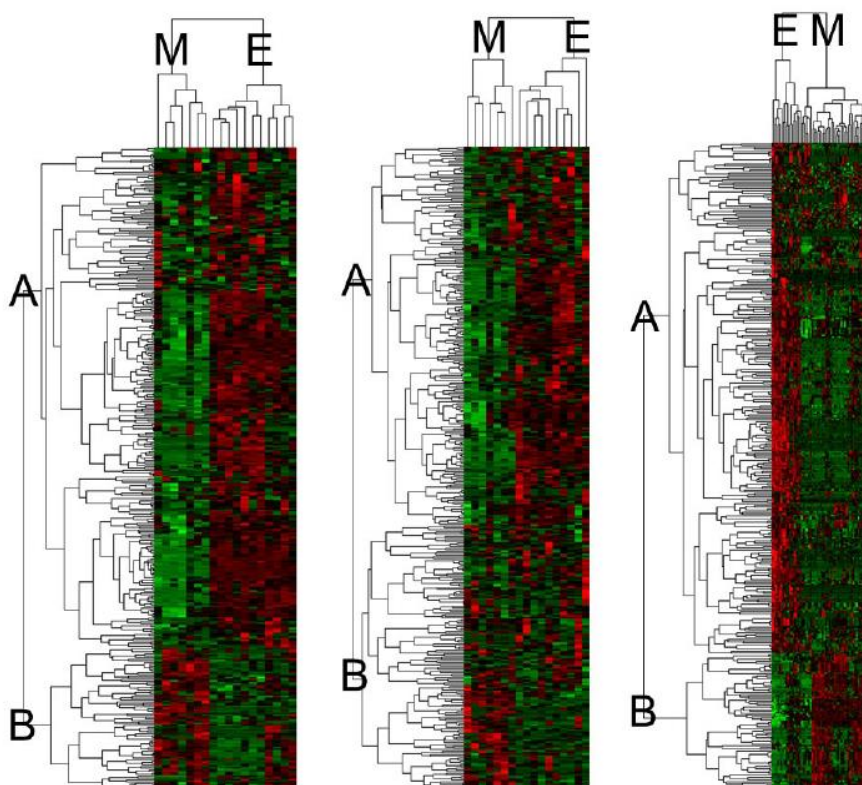


Figure 4.36: Hierarchical clustering analysis of gastric cancer cell lines with Tan et al. [1] generic EMT signature for cell lines. Left panel: gastric cancer cell lines in CCLE, middle panel: gastric cancer cell lines in CGP, right panel: all cancer cell lines in NCI60. A and B group genes are mostly comprised of epithelial and mesenchymal markers, respectively (**Table 4.23**). “E” represents cells with higher expression of epithelial and lower expression of mesenchymal markers and “M” is vice versa.

Table 4.23: Cell and gene distribution in hierarchical clustering analysis in terms of epithelial-mesenchymalness

Datasets	A group genes		B group genes		p (Fisher's exact)	Distribution of cells	
	E	M	E	M		E	M
CCLE (GC)	284	10	6	77	< 0.0001	11	7
CGP (GC)	221	1	69	86	< 0.0001	10	7
NCI60 (all cells)	290	0	8	79	< 0.0001	17	42

4.3. *IN SILICO* IDENTIFICATION OF COMPOUNDS WITH SELECTIVE ACTION ON EPITHELIAL AND MESENCHYMAL CELLS

The study of prognosis of two gastrointestinal cancer types, colon and gastric, led to an understanding of a major involvement of EMT in clinical outcome, as well as in molecular sub-group determination. However in current clinical practice EMT status is not a clinical determinant, and thus, tumors with various EMT phenotypes are treated with a single agent or a single combination of drugs. It is established that cancer cells with epithelial and mesenchymal characteristics have totally distinct biology and I therefore hypothesized that they could be targeted selectively using specific compounds.

Despite its exceptionally large drug cytotoxicity data including more than 50000 chemicals, FDA approved drugs, and natural compounds, the wide tissue heterogeneity of the NCI60 (National Cancer Institute 60) cell line panel [156,157] is considered a factor that hinders the utilization of this database. Here I describe a novel approach enabling the analysis of NCI60 data in order to identify drugs specifically effecting either epithelial or mesenchymal cells regardless of the tissue type. This demonstrates a novel use for the NCI60 database and might help us understand the major mechanisms that should be targeted for both types of cancer cells.

4.3.1. Generation of a in-house EMT scoring method

To understand the status of cancer cell lines in terms of their EMT status, I first developed a scoring that would assign a score for the mesenchymal (M) phenotype, and one for the E (epithelial). Then the combination of these would give the overall EMT score for a given cell line. For this purpose, we used a published EMT related gene list generated for cancer cell lines regardless of the tissue of origin as a reference [1]. This method relies on a complex scoring algorithm to determine the EMT status of tumors and cancer cell lines. The score is based on the evaluation of the expression of 315 genes for tumors, and 218 genes for cell lines and can successfully score tumor samples and cell lines. Utilizing this scoring method as reference, we generated an in-house 2 gene based EMT scoring method which can successfully define EMT status of cell lines to an extent that is as good as the original method described by Tan et al.

Tan et al.'s EMT scoring algorithm was applied for all NCI60 cell lines by the authors upon our request (Algorithm of the scoring is not publicly available) (APPENDIX K and **Figure 4.37A**). This scoring ranges from -1 to 1 (from epithelial to mesenchymal); cell lines are divided into epithelial and mesenchymal groups at the cut-off of 0. Since leukemia and ovarian cancer cells were all in

mesenchymal range (0 to 1) we excluded these cancer types from the further pooled analyses to avoid tissue-type based bias (**Figure 4.37A**).

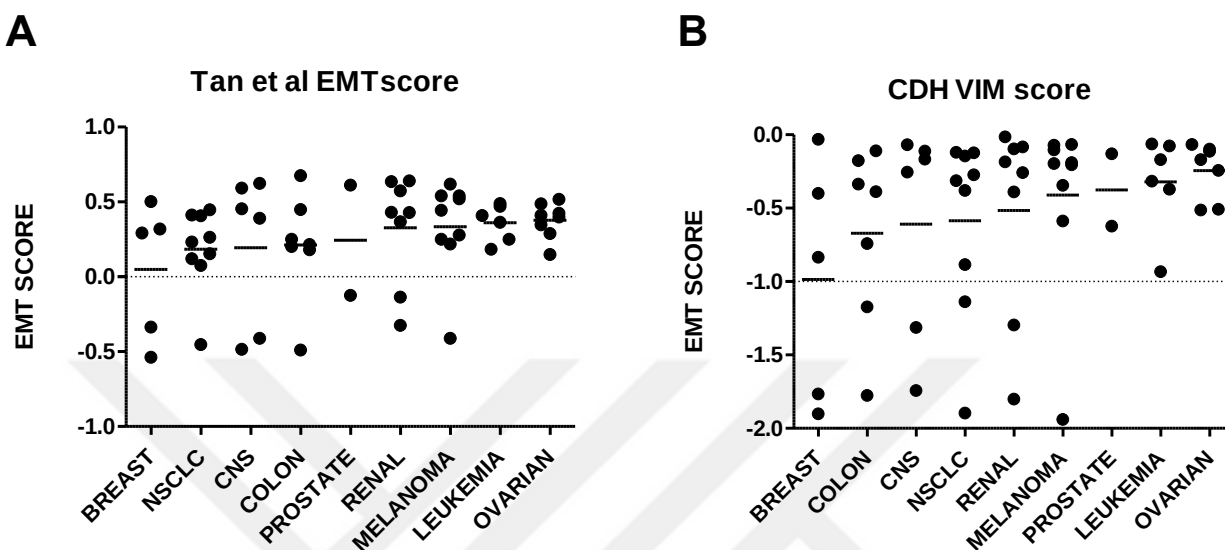


Figure 4.37: Scatter plot of EMT scoring of NCI60 cancer cell lines. Tan et al's EMT scoring [1] values in APPENDIX K (A), and in house CDH1-VIM gene expression based EMT scoring (B), for NCI60 cell lines were plotted. Horizontal line shows mean of each cancer type. EMT scores from min to max represents EMT status from epithelial to mesenchymal for both scoring methods.

Using breast, non-small cell lung cancer, central nervous system, colon, prostate, renal and melanoma tumor cell lines which have both epithelial and mesenchymal cells according to Tan et al. EMT score (**Figure 4.37A**), we pooled epithelial cells (with EMT score below 0) and mesenchymal cells (with EMT score above 0) separately. Using student's ttest, we identified 101 compounds which were differentially effective on epithelial and mesenchymal cell pools within the NCI60 database (p value<0.05).

I next aimed to select the genes which were highly correlated with sensitivity to these 101 compounds, because these genes would later be necessary for the development of our in-house EMT scoring method. The correlation analysis of the expression of each gene in Tan et al's generic EMT signature with sensitivity to these 101 drugs, revealed that CDH1 was the gene with the highest average correlation r value with 74 compounds which were selectively effective on epithelial cells (Average Pearson r=0.39, Pearson correlation r>0.30 for 73/74 compounds, Pearson correlation p value<0.05 for 65/74 compounds). On the opposite side, VIM gene had the highest average correlation r with 27 compounds which are selectively effective in mesenchymal cells (Average Pearson r=0.33, Pearson correlation r>0.30 for 20/27 compounds, Pearson correlation p value<0.05 for 19/27 compounds). As I observed that CDH1 and VIM expression had the highest correlation with sensitivity to these drugs I proceeded by generating a simplified version of the EMT scoring based on CDH1 and VIM expression

only (see methods). The CDH1-VIM score showed 93% correlation with the 218 gene based EMT score (Figures 4.37, 4.38 and 4.39). I thus decided to use this simpler and easily validatable 2 gene EMT score for the re-evaluation of the National cancer institute 60 data.



Figure 4.38: Correlation of a multi-gene EMT score and the CDH1-VIM score. 46 NCI60 cell lines (leukemia and ovarian cancer cell lines were excluded) were sorted according to their Tan et al. multi-gene EMT score [1] from min to max (upper panel), CDH1-VIM scores are shown below (lower panel). Green: epithelial, red: mesenchymal.

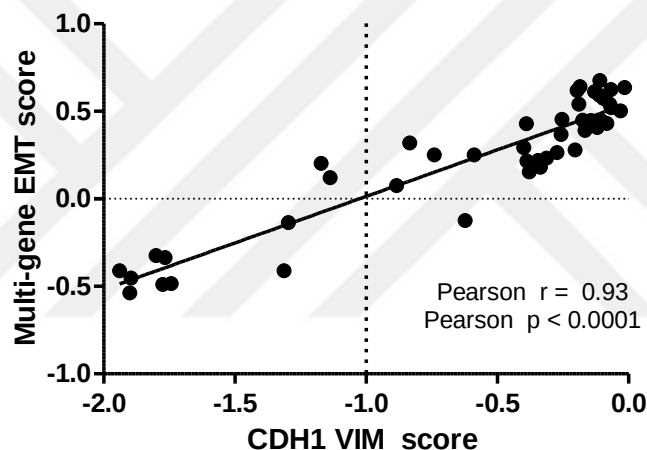


Figure 4.39: Correlation plot of the multi-gene EMT score and CDH1-VIM score. Tan et al.'s generic EMT score [1] is shown in y axis and our in-house CDH1-VIM based EMT scoring is shown on the x axis. 46 NCI60 cell lines (leukemia and ovarian cancer cell lines were excluded) were included. EMT scores from min to max represents EMT status from epithelial to mesenchymal for both scoring methods.

As mesenchymal status was widely associated with cancer stem cells, stem-like properties, and thus induction of stem cell marker expression [158], I wondered whether mesenchymal cells determined by CDH1-VIM scoring will relate to stem cell markers. An analysis of expression of CD44 and CD24 which are known stem and non-stem cell markers, respectively, with the EMT status revealed that the CDH1-VIM score is also related to stemness of NCI60 cancer cell lines (**Figure 4.40**).

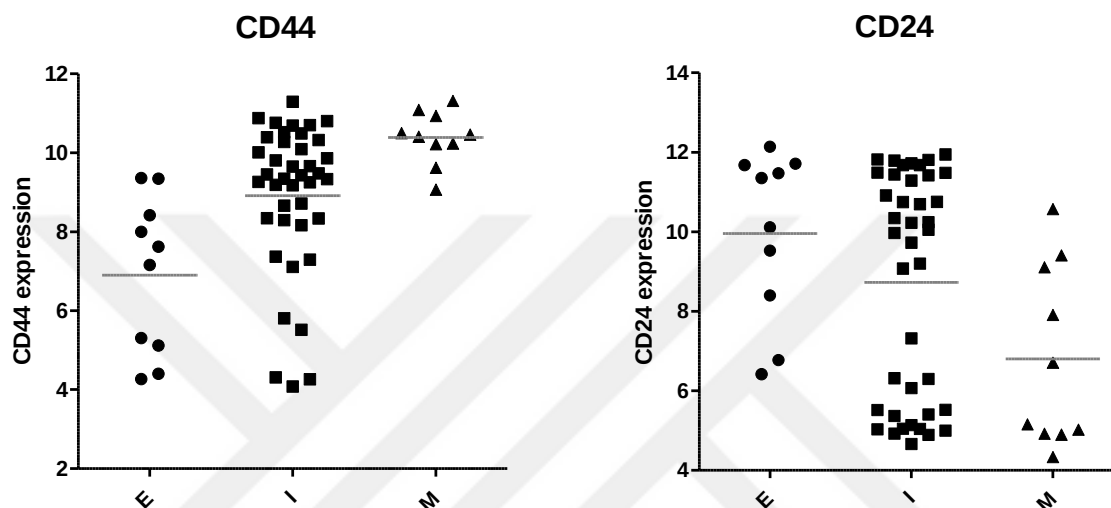


Figure 4.40: Stem cell marker expression in EMT groups determined by CDH1-VIM score. Microarray based mRNA level CD44 and CD24 expression in NCI60 cancer cell lines are plotted. Grouping is done as follows: Cell lines with highest 10 CDH1-VIM score: M(mesenchymal), lowest 10: E (epithelial) and the rest: I (intermediate). Medians are shown with red horizontal line.

4.3.2. Identification of novel compounds that are preferentially effective on epithelial or mesenchymal cells, independent of tissue-of-origin in NCI60

In order to identify drugs that selectively kill epithelial or mesenchymal cancer cells, using NCI60 data, I used 2 criteria. First set was for the quality check of drug sensitivity data and second for the identification of compounds with selective growth inhibition of epithelial or mesenchymal cells. (**Table 4.24**). To ensure data quality, I eliminated the drugs which had a sensitivity range below 1.2 M, the drugs for which more than 10 samples had GI50 at the maximum test dose, the drugs with more than 10 missing and 5 repeating GI50 values. Surprisingly only 299 compounds out of 50000 fulfilled our data quality criteria.

Table 4.24: Selection of compounds in NCI60

1. Data quality criteria
Drug sensitivity range (max-min logGI50 (M)) higher than 1.2 (NCI60's own quality criteria)
Maximum of 10 samples with GI50 values equal to the max test dose (elimination of totally inactive compounds)
Maximum of 10 missing values
Maximum of 5 repeating values



299 compounds

2. Drug Selection Criteria
Drug sensitivity correlated with CDH1 and VIM expression(abs. Pearson $r > 0.30$, $p < 0.05$)
T-test between the drug sensitivity of epithelial and mesenchymal samples (CDH1-VIM score above and below -1 respectively) $p < 0.05$



66 compounds

58 epithelial + 8 mesenchymal

For the selection of compounds with selective growth inhibition of epithelial and mesenchymal cells, a 30% significant ($p < 0.05$) linear correlation between the expression of CDH1 and VIM genes with drug cytotoxicity values was set as the first criteria. In addition to that a significant difference between the means of drug cytotoxicity values of epithelial and mesenchymal cells as determined by ttest ($p < 0.05$) was my second criteria. Out of these 299 compounds, sensitivity to 66 significantly correlated with my EMT score. 58 and 8 were effective on epithelial and mesenchymal cells, respectively (APPENDIX L).

There were no drugs with a name or known mode of action among the drugs targeting mesenchymal cells. The available data about these compounds were their NSC numbers-the NCI's internal ID number; and PubChem IDs for 6 of them (**Table 4.25**). The chemical composition of these compounds is shown in **Figure 4.41**.

Table 4.25: NSC codes and Pubchem IDs of compounds selectively effecting mesenchymal cells

NSC codes	Pubchem IDs	Name in Pubchem
700415	526946	3-ethyl-2-formylthiophene
766320	-	-
667529	512308	AC1LAE9M
732508	48429634	-
687578	521563	AC1LAVR8
635296	497572	AC1LAXR6
635297	497573	AC1LAXR8
699758	526589	2,6-Bis-(2-hydroxy-5-methyl-benzyl)-4-methyl-phenol, TMS

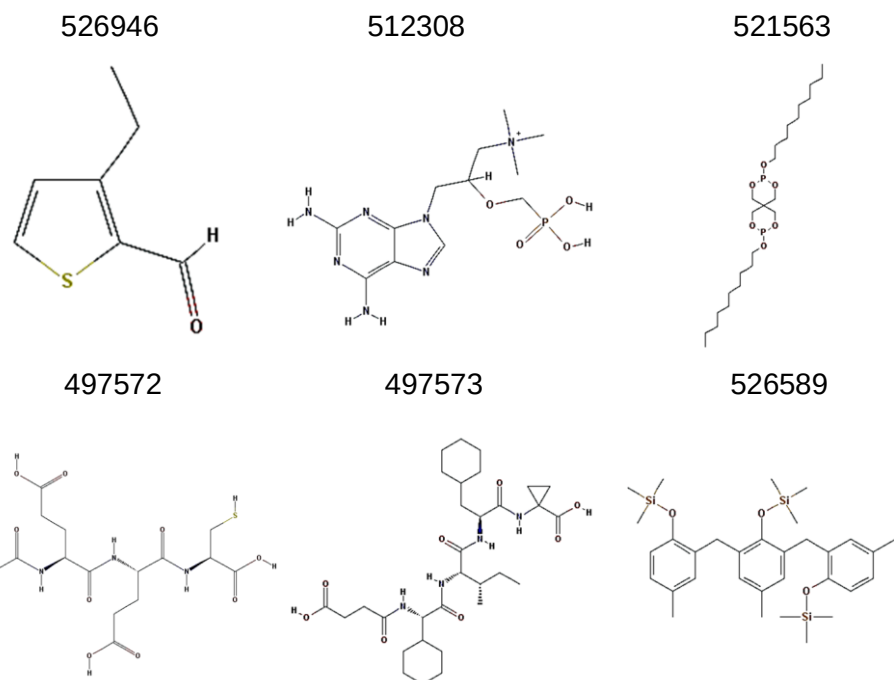


Figure 4.41: Pubchem 2D structure of compounds selectively effective on mesenchymal cells. Only those compounds with a Pubchem ID are shown. Pubchem ID's are shown above the structures.

We then tried to identify functional sub-clusters among the drugs that were found to selectively inhibit growth of epithelial cells. I therefore generated a correlation plot using cytotoxicity values of 58 `epithelial cell targeting` drugs in both rows and columns (**Figure 4.42**). This revealed 5 sub-clusters, the largest of which contained 3 berberine derivatives. These have been previously shown to downregulate EGFR and its downstream signaling through MEK-ERK, thus inhibiting proliferation in colon cancer, prostate cancer and glioblastoma separately [159-161]. Our data showed that epithelial cells are significantly more sensitive to berberine derivatives in NCI60 panel when all cancer types are pooled (**Figure 4.43**).

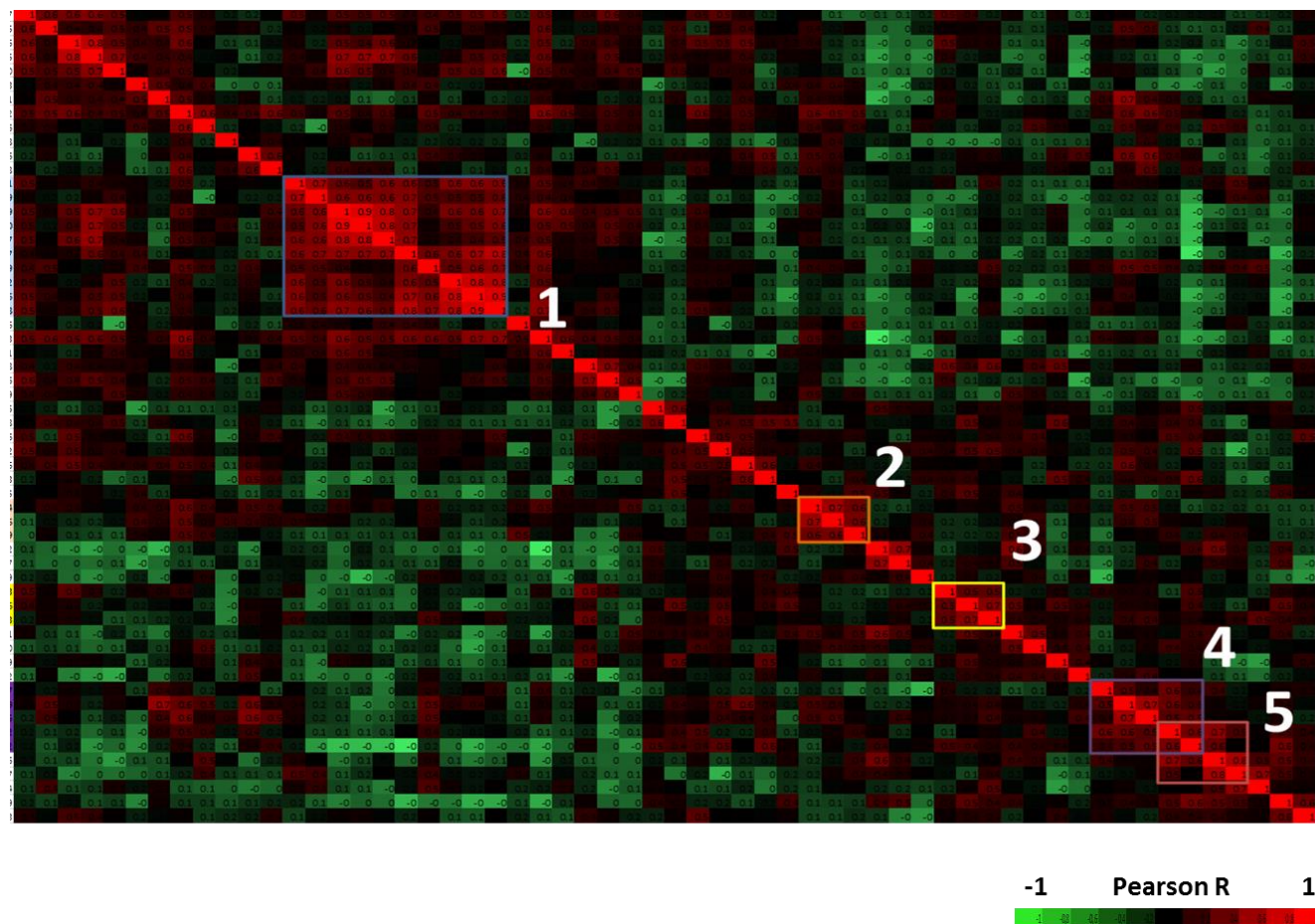


Figure 4.42: Correlation matrix of drugs which are effective on epithelial cancer cells. The compounds which inhibit the growth of epithelial cells listed in rows and columns in the same order. Pearson correlation test was performed between $-\log_{10} \text{GI}_{50}$ of each drug with all others'. Pearson r value of each test is plotted (see methods). Highly correlated sub-clusters are numbered from 1 to 5.

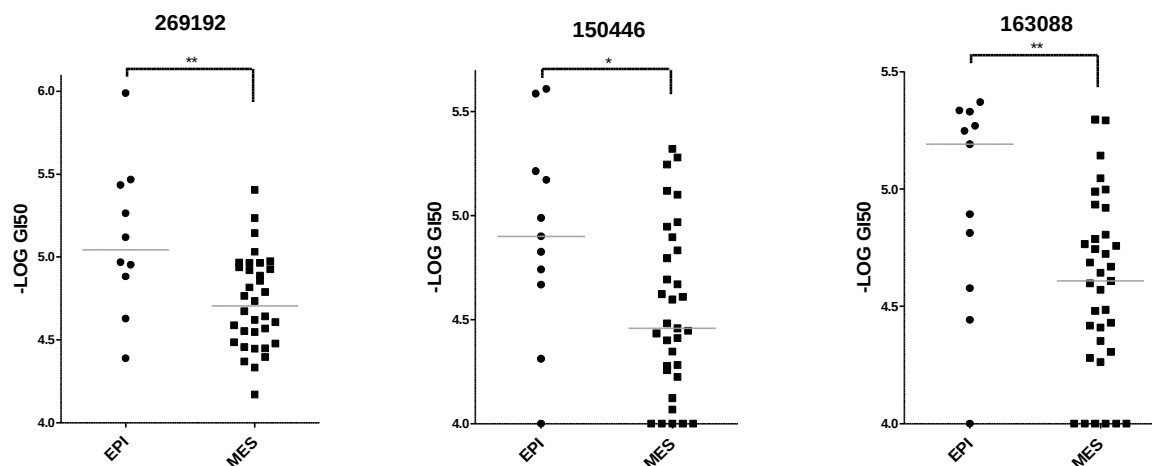


Figure 4.43: Sensitivity of epithelial and mesenchymal NCI60 cell lines to berberine derivatives. NCI's internal ID number of each berberine derivative is shown above the plots. A CDH1-VIM score of “-1” was used as the cut-off for the determination of epithelial (EPI) and mesenchymal (MES) groups. GI50: Growth inhibition of 50%. Medians are shown with horizontal line. Ttest is performed between drug sensitivity of EPI and MES groups. * $p < 0.05$, ** $p < 0.01$.

4.3.3. Identification of novel compounds that are preferentially effective on epithelial or mesenchymal cells, independent of tissue-of-origin in CCLE and CGP

We next aimed to perform a similar analysis using Cancer cell line encyclopedia and Cancer genome project databases [135,152]. The CDH1-VIM based scoring of all cancer cell lines in these databases revealed that different cancer types were preferentially more mesenchymal or epithelial, when compared with others (**Figures 4.44A, 4.45A**). As I was aiming to find drugs that would affect epithelial or mesenchymal cells preferentially, regardless of tissue of origin, I had to minimize tissue type based bias. I therefore, equalized the E and M-like cell ratios for each tumor type by randomly excluding cell lines from either the E or the M group (whichever contained the larger amount of cells) until the median CDH1-VIM score was around -1 (**Figures 4.44B, 4.45B**).

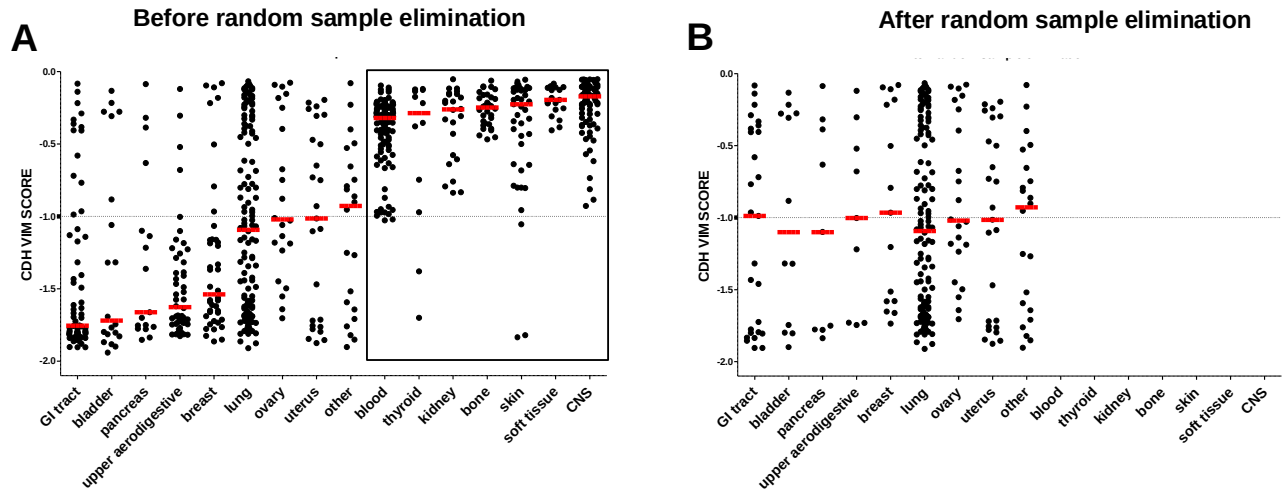


Figure 4.44: Equalization of distribution of epithelial and mesenchymal cells in CGP. Tissue specific distribution of CDH1-VIM score in CGP cancer cell lines with drug sensitivity data (A). Similar ratios of epithelial and mesenchymal cells were obtained by randomly eliminating cells in the predominating group (B) before determining differential drug effects on each cell type. Cell types for which equalization was not possible were eliminated (boxed). Medians are shown with red horizontal line. Black dashed line indicates a CDH1-VIM score of “-1”.

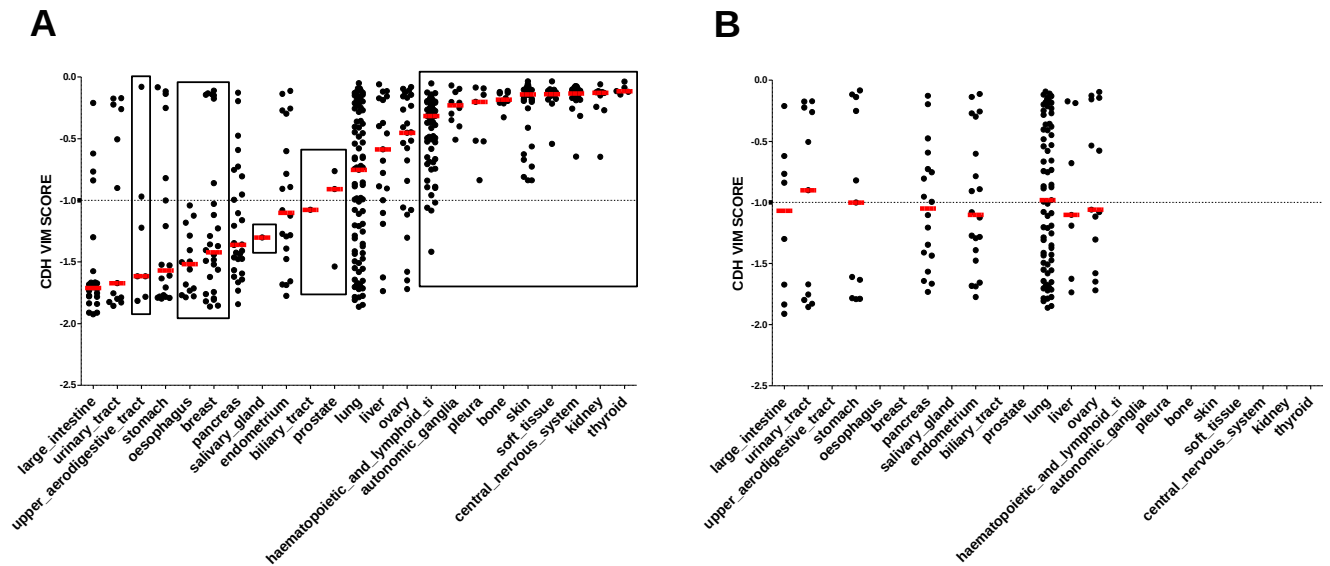


Figure 4.45: Equalization of distribution of epithelial and mesenchymal cells in CCLE. Tissue specific distribution of CDH1-VIM score in CCLE cancer cell lines with drug sensitivity data (A). Similar ratios of epithelial and mesenchymal cells were obtained by randomly eliminating cells in the predominating group (B) before determining differential drug effects on each cell type. Cell types for which equalization was not possible were eliminated (boxed). Medians are shown with red horizontal line. Black dashed line indicates a CDH1-VIM score of “-1”.

I thus generated a list of cell lines for each tumor that included an equal ratio of E and M cells. I then identified drugs which had a sensitivity profile that significantly and linearly correlated with the CDH1-VIM score as well as those that showed a significantly different mean when epithelial and mesenchymal cells were compared in both databases (corrected Pearson and t-test p values < 0.05, respectively). Only 3 drugs fulfilled our criteria, BIBW2992, Lapatinib, Erlotinib, all which are EGFR inhibitors, which were significantly more effective on epithelial cells (**Figure 4.46**).

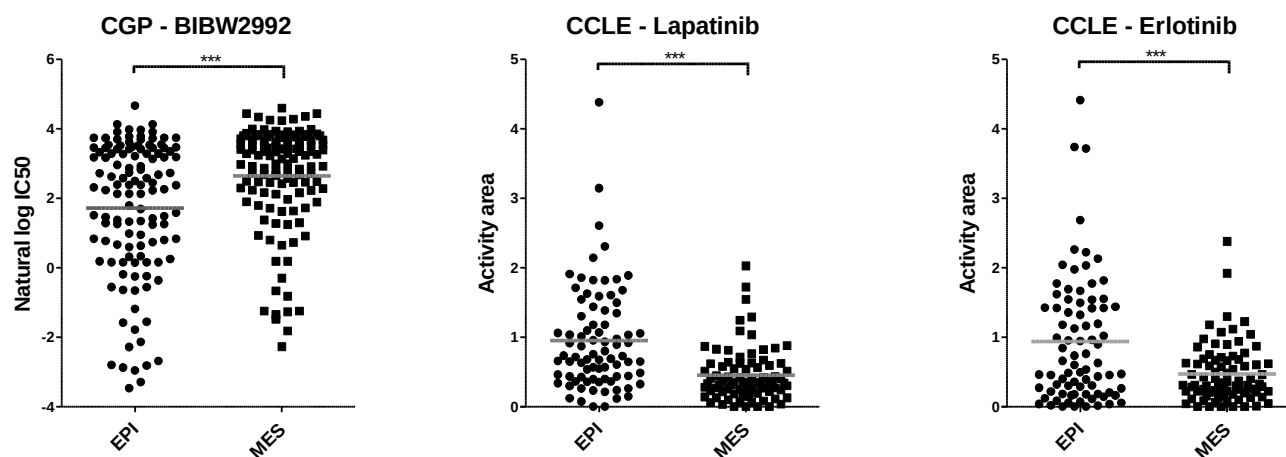


Figure 4.46: Sensitivity of CDH1-VIM based epithelial (EPI) and mesenchymal (MES) CGP and CCLE cell lines to EGFR inhibitors. CDH1-VIM score of -1 was used as the cut-off for the determination of epithelial (EPI) and mesenchymal (MES) groups. Cancer cell lines from all tissue types available after random sample elimination were pooled. The database and the name of the EGFR inhibitor are shown above each plot. Increasing IC50 correlates with resistance and increasing activity area correlates with sensitivity. IC50: Inhibitory concentration of 50%. Activity area: Area above dose response curve. Medians are shown with red horizontal line. . ***p < 0.0001

Therefore, my analysis of the three independent datasets, CCLE, CGP and NCI60, for a correlation between cytotoxicity and EMT score showed that epithelial cells are more sensitive to EGFR inhibitors when all cancer types were pooled. We next asked if the this observation held when each tumor type was analyzed individually. This was also important to perform because in both CCLE and CGP different tissues contributed different amounts of cell lines to the pooled analysis. For example, lung cancer cell lines constituted the majority of cell line pools, and therefore may have caused a bias towards the identification of drugs that are effective on not all epithelial cancer cells but epithelial lung cancer cells only. Strikingly, when I analyzed individual cancer types independently, I found that in each tumor type epithelial cancer cells were more sensitive to EGFR inhibitors in both CCLE and CGP databases (**Figures 4.47 and 4.48**).

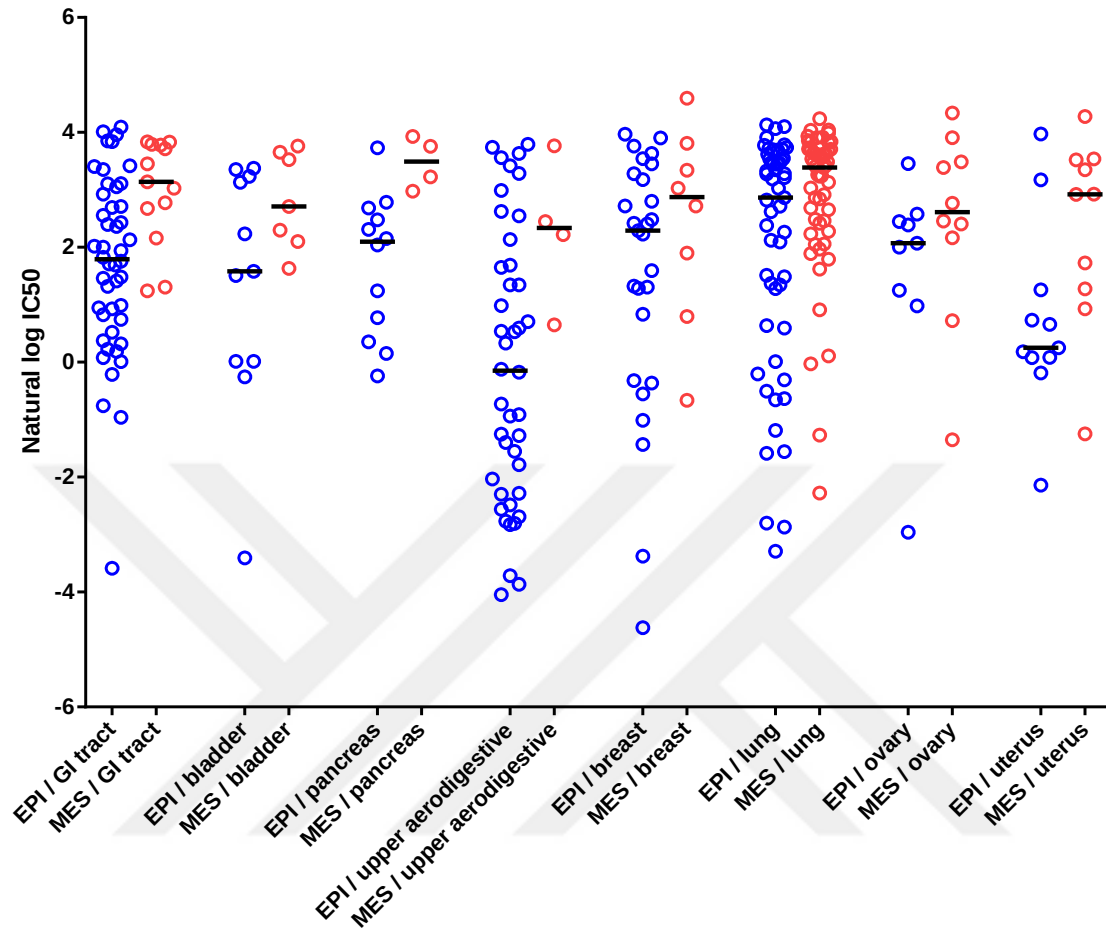


Figure 4.47: Tissue specific sensitivity of epithelial and mesenchymal CGP cell lines to BIBW2992, EGFR inhibitor. CDH1-VIM score -1 was used as the cut-off for the determination of epithelial (EPI) and mesenchymal (MES) groups. IC50: Inhibitory concentration of 50%. Medians are shown with a horizontal line.

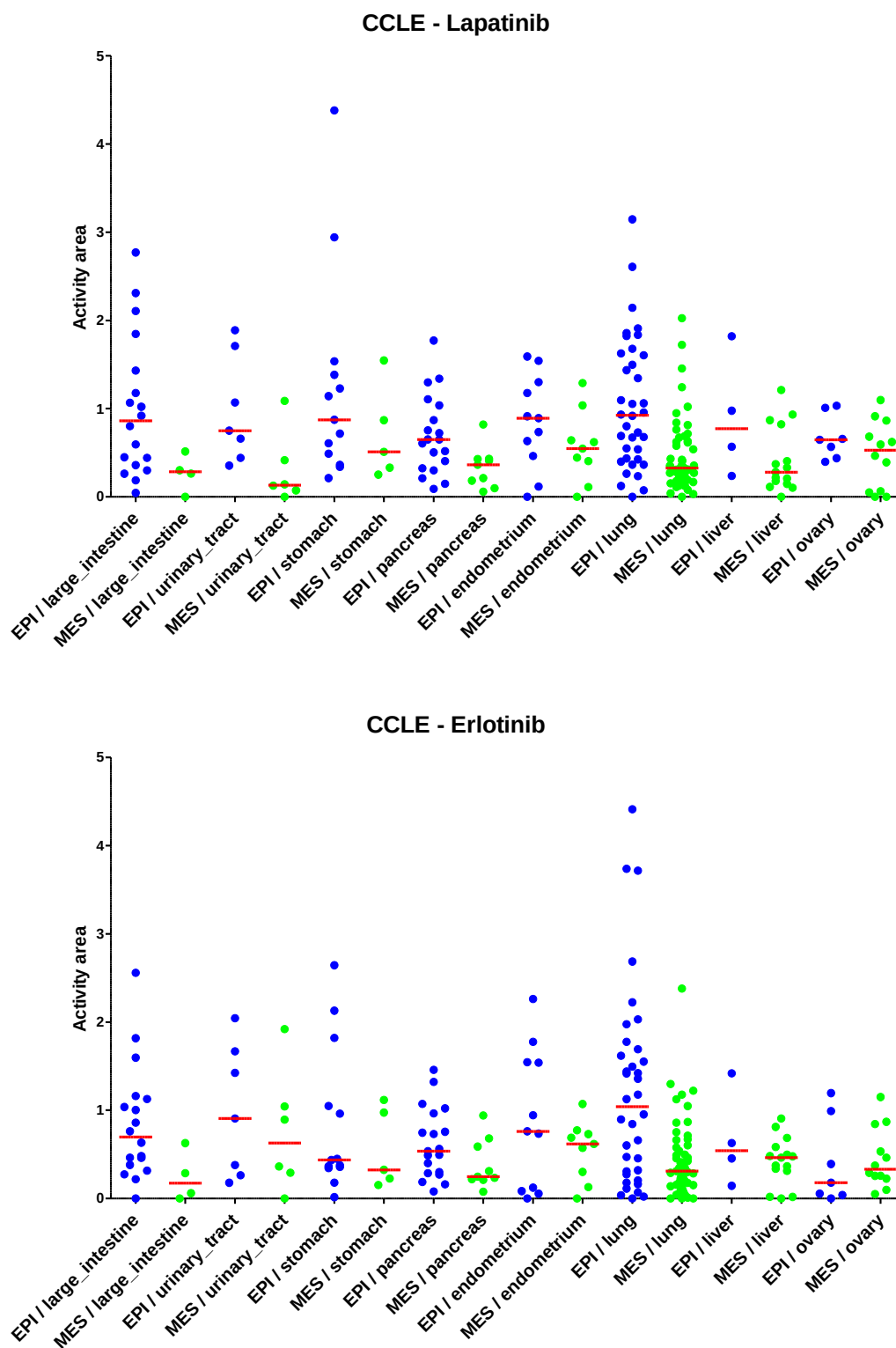


Figure 4.48: Tissue-specific sensitivity of epithelial and mesenchymal CCLE cell lines to Lapatinib and Erlotinib, EGFR inhibitors. CDH1-VIM score -1 was used as the cut-off for the determination of epithelial (EPI) and mesenchymal (MES) groups. Higher activity area indicates higher sensitivity. Medians are shown with a horizontal line.

4.3.4. Identification of novel and tissue-independent parameters in addition to EMT, which will improve the prediction of sensitivity to EGFR

Although I observed that epithelial cells are statistically more sensitive to EGFR inhibitors, the scatter plots demonstrated a wide range and overlapping sensitivity values for both E and M groups both in tumor specific and pooled analyses (**Figures 4.47 and 4.48**). This indicates that there should be other factors involved in determining sensitivity to EGFR inhibitors other than EMT. In this section, we aimed to identify additional novel and tissue-independent biomarkers which could be used for prediction of sensitivity to EGFR inhibitors. In order to do that we have analyzed specific parameters as listed in **Table 4.26**.

Table 4.26: List of *in silico* analyses performed

Analysis/Data type	Dataset
Global biological differences	CGP
Mutational profile	CGP
EGFR and EGFR ligand mRNA expression	NCI60
Novel coding and non-coding transcripts	CGP + E-MTAB2706
Novel proteins	TCPA

4.3.4.1. Global biological differences

I first asked if I could identify an obvious biological difference between EGFR inhibitor sensitive and resistant groups. For this purpose we performed GSEA analysis between top 1/3 EGFR inhibitor resistant and sensitive cells in 4 tissue types with the biggest sample size (breast, colon, GI tract, aerodigestive) separately in CGP. However this analysis neither identified highly significant biological differences for a given tumor type, nor a common biological process among different tumors that were analyzed. The result of GSEA in breast cancer is shown in **Table 4.27** as a representative. As can be seen, the FDR q values are quite high indicating an absence of clear biological differences between EGFR sensitive and resistant cells.

Table 4.27: GSEA analysis between cells lines which are sensitive and resistant to EGFR inhibitor, BIBW2992

Gene Ontology

NAME	ES	NES	NOM p-val	FDR q-val
INORGANIC_ANION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.621	1.891	0.000	0.135
GENERATION_OF_NEURONS	0.466	1.545	0.000	0.599
SKELETAL_DEVELOPMENT	0.489	1.548	0.000	0.625
PERINUCLEAR_REGION_OF_CYTOPLASM	0.457	1.554	0.024	0.640
INTRAMOLECULAR_OXIDOREDUCTASE_ACTIVITY	0.618	1.559	0.043	0.660
GLYCOSAMINOGLYCAN_BINDING	0.635	1.568	0.008	0.662
REGULATION_OF_HEART_CONTRACTION	0.510	1.527	0.038	0.662
AMINO_ACID_CATABOLIC_PROCESS	0.626	1.574	0.027	0.683
REGULATION_OF_CELL_DIFFERENTIATION	0.470	1.428	0.040	0.757

Oncologic Signatures

NAME	ES	NES	NOM p-val	FDR q-val
KRAS.KIDNEY_UP.V1_DN	-0.291	-1.094	0.298	0.741
HOXA9_DN.V1_DN	-0.331	-1.086	0.257	0.746
PTEN_DN.V2_DN	-0.327	-1.079	0.246	0.754
JNK_DN.V1_DN	-0.299	-1.147	0.180	0.756
ATM_DN.V1_UP	-0.285	-1.023	0.435	0.756
RELA_DN.V1_DN	-0.304	-1.095	0.275	0.765
P53_DN.V1_DN	-0.367	-1.099	0.247	0.775
PRC1_BMI_UP.V1_DN	-0.271	-1.023	0.392	0.775
GCNP_SHH_UP_EARLY.V1_UP	-0.256	-1.009	0.447	0.775
STK33_SKM_DN	-0.314	-1.055	0.328	0.780

4.3.4.2. Mutational profile

We next asked if mutational profiles are involved in sensitivity to EGFR inhibitors for E or M cells. For this purpose we analyzed CGP mutational data of cancer cell lines regardless of the tissue type. 208 and 342 cell lines which are epithelial and mesenchymal according to CDH1-VIM scoring in CGP database with available microarray expression, mutational profiles and sensitivity data for BIBW2992 were used. This revealed no significant associations between mutations in any gene analyzed and sensitivity to BIBW2992 among E cells. However, among M cells, various mutations in *PIK3CA* (p.E545K, p.H1047R, p.H1047L, p.D549N, p.P449T, p.E453K, p.P539R, p.Q546R, p.R88Q, p.K111E, p.P124L, p.A1066V, p.N345S, p.G1049R, p.E542K, p.K111N, p.G118D), and nonsense/ frame shift mutations (p.Y628fs*3, p.S2751fs*27, p.D2635fs*9, p.Q1336*, p.C167*, p.E1720*) and deletions in *NF1* occurred more frequently among BIBW2992 sensitive cells, whereas *NRAS* mutations (p.Q61R, p.Q61K, p.Q61L, p.G12C, p.G12D, p.G12S, p.G13D) were associated with drug resistance. I found no significant association between mutational status of *EGFR* and other proteins with sensitivity to BIBW2992 (**Figure 4.49**).

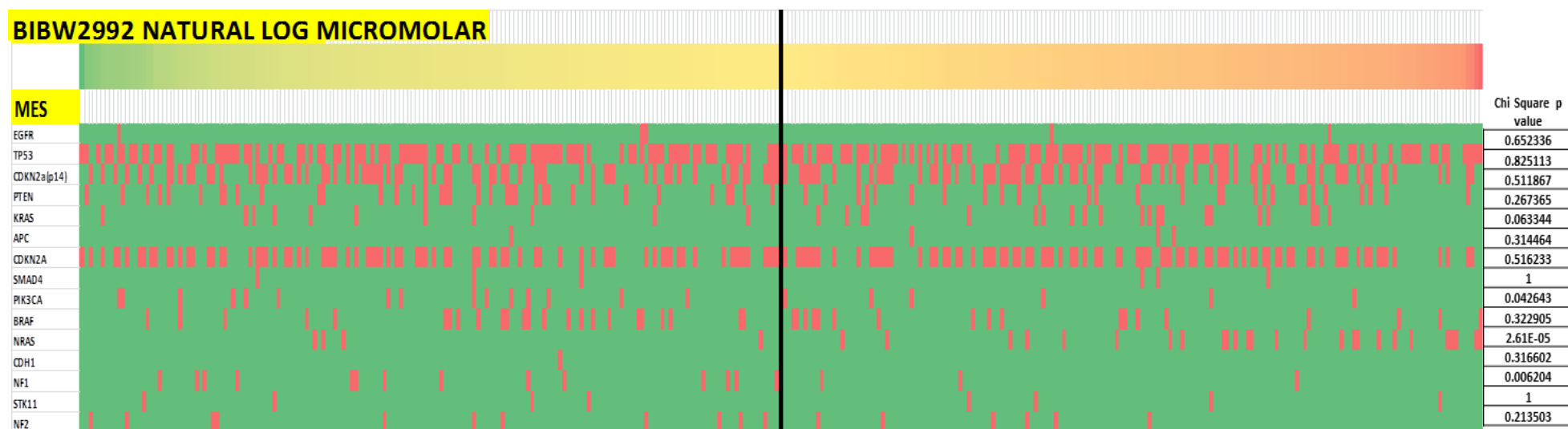
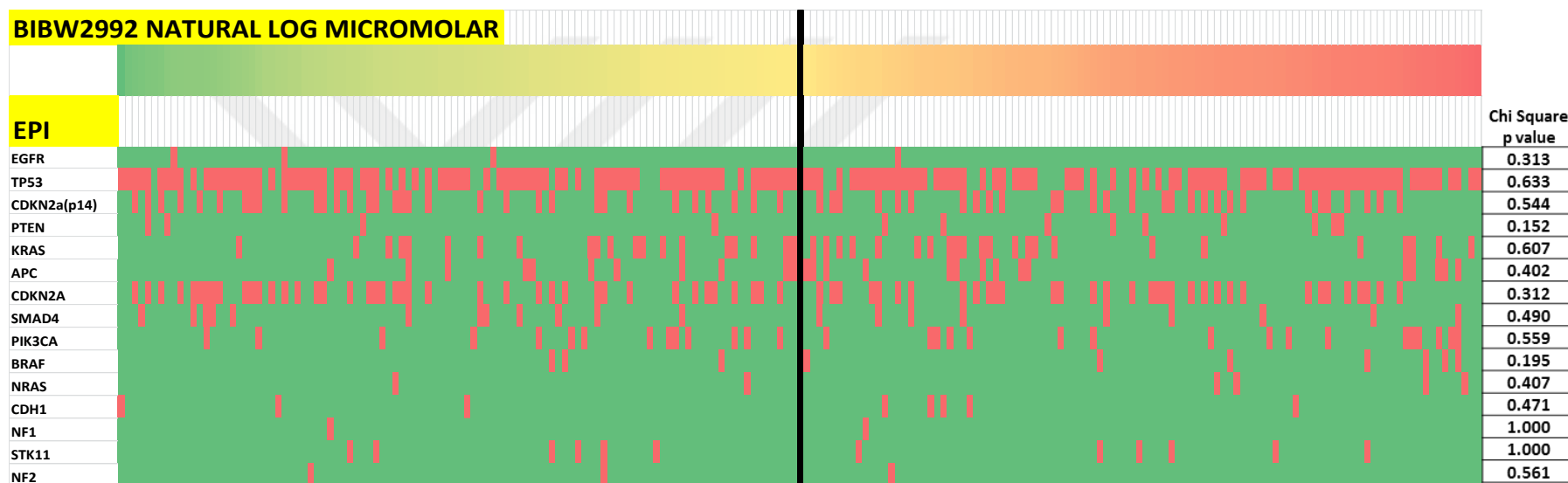


Figure 4.49: Analysis of mutational profile for associations with sensitivity to EGFR inhibitors. 84 gene mutations and fusions were evaluated. Among the genes listed on the left, EGFR is shown at the top since BIBW2992 is an EGFR inhibitor. The remaining 14 genes were the top mutations in terms of incidence (sorted from high to low). IC50 values of BIBW2992 (natural log micromolar) are sorted from smallest (green) to largest (red). Mutations are shown with red and wild types are shown with green rectangle. Chi square p values on the right show significance of the relationship between having a gene mutation and being sensitive or resistant to BIBW2992 (sensitivity threshold is shown with black horizontal line)

4.3.4.3. EGFR and EGFR ligand mRNA expression

Using NCI60's cancer cell lines as a pool, I next asked if EGFR or its ligands correlated with EGFR sensitivity. I found that EGFR mRNA levels are not statistically different in EGFR sensitive and resistant EPI and MES cells and thus EGFR mRNA levels cannot explain differential EGFR inhibitor sensitivity (**Figure 4.50**); despite the fact that this has been related to EGFR TKI sensitivity in cancer specific studies of lung cancer and endometrial cancer [162,163]. In the same way, we found that mRNA levels of EGFR ligands are not statistically different in EGFR inhibitor sensitive and resistant EPI and MES cells (**Figure 4.51**); in contrast to the differential response to EGFR targeted treatments observed in groups of tumors or cells with different EGFR ligand expression: Breast cancer cells with downregulation of HER-specific ligands exhibit intrinsic sensitivity to gefitinib [164] and *KRAS* WT colorectal cancer patients with low EGFR ligand expression were unresponsive to cetuximab [165]. Therefore we can conclude that the mRNA level expression of EGFR and its ligands were not tissue-type independent parameters in determination of sensitivity to EGFR inhibitors, whereas they could be one of the major determinants when individual cancer types were evaluated.

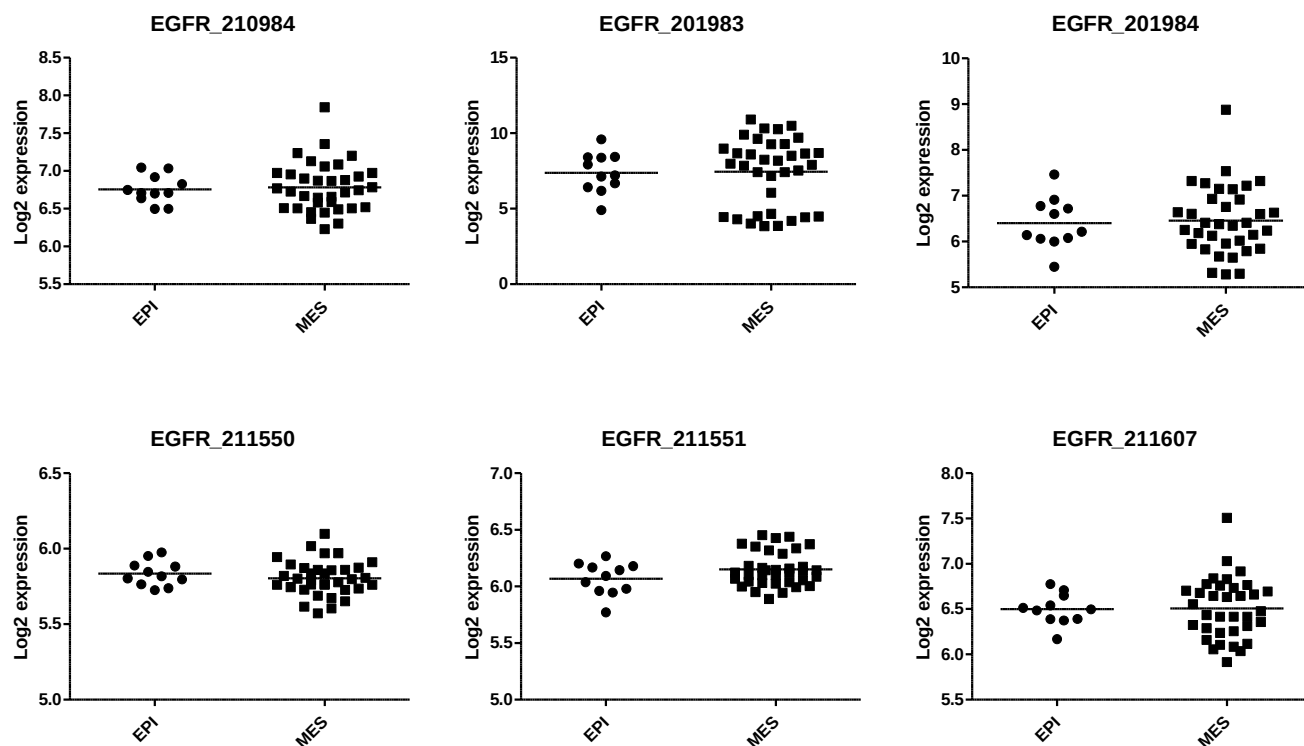


Figure 4.50: Microarray based expression level of EGFR in NCI60 cancer cell lines. Expression of EGFR probesets in 46 NCI60 cancer cell lines (leukemia and ovarian cells excluded) were plotted for epithelial and mesenchymal cells defined as CDH1-VIM score below and above “-1”, respectively. Horizontal lines show mean expression. Ttest p values for all probesets were above 0.05.

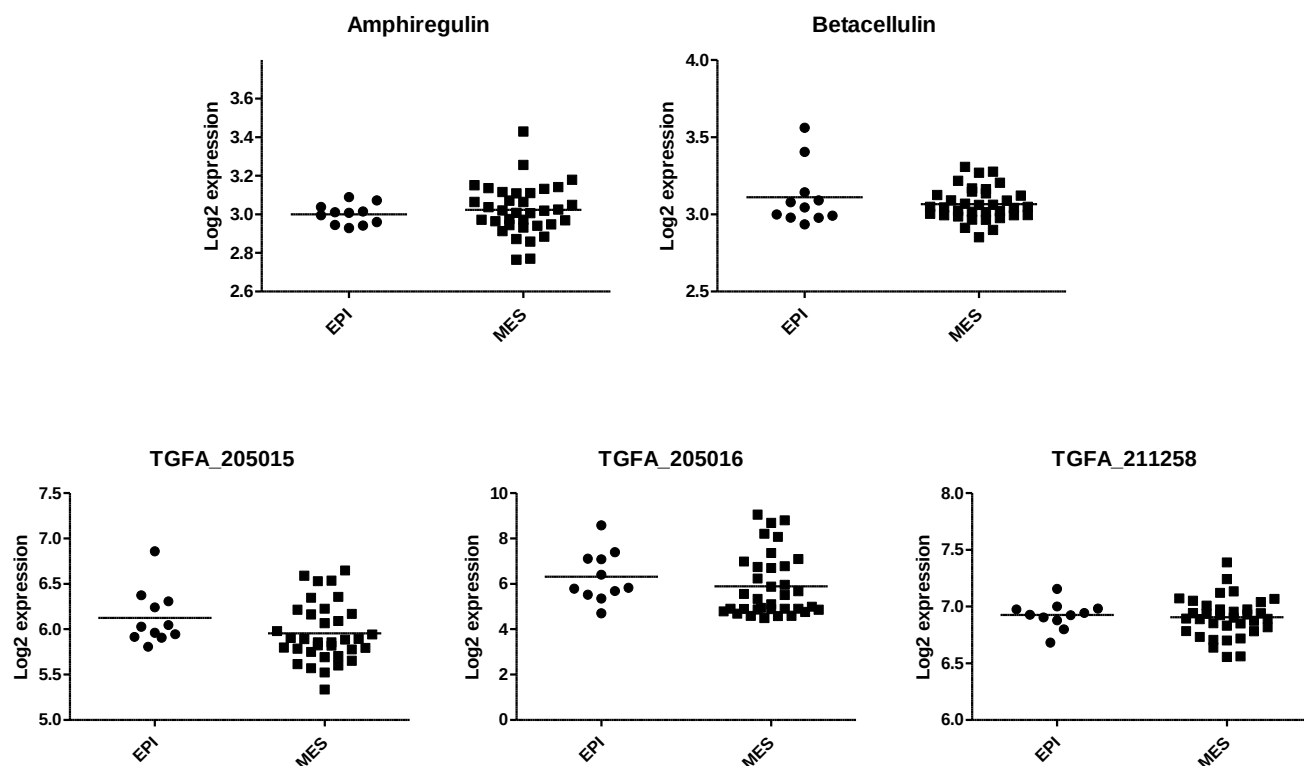


Figure 4.51: Microarray based expression level of EGFR ligands in NCI60 cancer cell lines. Expression of EGFR ligands' probesets in 46 NCI60 cancer cell lines (leukemia and ovarian cells excluded) were plotted for epithelial and mesenchymal cells defined as CDH1-VIM score below and above "-1", respectively. Horizontal lines show mean expression. Ttest p values for all probesets were above 0.05.

4.3.4.4. Novel coding and non-coding transcripts

Next we analyzed the microarray data of CGP cell lines in order to select the tissue-independent transcripts that linearly correlate with sensitivity to EGFR inhibitors. The transcripts with significant Pearson correlation ($r > 0.30$ and $p < 0.05$) within each cancer type are shown in the Venn diagram (**Figure 4.52**). As seen there are no common transcripts among 4 cancer types with the biggest sample size, and none of those transcripts in the diagram passed Benjamini Hochberg correction. We obtained similar results when we repeated the analysis for noncoding RNAs (**Figure 4.53**). Therefore we concluded that there is no tissue-independent coding and non-coding transcripts correlated with sensitivity to EGFR inhibitors in a tissue-independent fashion.

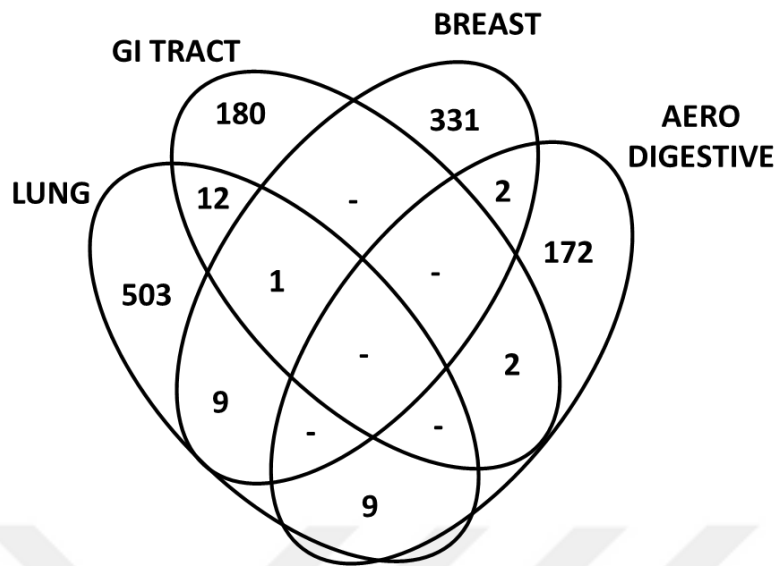


Figure 4.52: Venn diagram of probe sets which significantly linearly correlated with cytotoxicity of BIBW2992. Microarray based gene expression and BIBW2992 IC50 values of 4 cancer types with the biggest samples size, breast, lung, GI tract and aerodigestive were used. Transcripts with pearson r value of at least 0.3 and p value below 0.05 were shown in the diagram. Overlapping probe sets which are correlated in reverse directions were eliminated. None of these transcripts in the diagram passed Benjamini Hochberg correction.

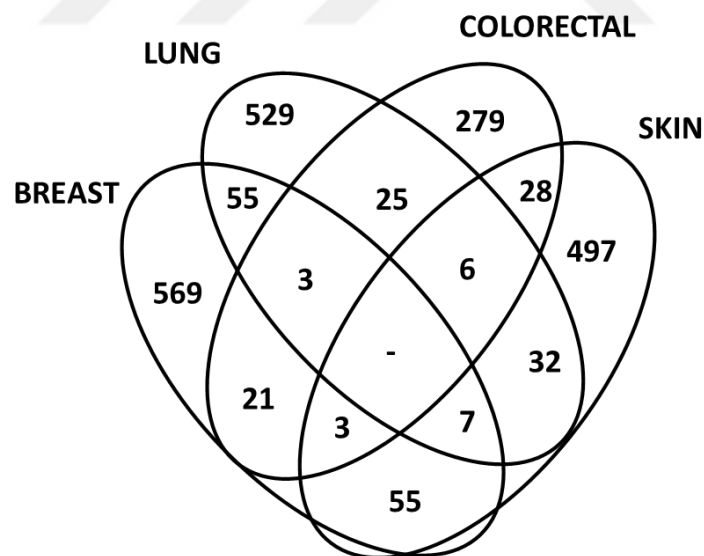


Figure 4.53: Venn diagram of non-coding transcripts which significantly linearly correlated with cytotoxicity of BIBW2992. RNAseq based gene expression of cancer cell lines in E-MTAB-2706 and their BIBW2992 IC50 values in CGP of including cancer types with the biggest samples size, breast, lung, colorectal and skin were used. Transcripts with pearson p value below 0.05 were shown in the diagram. None of these transcripts in lung and colorectal cancer and 6 and 173 transcripts in breast and skin cancers passed Benjamini Hochberg correction, respectively.

4.3.4.5. Novel proteins

Next we aimed to identify tissue independent markers for sensitivity to EGFR inhibitors at the protein-level. Cancer cell line RPPA (Reverse phase protein array) level 3 data were downloaded from TCGA database (see Methods). In order to perform linear correlation analysis, we tried to select the cancer types which have the highest number of cell lines with both the EGFR sensitivity data in CCLE or CGP database and TCGA proteomic data. 3 sets of protein expression data, S35, S40 and S51 were available and we used cancer types with sample size above 15; endometrial cancer (n=17) and breast cancer (n=16). As opposed to previous approaches, *in silico* analysis of RPPA data did reveal two proteins, cRAF and p38 MAPK, which significantly and linearly correlated with sensitivity to EGFR inhibitors in both cancer types. Decreasing cRAF protein level expression and increasing p38 MAPK phosphorylation were correlated with sensitivity to EGFR inhibitors in both breast and endometrium cancer cell lines (**Figure 4.54**).

It has been shown previously in breast epithelial cells that p38 MAPK enhances activation of EGFR and downstream signaling and thus leads to enhanced cell proliferation [166]. However this is the first data showing correlation of increasing p38 MAPK phosphorylation with sensitivity to EGFR inhibitors in two different tissues. Amplifications of cRAF are known to promote EGFR-TKI resistance [167]; and we found that increased cRAF protein expression also relates to resistance to EGFR inhibitors.

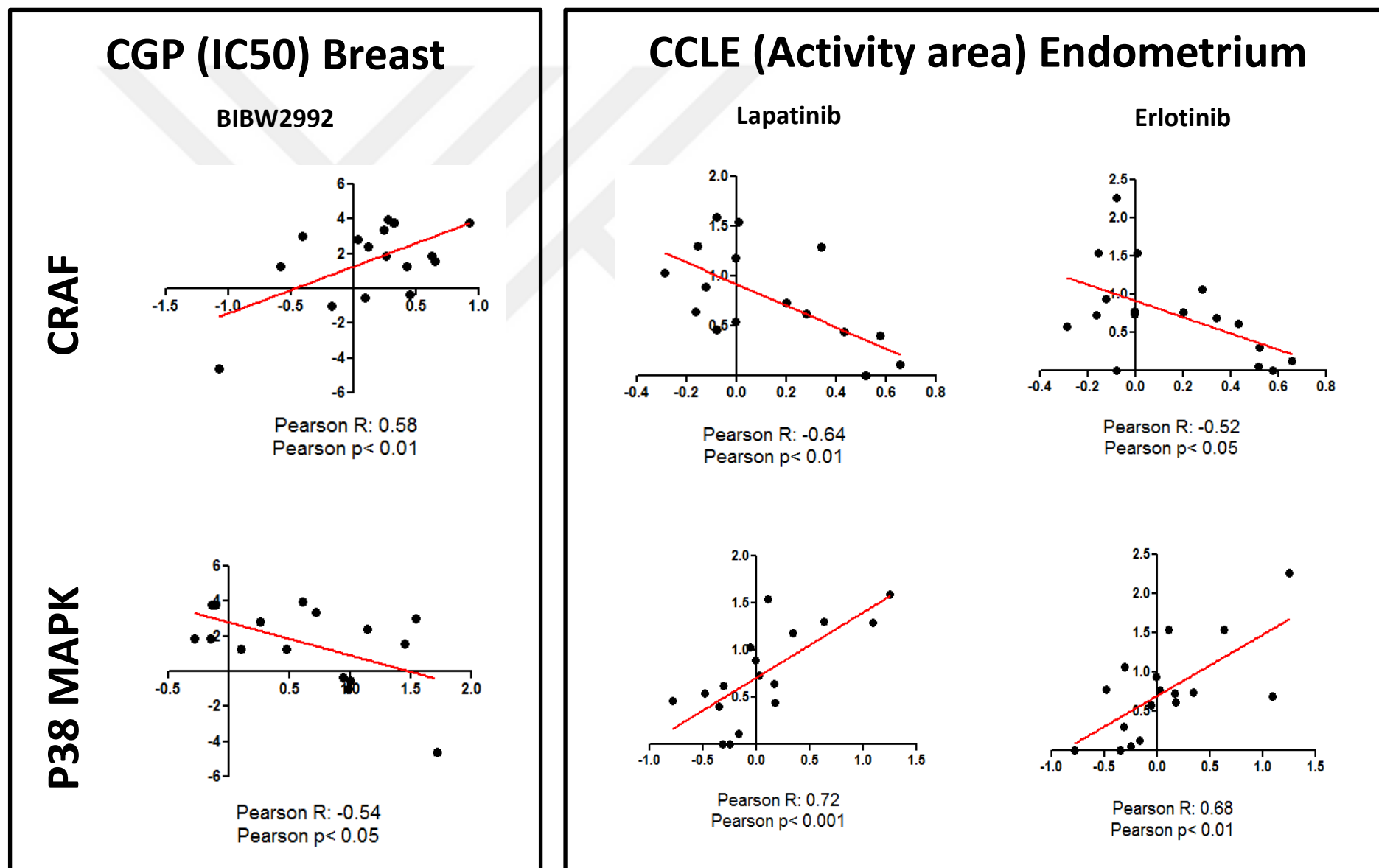


Figure 4.54: Correlation of cRAF protein level expression and EGFR inhibitors in CCLE and CGP databases. The cells which had protein level expression data and sensitivity data for BIBW2992, lapatinib and erlotinib in CGP and CCLE databases were used. X axis: protein expression/phosphorylation from TCPA S35 for CGP, and S51 for CCLE analysis; Y axis: logIC50 or activity area values for EGFR inhibitors. Pearson r and p values are shown.

We then generated a model using multiple linear regression analysis with cRAF and p38 MAPK protein-level expression values that could best explain BIBW2992 sensitivity. The regression value generated for each cell line could predict sensitivity to BIBW2992 better than any of the proteins used by themselves (**Figure 4.55**).

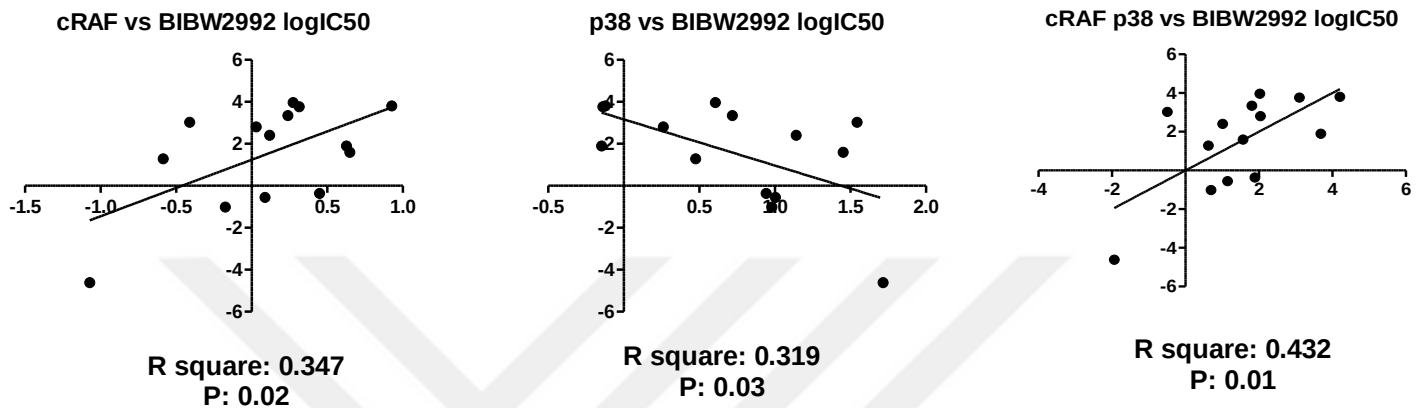


Figure 4.55: Linear regression analysis of cRAF and p38 MAPK for prediction of sensitivity to BIBW2992. The cells which have protein level expression data and sensitivity data for BIBW2992, in S35 dataset of TCGA breast cancer cells were included. X axis: Craf-p38 regression score; Y axis: Log IC50 of BIBW2992. Pearson r and p values are shown

To determine if the combined analysis of cRAF/p38 MAPK could explain sensitivity to other EGFR inhibitors as well, I generated a combined score with equal weight of both cRAF and p38 MAPK proteins. Cell lines were then classified into 4 groups according to their c-raf/p38 score and EMT score, using a single threshold for both parameters. EGFR inhibitor sensitivities of cells from those 4 groups in 2 datasets of TCGA for 2 different EGFR inhibitors, Lapatinib and Erlotinib are shown in (**Figure 4.56**). As we have shown before, overall, epithelial cells are more sensitive to EGFR inhibitors. Also cRAF and p38 MAPK expression can also divide epithelial and to a lesser extent mesenchymal groups further into EGFR inhibitor sensitive and resistant groups. This data is generated by using endometrial cancer cells, so we performed the same analysis with breast cancer cell lines.

For breast cancer, both EPI as well as MES cells could be further separated into sensitive and resistant cell line groups when c-raf/p38 scores were used in addition to the EMT classification (**Figure 4.57**). This was somewhat more obvious for the EPI group, but was observable with the MES group as well. We also wondered if EGFR inhibitor sensitivity of breast cancer cells are related to their molecular subtypes and labeled each cell according to the molecular subtype. In this rather small scale analysis, basal B cells were more resistant to EGFR inhibitors, as expected, as these cells were primarily of the MES type, and one which was of an EPI phenotype, was c-raf/p38 high (+), which predicts resistance to

EGFR inhibitors. Therefore, even though the samples used for this analysis is small, I predict the EMT and c-raf/p38 based scoring will possibly have the most clinical utility for Luminal and Basal A type cells.

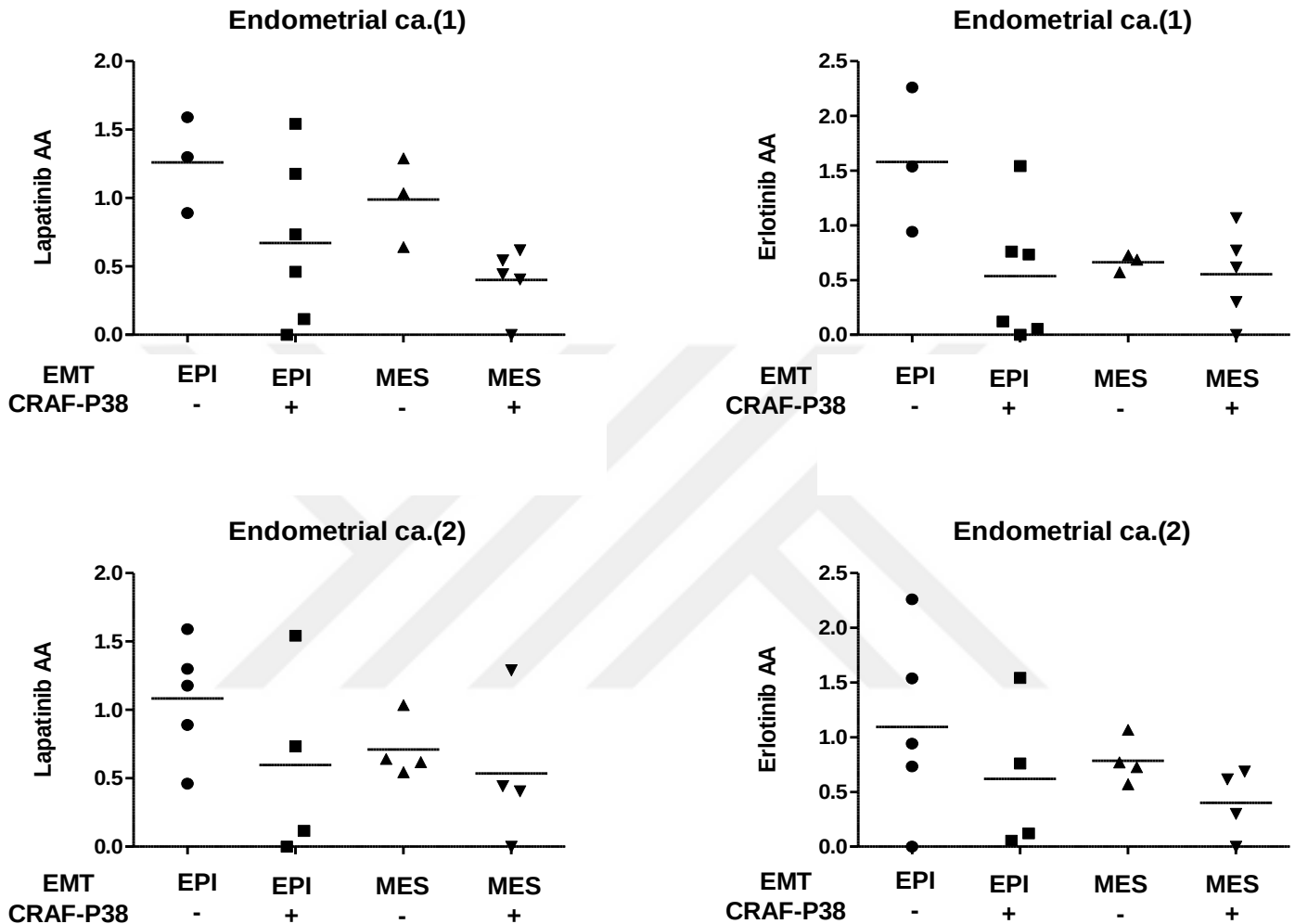


Figure 4.56: Sensitivity to EGFR inhibitors in 4 groups of endometrial cancer cells divided by EMT and cRAF-p38. Activity area of endometrial cancer cells was plotted for lapatinib and erlotinib. Crf-p38 score is calculated using protein level expression by subtracting p38 from cRAF. Samples with cRAF-p38 score above mean are shown with '+', below mean are shown with '-'. Endometrial ca.(1) and Endometrial ca.(2) are TCPA dataset S40 and S51, respectively. EMT grouping is determined by CDH1-VIM score above -1 (MES) and below -1 (EPI). Lapatinib and erlotinib cytotoxicity and microarray expression of cell lines were obtained from CCLE. Kruskal Wallis analyses failed to identify a difference which was statistically significant between groups.

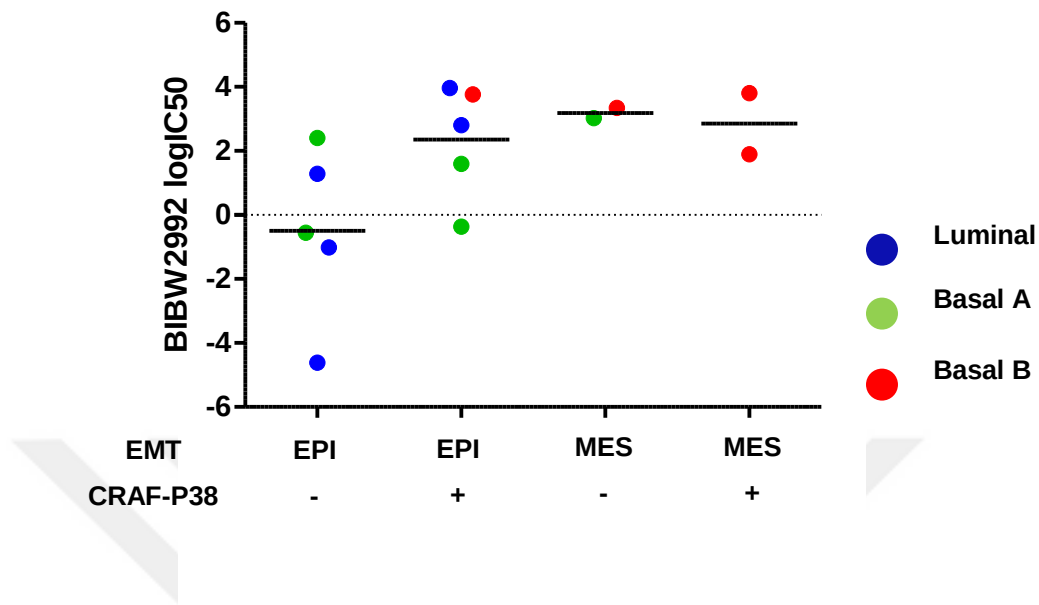


Figure 4.57: Sensitivity to EGFR inhibitors in 4 groups of breast cancer cells divided by EMT and cRAF-p38. LogIC50 of breast cancer cells plotted for BIBW2992. Crf-p38 score was calculated using TCPA protein level expression by subtracting p38 from cRAF. Samples with cRAF-p38 score above mean are shown with '+', below mean are shown with "-". EMT grouping is determined by CDH1-VIM score above -1 (MES) and below -1 (EPI). Lapatinib and erlotinib cytotoxicity and microarray expression of cell lines were obtained from CCLE.

CHAPTER 5

5.DISCUSSION AND CONCLUSION

5.1. Prognostic SU-GIB signature identified in colon cancer

In recent years, a large number of molecular tests designed for the assessment of colon cancer prognosis have been reported. Among these tests, very few have proceeded beyond initial discovery studies and involved in large clinical validations. The reasons underlying this fact were discussed in a comprehensive review of US Department of Health and Human Services which reviewed molecular diagnostic and prognostic tests in many cancer types [45]. They reported that the involvement of multiple patient cohorts was essential. We therefore, used multiple *in silico* cohorts, as well as an *ex vivo* validation cohort. In addition they strongly suggested the usage of MVAs with all confounding factors included and with multiple parameters of disease outcome: DFS, CSS and OS, as we performed for SU-GIB.

We discovered two novel mRNA level prognostic biomarkers, ULBP2 and SEMA5A, *in silico* utilizing multiple independent cohorts and survival outcomes, and validated these results *ex vivo* in fresh frozen tumor tissues. Further bioinformatic analyses showed that the patients with different prognostic profiles also show biological differences, regarding EMT, inflammation, and EGFR activity (**Table 5.1**). Therefore our next goal is to develop the SU-GIB classification as a molecular prognostic test to be used in clinics.

When the strength of SU-GIB test, which is comparable to an *in silico* adapted version of Oncotype Dx, is considered, we think that SU-GIB can be further developed into a clinical prognostic test. In order to achieve that, large scale retrospective and prospective validation studies are required. Within the scope of these studies, the optimal group of patients in which SU-GIB can define the best prognostic sub-groups should also be determined as has been done for other tests [168,169]. ASCO guidelines recommend adjuvant chemotherapy for stage II patients with T4 lesions, perforation, incompletely evaluated nodes and poor differentiation [13]. The clinical decision making route published by ASCO indicates that observation should be recommended for patients with low risk features and MSI tumors [170]. However assessment of risk in T3 and MSS tumors is unclear and essential and Oncotype Dx colon was shown to successfully predict risk groups and was offered to stage II patients with T3 and MSS tumors [171]. Our data showed that SU-GIB was superior to Oncotype Dx risk stratification in patients

with MSS tumors. Therefore, larger scale validation studies are justified for evaluation of prognostic risk groups defined by SU-GIB especially for patients with stage II, T3 and MSS disease.

Furthermore, our data demonstrated that SU-GIB is associated significantly with recurrence-free survival in distal tumors, but not in proximal tumors. Therefore tumor location seems to be an important parameter that would need to be included as a parameter in future validation studies. In this line, further biomarker discovery studies are required for discovery of a specific gene-set to evaluate prognosis for patients with proximal tumors.

Another critical factor for a molecular test to be used in clinics is how it changes treatment decisions [45]. This was assessed for the Oncotype Dx Colon © test resulting in 45% change in treatment decision for T3 MMR-P stage II patients [172,173]. Therefore the alterations in treatment decision with the usage of SU-GIB test should be evaluated in a comparative study with and without SU-GIB test results provided to clinical oncologists.

In MVA analyses, among confounding factors, microsatellite instability and Oncotype Dx scoring were evaluated by using published *in silico* approximations (except for GSE39582 which included original MSI status data). Therefore, for validation studies, the actual testing of Oncotype Dx Colon© and microsatellite status determination should be performed to be included in further studies. One cohort I studied (GSE39582) had clinical data involving KRAS and BRAF mutation and MSI status information. MVAs including these factors, showed very clearly that SU-GIB is an independent prognostic indicator.

SU-GIB can become an in-house alternative to Oncotype Dx Colon©, since MVA with Oncotype Dx Colon © *in silico* approximations showed that this test and SU-GIB identify overlapping patient groups, however SU-GIB has more statistical power. Another issue that needs to be resolved for effective clinical use of SU-GIB, is that it should be applicable to FFPE (formalin-fixed paraffin-embedded) tissue, which is used routinely in pathology. Therefore further studies should aim for applying SU-GIB to FFPE tissue.

The official price of Oncotype Dx in Turkey was fixed to 12,270 TL due to fluctuations in \$/TL ratio as we learned from the distributor, Genekor. The price also includes the transfer of the paraffin embedded tissue to USA. Upon our interviews with clinical oncologists, we learned that the usage of Oncotype Dx is quite rare due to its price which is not covered by government or insurance companies. Since Oncotype Dx includes 7 prognostic genes and 5 reference genes; and SU-GIB has only 2 prognostic genes with additional 2 or 3 reference genes, the end product of SU-GIB test will be cheaper. Apparently, the patient specimen will stay in Turkey for testing, reducing the transportation expenses. Altogether, these show that the SU-GIB test is a candidate standard test for the critical risk assessment of stage II colon cancer in the near future.

To my knowledge, this is the first report to associate SEMA5A and ULBP2 expression with clinical outcome in colon cancer. Semaphorins are involved in various and even contrasting cellular processes like tumor suppression, cell proliferation, apoptosis, oxidative stress, invasion, migration and evasion of tumor surveillance [50]. High expression of SEMA5A has been associated with favorable prognosis in NSCLC in non-smoking women [54], consistent with our findings. The reasons for this, has not been clarified in the literature, since studies indicated that some semaphorins promote tumorigenesis via contributing to migration and invasion whereas others block this process. The association of SEMA5A with a better prognostic profile could maybe rely on its possible role in the latter group.

SEMA5A can be cleaved from the cell surface by ADAM-17 and secreted to peripheral blood [174]. It can induce proinflammatory response in rheumatoid arthritis (RA) patients via activating T and NK cells [174]. The correlation of proinflammatory phenotype and expression or secretion of SEMA5A has not been studied in cancer cells to my knowledge. However my data showed that tumors with a proinflammatory gene expression profile transcribe relatively less SEMA5A in colon cancer tumors. The possible involvement of specifically activated T and NK cells in clinical outcome of colon cancer should be studied in fresh frozen tumors of colon.

ULBP2 protein has been well-studied for its involvement in natural killer cell mediated cytotoxicity [61]. Once cleaved from the cell surface, they block NKG2D receptors on natural killer cells, causing attenuation in natural cell mediated cancer cell death [68]. Consistent with this; ULBP2 protein expression and serum level was found as a poor prognostic marker in various cancer types ovarian cancer, NSCLC, melanoma and leukemia [69-71,73,74]. We have shown that high mRNA levels of ULBP2 have been associated with shorter survival *in silico* and *ex vivo*. In the light of these data, I think that the worse prognostic association of high ULBP2 expression in colon cancer patients can be most probably related to its role in blocking NK-cell mediated lysis and thus contribution to tumor immunosurveillance. Therefore future studies should include investigation of whether the plasma level of ULBP2 is associated with its expression in tumors and more importantly, the clinical outcome.

Further analyzing molecular features of the SU-GIB prognostic groups can improve understanding of how the expression of these genes may affect colon cancer prognosis. Major findings regarding this matter are summarized in **Table 5.1**. Our *in silico* analysis based on TCGA RNAseq and proteomic data revealed that patients with a better outcome have increased phosphorylation of EGFR at Tyrosine 1068 which is necessary for Grb2-EGFR binding and is strongly inhibited by Gefitinib [175,176]. It is surprising since we also found that patients with a 'Bad' SU-GIB signature are sensitive to a PI3K-mTOR inhibition, which plays a role in cell growth downstream of EGFR. In addition, mutational analysis of GSE39582 dataset showed that only 20% of SU-G patients contain *KRAS* or *BRAF* mutations, whereas tumors from B patients are mutated at a rate of 60%. These data suggest that patients with EGFR T1068

phosphorylation may not have *KRAS/BRAF* mutations and that NVP-BEZ235 may be more effective on tumors with *KRAS/BRAF* mutations, rather than those with EGFR autophosphorylation. An alternative scenario could be that SU-G patients have a basal level of activity in both arms of EGFR pathway triggered by phosphorylation of EGFR T1068, PI3K-mTOR and *KRAS-BRAF*, so that the NVP-BEZ235 mediated growth inhibition is negligible when compared to SU-Bs. Whereas in SU-B tumors, constitutive *KRAS/BRAF* mutations are quite common so that EGFR mediated signal induction is not needed, and PI3K-mTOR pathway is highly active so that NVP-BEZ235 mediated growth inhibition is effective.

Table 5.1: Characteristics associated with 'GOOD' and 'BAD' prognostic groups of SU-GIB

	ULBP2	SEMA5A	EMT	EGFR phosphorylation	<i>KRAS/BRAF</i> mutation	Resistance to PI3K/mTOR inhibition	Inflammation
Good Prognosis	Low	High	E	High	Low	High	Low
Bad Prognosis	High	Low	M	Low	High	Low	High

In respect to these, PI3K-MTOR pathway activity is associated with mesenchymal properties in the literature, similar to that of SU-B type colon cancer patients. It has been shown that increased transcription of mTOR, Raptor and Rictor was noted in advanced stages of CRC [177]. Inhibition or knock-down of Raptor and Rictor attenuated migration, invasion; and induced mesenchymal to epithelial transition (MET) and chemosensitivity of CRCs to oxaliplatin [177]. These data indicate that higher mTOR pathway activity is associated with migration, invasion, mesenchymal status and chemoresistance which are all hallmarks of a worse clinical outcome and therefore is reasonable since *in vitro* inhibition of PI3K-MTOR pathway using BEZ235 caused significantly higher growth inhibition in colon cancer cells with a 'bad' prognostic gene expression. Using EMT marker gene expression data from colon tumors, I clearly showed a more mesenchymal gene expression profile in SU-B patients, therefore it is consistent with the literature to obtain a more effective growth inhibition by BEZ235 as opposed to SU-G patients. *In vivo* treatment of mice with PIK3CA wild type containing colorectal cancer with BEZ235 resulted in a significant reduction in tumor volume, proliferation and angiogenesis when compared to untreated mice [178]. The latest "off the record" information related to clinical trials for BEZ235 in colon cancer is that they were terminated due to insufficient effect of the drug. However the SU-GIB signature can possibly define a group of patients who can benefit from BEZ235 treatment, and thus might help continue these trials.

In addition to finding increased Cleaved caspase 7 protein levels (associated with NFkB activity and inflammation [149,179]) in SU-B group cells, we also found elevated RNA levels of pro-inflammatory cytokines. In line with this finding, reducing inflammation using aspirin and non-steroidal anti-inflammatory drugs (NSAIDs) in cancer has been considered as a potential chemopreventive approach [180-182]. It has also been shown that NSAIDs reduce the risk of developing colorectal and other gastrointestinal cancers [181,183,184]. Thus, reducing the inflammatory response in SU-B patients via NSAID addition to adjuvant chemotherapy may lead to improvement in clinical outcome.

Inflammation has been related to poor outcome for cancer patients. During the initial phases of cancer, the tumor microenvironment is thought to include primarily M1 macrophages and Th1 type T helper cells which produce IL12 and IFN γ . Upon secretion of a chemokine CCL2 from cancer associated fibroblasts, monocytes are recruited. These monocytes then contribute to formation of a protumoral and invasive microenvironment via IL1 and IL6 secretion, a related transition from Th1 to Th2, and a decreased M1 to M2 ratio [185]. This leads to immunosuppression of cytotoxic T cells, increased angiogenesis, ECM remodeling and a poor prognostic profile [185-187]. Additionally, tumor associated macrophages of the M2 type also induce EMT via secretion of TGF β and TNF α [188]. In our study, we identified an upregulation of IL6, IL1 β , TGF β 1, TGF β 3, IL1R2, and TNF α in primary tumors of colon cancer patients with a bad prognostic profile, which is closely related to these observations.

Since inflammation is associated with a more mesenchymal phenotype [189], I performed an E-cadherin and Vimentin expression based analysis and observed that SU-B tumors are relatively mesenchymal and SU-Gs are epithelial.

Altogether, our patient stratification defines prognostic groups which are also biologically distinct (**Table 5.1**). Other biological differences and biomarkers could be further investigated as part of large-scale validation studies including analysis of high-throughput long-noncoding and miRNA expression data.

5.2. Molecular and prognostic sub-groups of gastric cancer

Gastric cancer is a major health problem in Turkey. As part of this thesis, a question I asked was to identify genes that could be used to determine the prognosis of gastric cancer patients as well as genes that could help generate a better molecular sub-grouping of this cancer; as the current classification of gastric cancer (diffuse vs. intestinal) is unable to generate clear prognostic or molecular subgroups, which would be expected to respond to different treatment options. The methodology I used for the *in silico* determination of genes related to survival for colon cancer was improved for those

analyses I performed for gastric cancer. Instead of performing MVAs after identifying genes, a MVA was added to the script to make sure that we select stage independent associations (mUSAT). I also evaluated every single data point as a cut-off for expression data as opposed to 7 thresholds used in the colon cancer study (SSAT). Furthermore, I did not aim a restriction in the number of genes identified as validation of these findings does not have to be necessarily performed via qPCR as RNAseq prices are getting lower each year.

In my first approach, I used two independent microarray cohorts, and ran mUSAT with whole transcriptome data, then I chose the top 20 genes with the lowest rank sum of Cox pH p values. The proteins coded by these 20 genes included cytoskeletal elements, actin binding proteins, extracellular matrix and cell adhesion related proteins with enrichment scores above 1 in DAVID. 5 of the 20 genes (ITGB5, MATN3, LOXL4, IGFBP7, TGFC2) identified had been associated with prognosis in gastric cancer previously [190-194] suggesting that the remaining 15 novel genes have a good chance of being validated in future studies.

This first approach identified patient subgroups with very distinct prognosis independent of known gastric cancer classifiers, mainly that of the diffuse/intestinal subtyping. Although a very clear ECM involvement was found for 'UP' group, no geneset enrichment in 'DOWN' group was obtained suggesting this latter group was biologically heterogeneous. This could also have been due the fact that the classification (especially for the "DOWN" group) was not truly successful. Therefore, as an alternative approach; I aimed to identify molecular sub-groups of gastric cancer based on differential gene expression only, without using prognosis data in the analysis. The evaluation of 134 microarray probesets selected due to their high variance in gene expression; enabled classification of gastric cancer tumors into three biological groups with 'High', 'Intermediate' and 'Low' expression consistently in multiple cohorts. The prognostic difference between groups was, however, not consistent among datasets, possibly due to variable patient characteristics of cohorts. In addition, only 5 out of 134 probesets were significantly associated with survival, supporting that these genes are indeed biological classifiers not prognostic indicators.

As part of approach 2, my goal was to find mRNA based prognostic biomarkers for each molecular subgroups identified. Interestingly mUSAT analysis performed separately for HIL groups resulted in only 1 gene that could be used for prognostic estimations in all three groups. This gene, Chemokine (C-X-C Motif) Ligand 5, CXCL5, which is a ligand of CXCR2; mediates neutrophil trafficking, induces migration, invasion and EMT [195,196] and promotes metastasis in colorectal cancer [197]. However, as I was interested in developing a prognostic test that did not have to be restricted by a given gene number, and which did not have to be identical for all subgroups, I did not pursue this single gene further. However it could be a future plan to validate the prognostic association of CXCR5 *ex vivo*.

In silico analysis showed a high overlap between prognostic UP-DOWN and molecular ‘HIL’ groups demonstrating that these two approaches identified similar sub-groups which shared biological properties like, EMT, ECM involvement, invasion and migration etc. However, the UP-DOWN classification could consistently identify patient sub-groups with clearly distinct prognosis. In contrast to the HIL classification which due its nature could not always separate prognostic subgroups. When we therefore, applied USAT to each of the H, I and L groups, we found distinct prognostic genes for each subgroup. Thus, this data suggests, the ideal scenario to be used for assessment of prognosis in gastric cancer should include a pre-test to define the molecular sub-group of the patients and then the application of a prognostic-test specific for the assigned molecular group. However, the *ex vivo* validation of this would require a large sample size. Given this, I would rather prefer to first validate UP-DOWN subtypes *ex vivo*, as this would generate larger samples in each subgroup, making it easier to validate.

This study is totally *in silico* based, and therefore small and large scale *ex vivo* validations are required. For validation purposes, I have worked with gastrointestinal surgeons in the Ankara Yuksek Ihtisas Training and Research Hospital and obtained the clinical data for 42 gastric cancer patients for which we had stored the fresh frozen primary tumors. This small cohort can be used for validation of UP-DOWN signature using qPCR or RNA sequencing, since further stratification is not preferable due to the sample size.

According to the Lauren classification, diffuse type gastric cancer has been associated with a more aggressive phenotype and with worse clinical outcome as opposed to the intestinal type [154]. Our prognostic UP-DOWN classification defines further prognostic stratification for both diffuse and intestinal type gastric cancer tumors raising the opportunity to be used for both types of tumors.

In all these analysis; multiple cohorts were utilized to eliminate the chance of obtaining cohort-specific results. Especially the GSEA analyses were highly consistent: The pattern of biology in enriched genesets within datasets and between datasets, showed highly related enrichment patterns in both prognostic subgroups identified by approach 1 and molecular sub-groups identified by approach 2. There were mainly two major biological classes, one related to ECM and collagen enrichment, and the other related to enrichment of proliferation and cell cycle related genesets, indeed supporting an EMT-like distinction. The EMT-like biology noted in GSEA analyses were indeed confirmed directly by hierarchical clustering analysis based on Tan et al.’s EMT biomarker list [1]. EMT phenotype in gastric cancer tumors has been widely studied [78]. However, EMT is known to associate with prognosis and drug sensitivity in many other tumors as well. These data therefore inspired me to study EMT further in other cancers, and to test if targeting epithelial and mesenchymal cells with selective compounds could be possible.

5.3. Selective targeting of epithelial and mesenchymal cells in cancers

To investigate the involvement of EMT in many other cancers, I utilized publically available drug cytotoxicity databases. I generated an in-house EMT scoring method which is applicable to any cancer type and is comparable to complex scoring algorithms that include hundreds of genes [1]. Analysis of three independent drug cytotoxicity databases revealed that epithelial cancer cell lines, as defined by my method, are more sensitive to EGFR inhibitors when tested for any cancer types available for analysis. In order to target mesenchymal cells, I identified 8 compounds, among which the chemical structure of 6 of them were available in Pubchem. To continue with the chemical domain based analysis, one could work with chemists in the near future. Otherwise there is no given name, mode of action or identified target available for these compounds.

Various inhibitors of the ERBB family have been approved by FDA in breast, colorectal, head and neck, lung, and pancreatic cancers [198]. Among these cancer types, breast, CRC and lung cancers were available in the NCI60 database and are in top 4 most epithelial type cancers according to my CDH1-VIM score. On the other hand, the fact that EMT is a determinant of sensitivity to EGFR inhibitors has been also observed in the literature in lung cancer patients [199], non-small-cell lung carcinoma cell lines and xenografts, in NSCLC patients [200,201], and in hepatocellular cancer cells [202]. To my knowledge, this is the first study demonstrating the tissue independent relationship between EMT and sensitivity to EGFR inhibitors.

An important clinical utilization of epithelialness can be as a marker of sensitivity to EGFR inhibitors in the treatment of metastatic CRC; the second most epithelial cancer type of NCI60. This clinical need originates from the insufficient benefit of mutation based treatment response associations for usage of EGFR inhibitors in mCRC. Large clinical trials; CO.17, CRYSTAL, OPUS, and PRIME, concluded that KRAS mutations are predictive of resistance to anti-EGFR mAbs in metastatic CRC [203]. Expert consensus guidelines agree on not using EGFR inhibitors for the treatment of patients with mutated KRAS [204]. However a meta-analysis of phase-III studies elucidated that the addition of targeted agents including cetuximab was not associated with significant improvement in DFS in patients with resected colorectal cancer ((HR=1.11, p=0.22) [205]. In line with this, a randomized controlled trial of 1630 patients demonstrated that in KRAS wild type, BRAF wild type and NRAS wild type CRC patients, addition of cetuximab did not result in survival benefit when compared to patients treated with standard therapy [133]. Therefore responders of EGFR therapy have not yet been completely defined for mCRC, and my data indicates that EMT could be a candidate mechanism for this purpose.

I therefore plan to work on identification of those patients who may benefit from EGFR targeted therapy via evaluating EMT marker expression. In addition to that, combination or sequential therapy with MET (mesenchymal-to-epithelial transition) inducing agents such as forskolin or all-trans retinoic acid (ATRA), may improve the growth inhibition effect of EGFR inhibitors. Forskolin directly activates the adenylate cyclase enzyme, that generates cAMP from ATP, thus raising intracellular cAMP concentrations [206]. Importantly, targeting cAMP levels and signaling has been shown to result in a number of relevant anticancer effects, such as the induction of MET, the enhancement of the sensitivity to conventional antineoplastic drugs, and the inhibition of proliferation, motility, and migration in many types of cancer cells [207]. Another inducer of MET, ATRA is a metabolite of vitamin A, and has shown to be involved in inhibition of migration and invasion in cancer cells and induction of differentiation of abnormal cells back to normal [208,209]. Evidence also demonstrated that ATRA treatment induces E-cadherin expression and down regulates the expression of mesenchymal markers in colon cancer cells [210], thus reversing EMT *in vitro* [211,212]. Therefore I would like to pursue *in vitro* experiments which will show MET induction by forskolin or ATRA and its possible synergistic effect when combined with EGFR inhibitors in colon cancer cells.

In addition to EMT, I investigated further predictors of response to EGFR targeted treatment of the DNA (mutation), coding and non-coding RNA and protein type data. As an outcome of these studies, I found that decreased cRAF protein level expression and increased p38 MAPK phosphorylation were correlated with sensitivity to EGFR inhibitors in both breast and endometrial cancer cell lines. Enhancement of EGFR and downstream signaling and thus increased cell proliferation via p38 MAPK has previously been shown in breast epithelial cells [166]. However, this is the first data demonstrating correlation of increased p38 MAPK phosphorylation with sensitivity to EGFR inhibitors. Amplification of cRAF is known to promote EGFR-TKI resistance [167]. I also find that increased cRAF protein expression relates to resistance to EGFR inhibitors.

An evaluation of the mutational profile of cancer cell lines in CGP separately for epithelial and mesenchymal cells revealed that, although there is no statistically significant difference between mutational frequency of 84 genes between epithelial cells which are sensitive and resistant to BIBW2992, an EGFR inhibitor, we found that various mutations of *PI3K* are more frequent in EGFR sensitive mesenchymal cells. In the literature, *PI3K* mutations have been found to be related to acquired resistance to EGFR-TKIs in NSCLC tumors [213] and another study showed that they have not been significantly beneficial for EGFR-TKI monotherapy in EGFR-mutant lung cancers [214]. Therefore a mutation and tissue specific approach should next be pursued especially for mesenchymal cells; since the tissue specific characteristics and the type of the studied mutation can cause major effects in terms of sensitivity to BIBW2992. My data also showed that nonsense/frameshift mutations and deletions of *NF1*, neurofibromin which is known to downregulate RAS [215], are significantly associated with sensitivity to

BIBW2992; in contrast there is evidence that reduced expression of NF1 is related to resistance to EGFR inhibition in lung cancer. Therefore I plan to perform mutational analysis tissue-specifically, since the cancer type and thus harboring other intrinsic mutations affect the associations between mutational profile and drug sensitivity, as is the case for PI3K results. Lastly, I found that multiple missense *NRAS* mutations in codon 12, 13 and 61 are related to resistance to BIBW2992 which is consistent with the literature as studies showed that many *NRAS* mutations; including those in codon 12, 13 and 61, cause constitutive activation of NRAS and were shown to occur more frequently in cetuximab and panitumumab non-responsive CRC patients [216,217].

The majority of the data used for studies related to EMT and selective targeting of epithelial or mesenchymal cells was from NCI60. Unfortunately the database harbors many obstacles with the quality of the cytotoxicity data. The biggest problems with the GI50 values were the presence of missing values, repeating values, and the narrow range of drug concentrations tested which generated quite a number of repeats in GI50 values at the maximum test dose. For my studies, I filtered out low-quality data and strikingly the number of compounds remaining was 299 out of more than 50000 compounds. While the data quality problem has restricted my analysis in terms of number of compounds included in the analysis, the tissue heterogeneity of this database was a unique feature enabling us to perform analyses in a tissue independent way.

Although I pursued three different approaches for studies related to colon, gastric cancers and EMT, my results arrived at one major conclusion: that EMT has a major role in prognosis, molecular classification and sensitivity to targeted agents in cancer. SU-GIB prognostic signature in colon cancer and molecular sub-groups identified in gastric cancer are related to EMT status of tumors. The *in silico* studies to identify agents that target epithelial and mesenchymal cells selectively, resulted in EGFR inhibitors with selective activity on epithelial cells regardless of the tissue type. Consistently, I also found higher EGFR phosphorylation in SU-G patients which have an epithelial gene expression profile. Overall, patients with bad prognosis are more mesenchymal, have higher ECM involvement and pro-inflammatory gene expression in our studies. Larger studies are required for validation of these observations *in vitro* in cell culture and *ex vivo* in fresh frozen tumor tissues.

6. BIBLIOGRAPHY

1. Tan TZ, Miow QH, Miki Y, Noda T, Mori S, et al. (2014) Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol Med* 6: 1279-1293.
2. Demirkol S, Gomceli I, İsbilen M, Dayanc BE, Tez M, et al. (2017) A Combined ULBP2 and SEMA5A Expression Signature as a Prognostic and Predictive Biomarker for Colon Cancer. *Journal of Cancer* 8: 1113-1122.
3. İsbilen M (2015) Identification of theranostic gene markers in cancers and prognostic validation in colorectal cancer: Bilkent University.
4. Reed KK, Wickham R (2009) Review of the gastrointestinal tract: from macro to micro. *Semin Oncol Nurs* 25: 3-14.
5. Griffin-Sobel JP (2017) Gastrointestinal Cancers: Screening and Early Detection. *Semin Oncol Nurs*.
6. Siegel RL, Miller KD, Jemal A (2015) Cancer statistics, 2015. *CA Cancer J Clin* 65: 5-29.
7. Tuncer AM (2007) Türkiye'de kanser kontrolü: T.C. Sağlık Bakanlığı.
8. (2014) Bowel cancer survival statistics. *Cancer Research UK*.
9. Young PE, Womeldorph CM, Johnson EK, Maykel JA, Brucher B, et al. (2014) Early detection of colorectal cancer recurrence in patients undergoing surgery with curative intent: current status and challenges. *J Cancer* 5: 262-271.
10. Armaghany T, Wilson JD, Chu Q, Mills G (2012) Genetic alterations in colorectal cancer. *Gastrointest Cancer Res* 5: 19-27.
11. Benson AB, 3rd, Venook AP, Bekaii-Saab T, Chan E, Chen YJ, et al. (2014) Colon cancer, version 3.2014. *J Natl Compr Canc Netw* 12: 1028-1059.
12. Gill S, Loprinzi CL, Sargent DJ, Thome SD, Alberts SR, et al. (2004) Pooled analysis of fluorouracil-based adjuvant therapy for stage II and III colon cancer: who benefits and by how much? *J Clin Oncol* 22: 1797-1806.
13. Benson AB, 3rd, Schrag D, Somerfield MR, Cohen AM, Figueredo AT, et al. (2004) American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. *J Clin Oncol* 22: 3408-3419.
14. Dienstmann R, Salazar R, Tabernero J (2015) Personalizing colon cancer adjuvant therapy: selecting optimal treatments for individual patients. *J Clin Oncol* 33: 1787-1796.
15. Grothey A (2010) Adjuvant chemotherapy in colon cancer - is it worth it? *Eur J Cancer* 46: 1768-1769.
16. De Divitiis C, Nasti G, Montano M, Fisichella R, Iaffaioli RV, et al. (2014) Prognostic and predictive response factors in colorectal cancer patients: between hope and reality. *World J Gastroenterol* 20: 15049-15059.
17. Lee MS, Menter DG, Kopetz S (2017) Right Versus Left Colon Cancer Biology: Integrating the Consensus Molecular Subtypes. *J Natl Compr Canc Netw* 15: 411-419.
18. Wong R (2010) Proximal tumors are associated with greater mortality in colon cancer. *J Gen Intern Med* 25: 1157-1163.
19. Phipps AI, Lindor NM, Jenkins MA, Baron JA, Win AK, et al. (2013) Colon and rectal cancer survival by tumor location and microsatellite instability: the Colon Cancer Family Registry. *Dis Colon Rectum* 56: 937-944.
20. Saridaki Z, Souglakos J, Georgoulas V (2014) Prognostic and predictive significance of MSI in stages II/III colon cancer. *World J Gastroenterol* 20: 6809-6814.
21. Sargent DJ, Marsoni S, Monges G, Thibodeau SN, Labianca R, et al. (2010) Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. *J Clin Oncol* 28: 3219-3226.
22. Popat S, Hubner R, Houlston RS (2005) Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol* 23: 609-618.
23. Jung B, Smith EJ, Doctolero RT, Gervaz P, Alonso JC, et al. (2006) Influence of target gene mutations on survival, stage and histology in sporadic microsatellite unstable colon cancers. *Int J Cancer* 118: 2509-2513.
24. Andreyev HJ, Norman AR, Cunningham D, Oates JR, Clarke PA (1998) Kirsten ras mutations in patients with colorectal cancer: the multicenter "RASCAL" study. *J Natl Cancer Inst* 90: 675-684.
25. Klump B, Nehls O, Okech T, Hsieh CJ, Gaco V, et al. (2004) Molecular lesions in colorectal cancer: impact on prognosis? Original data and review of the literature. *Int J Colorectal Dis* 19: 23-42.
26. Kihana T, Fujioka T, Hamada K, Kito K, Takahashi A, et al. (1998) Association of replication error positive phenotype with lymphocyte infiltration in endometrial cancers. *Jpn J Cancer Res* 89: 895-902.
27. Kim H, Jen J, Vogelstein B, Hamilton SR (1994) Clinical and pathological characteristics of sporadic colorectal carcinomas with DNA replication errors in microsatellite sequences. *Am J Pathol* 145: 148-156.
28. Pino MS, Kikuchi H, Zeng M, Herraiz MT, Sperduti I, et al. (2010) Epithelial to mesenchymal transition is impaired in colon cancer cells with microsatellite instability. *Gastroenterology* 138: 1406-1417.
29. Andreyev HJ, Norman AR, Cunningham D, Oates J, Dix BR, et al. (2001) Kirsten ras mutations in patients with colorectal cancer: the 'RASCAL II' study. *Br J Cancer* 85: 692-696.
30. Wilson PM, Lenz H-J Integrating Biomarkers Into Clinical Decision Making for Colorectal Cancer. *Clin Colorectal Cancer* 9: S16-S27.
31. Roth AD, Tejpar S, Delorenzi M, Yan P, Fiocca R, et al. (2010) Prognostic role of KRAS and BRAF in stage II and III resected colon cancer: results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00 trial. *J Clin Oncol* 28: 466-474.
32. Tol J, Nagtegaal ID, Punt CJ (2009) BRAF mutation in metastatic colorectal cancer. *N Engl J Med* 361: 98-99.
33. Clarke CN, Kopetz ES (2015) BRAF mutant colorectal cancer as a distinct subset of colorectal cancer: clinical characteristics, clinical behavior, and response to targeted therapies. *J Gastrointest Oncol* 6: 660-667.
34. Luo HY, Xu RH (2014) Predictive and prognostic biomarkers with therapeutic targets in advanced colorectal cancer. *World J Gastroenterol* 20: 3858-3874.
35. Schirripa M, Cremolini C, Loupakis F, Morvillo M, Bergamo F, et al. (2015) Role of NRAS mutations as prognostic and predictive markers in metastatic colorectal cancer. *Int J Cancer* 136: 83-90.

36. Molinari F, Frattini M (2013) Functions and Regulation of the PTEN Gene in Colorectal Cancer. *Front Oncol* 3: 326.
37. Goel G (2014) Evolving role of gene expression signatures as biomarkers in early-stage colon cancer. *J Gastrointest Cancer* 45: 399-404.
38. Salazar R, Roepman P, Capella G, Moreno V, Simon I, et al. (2011) Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 29: 17-24.
39. Kopetz S, Tabernero J, Rosenberg R, Jiang ZQ, Moreno V, et al. (2015) Genomic classifier ColoPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. *Oncologist* 20: 127-133.
40. Tan IB, Tan P (2011) Genetics: an 18-gene signature (ColoPrint(R)) for colon cancer prognosis. *Nat Rev Clin Oncol* 8: 131-133.
41. O'Connell MJ, Lavery I, Yothers G, Paik S, Clark-Langone KM, et al. (2010) 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. *J Clin Oncol* 28: 3937-3944.
42. Quasar Collaborative G, Gray R, Barnwell J, McConkey C, Hills RK, et al. (2007) Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. *Lancet* 370: 2020-2029.
43. Gray RG, Quirke P, Handley K, Lopatin M, Magill L, et al. (2011) Validation study of a quantitative multigene reverse transcriptase-polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer. *J Clin Oncol* 29: 4611-4619.
44. Di Narzo AF, Tejpar S, Rossi S, Yan P, Popovici V, et al. (2014) Test of four colon cancer risk-scores in formalin fixed paraffin embedded microarray gene expression data. *J Natl Cancer Inst* 106.
45. Meleth S, Reeder-Hayes K, Ashok M, Clark R, Funkhouser W, et al. (2014). Technology Assessment of Molecular Pathology Testing for the Estimation of Prognosis for Common Cancers. Rockville (MD).
46. He Z, Wang KC, Koprivica V, Ming G, Song HJ (2002) Knowing how to navigate: mechanisms of semaphorin signaling in the nervous system. *Sci STKE* 2002: re1.
47. Bahri SM, Chia W, Yang X (2001) Characterization and mutant analysis of the *Drosophila* sema 5c gene. *Dev Dyn* 221: 322-330.
48. Adams RH, Betz H, Puschel AW (1996) A novel class of murine semaphorins with homology to thrombospondin is differentially expressed during early embryogenesis. *Mech Dev* 57: 33-45.
49. Artigiani S, Conrotto P, Fazzari P, Gilestro GF, Barberis D, et al. (2004) Plexin-B3 is a functional receptor for semaphorin 5A. *EMBO Rep* 5: 710-714.
50. Rehman M, Tamagnone L (2013) Semaphorins in cancer: biological mechanisms and therapeutic approaches. *Semin Cell Dev Biol* 24: 179-189.
51. Sadanandam A, Varney ML, Singh S, Ashour AE, Moniaux N, et al. (2010) High gene expression of semaphorin 5A in pancreatic cancer is associated with tumor growth, invasion and metastasis. *Int J Cancer* 127: 1373-1383.
52. Pan G, Zhu Z, Huang J, Yang C, Yang Y, et al. (2013) Semaphorin 5A promotes gastric cancer invasion/metastasis via urokinase-type plasminogen activator/phosphoinositide 3-kinase/protein kinase B. *Dig Dis Sci* 58: 2197-2204.
53. Li X, Law JW, Lee AY (2012) Semaphorin 5A and plexin-B3 regulate human glioma cell motility and morphology through Rac1 and the actin cytoskeleton. *Oncogene* 31: 595-610.
54. Lu TP, Tsai MH, Lee JM, Hsu CP, Chen PC, et al. (2010) Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev* 19: 2590-2597.
55. Tamagnone L (2012) Emerging role of semaphorins as major regulatory signals and potential therapeutic targets in cancer. *Cancer Cell* 22: 145-152.
56. Zhou H, Wu A, Fu W, Lv Z, Zhang Z (2014) Significance of semaphorin-3A and MMP-14 protein expression in non-small cell lung cancer. *Oncol Lett* 7: 1395-1400.
57. Song X, Zhang W, Zhang Y, Zhang H, Fu Z, et al. (2012) Expression of semaphorin 3A and neuropilin 1 with clinicopathological features and survival in human tongue cancer. *Med Oral Patol Oral Cir Bucal* 17: e962-968.
58. Muller MW, Giese NA, Swiercz JM, Ceyhan GO, Esposito I, et al. (2007) Association of axon guidance factor semaphorin 3A with poor outcome in pancreatic cancer. *Int J Cancer* 121: 2421-2433.
59. Liu H, Yang Y, Xiao J, Yang S, Liu Y, et al. (2014) Semaphorin 4D expression is associated with a poor clinical outcome in cervical cancer patients. *Microvasc Res* 93: 1-8.
60. Karayan-Tapon L, Wager M, Guilhot J, Levillain P, Marquant C, et al. (2008) Semaphorin, neuropilin and VEGF expression in glial tumours: SEMA3G, a prognostic marker? *Br J Cancer* 99: 1153-1160.
61. Seaman WE, Sleisenger M, Eriksson E, Koo GC (1987) Depletion of natural killer cells in mice by monoclonal antibody to NK-1.1. Reduction in host defense against malignancy without loss of cellular or humoral immunity. *J Immunol* 138: 4539-4544.
62. Groh V, Rhinehart R, Secrist H, Bauer S, Grabstein KH, et al. (1999) Broad tumor-associated expression and recognition by tumor-derived gamma delta T cells of MICA and MICB. *Proc Natl Acad Sci U S A* 96: 6879-6884.
63. Cosman D, Mullberg J, Sutherland CL, Chin W, Armitage R, et al. (2001) ULBPs, novel MHC class I-related molecules, bind to CMV glycoprotein UL16 and stimulate NK cytotoxicity through the NKG2D receptor. *Immunity* 14: 123-133.
64. Ohashi M, Eagle RA, Trowsdale J (2010) Post-translational modification of the NKG2D ligand RAET1G leads to cell surface expression of a glycosylphosphatidylinositol-linked isoform. *J Biol Chem* 285: 16408-16415.
65. Hayakawa Y, Kelly JM, Westwood JA, Darcy PK, Diefenbach A, et al. (2002) Cutting edge: tumor rejection mediated by NKG2D receptor-ligand interaction is dependent upon perforin. *J Immunol* 169: 5377-5381.
66. Pende D, Rivera P, Marcenaro S, Chang CC, Biassoni R, et al. (2002) Major histocompatibility complex class I-related chain A and UL16-binding protein expression on tumor cell lines of different histotypes: analysis of tumor susceptibility to NKG2D-dependent natural killer cell cytotoxicity. *Cancer Res* 62: 6178-6186.

67. Guerra N, Tan YX, Joncker NT, Choy A, Gallardo F, et al. (2008) NKG2D-deficient mice are defective in tumor surveillance in models of spontaneous malignancy. *Immunity* 28: 571-580.
68. Waldhauer I, Steinle A (2006) Proteolytic release of soluble UL16-binding protein 2 from tumor cells. *Cancer Res* 66: 2520-2526.
69. Li K, Mandai M, Hamanishi J, Matsumura N, Suzuki A, et al. (2009) Clinical significance of the NKG2D ligands, MICA/B and ULBP2 in ovarian cancer: high expression of ULBP2 is an indicator of poor prognosis. *Cancer Immunol Immunother* 58: 641-652.
70. McGilvray RW, Eagle RA, Rolland P, Jafferji I, Trowsdale J, et al. (2010) ULBP2 and RAET1E NKG2D ligands are independent predictors of poor prognosis in ovarian cancer patients. *Int J Cancer* 127: 1412-1420.
71. Paschen A, Sucker A, Hill B, Moll I, Zapatka M, et al. (2009) Differential clinical significance of individual NKG2D ligands in melanoma: soluble ULBP2 as an indicator of poor prognosis superior to S100B. *Clin Cancer Res* 15: 5208-5215.
72. Chang YT, Wu CC, Shyr YM, Chen TC, Hwang TL, et al. (2011) Secretome-based identification of ULBP2 as a novel serum marker for pancreatic cancer detection. *PLoS One* 6: e20029.
73. Yamaguchi K, Chikumi H, Shimizu A, Takata M, Kinoshita N, et al. (2012) Diagnostic and prognostic impact of serum-soluble UL16-binding protein 2 in lung cancer patients. *Cancer Sci* 103: 1405-1413.
74. Nuckel H, Switala M, Sellmann L, Horn PA, Durig J, et al. (2010) The prognostic significance of soluble NKG2D ligands in B-cell chronic lymphocytic leukemia. *Leukemia* 24: 1152-1159.
75. Rothe A, Jachimowicz RD, Borchmann S, Madlener M, Kessler J, et al. (2014) The bispecific immunoligand ULBP2-aCEA redirects natural killer cells to tumor cells and reveals potent anti-tumor activity against colon carcinoma. *Int J Cancer* 134: 2829-2840.
76. Berdiel-Acer M, Berenguer A, Sanz-Pamplona R, Cuadras D, Sanjuan X, et al. (2014) A 5-gene classifier from the carcinoma-associated fibroblast transcriptomic profile and clinical outcome in colorectal cancer. *Oncotarget* 5: 6437-6452.
77. Eser S, Yakut C, Ozdemir R, Karakilinc H, Ozalan S, et al. (2010) Cancer incidence rates in Turkey in 2006: a detailed registry based estimation. *Asian Pac J Cancer Prev* 11: 1731-1739.
78. Huang L, Wu RL, Xu AM (2015) Epithelial-mesenchymal transition in gastric cancer. *Am J Transl Res* 7: 2141-2158.
79. Smalley SR, Benedetti JK, Haller DG, Hundahl SA, Estes NC, et al. (2012) Updated analysis of SWOG-directed intergroup study 0116: a phase III trial of adjuvant radiochemotherapy versus observation after curative gastric cancer resection. *J Clin Oncol* 30: 2327-2333.
80. Sunakawa Y, Lenz HJ (2015) Molecular classification of gastric adenocarcinoma: translating new insights from the cancer genome atlas research network. *Curr Treat Options Oncol* 16: 17.
81. Hudler P (2015) Challenges of deciphering gastric cancer heterogeneity. *World J Gastroenterol* 21: 10510-10527.
82. Hu B, El Hajj N, Sittler S, Lammert N, Barnes R, et al. (2012) Gastric cancer: Classification, histology and application of molecular pathology. *J Gastrointest Oncol* 3: 251-261.
83. Datta De D, Roychoudhury S (2015) To be or not to be: The host genetic factor and beyond in *Helicobacter pylori* mediated gastro-duodenal diseases. *World J Gastroenterol* 21: 2883-2895.
84. Katoh M (2005) Epithelial-mesenchymal transition in gastric cancer (Review). *Int J Oncol* 27: 1677-1683.
85. Oliveira C, Senz J, Kaurah P, Pinheiro H, Sanges R, et al. (2009) Germline CDH1 deletions in hereditary diffuse gastric cancer families. *Hum Mol Genet* 18: 1545-1555.
86. Tanabe S, Aoyagi K, Yokozaki H, Sasaki H (2014) Gene expression signatures for identifying diffuse-type gastric cancer associated with epithelial-mesenchymal transition. *Int J Oncol* 44: 1955-1970.
87. Kim C, Mulder K, Spratlin J (2014) How prognostic and predictive biomarkers are transforming our understanding and management of advanced gastric cancer. *Oncologist* 19: 1046-1055.
88. Gherardi E, Birchmeier W, Birchmeier C, Vande Woude G (2012) Targeting MET in cancer: rationale and progress. *Nat Rev Cancer* 12: 89-103.
89. Nakajima M, Sawada H, Yamada Y, Watanabe A, Tatsumi M, et al. (1999) The prognostic significance of amplification and overexpression of c-met and c-erb B-2 in human gastric carcinomas. *Cancer* 85: 1894-1902.
90. Kuniyasu H, Yasui W, Kitadai Y, Yokozaki H, Ito H, et al. (1992) Frequent amplification of the c-met gene in scirrhous type stomach cancer. *Biochem Biophys Res Commun* 189: 227-232.
91. Amemiya H, Kono K, Itakura J, Tang RF, Takahashi A, et al. (2002) c-Met expression in gastric cancer with liver metastasis. *Oncology* 63: 286-296.
92. Tokunaga A, Onda M, Okuda T, Teramoto T, Fujita I, et al. (1995) Clinical significance of epidermal growth factor (EGF), EGF receptor, and c-erbB-2 in human gastric cancer. *Cancer* 75: 1418-1425.
93. Yano T, Doi T, Ohtsu A, Boku N, Hashizume K, et al. (2006) Comparison of HER2 gene amplification assessed by fluorescence in situ hybridization and HER2 protein expression assessed by immunohistochemistry in gastric cancer. *Oncol Rep* 15: 65-71.
94. Kim MA, Lee HS, Lee HE, Jeon YK, Yang HK, et al. (2008) EGFR in gastric carcinomas: prognostic significance of protein overexpression and high gene copy number. *Histopathology* 52: 738-746.
95. Fassan M, Mastracci L, Grillo F, Zagonel V, Bruno S, et al. (2012) Early HER2 dysregulation in gastric and oesophageal carcinogenesis. *Histopathology* 61: 769-776.
96. Bang YJ, Van Cutsem E, Feyereislova A, Chung HC, Shen L, et al. (2010) Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet* 376: 687-697.
97. Gomez-Martin C, Lopez-Rios F, Aparicio J, Barriuso J, Garcia-Carbonero R, et al. (2014) A critical review of HER2-positive gastric cancer evaluation and treatment: from trastuzumab, and beyond. *Cancer Lett* 351: 30-40.

98. Kiyose S, Nagura K, Tao H, Igarashi H, Yamada H, et al. (2012) Detection of kinase amplifications in gastric cancer archives using fluorescence in situ hybridization. *Pathol Int* 62: 477-484.
99. Shi J, Yao D, Liu W, Wang N, Lv H, et al. (2012) Highly frequent PIK3CA amplification is associated with poor prognosis in gastric cancer. *BMC Cancer* 12: 50.
100. Singh SS, Yap WN, Arfuso F, Kar S, Wang C, et al. (2015) Targeting the PI3K/Akt signaling pathway in gastric carcinoma: A reality for personalized medicine? *World J Gastroenterol* 21: 12261-12273.
101. Xu DZ, Geng QR, Tian Y, Cai MY, Fang XJ, et al. (2010) Activated mammalian target of rapamycin is a potential therapeutic target in gastric cancer. *BMC Cancer* 10: 536.
102. Yu G, Wang J, Chen Y, Wang X, Pan J, et al. (2009) Overexpression of phosphorylated mammalian target of rapamycin predicts lymph node metastasis and prognosis of chinese patients with gastric cancer. *Clin Cancer Res* 15: 1821-1829.
103. Ohtsu A, Ajani JA, Bai YX, Bang YJ, Chung HC, et al. (2013) Everolimus for previously treated advanced gastric cancer: results of the randomized, double-blind, phase III GRANITE-1 study. *J Clin Oncol* 31: 3935-3943.
104. Shan YS, Hsu HP, Lai MD, Hung YH, Wang CY, et al. (2017) Cyclin D1 overexpression correlates with poor tumor differentiation and prognosis in gastric cancer. *Oncol Lett* 14: 4517-4526.
105. Miao Y, Sui J, Xu SY, Liang GY, Pu YP, et al. (2017) Comprehensive analysis of a novel four-lncRNA signature as a prognostic biomarker for human gastric cancer. *Oncotarget* 8: 75007-75024.
106. Zhu J, Wang H, Xu S, Hao Y (2017) Clinicopathological and prognostic significance of HMGA2 overexpression in gastric cancer: a meta-analysis. *Oncotarget* 8: 100478-100489.
107. Yin S, Li Z, Huang J, Miao Z, Zhang J, et al. (2017) Prognostic value and clinicopathological significance of proliferating cell nuclear antigen expression in gastric cancer: a systematic review and meta-analysis. *Onco Targets Ther* 10: 319-327.
108. Ye F, Jin P, Cai X, Cai P, Cai H (2017) High RNA-Binding Motif Protein 3 (RBM3) Expression is Independently Associated with Prolonged Overall Survival in Intestinal-Type Gastric Cancer. *Med Sci Monit* 23: 6033-6041.
109. Laitinen A, Bockelman C, Hagstrom J, Kokkola A, Kallio P, et al. (2017) High PROX1 expression in gastric cancer predicts better survival. *PLoS One* 12: e0183868.
110. Kalluri R, Weinberg RA (2009) The basics of epithelial-mesenchymal transition. *J Clin Invest* 119: 1420-1428.
111. Kalluri R, Neilson EG (2003) Epithelial-mesenchymal transition and its implications for fibrosis. *J Clin Invest* 112: 1776-1784.
112. Tsai JH, Yang J (2013) Epithelial-mesenchymal plasticity in carcinoma metastasis. *Genes Dev* 27: 2192-2206.
113. Heerboth S, Housman G, Leary M, Longacre M, Byler S, et al. (2015) EMT and tumor metastasis. *Clin Transl Med* 4: 6.
114. Singh A, Settleman J (2010) EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. *Oncogene* 29: 4741-4751.
115. Vega S, Morales AV, Ocana OH, Valdes F, Fabregat I, et al. (2004) Snail blocks the cell cycle and confers resistance to cell death. *Genes Dev* 18: 1131-1143.
116. Kajita M, McClintock KN, Wade PA (2004) Aberrant expression of the transcription factors snail and slug alters the response to genotoxic stress. *Mol Cell Biol* 24: 7559-7566.
117. Saxena M, Stephens MA, Pathak H, Rangarajan A (2011) Transcription factors that mediate epithelial-mesenchymal transition lead to multidrug resistance by upregulating ABC transporters. *Cell Death Dis* 2: e179.
118. Gupta PB, Onder TT, Jiang G, Tao K, Kuperwasser C, et al. (2009) Identification of selective inhibitors of cancer stem cells by high-throughput screening. *Cell* 138: 645-659.
119. Baranwal S, Alahari SK (2009) Molecular mechanisms controlling E-cadherin expression in breast cancer. *Biochem Biophys Res Commun* 384: 6-11.
120. Chen A, Beetham H, Black MA, Priya R, Telford BJ, et al. (2014) E-cadherin loss alters cytoskeletal organization and adhesion in non-malignant breast cells but is insufficient to induce an epithelial-mesenchymal transition. *BMC Cancer* 14: 552.
121. Pecina-Slaus N (2003) Tumor suppressor gene E-cadherin and its role in normal and malignant cells. *Cancer Cell Int* 3: 17.
122. Serrano-Gomez SJ, Maziveyi M, Alahari SK (2016) Regulation of epithelial-mesenchymal transition through epigenetic and post-translational modifications. *Mol Cancer* 15: 18.
123. Franke WW, Schmid E, Osborn M, Weber K (1978) Different intermediate-sized filaments distinguished by immunofluorescence microscopy. *Proc Natl Acad Sci U S A* 75: 5034-5038.
124. Dellagi K, Vainchenker W, Vinci G, Paulin D, Brouet JC (1983) Alteration of vimentin intermediate filament expression during differentiation of human hemopoietic cells. *EMBO J* 2: 1509-1514.
125. Ku NO, Liao J, Chou CF, Omary MB (1996) Implications of intermediate filament protein phosphorylation. *Cancer Metastasis Rev* 15: 429-444.
126. Eriksson JE, Dechat T, Grin B, Helfand B, Mendez M, et al. (2009) Introducing intermediate filaments: from discovery to disease. *J Clin Invest* 119: 1763-1771.
127. Ivaska J, Vuoriluoto K, Huovinen T, Izawa I, Inagaki M, et al. (2005) PKCepsilon-mediated phosphorylation of vimentin controls integrin recycling and motility. *EMBO J* 24: 3834-3845.
128. Eckes B, Dogic D, Colucci-Guyon E, Wang N, Maniotis A, et al. (1998) Impaired mechanical stability, migration and contractile capacity in vimentin-deficient fibroblasts. *J Cell Sci* 111 (Pt 13): 1897-1907.
129. Raymond WA, Leong AS (1989) Vimentin--a new prognostic parameter in breast carcinoma? *J Pathol* 158: 107-114.
130. Gilles C, Polette M, Piette J, Delvigne AC, Thompson EW, et al. (1996) Vimentin expression in cervical carcinomas: association with invasive and migratory potential. *J Pathol* 180: 175-180.
131. Kidd ME, Shumaker DK, Ridge KM (2014) The role of vimentin intermediate filaments in the progression of lung cancer. *Am J Respir Cell Mol Biol* 50: 1-6.

132. Ngan CY, Yamamoto H, Seshimo I, Tsujino T, Man-i M, et al. (2007) Quantitative evaluation of vimentin expression in tumour stroma of colorectal cancer. *Br J Cancer* 96: 986-992.
133. Maughan TS, Adams RA, Smith CG, Meade AM, Seymour MT, et al. (2011) Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *Lancet* 377: 2103-2114.
134. Li Z, Tian T, Lv F, Chang Y, Wang X, et al. (2013) Six1 promotes proliferation of pancreatic cancer cells via upregulation of cyclin D1 expression. *PLoS One* 8: e59203.
135. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483: 603-607.
136. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, et al. (2012) CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res* 72: 3499-3511.
137. Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, et al. (2010) Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* 138: 958-968.
138. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, et al. (2013) Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med* 10: e1001453.
139. Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, et al. (2015) Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat Med* 21: 449-456.
140. Ooi CH, Ivanova T, Wu J, Lee M, Tan IB, et al. (2009) Oncogenic pathway combinations predict clinical prognosis in gastric cancer. *PLoS Genet* 5: e1000676.
141. Wang G, Hu N, Yang HH, Wang L, Su H, et al. (2013) Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in china. *PLoS One* 8: e63826.
142. Wu Y, Grabsch H, Ivanova T, Tan IB, Murray J, et al. (2013) Comprehensive genomic meta-analysis identifies intra-tumoural stroma as a predictor of survival in patients with gastric cancer. *Gut* 62: 1100-1111.
143. Forster S, Gretscher S, Jons T, Yashiro M, Kemmner W (2011) THBS4, a novel stromal molecule of diffuse-type gastric adenocarcinomas, identified by transcriptome-wide expression profiling. *Mod Pathol* 24: 1390-1403.
144. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25: 402-408.
145. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, et al. (2006) Design and Analysis of DNA Microarray Investigations: Springer New York.
146. Tian S, Roepman P, Popovici V, Michaut M, Majewski I, et al. (2012) A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *J Pathol* 228: 586-595.
147. Therneau TM, Grambsch PM (2000) Modeling Survival Data: Extending the Cox Model: Springer.
148. Zhang J, Zheng J, Yang Y, Lu J, Gao J, et al. (2015) Molecular spectrum of KRAS, NRAS, BRAF and PIK3CA mutations in Chinese colorectal cancer patients: analysis of 1,110 cases. *Sci Rep* 5: 18678.
149. Erener S, Petrilli V, Kassner I, Minotti R, Castillo R, et al. (2012) Inflammasome-activated caspase 7 cleaves PARP1 to enhance the expression of a subset of NF-kappaB target genes. *Mol Cell* 46: 200-211.
150. Lopez-Novoa JM, Nieto MA (2009) Inflammation and EMT: an alliance towards organ fibrosis and cancer progression. *EMBO Mol Med* 1: 303-314.
151. Wang H, Wang HS, Zhou BH, Li CL, Zhang F, et al. (2013) Epithelial-mesenchymal transition (EMT) induced by TNF-alpha requires AKT/GSK-3beta-mediated stabilization of snail in colorectal cancer. *PLoS One* 8: e56664.
152. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, et al. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 483: 570-575.
153. Poland JA, Rife TW (2012) Genotyping-by-Sequencing for Plant Breeding and Genetics. *The Plant Genome* 5: 92-102.
154. Petrelli F, Berenato R, Turati L, Mennitto A, Steccanella F, et al. (2017) Prognostic value of diffuse versus intestinal histotype in patients with gastric cancer: a systematic review and meta-analysis. *J Gastrointest Oncol* 8: 148-163.
155. Lamouille S, Xu J, Derynck R (2014) Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol* 15: 178-196.
156. Shoemaker RH (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer* 6: 813-823.
157. Shankavaram UT, Reinhold WC, Nishizuka S, Major S, Morita D, et al. (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Mol Cancer Ther* 6: 820-832.
158. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, et al. (2008) The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* 133: 704-715.
159. Wang L, Cao H, Lu N, Liu L, Wang B, et al. (2013) Berberine inhibits proliferation and down-regulates epidermal growth factor receptor through activation of Cbl in colon tumor cells. *PLoS One* 8: e56666.
160. Liu Q, Xu X, Zhao M, Wei Z, Li X, et al. (2015) Berberine induces senescence of human glioblastoma cells by downregulating the EGFR-MEK-ERK signaling pathway. *Mol Cancer Ther* 14: 355-363.
161. Huang ZH, Zheng HF, Wang WL, Wang Y, Zhong LF, et al. (2015) Berberine targets epidermal growth factor receptor signaling to suppress prostate cancer proliferation in vitro. *Mol Med Rep* 11: 2125-2128.
162. Dziadziuszko R, Witta SE, Cappuzzo F, Park S, Tanaka K, et al. (2006) Epidermal growth factor receptor messenger RNA expression, gene dosage, and gefitinib sensitivity in non-small cell lung cancer. *Clin Cancer Res* 12: 3078-3084.
163. Nishimura T, Nakamura K, Yamashita S, Ikeda S, Kigure K, et al. (2015) Effect of the molecular targeted drug, erlotinib, against endometrial cancer expressing high levels of epidermal growth factor receptor. *BMC Cancer* 15: 957.
164. Ferrer-Soler L, Vazquez-Martin A, Brunet J, Menendez JA, De Llorens R, et al. (2007) An update of the mechanisms of resistance to EGFR-tyrosine kinase inhibitors in breast cancer: Gefitinib (Iressa) -induced changes in the expression and nucleo-cytoplasmic trafficking of HER-ligands (Review). *Int J Mol Med* 20: 3-10.

165. Khambata-Ford S, Garrett CR, Meropol NJ, Basik M, Harbison CT, et al. (2007) Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab. *J Clin Oncol* 25: 3230-3237.
166. Xu P, Derynck R (2010) Direct activation of TACE-mediated ectodomain shedding by p38 MAP kinase regulates EGF receptor-dependent cell proliferation. *Mol Cell* 37: 551-566.
167. Ma P, Fu Y, Chen M, Jing Y, Wu J, et al. (2016) Adaptive and Acquired Resistance to EGFR Inhibitors Converge on the MAPK Pathway. *Theranostics* 6: 1232-1243.
168. Yothers G, O'Connell MJ, Lee M, Lopatin M, Clark-Langone KM, et al. (2013) Validation of the 12-gene colon cancer recurrence score in NSABP C-07 as a predictor of recurrence in patients with stage II and III colon cancer treated with fluorouracil and leucovorin (FU/LV) and FU/LV plus oxaliplatin. *J Clin Oncol* 31: 4512-4519.
169. You YN, Rustin RB, Sullivan JD (2015) Oncotype DX((R)) colon cancer assay for prediction of recurrence risk in patients with stage II and III colon cancer: A review of the evidence. *Surg Oncol* 24: 61-66.
170. Kannarkatt J, Joseph J, Kurniali PC, Al-Janadi A, Hrinchenko B (2017) Adjuvant Chemotherapy for Stage II Colon Cancer: A Clinical Dilemma. *Journal of Oncology Practice* 13: 233-241.
171. Genomic Health Colon Recurrence Score Result, MMR Status, and T-Stage.
172. Srivastava G, Renfro LA, Behrens RJ, Lopatin M, Chao C, et al. (2014) Prospective multicenter study of the impact of oncotype DX colon cancer assay results on treatment recommendations in stage II colon cancer patients. *Oncologist* 19: 492-497.
173. Renfro LA, Zhang N, Lopatin M, Chao C, Alberts SR (2017) Prospective Evaluation of a 12-Gene Assay on Patient Treatment Decisions and Physician Confidence in Mismatch Repair Proficient Stage IIA Colon Cancer. *Clin Colorectal Cancer* 16: 23-30.
174. Gras C, Eiz-Vesper B, Jaimes Y, Immenschuh S, Jacobs R, et al. (2014) Secreted semaphorin 5A activates immune effector cells and is a biomarker for rheumatoid arthritis. *Arthritis Rheumatol* 66: 1461-1471.
175. Batzer AG, Rotin D, Urena JM, Skolnik EY, Schlessinger J (1994) Hierarchy of binding sites for Grb2 and Shc on the epidermal growth factor receptor. *Mol Cell Biol* 14: 5192-5201.
176. Phuchareon J, McCormick F, Eisele DW, Tetsu O (2015) EGFR inhibition evokes innate drug resistance in lung cancer cells by preventing Akt activity and thus inactivating Ets-1 function. *Proc Natl Acad Sci U S A* 112: E3855-3863.
177. Gulhati P, Bowen KA, Liu J, Stevens PD, Rychahou PG, et al. (2011) mTORC1 and mTORC2 regulate EMT, motility, and metastasis of colorectal cancer via RhoA and Rac1 signaling pathways. *Cancer Res* 71: 3246-3256.
178. Roper J, Richardson MP, Wang WV, Richard LG, Chen W, et al. (2011) The dual PI3K/mTOR inhibitor NVP-BE235 induces tumor regression in a genetically engineered mouse model of PIK3CA wild-type colorectal cancer. *PLoS One* 6: e25132.
179. Lamkanfi M, Kanneganti TD (2010) Caspase-7: a protease involved in apoptosis and inflammation. *Int J Biochem Cell Biol* 42: 21-24.
180. Harris RE, Beebe-Donk J, Doss H, Burr Doss D (2005) Aspirin, ibuprofen, and other non-steroidal anti-inflammatory drugs in cancer prevention: a critical review of non-selective COX-2 blockade (review). *Oncol Rep* 13: 559-583.
181. Cuzick J, Otto F, Baron JA, Brown PH, Burn J, et al. (2009) Aspirin and non-steroidal anti-inflammatory drugs for cancer prevention: an international consensus statement. *Lancet Oncol* 10: 501-507.
182. Elwood PC, Gallagher AM, Duthie GG, Mur LA, Morgan G (2009) Aspirin, salicylates, and cancer. *Lancet* 373: 1301-1309.
183. Ulrich CM, Bigler J, Potter JD (2006) Non-steroidal anti-inflammatory drugs for cancer prevention: promise, perils and pharmacogenetics. *Nature Reviews Cancer* 6: 130-140.
184. Gonzalez-Perez A, Rodriguez LAG, Lopez-Ridaura R (2003) Effects of non-steroidal anti-inflammatory drugs on cancer sites other than the colon and rectum: a meta-analysis. *BMC Cancer* 3.
185. Drak Alsibai K, Meseure D (2017) Tumor microenvironment and noncoding RNAs as co-drivers of epithelial-mesenchymal transition and cancer metastasis. *Dev Dyn*.
186. Hu H, Hang JJ, Han T, Zhuo M, Jiao F, et al. (2016) The M2 phenotype of tumor-associated macrophages in the stroma confers a poor prognosis in pancreatic cancer. *Tumour Biol* 37: 8657-8664.
187. Zhang B, Yao G, Zhang Y, Gao J, Yang B, et al. (2011) M2-polarized tumor-associated macrophages are associated with poor prognoses resulting from accelerated lymphangiogenesis in lung adenocarcinoma. *Clinics (Sao Paulo)* 66: 1879-1886.
188. Son B, Lee S, Youn H, Kim E, Kim W, et al. (2017) The role of tumor microenvironment in therapeutic resistance. *Oncotarget* 8: 3933-3945.
189. Lopez-Novoa JM, Nieto MA (2009) Inflammation and EMT: an alliance towards organ fibrosis and cancer progression. *EMBO Mol Med* 1: 303-314.
190. Xu ZY, Chen JS, Shu YQ (2010) Gene expression profile towards the prediction of patient survival of gastric cancer. *Biomed Pharmacother* 64: 133-139.
191. Lee J, Sohn I, Do IG, Kim KM, Park SH, et al. (2014) Nanostring-based multigene assay to predict recurrence for gastric cancer patients after surgery. *PLoS One* 9: e90133.
192. Li RK, Zhao WY, Fang F, Zhuang C, Zhang XX, et al. (2015) Lysyl oxidase-like 4 (LOXL4) promotes proliferation and metastasis of gastric cancer via FAK/Src pathway. *J Cancer Res Clin Oncol* 141: 269-281.
193. Sato Y, Inokuchi M, Takagi Y, Otsuki S, Fujimori Y, et al. (2015) Relationship between expression of IGFBP7 and clinicopathological variables in gastric cancer. *J Clin Pathol* 68: 795-801.
194. Vagenas K, Spyropoulos C, Gavala V, Tsamandas AC (2007) TGFbeta1, TGFbeta2, and TGFbeta3 protein expression in gastric carcinomas: correlation with prognostics factors and patient survival. *J Surg Res* 139: 182-188.

195. O'Hayre M, Salanga CL, Handel TM, Allen SJ (2008) Chemokines and cancer: migration, intracellular signalling and intercellular communication in the microenvironment. *Biochem J* 409: 635-649.
196. Yu PF, Huang Y, Han YY, Lin LY, Sun WH, et al. (2017) TNF α -activated mesenchymal stromal cells promote breast cancer metastasis by recruiting CXCR2⁺ neutrophils. *Oncogene* 36: 482-490.
197. Zhao J, Ou B, Han D, Wang P, Zong Y, et al. (2017) Tumor-derived CXCL5 promotes human colorectal cancer metastasis through activation of the ERK/Elk-1/Snail and AKT/GSK3 β /beta-catenin pathways. *Mol Cancer* 16: 70.
198. National Cancer Institute (April 17, 2017) Targeted Cancer Therapies. NIH.
199. Yauch RL, Januario T, Eberhard DA, Cavet G, Zhu W, et al. (2005) Epithelial versus mesenchymal phenotype determines in vitro sensitivity and predicts clinical activity of erlotinib in lung cancer patients. *Clin Cancer Res* 11: 8686-8698.
200. Thomson S, Buck E, Petti F, Griffin G, Brown E, et al. (2005) Epithelial to mesenchymal transition is a determinant of sensitivity of non-small-cell lung carcinoma cell lines and xenografts to epidermal growth factor receptor inhibition. *Cancer Res* 65: 9455-9462.
201. Yoshida T, Song L, Bai Y, Kinose F, Li J, et al. (2016) ZEB1 Mediates Acquired Resistance to the Epidermal Growth Factor Receptor-Tyrosine Kinase Inhibitors in Non-Small Cell Lung Cancer. *PLoS One* 11: e0147344.
202. Fuchs BC, Fujii T, Dorfman JD, Goodwin JM, Zhu AX, et al. (2008) Epithelial-to-mesenchymal transition and integrin-linked kinase mediate sensitivity to epidermal growth factor receptor inhibition in human hepatoma cells. *Cancer Res* 68: 2391-2399.
203. Waring P, Tie J, Maru D, Karapetis CS (2016) RAS Mutations as Predictive Biomarkers in Clinical Management of Metastatic Colorectal Cancer. *Clin Colorectal Cancer* 15: 95-103.
204. Lou E, D'Souza D, Nelson AC (2017) Therapeutic Response of Metastatic Colorectal Cancer Harboring a KRAS Missense Mutation After Combination Chemotherapy With the EGFR Inhibitor Panitumumab. *J Natl Compr Canc Netw* 15: 427-432.
205. Kim BJ, Jeong JH, Kim JH, Kim HS, Jang HJ (2017) The role of targeted agents in the adjuvant treatment of colon cancer: a meta-analysis of randomized phase III studies and review. *Oncotarget*.
206. Seamon K, Daly JW (1981) Activation of adenylate cyclase by the diterpene forskolin does not require the guanine nucleotide regulatory protein. *J Biol Chem* 256: 9799-9801.
207. Sapio L, Gallo M, Illiano M, Chiosi E, Naviglio D, et al. (2017) The Natural cAMP Elevating Compound Forskolin in Cancer Therapy: Is It Time? *J Cell Physiol* 232: 922-927.
208. Hansen LA, Sigman CC, Andreola F, Ross SA, Kelloff GJ, et al. (2000) Retinoids in chemoprevention and differentiation therapy. *Carcinogenesis* 21: 1271-1279.
209. Adachi Y, Itoh F, Yamamoto H, Iku S, Matsuno K, et al. (2001) Retinoic acids reduce matrilysin (matrix metalloproteinase 7) and inhibit tumor cell invasion in human colon cancer. *Tumour Biol* 22: 247-253.
210. Woo YJ, Jang KL (2012) All-trans retinoic acid activates E-cadherin expression via promoter hypomethylation in the human colon carcinoma HCT116 cells. *Biochem Biophys Res Commun* 425: 944-949.
211. Cui J, Gong M, He Y, Li Q, He T, et al. (2016) All-trans retinoic acid inhibits proliferation, migration, invasion and induces differentiation of hepa1-6 cells through reversing EMT in vitro. *Int J Oncol* 48: 349-357.
212. Guan J, Zhang H, Wen Z, Gu Y, Cheng Y, et al. (2014) Retinoic acid inhibits pancreatic cancer cell migration and EMT through the downregulation of IL-6 in cancer associated fibroblast cells. *Cancer Lett* 345: 132-139.
213. Sequist LV, Waltman BA, Dias-Santagata D, Digumarthy S, Turke AB, et al. (2011) Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors. *Sci Transl Med* 3: 75ra26.
214. Eng J, Woo KM, Sima CS, Plodkowski A, Hellmann MD, et al. (2015) Impact of Concurrent PIK3CA Mutations on Response to EGFR Tyrosine Kinase Inhibition in EGFR-Mutant Lung Cancers and on Prognosis in Oncogene-Driven Lung Adenocarcinomas. *J Thorac Oncol* 10: 1713-1719.
215. Philpott C, Tovell H, Frayling IM, Cooper DN, Upadhyaya M (2017) The NF1 somatic mutational landscape in sporadic human cancers. *Hum Genomics* 11: 13.
216. Hsu HC, Thiam TK, Lu YJ, Yeh CY, Tsai WS, et al. (2016) Mutations of KRAS/NRAS/BRAF predict cetuximab resistance in metastatic colorectal cancer patients. *Oncotarget* 7: 22257-22270.
217. Douillard JY, Siena S, Cassidy J, Tabernero J, Burkes R, et al. (2014) Final results from PRIME: randomized phase III study of panitumumab with FOLFOX4 for first-line treatment of metastatic colorectal cancer. *Ann Oncol* 25: 1346-1355.

7.APPENDIX

7.1. APPENDIX A

TCGA Proteomic analysis comparing BAD vs GOOD prognosis groups

Protein	Ttest p values	log fold (bad/good)
Shc_pY317-R-NA	0.000	-0.247
EGFR_pY1068-R-V	0.000	-0.458
Caspase-7_cleavedD198-R-C	0.000	1.244
Rad51-M-C	0.001	0.157
MEK1-R-V	0.003	0.216
MIG-6-M-V	0.004	0.084
HER2_pY1248-R-V	0.004	-0.166
NF2-R-C	0.004	0.191
c-Kit-R-V	0.006	-0.634
Src-M-V	0.008	-0.195
YAP_pS127-R-C	0.009	-0.214
ERK2-R-NA	0.011	0.189
Chk2_pT68-R-C	0.012	-0.150
AIB1-M-V	0.013	-0.168
Bcl-2-R-NA	0.013	0.286
PI3K-p85-R-V	0.015	0.118
GSK3_pS9-R-V	0.015	-0.290
c-Jun_pS73-R-C	0.017	-0.143
Src_pY527-R-V	0.017	-0.294
p90RSK_pT359_S363-R-C	0.017	-0.142
4E-BP1-R-V	0.021	0.164
DJ-1-R-C	0.025	0.159
Lck-R-V	0.026	0.268
Annexin_I-R-V	0.028	0.301
EGFR_pY992-R-V	0.029	-0.215
P-Cadherin-R-C	0.032	0.132
p38_MAPK-R-C	0.033	0.140
NF-kB-p65_pS536-R-C	0.033	-0.412
GSK3-alpha-beta_pS21_S9-R-V	0.041	-0.235
Cyclin_E1-M-V	0.041	0.192
Rad50-M-C	0.044	-0.169
mTOR-R-V	0.048	0.117
LKB1-M-NA	0.049	0.070
GSK3-alpha-beta-M-V	0.049	0.098
p21-R-C	0.050	0.098

7.2. APPENDIX B

63 gensesets enriched in 'UP' group of GSE15459 (FDR q value<0.25)

NAME	FDR q-val
REGULATION_OF_TRANSCRIPTION	0.169
SKELETAL_DEVELOPMENT	0.170
NERVOUS_SYSTEM_DEVELOPMENT	0.170
NEGATIVE_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	0.171
VESICLE_MEDIATED_TRANSPORT	0.171
POSITIVE_REGULATION_OF_TRANSCRIPTION	0.172
ENZYME_LINKED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	0.174
REGULATION_OF_GROWTH	0.174
NEURON_DIFFERENTIATION	0.174
GROWTH	0.175
GENERATION_OF_NEURONS	0.176
GTPASE_ACTIVATOR_ACTIVITY	0.177
POSITIVE_REGULATION_OF_NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_M	0.179
NEURITE_DEVELOPMENT	0.181
TRANSCRIPTION_COFACTOR_ACTIVITY	0.181
CELL_PROJECTION	0.186
PROTEIN_IMPORT	0.186
PYROPHOSPHATASE_ACTIVITY	0.187
ANATOMICAL_STRUCTURE_MORPHOGENESIS	0.187
TRANSCRIPTION_COREPRESSOR_ACTIVITY	0.187
SMALL_GTPASE_REGULATOR_ACTIVITY	0.187
SMALL_GTPASE_MEDIATED_SIGNAL_TRANSDUCTION	0.187
ACTIN_FILAMENT_BASED_PROCESS	0.187
REGULATION_OF_TRANSPORT	0.187
NUCLEOSIDE_TRIPHOSPHATASE_ACTIVITY	0.187
NEGATIVE_REGULATION_OF_TRANSCRIPTION	0.188
NEGATIVE_REGULATION_OF_NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_	0.188
CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.189
NEGATIVE_REGULATION_OF_CELL_CYCLE	0.190
CELL_SURFACE	0.190
REGULATION_OF_RNA_METABOLIC_PROCESS	0.191
NEGATIVE_REGULATION_OF_CELLULAR_METABOLIC_PROCESS	0.191
NEGATIVE_REGULATION_OF_METABOLIC_PROCESS	0.192
TRANSCRIPTION_REPRESSOR_ACTIVITY	0.192
ACTIN_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.192
REGULATION_OF_TRANSCRIPTIONDNA_DEPENDENT	0.195
HYDROLASE_ACTIVITY_ACTING_ON_ACID_ANHYDRIDES	0.195
PROTEIN_TYROSINE_KINASE_ACTIVITY	0.195

TRANSCRIPTION_FACTOR_BINDING	0.197
ACTIN_BINDING	0.197
TRANSCRIPTION_COACTIVATOR_ACTIVITY	0.201
REGULATION_OF_SIGNAL_TRANSDUCTION	0.201
EXTRACELLULAR_MATRIX	0.204
PROTEINACEOUS_EXTRACELLULAR_MATRIX	0.211
CYTOSKELETAL_PROTEIN_BINDING	0.213
GTPASE_REGULATOR_ACTIVITY	0.213
REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	0.220
NEURON_DEVELOPMENT	0.222
EXTRACELLULAR_MATRIX_PART	0.224
REGULATION_OF_BODY_FLUID_LEVELS	0.226
REGULATION_OF_CELL_DIFFERENTIATION	0.227
NEUROGENESIS	0.228
RAS_PROTEIN_SIGNAL_TRANSDUCTION	0.228
ORGAN_MORPHOGENESIS	0.228
POSITIVE_REGULATION_OF_TRANSCRIPTIONDNA_DEPENDENT	0.230
NUCLEAR_TRANSPORT	0.231
NEGATIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	0.233
ENZYME_ACTIVATOR_ACTIVITY	0.234
NUCLEOCYTOPLASMIC_TRANSPORT	0.235
CELL_CYCLE_ARREST_GO_0007050	0.235
POSITIVE_REGULATION_OF_RNA_METABOLIC_PROCESS	0.237
WOUND_HEALING	0.239
TRANSMEMBRANE_RECEPTOR_PROTEIN_KINASE_ACTIVITY	0.241

7.3. APPENDIX C

188 gensesets enriched in 'UP' group of GSE62254 (FDR q value<0.25)

NAME	FDR q-val
NEGATIVE_REGULATION_OF_METABOLIC_PROCESS	0.060
EXTRACELLULAR_REGION	0.061
ACTIN_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.061
ACTIN_BINDING	0.061
CENTRAL_NERVOUS_SYSTEM_DEVELOPMENT	0.062
POSITIVE_REGULATION_OF_TRANSCRIPTIONDNA_DEPENDENT	0.063
NEGATIVE_REGULATION_OF_CELLULAR_METABOLIC_PROCESS	0.063
PROTEIN_TYROSINE_KINASE_ACTIVITY	0.064
MEMBRANE_ORGANIZATION_AND_BIOGENESIS	0.064
GROWTH	0.064
POSITIVE_REGULATION_OF_NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_ME TABOLIC_PROCESS	0.064
NEGATIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	0.064
ENZYME_REGULATOR_ACTIVITY	0.065
CELL_CYCLE_ARREST_GO_0007050	0.065
SMALL_GTPASE_REGULATOR_ACTIVITY	0.065
ACTIN_FILAMENT_BASED_PROCESS	0.065
NEGATIVE_REGULATION_OF_CELL_CYCLE	0.065
ENZYME_ACTIVATOR_ACTIVITY	0.065
REGULATION_OF_GROWTH	0.065
SYNAPTIC_TRANSMISSION	0.065
ENZYME_LINKED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	0.065
NERVOUS_SYSTEM_DEVELOPMENT	0.065
NEURITE_DEVELOPMENT	0.066
GTPASE_REGULATOR_ACTIVITY	0.066
NEGATIVE_REGULATION_OF_TRANSCRIPTION	0.066
CYTOSKELETON	0.066
ORGAN_MORPHOGENESIS	0.066
POSITIVE_REGULATION_OF_RNA_METABOLIC_PROCESS	0.066
EXTRACELLULAR_REGION_PART	0.067
NUCLEOSIDE_TRIPHOSPHATASE_ACTIVITY	0.067
CELL_PROJECTION	0.067
NEUROGENESIS	0.067
GTPASE_ACTIVATOR_ACTIVITY	0.068
CYTOSKELETAL_PROTEIN_BINDING	0.068
ACTIN_CYTOSKELETON	0.068
GENERATION_OF_NEURONS	0.068
ANATOMICAL_STRUCTURE_MORPHOGENESIS	0.069
CELL_SURFACE	0.069

NEURON_DEVELOPMENT	0.071
NEURON_DIFFERENTIATION	0.071
POSITIVE_REGULATION_OF_TRANSCRIPTION	0.072
NEGATIVE_REGULATION_OF_NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_M ETABOLIC_PROCESS	0.076
ANATOMICAL_STRUCTURE_FORMATION	0.077
TRANSMEMBRANE_RECEPTOR_PROTEIN_KINASE_ACTIVITY	0.077
TRANSMISSION_OF_NERVE_IMPULSE	0.079
REGULATION_OF_TRANSCRIPTION	0.080
PROTEINACEOUS_EXTRACELLULAR_MATRIX	0.083
CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.084
PYROPHOSPHATASE_ACTIVITY	0.085
NEUROLOGICAL_SYSTEM_PROCESS	0.085
ANTI_APOPTOSIS	0.086
PHOSPHOPROTEIN_PHOSPHATASE_ACTIVITY	0.086
EXTRACELLULAR_MATRIX	0.086
REGULATION_OF_TRANSCRIPTIONDNA_DEPENDENT	0.086
ION_BINDING	0.086
REGULATION_OF_SIGNAL_TRANSDUCTION	0.087
EXTRACELLULAR_MATRIX_PART	0.087
HYDROLASE_ACTIVITY_ACTING_ON_ACID_ANHYDRIDES	0.087
REGULATION_OF_BIOLOGICAL_QUALITY	0.087
REGULATION_OF_CELLULAR_COMPONENT_ORGANIZATION_AND_BIOGENESIS	0.087
REGULATION_OF_RNA_METABOLIC_PROCESS	0.088
REGULATION_OF_DEVELOPMENTAL_PROCESS	0.088
WOUND_HEALING	0.088
CELL_MIGRATION	0.088
RECEPTOR_BINDING	0.089
REGULATION_OF_BODY_FLUID_LEVELS	0.090
REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	0.092
CELLULAR_COMPONENT_ASSEMBLY	0.092
CATION_BINDING	0.093
PROTEIN_TYROSINE_PHOSPHATASE_ACTIVITY	0.093
PROTEIN_KINASE_ACTIVITY	0.094
NEGATIVE_REGULATION_OF_PROGRAMMED_CELL_DEATH	0.094
MACROMOLECULAR_COMPLEX_ASSEMBLY	0.094
REGULATION_OF_CELL_DIFFERENTIATION	0.094
REGULATION_OF_TRANSPORT	0.094
VESICLE_MEDIATED_TRANSPORT	0.095
NEGATIVE_REGULATION_OF_CELL_PROLIFERATION	0.095
PROTEIN_COMPLEX_BINDING	0.097
TRANSCRIPTION_COREPRESSOR_ACTIVITY	0.100
TRANSMEMBRANE_RECEPTOR_ACTIVITY	0.100
STRUCTURAL_MOLECULE_ACTIVITY	0.101
TRANSITION_METAL_ION_BINDING	0.102

NEGATIVE_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	0.102
TRANSCRIPTION_REPRESSOR_ACTIVITY	0.103
TRANSMEMBRANE_RECEPTOR_PROTEIN_TYROSINE_KINASE_SIGNALING_PATHWAY	0.103
DEPHOSPHORYLATION	0.103
POSITIVE_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	0.103
REGULATION_OF_CELL_PROLIFERATION	0.103
ENZYME_INHIBITOR_ACTIVITY	0.103
CELL_CELL_SIGNALING	0.104
NEGATIVE_REGULATION_OF_APOPTOSIS	0.104
MAPKKK_CASCADE_GO_0000165	0.105
SMALL_GTPASE_MEDIATED_SIGNAL_TRANSDUCTION	0.105
POSITIVE_REGULATION_OF_CELLULAR_METABOLIC_PROCESS	0.105
PROTEIN_AMINO_ACID_DEPHOSPHORYLATION	0.105
PHOSPHORIC_ESTER_HYDROLASE_ACTIVITY	0.105
POSITIVE_REGULATION_OF_METABOLIC_PROCESS	0.106
REGULATION_OF_MULTICELLULAR_ORGANISMAL_PROCESS	0.106
G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	0.106
RAS_PROTEIN_SIGNAL_TRANSDUCTION	0.107
CYTOPLASMIC_MEMBRANE_BOUND_VESICLE	0.109
EXTRACELLULAR_SPACE	0.109
MUSCLE_DEVELOPMENT	0.110
POSITIVE_REGULATION_OF_CELL_PROLIFERATION	0.114
RESPONSE_TO_EXTERNAL_STIMULUS	0.118
TRANSCRIPTION_ACTIVATOR_ACTIVITY	0.119
CYTOPLASMIC_VESICLE	0.122
MEMBRANE_BOUND_VESICLE	0.122
CARBOHYDRATE_BINDING	0.122
RESPONSE_TO_WOUNDING	0.127
ZINC_ION_BINDING	0.130
REGULATION_OF_RESPONSE_TO_STIMULUS	0.131
SKELETAL_DEVELOPMENT	0.133
IDENTICAL_PROTEIN_BINDING	0.137
TRANSCRIPTION_FACTOR_BINDING	0.137
RECEPTOR_COMPLEX	0.138
CELL_PROLIFERATION_GO_0008283	0.145
NEGATIVE_REGULATION_OF_RNA_METABOLIC_PROCESS	0.148
PROTEIN_DOMAIN_SPECIFIC_BINDING	0.148
KINASE_ACTIVITY	0.149
POST_TRANSLATIONAL_PROTEIN_MODIFICATION	0.150
PHOSPHOTRANSFERASE_ACTIVITY_ALCOHOL_GROUP_AS_ACCEPTOR	0.153
TRANSCRIPTION_COFACTOR_ACTIVITY	0.154
TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	0.154
VESICLE	0.156
CELL_JUNCTION	0.156
PHOSPHORIC_MONOESTER_HYDROLASE_ACTIVITY	0.157

TRANSCRIPTION_FACTOR_ACTIVITY	0.158
HEMOPOIESIS	0.159
CALCIUM_ION_BINDING	0.159
PERINUCLEAR_REGION_OF_CYTOPLASM	0.160
CELL_CELL_ADHESION	0.160
METAL_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.166
G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY	0.166
SECOND_MESSENGER_MEDIATED_SIGNALING	0.167
CAMP_MEDIATED_SIGNALING	0.169
PURINE_NUCLEOTIDE_BINDING	0.173
G_PROTEIN_SIGNALING_COUPLED_TO_CAMP_NUCLEOTIDE_SECOND_MESSENGER	0.174
HEMOPOIETIC_OR_LYMPHOID_ORGAN_DEVELOPMENT	0.174
PROTEIN_KINASE_CASCADE	0.181
CATION_CHANNEL_ACTIVITY	0.182
BEHAVIOR	0.182
RESPONSE_TO_STRESS	0.188
ESTABLISHMENT_OF_CELLULAR_LOCALIZATION	0.190
SENSORY_PERCEPTION	0.193
GATED_CHANNEL_ACTIVITY	0.196
IMMUNE_SYSTEM_DEVELOPMENT	0.197
HYDROLASE_ACTIVITY_ACTING_ON_ESTER_BONDS	0.198
CELLULAR_LOCALIZATION	0.198
CELLULAR_HOMEOSTASIS	0.199
CATION_TRANSPORT	0.202
HOMEOSTATIC_PROCESS	0.216
CELLULAR_CATION_HOMEOSTASIS	0.217
GTPASE_ACTIVITY	0.217
REGULATION_OF_PROGRAMMED_CELL_DEATH	0.218
CYTOSOL	0.220
NEGATIVE_REGULATION_OF_TRANSCRIPTION_DNA_DEPENDENT	0.222
PROTEIN_AMINO_ACID_PHOSPHORYLATION	0.226
ION_HOMEOSTASIS	0.226
ORGANELLE_ORGANIZATION_AND_BIOGENESIS	0.227
RHODOPSIN_LIKE_RECEPTOR_ACTIVITY	0.228
CHEMICAL_HOMEOSTASIS	0.229
LOCOMOTORY_BEHAVIOR	0.232
PURINE_RIBONUCLEOTIDE_BINDING	0.237
REGULATION_OF_APOPTOSIS	0.238
RESPONSE_TO_CHEMICAL_STIMULUS	0.239
ENDOPLASMIC_RETICULUM	0.239
CATION_HOMEOSTASIS	0.240
ENDOMEMBRANE_SYSTEM	0.240
CYTOSKELETAL_PART	0.240
INTRACELLULAR_TRANSPORT	0.240
POSITIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	0.241

TISSUE_DEVELOPMENT	0.241
METAL_ION_TRANSPORT	0.242
ELECTRON_TRANSPORT_GO_0006118	0.242
CYCLIC_NUCLEOTIDE_MEDIATED_SIGNALING	0.242
EMBRYONIC_DEVELOPMENT	0.242
G_PROTEIN_SIGNALING_COUPLED_TO_CYCLIC_NUCLEOTIDE_SECOND_MESSENGER	0.242
TRANSFERASE_ACTIVITY_TRANSFERRING_PHOSPHORUS_CONTAINING_GROUPS	0.242
POSITIVE_REGULATION_OF_CATALYTIC_ACTIVITY	0.243
REPRODUCTION	0.243
RNA_POLYMERASE_II_TRANSCRIPTION_FACTOR_ACTIVITY	0.243
PROTEOLYSIS	0.244
INFLAMMATORY_RESPONSE	0.244
PHOSPHORYLATION	0.245
PROGRAMMED_CELL_DEATH	0.245
NUCLEOTIDE_BINDING	0.246
PROTEIN_COMPLEX_ASSEMBLY	0.249

7.4. APPENDIX D

List of 134 probesets used for molecular classification of gastric cancer tumor datasets

Gene Symbol	Probeset
CDC6	203968_s_at
TNC	216005_at
MYH11	201495_x_at
PRUNE2	212806_at
TCEAL2	211276_at
HAND2-AS1	219791_s_at
PLN	204940_at
PLN	204938_s_at
PLN	204939_s_at
FHL1	201539_s_at
FHL1	210298_x_at
FHL1	210299_s_at
SORBS1	218087_s_at
PRUNE2	212805_at
ACTG2	202274_at
MYH11	201496_x_at
DES	202222_s_at
SYNM	212730_at
CNN1	203951_at
	214027_x_at
LMOD1	203766_s_at
MYL9	201058_s_at
TPM2	204083_s_at
SCRG1	205475_at
CALD1	201617_x_at
FERMT2	209209_s_at
DMD	203881_s_at
PDZRN3	212915_at
SYNC	221276_s_at
EML1	204796_at
TMEM100	219230_at
PTN	211737_x_at
EPHA3	206070_s_at
BAG2	209406_at
TAC1	206552_s_at
AGTR1	205357_s_at
TSPYL5	213122_at
GPM6B	209170_s_at
GPM6B	209167_at
BCHE	205433_at
ATP1A2	203296_s_at
COL4A5	213110_s_at
COL4A6	213992_at
NBEA	221207_s_at
NFASC	213438_at
LDB3	213371_at
GREM1	218468_s_at
GREM1	218469_at

Gene Symbol	Probeset
ZFPM2	219778_at
ANK2	202920_at
C7	202992_at
OGN	218730_s_at
IGF1	209541_at
IGF1	209540_at
SPON1	213994_s_at
COL14A1	212865_s_at
MEOX2	206201_s_at
LEPR	209894_at
KLF9	203543_s_at
CAV2	203324_s_at
SEMA3C	203789_s_at
ID4	209292_at
HTR2B	206638_at
	202350_s_at
APOD	201525_at
MT1M	217546_at
C8orf4	218541_s_at
MS4A4A	219607_s_at
CYP1B1	202437_s_at
CYP1B1	202436_s_at
ITM2A	202746_at
IFI44L	204439_at
NTRK2	221796_at
ADIPOQ	207175_at
LPL	203548_s_at
FABP4	203980_at
CD36	206488_s_at
CD36	209555_s_at
PCOLCE2	219295_s_at
PTX3	206157_at
LYVE1	220037_s_at
EPB41L3	212681_at
IL33	209821_at
SFRP1	202037_s_at
SFRP1	202036_s_at
EDNRB	204273_at
CHI3L1	209395_at
CHI3L1	209396_s_at
RARRES1	206392_s_at
RARRES1	221872_at
EPYC	206439_at
RBP4	219140_s_at
PLA2G2A	203649_s_at
	208607_s_at
	214456_x_at
	208470_s_at

AHNAK2	212992_at
XIST	214218_s_at
XIST	221728_x_at
CCL18	32128_at
RGS13	210258_at
TOX	204529_s_at
HLA-DRB4	209728_at
ADH1B	209613_s_at
ADH1B	209612_s_at
GHR	205498_at
SRPX	204955_at
ABCA8	204719_at
CHRD1	209763_at
CFD	205382_s_at
CLU	222043_at
DPT	207977_s_at
DPT	213068_at
PRKAR2B	203680_at
CXCL12	209687_at

HP	206697_s_at
IGF2	202409_at
MFAP5	213764_s_at
MFAP5	213765_at
LOX	215446_s_at
GAS1	204457_s_at
PDGFRL	205226_at
COMP	205713_s_at
SFRP4	204052_s_at
ITGBL1	205422_s_at
PTGS2	204748_at
TNFAIP6	206025_s_at
SELE	206211_at
EGR3	206115_at
FOSB	202768_at
NR4A2	216248_s_at
SERPINE1	202628_s_at
ACKR3	212977_at
HBB	209116_x_at

7.5. APPENDIX E

134 genesets enriched in HIGH group in GSE29272 dataset (FDR q value< 0.25)

NAME	FDR q-val
PROTEINACEOUS_EXTRACELLULAR_MATRIX	0
EXTRACELLULAR_MATRIX	0
EXTRACELLULAR_MATRIX_PART	0
EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT	1.68E-04
BASEMENT_MEMBRANE	2.10E-04
INTEGRIN_BINDING	7.17E-04
STRUCTURAL_CONSTITUENT_OF_MUSCLE	0.001
AXONOGENESIS	0.001
MUSCLE_DEVELOPMENT	0.001
EXTRACELLULAR_REGION_PART	0.002
CELLULAR_MORPHOGENESIS_DURING_DIFFERENTIATION	0.003
AXON_GUIDANCE	0.004
BASAL_LAMINA	0.005
EXTRACELLULAR_REGION	0.005
CELL_SUBSTRATE_ADHESION	0.005
NEURITE_DEVELOPMENT	0.005
CONTRACTILE_FIBER	0.006
INTERLEUKIN_BINDING	0.006
CELL_MATRIX_ADHESION	0.008
PROTEIN_COMPLEX_BINDING	0.008
CONTRACTILE_FIBER_PART	0.010
REGULATION_OF_HEART_CONTRACTION	0.015
BASOLATERAL_PLASMA_MEMBRANE	0.015
ACTIN_CYTOSKELETON	0.016
INTERLEUKIN_RECEPTOR_ACTIVITY	0.017
ACTIN_BINDING	0.018
HEMATOPOIETIN_INTERFERON_CLASSD200_DOMAIN_CYTOKINE_RECEPTOR_ACTIVITY	0.018
G_PROTEIN_SIGNALING_COUPLED_TO_IP3_SECOND_MESSENGERPHOSPHOLIPASE_C_ACTIVATI	0.019
COAGULATION	0.021
MYOFIBRIL	0.021
BLOOD_COAGULATION	0.023
CELL_MIGRATION	0.026
LIPID_RAFT	0.027
LIPOPROTEIN_BINDING	0.028
G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY	0.030
PHAGOCYTOSIS	0.030
CYTOSOLIC_PART	0.033
COLLAGEN	0.042
NEURON_DEVELOPMENT	0.043
SYSTEM_PROCESS	0.043
TRANSMEMBRANE_RECEPTOR_ACTIVITY	0.043

ACTIN_POLYMERIZATION_AND_OR_DEPOLYMERIZATION	0.044
HEMOSTASIS	0.045
RHODOPSIN_LIKE_RECEPTOR_ACTIVITY	0.045
GROWTH_FACTOR_BINDING	0.045
WOUND_HEALING	0.045
PHOSPHORIC_DIESTER_HYDROLASE_ACTIVITY	0.046
INSULIN_RECEPTOR_SIGNALING_PATHWAY	0.046
NEURON_DIFFERENTIATION	0.050
REGULATION_OF_MUSCLE_CONTRACTION	0.055
STRUCTURAL_MOLECULE_ACTIVITY	0.055
SKELETAL_DEVELOPMENT	0.059
PHOSPHOINOSITIDE_MEDIATED_SIGNALING	0.059
REGULATION_OF_MULTICELLULAR_ORGANISMAL_PROCESS	0.063
CYTOSKELETAL_PROTEIN_BINDING	0.065
REGULATION_OF_BODY_FLUID_LEVELS	0.074
PATTERN_BINDING	0.075
CYTOKINE_BINDING	0.082
CAMP_MEDIATED_SIGNALING	0.083
CALCIUM_ION_TRANSPORT	0.085
GENERATION_OF_NEURONS	0.087
CELLULAR_CATION_HOMEOSTASIS	0.087
GLYCOSAMINOGLYCAN_BINDING	0.096
NEUROGENESIS	0.096
ADHERENS_JUNCTION	0.096
NERVOUS_SYSTEM_DEVELOPMENT	0.096
CELL_SURFACE	0.101
ACTIN_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.107
HEPARIN_BINDING	0.107
IMMUNE_RESPONSE	0.109
RECEPTOR_BINDING	0.109
CALCIUM_CHANNEL_ACTIVITY	0.109
G_PROTEIN_SIGNALING_COUPLED_TO_CAMP_NUCLEOTIDE_SECOND_MESSENGER	0.110
ORGAN_DEVELOPMENT	0.111
CATION_HOMEOSTASIS	0.111
CARBOHYDRATE_BINDING	0.111
POLYSACCHARIDE_BINDING	0.111
TRANSFORMING_GROWTH_FACTOR_BETA_RECEPTOR_SIGNALING_PATHWAY	0.111
RESPONSE_TO_VIRUS	0.112
TRANSMEMBRANE_RECEPTOR_PROTEIN_TYROSINE_KINASE_ACTIVITY	0.115
TRANSMEMBRANE_RECEPTOR_PROTEIN_KINASE_ACTIVITY	0.137
REGULATION_OF_TRANSPORT	0.139
DI_TRI_VALENT_INORGANIC_CATION_TRANSPORT	0.141
PHOSPHOLIPASE_ACTIVITY	0.142
MEMBRANE_ORGANIZATION_AND_BIOGENESIS	0.146
SECOND_MESSENGER_MEDIATED_SIGNALING	0.149
EXTRACELLULAR_SPACE	0.159
RESPONSE_TO_WOUNDING	0.160

CELLULAR_DEFENSE_RESPONSE	0.162
LIPASE_ACTIVITY	0.163
MESODERM_DEVELOPMENT	0.163
TRANSMEMBRANE_RECEPTOR_PROTEIN_SERINE_THREONINE_KINASE_SIGNALING_PATHWAY	0.168
REGULATION_OF_G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	0.169
G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	0.171
CATION_CHANNEL_ACTIVITY	0.174
RAS_PROTEIN_SIGNAL_TRANSDUCTION	0.176
CELL_JUNCTION	0.180
ION_HOMEOSTASIS	0.181
CHEMICAL_HOMEOSTASIS	0.182
GTPASE_ACTIVITY	0.183
REGULATION_OF_CELL_MIGRATION	0.183
ADENYLATE_CYCLASE_ACTIVATION	0.185
G_PROTEIN_SIGNALING_ADENYLATE_CYCLASE_ACTIVATING_PATHWAY	0.185
BEHAVIOR	0.186
SKELETAL_MUSCLE_DEVELOPMENT	0.187
IMMUNE_SYSTEM_PROCESS	0.190
RESPONSE_TO_EXTERNAL_STIMULUS	0.192
REGULATION_OF_BLOOD_PRESSURE	0.195
CARBOXYLESTERASE_ACTIVITY	0.199
HORMONE_ACTIVITY	0.200
REGULATION_OF_CELLULAR_COMPONENT_ORGANIZATION_AND_BIOGENESIS	0.200
PEPTIDYL_TYROSINE_MODIFICATION	0.203
REGULATION_OF_PROTEIN_AMINO_ACID_PHOSPHORYLATION	0.204
VOLTAGE_GATED_POTASSIUM_CHANNEL_COMPLEX	0.205
CYCLIC_NUCLEOTIDE_MEDIATED_SIGNALING	0.205
ACTIN_FILAMENT_BASED_PROCESS	0.205
REGULATION_OF_SIGNAL_TRANSDUCTION	0.206
PROTEIN_TYROSINE_KINASE_ACTIVITY	0.206
STRIATED_MUSCLE_DEVELOPMENT	0.207
RECEPTOR_MEDIATED_ENDOCYTOSIS	0.207
RECEPTOR_COMPLEX	0.207
MOLECULAR_ADAPTOR_ACTIVITY	0.207
TRANSCRIPTION_REPRESSOR_ACTIVITY	0.208
LIGAND_DEPENDENT_NUCLEAR_RECEPTOR_ACTIVITY	0.209
GROWTH	0.220
REGULATION_OF_BIOLOGICAL_QUALITY	0.220
RESPONSE_TO_BIOTIC_STIMULUS	0.221
BONE_REMODELING	0.222
REGULATION_OF_ANGIOGENESIS	0.223
SMALL_GTPASE_MEDIATED_SIGNAL_TRANSDUCTION	0.223
DEFENSE_RESPONSE	0.225
RESPONSE_TO_OTHER_ORGANISM	0.245
EXTRACELLULAR_STRUCTURE_ORGANIZATION_AND_BIOGENESIS	0.247
G_PROTEIN_SIGNALING_COUPLED_TO_CYCLIC_NUCLEOTIDE_SECOND_MESSENGER	0.249

7.6. APPENDIX F

187 Genesets enriched in LOW group in GSE29272 dataset (FDR q value< 0.25)

NAME	FDR q-val
CELL_CYCLE_PROCESS	0
MITOTIC_CELL_CYCLE	0
CELL_CYCLE_PHASE	0
MITOSIS	0
M_PHASE	0
M_PHASE_OF_MITOTIC_CELL_CYCLE	0
CHROMOSOME	0
SPINDLE	0
CHROMOSOMAL_PART	0
CELL_CYCLE_CHECKPOINT_GO_0000075	0
CELL_CYCLE_GO_0007049	0
MICROTUBULE_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	8.41E-05
CHROMOSOMEPERICENTRIC_REGION	9.06E-05
REGULATION_OF_MITOSIS	9.81E-05
RIBONUCLEOPROTEIN_COMPLEX	1.60E-04
KINETOCHORE	3.07E-04
RESPONSE_TO_DNA_DAMAGE_STIMULUS	3.67E-04
MICROTUBULE_CYTOSKELETON	4.19E-04
RESPONSE_TO_ENDOGENOUS_STIMULUS	5.62E-04
DNA_REPAIR	5.91E-04
CHROMOSOME_SEGREGATION	6.20E-04
DNA_METABOLIC_PROCESS	6.49E-04
INTERPHASE	7.03E-04
MICROTUBULE_MOTOR_ACTIVITY	8.24E-04
CONDENSED_CHROMOSOME	0.002
SPINDLE_MICROTUBULE	0.002
NUCLEAR_CHROMOSOME	0.002
MICROTUBULE	0.002
DNA_REPLICATION	0.002
PROTEIN_HETERODIMERIZATION_ACTIVITY	0.002
MITOTIC_CELL_CYCLE_CHECKPOINT	0.002
INTERPHASE_OF_MITOTIC_CELL_CYCLE	0.002
NUCLEAR_MEMBRANE_PART	0.003
REGULATION_OF_CELL_CYCLE	0.003
MOTOR_ACTIVITY	0.004
DNA_DEPENDENT_DNA_REPLICATION	0.004
DNA_INTEGRITY_CHECKPOINT	0.005
HYDROLASE_ACTIVITY_ACTING_ON_CARBON_NITROGEN_NOT_PEPTIDEBON DSIN_CYCLIC_AMIDINES	0.010
NUCLEOLUS	0.010

NUCLEAR_PART	0.011
NUCLEAR_MEMBRANE	0.012
ORGANELLE_LOCALIZATION	0.012
MITOCHONDRIAL_RIBOSOME	0.012
SPINDLE_POLE	0.012
STEROID_BIOSYNTHETIC_PROCESS	0.014
NUCLEAR_PORE	0.015
ORGANELLAR_RIBOSOME	0.015
DNA_RECOMBINATION	0.015
MRNA_METABOLIC_PROCESS	0.015
REPLICATION_FORK	0.016
SPLICEOSOME	0.016
MICROTUBULE_BASED_PROCESS	0.016
RIBOSOME	0.019
ESTABLISHMENT_OF_ORGANELLE_LOCALIZATION	0.021
CHROMATIN	0.022
CENTROSOME	0.024
TRANSFERASE_ACTIVITY_TRANSFERRING_GLYCOSYL_GROUPS	0.028
ENVELOPE	0.028
DNA_DEPENDENT_ATPASE_ACTIVITY	0.030
RNA_PROCESSING	0.031
MRNA_PROCESSING_GO_0006397	0.031
SMALL_NUCLEAR_RIBONUCLEOPROTEIN_COMPLEX	0.031
NUCLEAR_ENVELOPE	0.031
COENZYME_METABOLIC_PROCESS	0.031
AMINO_ACID_AND_DERIVATIVE_METABOLIC_PROCESS	0.032
AMINO_ACID_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.032
REGULATION_OF_MITOTIC_CELL_CYCLE	0.032
CHROMATIN_BINDING	0.032
MICROTUBULE_ORGANIZING_CENTER	0.032
AMINO_ACID_METABOLIC_PROCESS	0.032
ORGANELLE_ENVELOPE	0.032
INTERMEDIATE_FILAMENT_CYTOSKELETON	0.032
INTERMEDIATE_FILAMENT	0.033
G1_S_TRANSITION_OF_MITOTIC_CELL_CYCLE	0.033
TRANSFERASE_ACTIVITY_TRANSFERRING_PENTOSYL_GROUPS	0.039
REGULATION_OF_CYCLIN_DEPENDENT_PROTEIN_KINASE_ACTIVITY	0.039
BASE_EXCISION_REPAIR	0.042
DEOXYRIBONUCLEASE_ACTIVITY	0.045
TRANSFERASE_ACTIVITY_TRANSFERRING_HEXOSYL_GROUPS	0.049
RNA_SPLICING	0.049
NEGATIVE_REGULATION_OF_DNA_METABOLIC_PROCESS	0.049
NUCLEAR_ORGANIZATION_AND_BIOGENESIS	0.050
CYTOSKELETAL_PART	0.050
DAMAGED_DNA_BINDING	0.050

METALLOENDOPEPTIDASE_ACTIVITY	0.051
ECTODERM_DEVELOPMENT	0.053
MITOCHONDRION_ORGANIZATION_AND_BIOGENESIS	0.053
MEIOSIS_I	0.053
CHROMOSOME_ORGANIZATION_AND_BIOGENESIS	0.053
CARBOHYDRATE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.054
REGULATION_OF_DNA_REPLICATION	0.055
AMINE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.059
MICROTUBULE_ORGANIZING_CENTER_PART	0.059
UDP_GLYCOSYLTRANSFERASE_ACTIVITY	0.060
MEMBRANE_ENCLOSED_LUMEN	0.062
ORGANELLE_LUMEN	0.062
STEROID_METABOLIC_PROCESS	0.064
DOUBLE_STRAND_BREAK_REPAIR	0.067
MITOCHONDRION	0.074
DNA_DAMAGE_RESPONSESIGNAL_TRANSDUCTION	0.079
COFACTOR_METABOLIC_PROCESS	0.080
LIGASE_ACTIVITY	0.081
MEIOTIC_CELL_CYCLE	0.082
ORGANELLE_INNER_MEMBRANE	0.083
ORGANELLE_MEMBRANE	0.083
CARBON_CARBON_LYASE_ACTIVITY	0.084
NUCLEAR_CHROMOSOME_PART	0.087
EPIDERMIS_DEVELOPMENT	0.090
AROMATIC_COMPOUND_METABOLIC_PROCESS	0.091
RNA_HELICASE_ACTIVITY	0.100
DNA_DAMAGE_CHECKPOINT	0.100
CELLULAR_COMPONENT_DISASSEMBLY	0.101
SINGLE_STRANDED_DNA_BINDING	0.104
ENDOMEMBRANE_SYSTEM	0.112
NUCLEOTIDE_METABOLIC_PROCESS	0.112
RESPONSE_TO_HORMONE_STIMULUS	0.116
PROTEASOME_COMPLEX	0.117
NUCLEAR_LUMEN	0.119
ALCOHOL_METABOLIC_PROCESS	0.119
MITOCHONDRIAL_INNER_MEMBRANE	0.119
BIOSYNTHETIC_PROCESS	0.119
REGULATION_OF_DNA_METABOLIC_PROCESS	0.121
NUCLEOTIDYLTRANSFERASE_ACTIVITY	0.126
APOPTOTIC_NUCLEAR_CHANGES	0.126
MITOCHONDRIAL_PART	0.127
TRNA_METABOLIC_PROCESS	0.128
APOPTOTIC_PROGRAM	0.129
DIGESTION	0.132
L_AMINO_ACID_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.132

NUCLEAR_ENVELOPE_ENDOPLASMIC_RETICULUM_NETWORK	0.133
MITOCHONDRIAL_MATRIX	0.138
TRANSLATION_INITIATION_FACTOR_ACTIVITY	0.138
GLYCOPROTEIN_BIOSYNTHETIC_PROCESS	0.139
NITROGEN_COMPOUND_METABOLIC_PROCESS	0.139
MICROBODY	0.153
ORGANIC_ACID_METABOLIC_PROCESS	0.157
CARBON_OXYGEN_LYASE_ACTIVITY	0.158
NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_TRANSPORT	0.158
RNA_BINDING	0.158
MITOCHONDRIAL_LUMEN	0.159
CELL_DIVISION	0.160
NUCLEOBASENUCLEOSIDE_AND_NUCLEOTIDE_METABOLIC_PROCESS	0.162
ENDOPLASMIC_RETICULUM_PART	0.163
ONE_CARBON_COMPOUND_METABOLIC_PROCESS	0.163
CARBOXYLIC_ACID_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.164
DNA_PACKAGING	0.164
CARBOHYDRATE_METABOLIC_PROCESS	0.165
ORGANIC_ACID_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.165
PEROXISOME	0.169
GALACTOSYLTRANSFERASE_ACTIVITY	0.169
RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS_AND_ASSEMBLY	0.172
RNA_EXPORT_FROM_NUCLEUS	0.172
CARBOXYLIC_ACID_METABOLIC_PROCESS	0.178
MITOCHONDRIAL_TRANSPORT	0.184
GOLGI_ASSOCIATED_VESICLE	0.184
TRANSCRIPTION_INITIATION_FROM_RNA_POLYMERASE_II_PROMOTER	0.188
CYTOKINESIS	0.191
DNA_DIRECTED_RNA_POLYMERASEII_HOLOENZYME	0.195
TRANSLATION_FACTOR_ACTIVITY_NUCLEIC_ACID_BINDING	0.197
TRANSFERASE_ACTIVITY_TRANSFERRING_ONE_CARBON_GROUPS	0.197
NITROGEN_COMPOUND_CATABOLIC_PROCESS	0.198
ENDONUCLEASE_ACTIVITY	0.201
GLYCOPROTEIN_METABOLIC_PROCESS	0.201
METALLOPEPTIDASE_ACTIVITY	0.201
MACROMOLECULE_BIOSYNTHETIC_PROCESS	0.204
POSITIVE_REGULATION_OF_BINDING	0.206
CYTOSKELETON	0.206
METHYLTRANSFERASE_ACTIVITY	0.207
ORGANELLE_ORGANIZATION_AND_BIOGENESIS	0.208
PROTEIN_DIMERIZATION_ACTIVITY	0.208
CHROMATIN_REMODELING	0.210
LIGASE_ACTIVITY_FORMING_CARBON_NITROGEN_BONDS	0.211
DNA_HELICASE_ACTIVITY	0.212
RNA_SPLICINGVIA_TRANSESTERIFICATION_REACTIONS	0.213

CARBOHYDRATE_BIOSYNTHETIC_PROCESS	0.215
ENDOPLASMIC_RETICULUM_MEMBRANE	0.216
NUCLEASE_ACTIVITY	0.217
POSITIVE_REGULATION_OF_DNA_BINDING	0.218
RNA_CATABOLIC_PROCESS	0.219
COFACTOR_BIOSYNTHETIC_PROCESS	0.230
HYDRO_LYASE_ACTIVITY	0.234
LIPID_BIOSYNTHETIC_PROCESS	0.235
CELLULAR_BIOSYNTHETIC_PROCESS	0.235
NEGATIVE_REGULATION_OF_CATALYTIC_ACTIVITY	0.236
AMINO_ACID_TRANSPORT	0.242
AMINE_METABOLIC_PROCESS	0.246
LIPOPROTEIN_BIOSYNTHETIC_PROCESS	0.248

7.7. APPENDIX G

138 Genesets enriched in HIGH group in GSE62254 dataset (FDR q value< 0.25)

NAME	FDR q-val
CALCIUM_CHANNEL_ACTIVITY	0.216
CELL_MATRIX_JUNCTION	0.217
IMMUNE_SYSTEM_DEVELOPMENT	0.217
TRANSMEMBRANE_RECEPTOR_PROTEIN_TYROSINE_KINASE_ACTIVITY	0.218
BASOLATERAL_PLASMA_MEMBRANE	0.218
PYROPHOSPHATASE_ACTIVITY	0.218
TISSUE_REMODELING	0.218
POLYSACCHARIDE_BINDING	0.218
ATPASE_ACTIVITY_COUPLED_TO_TRANSMEMBRANE_MOVEMENT_OF_IONS	0.218
EXTRACELLULAR_STRUCTURE_ORGANIZATION_AND_BIOGENESIS	0.218
DI__TRI_VALENT_INORGANIC_CATION_TRANSPORT	0.218
PROTEIN_KINASE_REGULATOR_ACTIVITY	0.218
SKELETAL_DEVELOPMENT	0.219
ADENYLATE_CYCLASE_ACTIVATION	0.219
BRAIN_DEVELOPMENT	0.219
CALCIUM_ION_TRANSPORT	0.219
SUBSTRATE_SPECIFIC_CHANNEL_ACTIVITY	0.219
NEGATIVE_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	0.219
CARBOHYDRATE_BINDING	0.220
CELL_RECOGNITION	0.220
BASEMENT_MEMBRANE	0.220
TRANSFORMING_GROWTH_FACTOR_BETA_RECEPTOR_SIGNALING_PATHWAY	0.220
COLLAGEN	0.220
CELL_PROJECTION	0.220
PATTERN_BINDING	0.220
TRANSMEMBRANE_RECEPTOR_PROTEIN_SERINE_THREONINE_KINASE_SIGNALING_PATHWAY	0.221
ACTIN_FILAMENT_BASED_PROCESS	0.221
ADHERENS_JUNCTION	0.221
DENDRITE	0.221
INORGANIC_CATION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.222
TRANSCRIPTION_REPRESSOR_ACTIVITY	0.222
CELL_JUNCTION	0.222
REGULATION_OF_CELL_DIFFERENTIATION	0.223
CYCLIC_NUCLEOTIDE_MEDIATED_SIGNALING	0.223
GLYCOSAMINOGLYCAN_BINDING	0.223
G_PROTEIN_SIGNALING_COUPLED_TO_CYCLIC_NUCLEOTIDE_SECOND_MESSENGER	0.223
POSITIVE_REGULATION_OF_SECRETION	0.223
ACTIN_FILAMENT_BINDING	0.223
SYNAPSE_ORGANIZATION_AND_BIOGENESIS	0.223
PHOSPHOINOSITIDE_MEDIATED_SIGNALING	0.224

AMINE_RECEPTOR_ACTIVITY	0.224
POSITIVE_REGULATION_OF_CELL_DIFFERENTIATION	0.224
NUCLEOSIDE_TRIPHOSPHATASE_ACTIVITY	0.224
REGULATION_OF_ACTION_POTENTIAL	0.224
HEPARIN_BINDING	0.225
MOLECULAR_ADAPTOR_ACTIVITY	0.225
MAPKKK_CASCADE_GO_0000165	0.225
GROWTH	0.226
REGULATION_OF_SIGNAL_TRANSDUCTION	0.226
ION_TRANSPORT	0.226
RUFFLE	0.226
RHO_GTPASE_ACTIVATOR_ACTIVITY	0.227
POSITIVE_REGULATION_OF_TRANSCRIPTION	0.227
ACTIVATION_OF_PROTEIN_KINASE_ACTIVITY	0.227
HYDROLASE_ACTIVITY_ACTING_ON_ACID_ANHYDRIDESCATALYZING_TRANSMEMBRANE_MOVE	0.228
TRANSMEMBRANE_RECEPTOR_PROTEIN_KINASE_ACTIVITY	0.228
PHOSPHORIC_MONOESTER_HYDROLASE_ACTIVITY	0.229
REGULATION_OF_TRANSPORT	0.229
VOLTAGE_GATED_CATION_CHANNEL_ACTIVITY	0.229
GLUTAMATE_RECEPTOR_ACTIVITY	0.229
CHROMATIN_ASSEMBLY_OR_DISASSEMBLY	0.229
TRANSPORT_VESICLE	0.230
CYTOSKELETON_DEPENDENT_INTRACELLULAR_TRANSPORT	0.230
HYDROLASE_ACTIVITY_ACTING_ON_ACID_ANHYDRIDES	0.230
HYDROGEN_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.230
CATION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.231
NEGATIVE_REGULATION_OF_RNA_METABOLIC_PROCESS	0.231
REGULATION_OF_BIOLOGICAL_QUALITY	0.231
PROTEIN_SECRETION	0.231
LIPOPROTEIN_BINDING	0.231
CELL_SURFACE	0.231
REGULATION_OF_SMALL_GTPASE_MEDIATED_SIGNAL_TRANSDUCTION	0.232
ENDOSOME_TRANSPORT	0.232
RAS_GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY	0.232
PRIMARY_ACTIVE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.232
SECOND_MESSENGER_MEDIATED_SIGNALING	0.233
KINASE_REGULATOR_ACTIVITY	0.233
ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	0.233
EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT	0.233
PHOSPHATASE_REGULATOR_ACTIVITY	0.233
ATPASE_ACTIVITY_COUPLED_TO_TRANSMEMBRANE_MOVEMENT_OF_IONS_PHOSPHORYLATIV	0.233
REGULATION_OF_ANATOMICAL_STRUCTURE_MORPHOGENESIS	0.233
HEART_DEVELOPMENT	0.233
CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.233
SENSORY_PERCEPTION_OF_CHEMICAL_STIMULUS	0.233
HEMOPOIETIC_OR_LYMPHOID_ORGAN_DEVELOPMENT	0.234
PROTEIN_TYROSINE_KINASE_ACTIVITY	0.234

NEGATIVE_REGULATION_OF_TRANSCRIPTION_DNA_DEPENDENT	0.234
ENZYME_LINKED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	0.234
RAS_GTPASE_ACTIVATOR_ACTIVITY	0.234
PROTEIN_POLYMERIZATION	0.234
REGULATION_OF_MYELOID_CELL_DIFFERENTIATION	0.234
BASAL_LAMINA	0.234
ATPASE_ACTIVITY_COUPLED_TO_MOVEMENT_OF_SUBSTANCES	0.234
MEMBRANE_ORGANIZATION_AND_BIOGENESIS	0.234
MUSCLE_CELL_DIFFERENTIATION	0.234
POSITIVE_REGULATION_OF_NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_MET	0.234
ENZYME_ACTIVATOR_ACTIVITY	0.234
CONTRACTILE_FIBER_PART	0.234
MACROMOLECULAR_COMPLEX_ASSEMBLY	0.234
POSITIVE_REGULATION_OF_TRANSCRIPTIONDNA_DEPENDENT	0.234
SH3_SH2_ADAPTOR_ACTIVITY	0.235
GUANYL_NUCLEOTIDE_EXCHANGE_FACTOR_ACTIVITY	0.235
TRANSCRIPTION_ACTIVATOR_ACTIVITY	0.235
DETECTION_OF_STIMULUS	0.235
RECEPTOR_COMPLEX	0.235
CONTRACTILE_FIBER	0.235
CELLULAR_COMPONENT_ASSEMBLY	0.235
PHOSPHORIC_DIESTER_HYDROLASE_ACTIVITY	0.235
RECEPTOR_BINDING	0.235
LIPID_RAFT	0.235
MYELOID_CELL_DIFFERENTIATION	0.236
REGULATION_OF_NEUROTRANSMITTER_LEVELS	0.236
POTASSIUM_CHANNEL_ACTIVITY	0.236
EXTRACELLULAR_MATRIX	0.236
NEURON_PROJECTION	0.237
DEPHOSPHORYLATION	0.238
LEUKOCYTE_DIFFERENTIATION	0.240
PROTEIN_AMINO_ACID_DEPHOSPHORYLATION	0.240
SYNAPTOGENESIS	0.241
EXTRACELLULAR_REGION	0.243
HEMOPOIESIS	0.243
ENZYME_REGULATOR_ACTIVITY	0.244
G_PROTEIN_SIGNALING_ADENYLATE_CYCLASE_ACTIVATING_PATHWAY	0.244
GENERATION_OF_PRECURSOR_METABOLITES_AND_ENERGY	0.244
ORGAN_MORPHOGENESIS	0.245
RAS_PROTEIN_SIGNAL_TRANSDUCTION	0.245
POSITIVE_REGULATION_OF_RNA_METABOLIC_PROCESS	0.246
NEGATIVE_REGULATION_OF_CELL_DIFFERENTIATION	0.246
EXTRACELLULAR_REGION_PART	0.246
PROTEIN_DOMAIN_SPECIFIC_BINDING	0.247
VESICLE_MEDIATED_TRANSPORT	0.247
REGULATION_OF_BLOOD_PRESSURE	0.247
TRANSCRIPTION_FACTOR_BINDING	0.248

POSITIVE_REGULATION_OF_CELL_PROLIFERATION	0.248
ION_CHANNEL_ACTIVITY	0.248
AUXILIARY_TRANSPORT_PROTEIN_ACTIVITY	0.248
REGULATION_OF_RESPONSE_TO_STIMULUS	0.248



7.8. APPENDIX H

53 Genesets enriched in LOW group in GSE62254 dataset (FDR q value< 0.25)

NAME	FDR q-val
KINETOCHORE	0.181
MITOTIC_CELL_CYCLE_CHECKPOINT	0.183
RESPONSE_TO_DNA_DAMAGE_STIMULUS	0.185
MITOCHONDRIAL_PART	0.186
CHROMOSOMEPERICENTRIC_REGION	0.186
CELL_CYCLE_GO_0007049	0.187
NEGATIVE_REGULATION_OF_DNA_METABOLIC_PROCESS	0.188
MITOCHONDRIAL_LUMEN	0.190
INTERPHASE	0.190
DNA_REPAIR	0.190
SPINDLE_POLE	0.191
RIBONUCLEOPROTEIN_COMPLEX	0.192
CELL_CYCLE_CHECKPOINT_GO_0000075	0.192
MITOCHONDRIAL_RIBOSOME	0.192
DNA_RECOMBINATION	0.193
CONDENSED_NUCLEAR_CHROMOSOME	0.194
DNA_METABOLIC_PROCESS	0.194
REGULATION_OF_MITOTIC_CELL_CYCLE	0.195
NUCLEAR_CHROMOSOME	0.195
MITOCHONDRIAL_MATRIX	0.195
CONDENSED_CHROMOSOME	0.198
MITOCHONDRIAL_INNER_MEMBRANE	0.198
INTERPHASE_OF_MITOTIC_CELL_CYCLE	0.199
TRANSFERASE_ACTIVITY_TRANSFERRING_GROUPS_OTHER_THAN_AMINO_ACYL_GROUPS	0.199
DNA_REPLICATION	0.199
REGULATION_OF_MITOSIS	0.200
RIBOSOMAL_SUBUNIT	0.201
ORGANELLAR_RIBOSOME	0.202
LIGASE_ACTIVITY	0.202
M_PHASE_OF_MITOTIC_CELL_CYCLE	0.203
COENZYME_METABOLIC_PROCESS	0.204
CHROMOSOME_SEGREGATION	0.207
MITOSIS	0.208
MRNA_METABOLIC_PROCESS	0.210
CENTROSOME	0.210
CARBOHYDRATE_KINASE_ACTIVITY	0.214
CHROMOSOMAL_PART	0.216
NUCLEAR_MEMBRANE_PART	0.217
MICROTUBULE_ORGANIZING_CENTER_ORGANIZATION_AND_BIOGENESIS	0.219
MITOCHONDRION_ORGANIZATION_AND_BIOGENESIS	0.220
ENDONUCLEASE_ACTIVITY	0.223

MICROTUBULE_CYTOSKELETON	0.224
NUCLEOBASENUCLEOSIDENUCLEOTIDE_AND_NUCLEIC_ACID_TRANSPORT	0.225
COFACTOR_METABOLIC_PROCESS	0.229
ORGANELLE_INNER_MEMBRANE	0.229
MITOCHONDRIAL_ENVELOPE	0.232
REGULATION_OF_DNA_REPLICATION	0.232
N_ACETYLTRANSFERASE_ACTIVITY	0.234
COATED_MEMBRANE	0.244
DNA_DEPENDENT_DNA_REPLICATION	0.247
MEMBRANE_COAT	0.247
MICROTUBULE_ORGANIZING_CENTER	0.249
ORGANELLE_ENVELOPE	0.250



7.9. APPENDIX I

62 Genesets enriched in LOW group in GSE35809 dataset (FDR q value< 0.25)

NAME	FDR q-val
PROTEIN_FOLDING	0.188
CENTROSOME	0.190
REGULATION_OF_CELL_CYCLE	0.190
REGULATION_OF_DNA_METABOLIC_PROCESS	0.190
PIGMENT_BIOSYNTHETIC_PROCESS	0.191
DNA_METABOLIC_PROCESS	0.193
DNA_DAMAGE_RESPONSESIGNAL_TRANSDUCTION	0.193
MITOCHONDRION_ORGANIZATION_AND_BIOGENESIS	0.193
NEGATIVE_REGULATION_OF_DNA_METABOLIC_PROCESS	0.196
INTERPHASE_OF_MITOTIC_CELL_CYCLE	0.197
PROTEASOME_COMPLEX	0.198
MICROTUBULE_ORGANIZING_CENTER_PART	0.198
MITOSIS	0.198
TRNA_METABOLIC_PROCESS	0.199
NUCLEAR_PORE	0.199
MICROTUBULE_ORGANIZING_CENTER	0.199
REPLICATION_FORK	0.200
RNA_HELICASE_ACTIVITY	0.200
CHROMOSOMAL_PART	0.200
PIGMENT_METABOLIC_PROCESS	0.200
CHROMOSOME	0.200
DNA_INTEGRITY_CHECKPOINT	0.201
NUCLEAR_MEMBRANE	0.202
M_PHASE_OF_MITOTIC_CELL_CYCLE	0.202
REGULATION_OF_MITOTIC_CELL_CYCLE	0.203
ONE_CARBON_COMPOUND_METABOLIC_PROCESS	0.204
MITOTIC_CELL_CYCLE	0.207
CELL_CYCLE_CHECKPOINT_GO_000075	0.208
MRNA_PROCESSING_GO_0006397	0.208
REGULATION_OF_MITOSIS	0.208
SPLICEOSOME	0.210
MITOCHONDRIAL_RIBOSOME	0.211
TRANSCRIPTION_FROM_RNA_POLYMERASE_III_PROMOTER	0.214
CONDENSED_CHROMOSOME	0.215
CELL_CYCLE_PHASE	0.217
M_PHASE	0.217
NUCLEAR_MEMBRANE_PART	0.217
LIGASE_ACTIVITY	0.217
MEIOSIS_I	0.218
MICROTUBULE_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	0.218
CELL_CYCLE_GO_0007049	0.220

NUCLEOTIDYLTRANSFERASE_ACTIVITY	0.220
SPINDLE_POLE	0.221
MICROTUBULE	0.221
SPINDLE	0.221
ORGANELLAR_RIBOSOME	0.222
CHROMATIN	0.222
GOLGI_ASSOCIATED_VESICLE	0.223
DNA_DAMAGE_CHECKPOINT	0.224
NUCLEAR_PART	0.224
NUCLEAR_CHROMOSOME	0.225
INTERPHASE	0.230
CELL_CYCLE_PROCESS	0.231
G1_S_TRANSITION_OF_MITOTIC_CELL_CYCLE	0.236
NUCLEOLUS	0.243
REGULATION_OF_DNA_REPLICATION	0.246
CARBOHYDRATE_KINASE_ACTIVITY	0.246
RNA_BINDING	0.246
MICROTUBULE_CYTOSKELETON	0.247
DNA_DEPENDENT_DNA_REPLICATION	0.249
DOUBLE_STRAND_BREAK_REPAIR	0.249
RESPONSE_TO_DNA_DAMAGE_STIMULUS	0.250

7.10. APPENDIX J

Probesets for generic EMT signature of tumor tissues obtained from Tan et al.[1] –requested from authors: A total of 315 genes (Epithelial: 145, Mesenchymal:170)

Epithelial	
Gene	Probe
KRT19	201650_at
AGR2	209173_at
RAB25	218186_at
CDH1	201131_s_at
ERBB3	202454_s_at
FXYD3	202489_s_at
SLC44A4	205597_at
S100P	204351_at
SCNN1A	203453_at
GALNT3	203397_s_at
PRSS8	202525_at
ELF3	210827_s_at
CEACAM6	203757_s_at
TMPRSS4	218960_at
CLDN7	202790_at
TACSTD2	202286_s_at
CLDN3	203953_s_at
EPCAM	201839_s_at
SPINT1	202826_at
TSPAN1	209114_at
PLS1	205190_at
TMEM30B	213285_at
PRR15L	219127_at
KRT8	209008_x_at
ST14	216905_s_at
RBM47	218035_s_at
S100A14	218677_at
C1orf106	219010_at
NQO1	201468_s_at
TOX3	214774_x_at
PTK6	206482_at
TFF1	205009_at
CLDN4	201428_at
GPRC5A	203108_at
TJP3	35148_at
KRT18	201596_x_at
MAP7	202890_at

Epithelial	
Gene	Probe
EHF	219850_s_at
AP1M2	218261_at
CEACAM5	201884_at
LAD1	203287_at
ARHGAP8	37117_at
TFF3	204623_at
JUP	201015_s_at
CD24	208650_s_at
TMC5	219580_s_at
MLPH	218211_s_at
ELMO3	219411_at
ERBB2	216836_s_at
LLGL2	203713_s_at
DDR1	207169_x_at
FA2H	219429_at
CBLC	220638_s_at
TMPRSS2	211689_s_at
LSR	208190_s_at
PERP	217744_s_at
POF1B	219756_s_at
MYO5C	218966_at
RAB11FIP1	219681_s_at
MAPK13	210058_at
KRT7	209016_s_at
CEACAM1	209498_at
CXADR	203917_at
ATP2C2	206043_s_at
RNF128	219263_at
MPZL2	203780_at
EPS8L1	221665_s_at
GALNT7	218313_s_at
CORO2A	205538_at
BCAS1	204378_at
TPD52	201689_s_at
ARHGAP32	203431_s_at
FUT2	210608_s_at
OR7E14P	217551_at

Epithelial	
Gene	Probe
PSCA	205319_at
EZR	217234_s_at
TJP2	202085_at
FGFR3	204379_s_at
FUT3	214088_s_at
BSPRY	218792_s_at
TOM1L1	204485_s_at
IRF6	202597_at
EPB41L4B	220161_s_at
SPDEF	220192_x_at
OCLN	209925_at
LRRC1	218816_at
C19orf21	212925_at
ABHD11	221927_s_at
EPS8L2	218180_s_at
MYO6	203216_s_at
TSPAN8	203824_at
MST1R	205455_at
SLC16A5	206600_s_at
GPR56	212070_at
AZGP1	209309_at
TOB1	202704_at
SLC35A3	206770_s_at
TRPM4	219360_s_at
PHLDA2	209803_s_at
VAMP8	202546_at
SLC22A18	204981_at
AKR1B10	206561_s_at
VAV3	218806_s_at
SPAG1	210117_at
ABCC3	208161_s_at
SYNGR2	201079_at
STAP2	221610_s_at
C4orf19	219450_at
PPL	203407_at
PLL	204519_s_at
DSG2	217901_at

Epithelial	
Gene	Probe
MANSC1	220945_x_at
DHCR24	200862_at
EPN3	220318_at
TUFT1	205807_s_at
GMDS	204875_s_at
EXPH5	214734_at
DSP	200606_at
SDC4	202071_at
IL20RA	219115_s_at
FAM174B	51158_at
PTPRF	200637_s_at
SORD	201563_at
MUC1	207847_s_at
SPINT2	210715_s_at
ESRP2	219395_at
CDS1	205709_s_at
PPAP2C	209529_at
CEACAM7	206199_at
TTC39A	210652_s_at
OVOL2	211778_s_at
BIK	205780_at
RAPGEFL1	218657_at
STYK1	220030_at
F11R	221664_s_at
PKP3	209873_s_at
CYB561	209163_at
SH3YL1	204019_s_at
GDF15	221577_x_at
CKMT1A	202712_s_at
ESRP1	219121_s_at
GALE	202528_at
GRHL2	219388_at
HDHD3	221256_s_at
CD2AP	203593_at

Mesenchymal	
Gene	Probe
GAS1	204457_s_at
CXCL12	209687_at
ZEB1	212764_at
GLYR1	212414_s_at
FHL1	201540_at
FERMT2	209210_s_at
C1S	208747_s_at
FYN	210105_s_at
WIPF1	202664_at
CYP1B1	202437_s_at
SERPING1	200986_at
SERPINF1	202283_at
VCAM1	203868_s_at
MAP1B	212233_at
TCF4	213891_s_at
SRPX	204955_at
EMP3	203729_at
DPT	213068_at
CALD1	201616_s_at
PTGIS	208131_s_at
VIM	201426_s_at
CD163	215049_x_at
C1R	212067_s_at
FBN1	202766_s_at
FN1	212464_s_at
FXVD6	217897_at
IGF1	209541_at
NAP1L3	204749_at
MRC1	204438_at
QKI	212636_at
MS4A4A	219607_s_at
DCN	211896_s_at
LOX	215446_s_at
RECK	205407_at
ANK2	202920_at
LY96	206584_at
ZFPM2	219778_at
CSRP2	207030_s_at
EFEMP1	201842_s_at
RARRES2	209496_at
PTPRC	212588_at
PLEKHO1	218223_s_at

Mesenchymal	
Gene	Probe
RGS2	202388_at
F13A1	203305_at
JAM2	219213_at
CHRD1	209763_at
TUBA1A	209118_s_at
AP1S2	203300_x_at
MYLK	202555_s_at
DDR2	205168_at
DSE	218854_at
SACS	213262_at
GLIPR1	204222_s_at
CXCL13	205242_at
FLRT2	204359_at
PTX3	206157_at
AKT3	212607_at
COL6A2	209156_s_at
DPYSL3	201431_s_at
CDH11	207173_x_at
PDZRN3	212915_at
ZEB2	203603_s_at
CCL2	216598_s_at
MAFB	218559_s_at
SFRP1	202037_s_at
C1orf139	219563_at
MFAP4	212713_at
MAF	209348_s_at
UCHL1	201387_s_at
TUBB6	209191_at
SRGN	201858_s_at
HEG1	213069_at
KCNJ8	205303_at
AKAP12	210517_s_at
EVI2A	204774_at
COL14A1	212865_s_at
AXL	202686_s_at
ECM2	206101_at
FSTL1	208782_at
PLN	204939_s_at
MYL9	201058_s_at
OLFML3	218162_at
STON1	213413_at
SLIT2	209897_s_at

Mesenchymal	
Gene	Probe
BICC1	213429_at
SOBP	218974_at
CLIC4	201560_at
ENPP2	209392_at
SAMSN1	220330_s_at
TPM2	204083_s_at
ASPN	219087_at
COL6A1	213428_s_at
IGFBP5	211959_at
MOXD1	209708_at
AKAP2	202760_s_at
SLC2A3	202499_s_at
OLFML2B	213125_at
ANGPTL2	213001_at
PCOLCE	202465_at
COLEC12	221019_s_at
CTSK	202450_s_at
TAGLN	205547_s_at
CDH2	203440_at
IL10RA	204912_at
C1orf54	219506_at
CEP170	207719_x_at
TNS1	221748_s_at
CLEC2B	209732_at
JAM3	212813_at
6-Sep	214298_x_at
GREM1	218469_at
VCAN	204620_s_at
ZCCHC24	212419_at
CRYAB	209283_at
SFRP4	204051_s_at
RUNX1T1	205529_s_at
FGL2	204834_at
MS4A6A	219666_at
PTRF	208789_at
GIMAP4	219243_at
TWIST1	213943_at
GFPT2	205100_at
LHFP	218656_s_at
CXCR4	217028_at
SPOCK1	202363_at
SPARC	212667_at

Mesenchymal	
Gene	Probe
VSIG4	204787_at
GPM6B	209167_at
TRPC1	205803_s_at
SNAI2	213139_at
GUCY1B3	203817_at
PLXNC1	213241_at
SYT11	209197_at
FLI1	204236_at
MYH10	212372_at
CSF2RB	205159_at
TNC	201645_at
PMP22	210139_s_at
COL5A2	221730_at
MMP2	201069_at
GNG11	204115_at
CAV1	212097_at
CDK14	204604_at
SDC2	212158_at
PTGDS	211748_x_at
NR3C1	216321_s_at
SYNM	212730_at
FAP	209955_s_at
NUAK1	204589_at
WWTR1	202133_at
FBLN1	202995_s_at
MPDZ	213306_at
SYNE1	209447_at
EFEMP2	206580_s_at
GIMAP6	219777_at
KIAA1462	213316_at
CCL8	214038_at
COL15A1	203477_at
CHN1	212624_s_at
CRISPLD2	221541_at
PDGFC	218718_at
GEM	204472_at
ISLR	207191_s_at
GZMK	206666_at
SPARCL1	200795_at
BNC2	220272_at
BGN	213905_x_at
MEOX2	206201_s_at
ITM2A	202746_at
IFFO1	36030_at

7.11. APPENDIX K

EMT scores of NCI60 cancer cell lines obtained from authors of Tan et al. [1]

Cancer type/Tissue	Cell Line	Generic EMT Score
Breast	BT_549	-0.336
	HS578T	0.503
	MCF7	0.319
	MDA_MB_231	-0.537
	T47D	0.293
Central Nervous System	SF_268	-0.485
	SF_295	0.391
	SF_539	0.624
	SNB_19	0.593
	SNB_75	0.455
	U251	-0.410
Colon	COLO205	-0.490
	HCC_2998	0.216
	HCT_116	0.251
	HCT_15	0.182
	HT29	0.677
	KM12	0.202
	SW_620	0.449
Leukemia	CCRF_CEM	0.251
	HL_60	0.185
	K_562	0.489
	MOLT_4	0.409
	RPMI_8226	0.364
	SR	0.471
Melanoma	LOXIMVI	0.540
	M14	0.541
	MALME_3M	0.250
	MDA_MB_435	-0.410
	SK_MEL_2	0.220
	SK_MEL_28	0.522
	SK_MEL_5	0.444
	UACC_257	0.279
	UACC_62	0.619

Cancer type/Tissue	Cell Line	Generic EMT Score
Lung	A549	0.412
	EKVX	0.155
	HOP_62	-0.452
	HOP_92	0.448
	NCI_H226	0.407
	NCI_H23	0.233
	NCI_H322M	0.263
	NCI_H460	0.120
Ovary	NCI_H522	0.076
	IGROV1_1	0.289
	IGROV1_2	0.517
	NCI_ADR_RES	0.425
	OVCAR_3	0.150
	OVCAR_4	0.400
	OVCAR_5	0.411
	OVCAR_8	0.488
Prostate	SK_OV_3	0.348
	DU_145	0.613
	PC_3	-0.125
Renal	786_0	0.429
	A498	0.636
	ACHN	0.368
	CAKI_1	-0.135
	RXF_393	0.430
	SN12C	-0.324
	TK_10	0.573
	UO_31	0.641

7.12. APPENDIX L

58 compounds from NCI60 database which were selectively effective on epithelial cells

NSC #	Drug name	PubChem Id
729192	antineoplastic-672553	-
646702	0	581769
645306	0	581696
762346	#N/A	#N/A
403831	0	579004
650259	0	504774
645317	sanguilutine pseudobase	502491
33419	sn 967	543938
73495	chelocardin dodecyl sulfate	550460
349155	margetine	575919
3093	crude pyronine b	538594
645305	0	581695
22348	0	84524
269192	8-trichloromethyldihydroberberine	140807
633093	6-o[n-1'-nitroacridyl-9']-.beta.-alanyl]muramyl dipeptide	496234
706019	0	528980
724006	0	537280
46085	0	546811
651646	0	581829
674509	0	515676
729168	0	48428113
706491	0	529186
144303	0	428565
642492	O-Chloroacetylcarbamoylefumagillol	501161
676006	0	516216
709477	0	530308
720294	0	535389
720306	0	535401
720431	0	535495
693573	2-(4-methylthio-benzylidene)-2-(2-pyridyl)acetonitrile	-
646127	0	502909
728666	0	48427947
729052	0	48428057
728667	0	48427948
710746	0	530954
730869	(e)-6-(7-hydroxy-5-hepten-1,3-diyn-1-yl)-2h-pyran-2-one	48428820
78627	naphtho[2,1-b]quinolizinium, 7,10-dimethyl-perchlorate	551192

645300	0	502477
650744	1,1'-biphenyl, 4,4'-bis(thiomethyl)-	504950
720329	0	535418
737796	0	48431592
647889	0	503737
729133	0	48428083
729109	trans-n-(6-chrysenyl)-3-mesyloxy-4-phenyl-2-azetidinone	48428066
699478	0	526491
730922	0	91145667
728512	0	48427885
679091	2-(5-bromouracil)-4'-phenylacetophenone	517910
698945	0	526163
711628	0	531364
650260	0	504775
328567	0	574711
150446	berberine iodide	560320
163088	berberine chloride	561883
625456	0	492395
732820	0	48429795
641459	0	581652
628503	Docetaxel	494081

7.13. APPENDIX M

Patient characteristics	CSS 1: GSE17536 (n = 177)			CSS 2: GSE17537 (n = 55)			OS: Ankara cohort (n = 48)		
	Nr. ^{\$}	HR	P*	Nr.	HR	P*	Nr.	HR	P*
Age									
≤65 (ref.)	83	0.808	0.432	33	0.992	0.990	24	2.173	0.104
>65	94			22			22		
Unknown	-			-			2		
Gender									
Male (ref.)	96	0.918	0.534	26	2.003	0.269	24	0.857	0.740
Female	81			29			24		
TNM Stage**									
1	24	3.623	<0.001	4	14.085	0.001	3	4.434	0.001
2	57			15			0		
3	57			19			38		
4	39			17			5		
Unknown	-			-			2		
Recurrence									
No (ref.)	109	42.869	<0.001	36	364.058	<0.001	NA		
Yes	36			19			NA		
Other/Unknown	32			-			NA		
Grade									
Well Differentiated	16	2.141	0.005	1	3.771	0.119	13	1.279	0.578
Moderately Differentiated	134			32			32		
Poorly Differentiated	27			3			3		
Other/Unknown	-			19			-		
Perineural Invasion									
0 (ref.)	NA			NA			31	4.131	0.003
1	NA			NA			15		
Other/Unknown	NA			NA			2		
Vascular Invasion									
0 (ref.)	NA			NA			23	3.739	0.012
1	NA			NA			23		
Other/Unknown	NA			NA			2		
Microsatellite									
Low or stable (ref.)	146	1.319	0.415	NA			NA		
High	31			NA			NA		

*Cox proportional hazards regression

**Stage: For Ankara cohort - treated as a continuous variable (1: stage 1, 2: stage 3A and 3B, 3: stage 3 and 4. For GSE17536 and GSE17537 - treated as a continuous variable (1, 2, 3, 4)

***MSI status: Determined *in silico* (see Methods) \$Nr.: case numbers

Table obtained with permission from Demirkol et al [2], also reported by Murat İşbilen [3].

7.14. APPENDIX N

MVA-1 (GSE17536) - CSS

Parameters	Hazard Ratio	95%CI	<i>P</i> *
TNM Stage**	5.648	3.48-9.17	<0.001
ULBP2 (3-8 vs. 1-2)	2.682	1.41-5.09	0.003
SEMA5A (4-8 vs. 1-3)	0.459	0.24-0.88	0.019
PCDH7 (4-8 vs. 1-3)	2.765	1.20-6.35	0.017

MVA-2 (GSE17537) - CSS

Parameters	Hazard Ratio	95%CI	<i>P</i> *
TNM Stage**	13.829	2.78-68.87	0.001
ULBP2 (3-8 vs. 1-2)	5.808	1.46-23.09	0.012
SEMA5A (4-8 vs. 1-3)	0.257	0.07-0.95	0.042

*Cox proportional hazards regression

**TNM Stage: Treated as a continuous variable (1, 2, 3, 4)

Table obtained with permission from Demirkol et al [2], also reported by Murat İşbilen [3].

7.15. APPENDIX O

Copyright permission

Copyright permissions for figures and tables taken or reproduced from Demirkol *et al.* [2] <http://www.jcancer.org/v08p1113.htm> (Tables 1.2, 1.3, 3.3, 4.1, 4.2, 4.4, 4.5, 4.6, 4.8, 4.9, 4.10; Figures 1.2, 1.3, 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.12, 4.13, 4.14, 4.15A; APPENDIX M and N). License: Creative Commons Attribution 4.0 International. URL: <http://creativecommons.org/licenses/by/4.0/>.



IVYSPRING
INTERNATIONAL PUBLISHER



Home About License Contact

Int. J. Biol. Sci.
Int. J. Med. Sci.
Journal of Cancer
Theranostics
Journal of Genomics
J. Bone Joint Infect.
Journal of Biomedicine

License

Articles in our journals are distributed under the [Creative Commons Attribution-NonCommercial 4.0 International License](#). We also use the [CC-BY license](#) for authors funded by Wellcome Trust, RCUK and other funding agencies if required, or if authors request.

Articles published by Ivyspring International Publisher are made available in PubMed Central, the U.S. National Institutes of Health (NIH) digital archive of biomedical journal literature. Authors are also encouraged to self-archive the published articles in their own or institutional websites as well as governmental websites, library websites, without permissions required.



Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

You are free to:

- Share** — copy and redistribute the material in any medium or format
- Adapt** — remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

-  **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
-  **NonCommercial** — You may not use the material for [commercial purposes](#).

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

Konu **RE: copyright permission**
Gönderen Journal of Cancer Publishing Team <publish@jcancer.org>
Alıcı 'Seçil Demirkol' <secil.demirkol@bilkent.edu.tr>
Tarih 2017-05-18 12:02



Thanks for your email. As an author or co-author of the article, you are free to include or reuse the materials in other articles or thesis or other publications that you write, with citation of the original source. Permission is automatically granted.

Kind regards,
Publishing Team
Journal of Cancer
<http://www.jcancer.org>

-----Original Message-----
From: Seçil Demirkol [mailto:secil.demirkol@bilkent.edu.tr]
Sent: Friday, May 12, 2017 12:04 AM
To: publish@jcancer.org
Cc: hoca <agure@bilkent.edu.tr>
Subject: copyright permission

Dear Sir or Madam,

I would like to get permission for the attached article titled 'A Combined ULBP2 and SEMA5A Expression Signature as a Prognostic and Predictive Biomarker for Colon Cancer', published in April 2017, to be used in my PhD thesis. I would be glad if you could provide me an official document for the permission.

My full name, affiliation and title: Seçil Demirkol, Bilkent University, PhD student
The figures and tables will be used in 6 hard copies and soft copies of my thesis.

Best regards

--
Secil Demirkol
Phd candidate
Bilkent University
Ankara/TURKEY
Phone:+903122903319

Research Paper

A Combined ULBP2 and SEMA5A Expression Signature as a Prognostic and Predictive Biomarker for Colon Cancer

Secil Demirkol¹, Ismail Gomceli², Murat Isbilen¹, Baris Emre Dayanc³, Mesut Tez⁴, Erdal Birol Bostanci⁵, Nesrin Turhan⁶, Musa Akoglu⁵, Ezgi Ozyerli¹, Sevi Durdu¹, Ozlen Konu¹, Aviram Nissan⁶, Mithat Gonen⁸, Ali Osmay Gure¹ 

1. Department of Molecular Biology and Genetics, Bilkent University, Cankaya, Ankara, Turkey;
2. Department of Gastroenterological Surgery, Antalya Education and Research Hospital, Antalya, Turkey;
3. Faculty of Medicine, Izmir University of Economics, Izmir, Turkey;
4. Fifth Department of Surgery, Ankara Numune Training and Research Hospital, Ankara, Turkey;
5. Department of Gastroenterological Surgery, Yuksek Ihtisas Training and Research Hospital, Ankara, Turkey;
6. Department of General and Oncological Surgery - Surgery C, The Chaim Sheba Medical Center, Tel Hashomer, Israel;
7. Department of Pathology, Yuksek Ihtisas Training and Research Hospital, Ankara, Turkey;
8. Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA.

 Corresponding author: agure@bilkent.edu.tr

© Ivyspring International Publisher. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY-NC) license (<https://creativecommons.org/licenses/by-nc/4.0/>). See <http://ivyspring.com/terms> for full terms and conditions.

Received: 2016.10.11; Accepted: 2017.03.05; Published: 2017.04.09

Abstract

Background: Prognostic biomarkers for cancer have the power to change the course of disease if they add value beyond known prognostic factors, if they can help shape treatment protocols, and if they are reliable. The aim of this study was to identify such biomarkers for colon cancer and to understand the molecular mechanisms leading to prognostic stratifications based on these biomarkers.

Methods and Findings: We used an in house R based script (SSAT) for the *in silico* discovery of stage-independent prognostic biomarkers using two cohorts, GSE17536 and GSE17537, that include 177 and 55 colon cancer patients, respectively. This identified 2 genes, ULBP2 and SEMA5A, which when used jointly, could distinguish patients with distinct prognosis. We validated our findings using a third cohort of 48 patients *ex vivo*. We find that in all cohorts, a combined ULBP2/SEMA5A classification (SU-GIB) can stratify distinct prognostic sub-groups with hazard ratios that range from 2.4 to 4.5 ($p \leq 0.01$) when overall- or cancer-specific survival is used as an end-measure, independent of confounding prognostic parameters. In addition, our preliminary analyses suggest SU-GIB is comparable to Oncotype DX colon(®) in predicting recurrence in two different cohorts (HR: 1.5-2; $p \leq 0.02$). SU-GIB has potential as a companion diagnostic for several drugs including the PI3K/mTOR inhibitor BEZ235, which are suitable for the treatment of patients within the bad prognosis group. We show that tumors from patients with worse prognosis have low EGFR autophosphorylation rates, but high caspase 7 activity, and show upregulation of pro-inflammatory cytokines that relate to a relatively mesenchymal phenotype.

Conclusions: We describe two novel genes that can be used to prognosticate colon cancer and suggest approaches by which such tumors can be treated. We also describe molecular characteristics of tumors stratified by the SU-GIB signature.

Key words: Colon Cancer, Prognosis, Biomarker.

Introduction

Foreseeing the prognosis of colon cancer patients, especially those with stage II or III disease is critical as it will change treatment decisions. In addition to known prognostic factors, such as stage, mismatch repair proficiency, Kras and Braf mutation

status, various prognostic assays based on molecular signatures have been developed as clinical tests (e.g. Oncotype DX colon, ColoPrint). A recent report that evaluated most oncological prognostic tests, including those for colon cancer, highlighted various

deficiencies in discovery and validation studies [1]. These include the lack of complete multivariate analyses where all prognostic factors are considered, lack of an evaluation of all survival measures, including overall survival, disease-free survival and response to therapy, and lack of validation in multiple cohorts. These criteria, as well as the possibility of using the test "in house" at minimum cost are considered essential in expediting the utilization of a molecular test in the clinic. We planned this study aiming to address most of these issues utilizing an *in silico* approach to identify a minimum number of prognostic genes and performed *ex vivo* as well as *in silico* validation studies with multiple cohorts. Our results suggest that the joint evaluation mRNA expression of two novel genes, ULBP2 and SEMA5A in colon cancer tumor tissue can identify patients with good, intermediate and bad prognosis, independent of stage, K-ras/B-raf mutation, and mismatch repair status. As prognostic signatures are considered valuable especially if they can predict response to therapy, we searched for therapeutic agents that would be suitable for the treatment of those patients with a predicted bad prognosis. Our results indicate that the PI3K/mTOR inhibitor NVP-BEZ235 is such a drug. At a molecular level, we find that tumors from patients with bad prognosis have decreased EGFR-Y1068 phosphorylation and increased Caspase 7 cleavage, together with up-regulation of inflammatory cytokines; in line with our findings showing such tumors are more mesenchymal, as opposed to epithelial in character.

Methods

Study cohorts

Patient characteristics of the 4 cohorts used in this study are summarized in Table S1 and Table S2. Datasets with gene expression data from 177 and 55 tumor tissues (GSE17536 and GSE17537) were used as discovery cohorts [2]. Median survival times for patients within the GSE17536 and -37 cohorts were 42.2 (0.9-142.6) and 41.7 (0.4-70.5) months respectively. For *ex vivo* validation of biomarkers of overall survival (OS) we used tumor tissues obtained from a third cohort (Ankara) consisting of 46 patients with a pathological diagnosis of colon cancer, collected at the Department of Gastroenterological Surgery, Yuksek İhtisas Training and Research Hospital, Ankara, Turkey, following informed consent obtained from all patients. Information regarding overall survival time, follow-up status, age, gender, TNM stage, grade, perineural invasion and vascular invasion were available for all patients. Representative images from tumor sections are shown

in Figure S1. Follow up times ranged from 1 to 49 months. Median survival was 23 months for this cohort, as most patients had stage III disease. For genes that associated with OS, association with disease-free survival (DFS) was assessed *in silico* using the aforementioned datasets in addition to a third: GSE39582, which included data from 566 patients [3].

In silico methods

Microarray data analysis

Gene expression data of tumor datasets were downloaded from GEO (<http://www.ncbi.nlm.nih.gov/geo/>) and GC-RMA normalized without baseline transformation using GeneSpring v12.0. Clinical data related to these were downloaded from ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>). For cell line gene expression analysis GSE36133 (Cancer Cell Line Encyclopedia) [4] and E-MTAB-783 (Cancer Genome Project) [5] datasets were downloaded from ArrayExpress and RMA normalized using GeneSpring v12.0.

Semisupervised survival analysis tool (SSAT)

This analysis aims the identification of probesets that can stratify patients into 2 groups with significantly different survival times as assessed by the log-rank test when a cut-off value of either 2, 4, 6, 8, 10, 12 or 14 (log2 expression data) is used. Thus, 7 different log-rank tests were performed for each probeset using an R (Bioconductor) based script. For a given probeset, the cut-off value with the lowest p value was used for further analyses. For GSE17536, SSAT could identify 24 of 48 genes that were previously validated as prognostic biomarkers for colon cancer [6]. Among significant genes, those that gave hazard ratios (HR) in opposite direction between two analyses (GSE17536 and GSE17527) were 0%, 5%, and 12.5% for the first 10, 20, and 40 most significant genes, respectively.

Oncotype Dx colon risk group assignment

Recurrence risk was calculated based on the expression of 7 test and 5 control genes, and each sample assigned to either a low or high risk group as previously described [3].

Microarray based microsatellite instability determination

Affymetrix probesets corresponding to a genomic signature that identifies colorectal cancer patients with MSI [7] available in the HGU 133A platform were identified. The median expression values of these probesets were calculated for samples within the GSE41258 human colon cancer dataset for MSI-high or -low, versus MSS subgroups, as the MSI

status of samples in this dataset are known. These median values were then used as reference centroids for the classification of samples in GSE26682 (HGU133 Plus 2.0) and GSE17536 (HGU133A) datasets. A correlation based classification was performed using the reference median values (MSI and MSS median values of 46 probesets) and the expression values for each test sample. The difference between MSI and MSS correlation values were used to generate a MSI index value for each sample. A threshold index value (0.45) that gave the maximum accuracy in both platforms was determined based on a ROC analysis. Samples with index values above 0.45 were classified as MSI, or MSS if the value was equal or below this threshold.

TCGA based expression and proteomic analyses

Level 3 RNA seq. and level 3 proteome data for 132 colon cancer primary tumor tissues were downloaded from cancergenome.nih.gov via the TCGA data portal. Tumors were classified according to the SU-GIB signature using median gene expression values as thresholds for both ULBP2 and SEMA5A, yielding 37, 55 and 40 samples with good (G), intermediate (I) and bad (B) prognosis, respectively. ANOVA test was used to identify proteins with significantly different expression between groups, based on the RPPA data. ANOVA based differences in expression of inflammatory cytokines were identified using level 3 RNA-seq. data of TCGA.

RT-PCR

Total RNA was extracted from either snap-frozen tissue using Trizol reagent (Ambion, Foster City, CA, USA) following the manufacturers protocol. Following DNase I treatment using the DNA-freeTM Kit (Ambion, Foster City, CA, USA), total RNA concentration and purity were measured using Nanodrop spectrophotometer ND-1000 (Thermo Scientific, MA, USA). 500 ng RNA was reverse transcribed into cDNA using Revert-aid first strand cDNA synthesis kit (Thermo Fisher Scientific, Boston, MA, USA) using random primers following the suggested protocol.

All Quantitative-PCR experiments except for ULBP2 were performed using SYBR Green technology (Applied Biosystems, CA, USA). Custom designed Taqman primers & probes were used for quantification of ULBP2 in frozen tissues (Hs00607609_mH). GAPDH and ACTB were used as reference genes (4352934E, 4352935E) [8] [9]. Primers used for SYBR Green experiments are given in Table S3. Q-PCR was performed in a ABI 7500 Real Time PCR System (Applied Biosystems, Carlsbad, CA,

USA) using either Taqman universal master mix (Applied Biosystems, Carlsbad, CA, USA) or Power SYBR Green master mix (Applied Biosystems, Carlsbad, CA, USA), according to manufacturer's protocols. Thermal conditions were 50°C for 2 min for UNG enzyme activity, an initial denaturation stage of 95°C for 10 min., followed by 45 cycles of 95°C 15 sec. denaturation, 60°C 1 min. of annealing and extension. Data was analyzed using ABI 7500 Software v2.0.4 and gene expression normalized to two control genes GAPDH, ACTB for Taqman experiments and three control genes GAPDH, ACTB, 18S for SYBR Green. All experiments were repeated at least 3 times for each reaction. Gene expression was quantified using the $2^{-\Delta\Delta CT}$ method [10].

Statistical methods

Statistical analyses were performed using Graph Pad Prism 6 (GraphPad Prism 6 Software, San Diego, CA, USA) or SPSS Statistics v.19 (IBM, 2010, Chicago, IL, USA). P values below 0.05 were considered statistically significant. ANOVA tests were used to determine differences among multiple groups when comparing gene expression, or cytotoxicity. Kaplan-Meier plots and the log-rank test were utilized to determine significance in survival differences among prognostic groups. Univariate and multivariate cox proportional hazards regression analysis was performed using SPSS Statistics v.19.

Results

Identifying prognostic biomarkers for colon cancer with cancer-specific (CSS) or overall survival (OS) as end-point measures

We generated an R based tool (semi-supervised survival analysis tool: SSAT) to discover genes whose expression correlated with survival (see Methods). SSAT analysis of two colon cancer datasets, GSE17536 and GSE17527, revealed 400 and 269 genes respectively, which could generate a statistically significant stratification using at least at one of the 7 SSAT defined thresholds. Among these, 64 probesets were common to both datasets, had HRs in the same direction and shared the same threshold value for both cohorts and therefore were studied further. Probesets were ranked from the most to the least significant and 20 with the lowest weighted ranksum values were included in a multivariate model (MVA) that included stage as a confounding factor using GSE17536 cancer-specific survival data (Table S4). The MVA identified 3 probesets/genes as independent prognostic biomarkers of cancer-specific survival (Table S5). The same analysis was then performed using these 3 probesets and stage for

GSE17537; which revealed that other than stage, ULBP2 and SEMA5A were independent prognostic biomarkers for colon cancer when all stages were included in the analysis (Figure S2, Figure S3 and Table S5) as well as when only stage 2 and 3 patients were evaluated (Figure S4). ULBP2 and SEMA5A gene expression were also found to be linearly associated with cancer specific survival (Table S6). To generate a combined gene signature for prognosis prediction we used SSAT identified cut-off values and classified SEMA5A-high and ULBP2-low patients as those with good prognosis (G), while the reverse predicted bad (B), and the rest patients with intermediary (I) prognosis. This "SU-GIB" stratification was a highly significant prognostic classifier for both cohorts (Figure 1), independent of confounding factors (Table 1). To validate these findings *ex vivo* we used tumor tissue from a third cohort of patients with colon cancer for which overall survival time was available as an end-point measure (Table S1). The median expression value for SEMA5A and ULBP2 mRNA levels, as determined by Q-PCR was used to stratify these patients into good and bad prognosis groups. Similar to our *in silico* findings, this identified distinct prognostic groups when all stages were included in the analysis (Figure S5), as well as for patients with stage 2 and 3 disease (Figure S6); and the SU-GIB stratification revealed significantly distinct prognostic subgroups when qPCR data was used (Figure 1). Finally, a MVA that included grade, stage, age, gender and the SU-GIB classification showed that the SEMA5A/ULBP2 based stratification could predict prognosis independent of confounding factors (Table 1). The stage independence of SU-GIB

classification is preserved when only stage II&III patients were included into analysis (Table S7).

As DNA mismatch repair (MMR) status of colon cancer can be a prognostic factor, we asked if SU-GIB would further stratify patients with and without microsatellite instability (MSI). For this purpose we predicted the MSI status of samples within the GSE17536 dataset and performed SU-GIB stratification for both MSI and MSS samples. This revealed a significant difference between prognostic groups in MSS patients. Although the trend was in the expected direction for the MSI group, log-rank test did not show a significance, probably due to a low sample size (Figure S7). In summary, our results suggest that SU-GIB stratification can predict subgroups of patients with differing prognosis, independent of confounding factors.

SU-GIB and disease-free survival (DFS)

SU-GIB stratification identified 3 significantly distinct prognostic groups when DFS was used as an end-point measure in both GSE17536 and GSE17537 datasets (Figure S8). When MSI and MSS stable groups were analyzed separately a significant stratification for patients especially with MSS tumors could be seen (Figure 2). This was significant for stage 2 patients, but was also evident (albeit not significant) for stage 3 patients (Figure 3). As the GSE39582 dataset contains *KRAS* and *BRAF* mutation data, we included this information in a MVA that included additionally SU-GIB and stage. In this model, mutational status was not an independent prognostic predictor among patients with MSS tumors (Table S8).

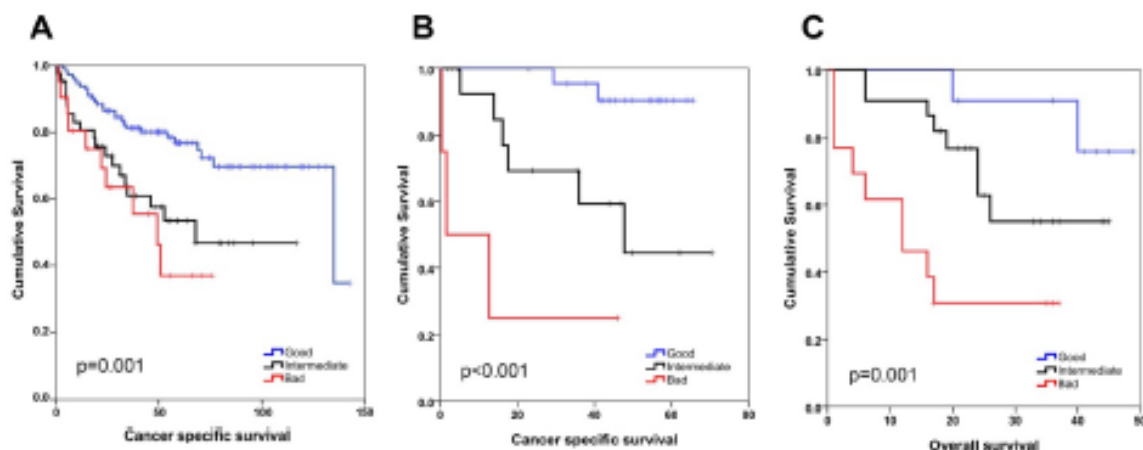


Figure 1. SEMA5A/ULBP2 (SU) gene signature for colon cancer prognostication. Kaplan-Meier plots based on the SU signature for GSE17536 (A), GSE17537 (B) and the Ankara cohort (C), and their respective log-rank *p* values are shown. Survival times are in months. A combined score based on ULBP2 and SEMA5A expression that separates patients into good (SEMA5A high, ULBP2 low), bad (SEMA5A low, ULBP2 high) and intermediate (patients with both SEMA5A and ULBP2 high or both low) groups results in better stratification of colon cancer patients. ULBP2 and SEMA5A cut-off values were 4 and 6, respectively for *in silico* analyses. For the Ankara cohort, cut-off values correspond to the median expression value for both genes.

Table 1. Multivariate analysis of clinicopathological parameters and SU-GIB

GSE17530 - CSS	Hazard Ratio	95%CI	P*
Grade**	1.165	0.685 - 1.979	0.573
Stage†	5.028	3.133 - 8.068	<0.001
Age (above 65 vs equal to or below 65)	0.818	0.455 - 1.467	0.500
Gender (female vs. male)	1.018	0.567 - 1.831	0.951
MSI_transcription based (stable vs. instable)	1.319	0.558 - 3.119	0.529
SU-GIB*	2.452	1.616 - 3.720	<0.001
GSE17537 - CSS	Hazard ratio	95% CI	P*
Stage†	13.662	2.769 - 67.400	0.001
Age (equal to or below 65 vs. above 65)	1.370	0.292 - 6.434	0.690
Gender (female vs. male)	0.924	0.183 - 4.668	0.924
SU-GIB*	4.502	1.557 - 13.017	0.005
Arizono Cohort - OS	Hazard ratio	95% CI	P*
Grade**	1.171	0.409 - 3.358	0.769
Stage‡	4.481	1.597 - 12.574	0.004
Age (above 65 vs equal to or below 65)	2.300	0.857 - 6.174	0.098
Gender (female vs. Male)	1.062	0.407 - 2.775	0.902
SU-GIB*	3.481	1.514 - 8.004	0.003

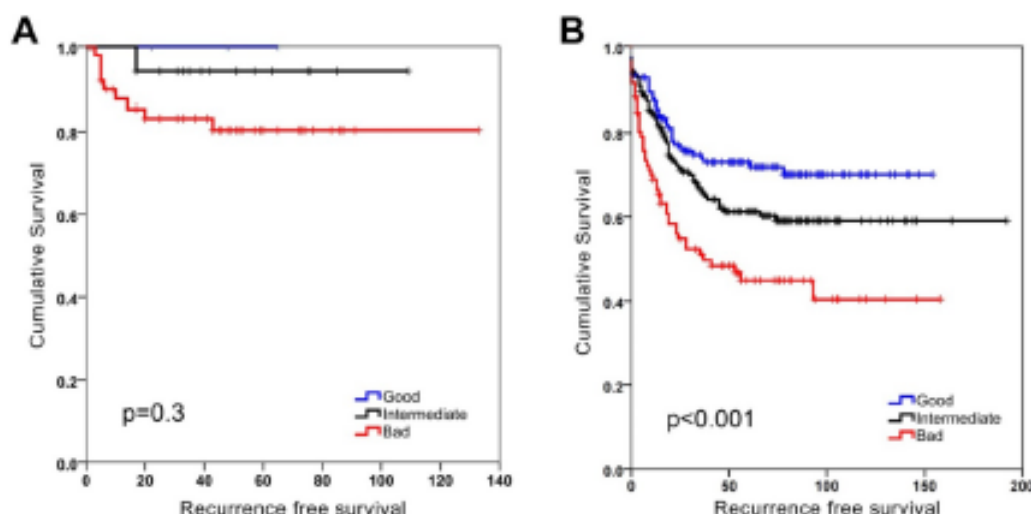
*Cox proportional hazards regression

**Grade: Treated as a continuous variable; Poorly differentiated (1), moderately differentiated (2), well differentiated (3)

†Stage: Treated as a continuous variable (1, 2, 3, 4)

‡Stage: Treated as a continuous variable (1: stage 1, 2: stage 3A and 3B, 3: stage 3C and 4: stage 4)

*SU-GIB: Treated as a continuous variable (1: Good, 2: Intermediate, 3: Bad)

**Figure 2.** SU-GIB can predict recurrence-free survival especially for microsatellite stable patients in GSE39582. Kaplan-Meier plots of microsatellite instable (MSI, A) and stable (MSS, B) patients stratified independently using SU-GIB, and log-rank p values are shown. Cut-off values for both series were the most significant within the 25% and 75% interquartile range.

SU-GIB prediction compared to other molecular classifiers

Among several prognostic molecular classifiers for colon cancer, Oncotype DX colon cancer (Oncotype) developed to predict recurrence [6], is one that can be applied to microarray expression data [3]. We, therefore, tested how the SU-GIB classification would compare to the Oncotype score for colon cancer patients with stage II or III disease. A MVA that included age, gender, stage, tumor grade, MSI

status and the prognostic tests revealed that both SU-GIB or Oncotype were independent prognostic classifiers for DFS in cohorts GSE17536 and GSE39582 (Table S9). When SU-GIB and Oncotype were both included in the MVA for GSE39582, SU-GIB was still an independent significant prognostic parameter when Oncotype was not, suggesting that the two tests identify mostly overlapping prognostic sub-groups (Table S10). We also asked if the SU-GIB classifier could stratify colon cancer patient groups as identified by Marisa et al., which is a functional

classification with 6 molecularly distinct subtypes (C1-6). We find that SU-GIB can distinguish 3 groups within all categories. However, C1 (DMMR) and C5 (CINwntup) contained the largest number for G tumors, whereas the CSC group (C2) contained the largest number of B samples.

Overall, we find that the SU-GIB classification is as powerful as Oncotype in predicting patients with recurrence and could be useful even for colon cancer subtypes sub-grouped based on given gene expression signatures.

Molecular characteristics of G, I and B tumors

The Cancer Genome Atlas (TCGA) database contains protein as well as RNA based expression data for tumors. We, therefore, asked if the SU defined prognostic groups differed in terms of signaling pathways as could be determined by an analysis of the TCGA reverse-phase protein expression data (RPPA). When tumors corresponding to SU-G, I and B groups were compared, we found that "G" tumors showed significantly increased EGFR Y1068, Y992 and Shc Y317 phosphorylation, while

cleaved (activated) Caspase 7 levels were increased in "B" tumors (Figure 4 and S9). As Caspase 7 activation promotes NF κ B target activation [11], we asked if inflammatory cytokines were transcriptionally upregulated in such tumors. Indeed, several inflammatory cytokines, including IL6, IL1 β , TGF β 1, TGF β 3, IL1R2, and TNF α were upregulated among B tumors (Figure 5 and S10). Inflammation has been associated with epithelial-to-mesenchymal transition (EMT) in several tumors including colon cancer [12]. When tumors with G, I and B genotypes as determined for the GSE36582 dataset were analyzed for Vimentin and E-cadherin expression (mesenchymal and epithelial markers, respectively), we observed that the majority of "G" tumors are epithelial in nature while "B" tumors are mostly mesenchymal when "I" tumors show an intermediary phenotype (Figure S11). We, therefore, conclude that the SU-GIB classification identifies tumor subgroups that differ in terms of EGFR signaling, inflammation and EMT-related characteristics.

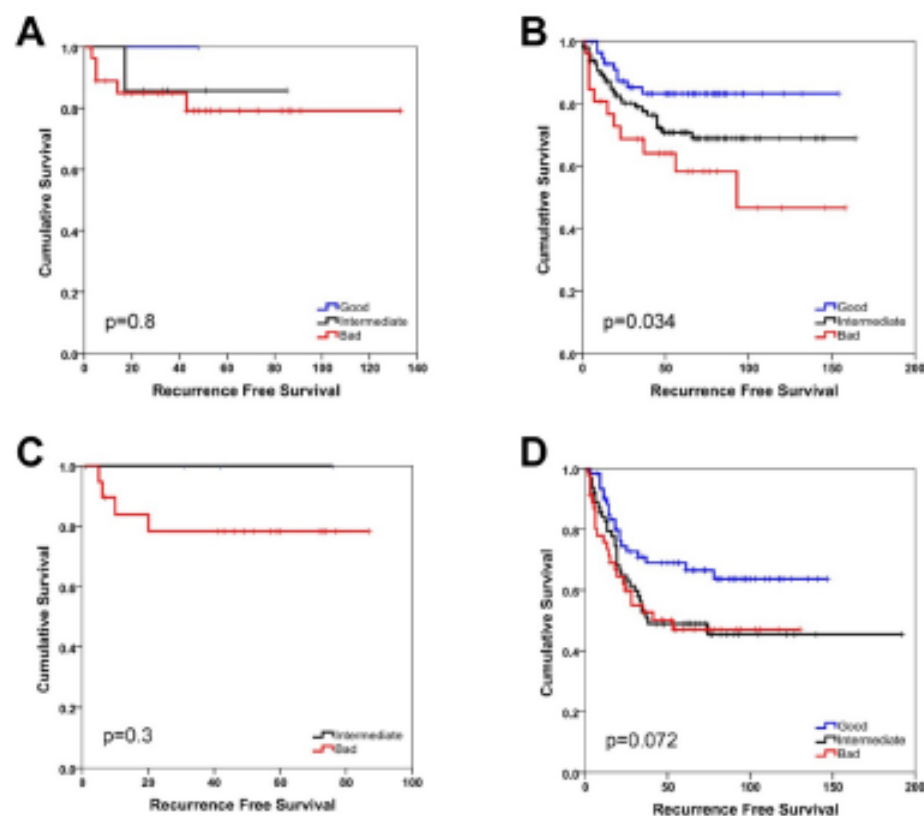


Figure 3: SU-GIB can predict recurrence-free survival for microsatellite stable patients with stage 2 or 3 disease in GSE39582. Kaplan-Meier graphs and log-rank p values are shown for stage 2 (A, B) and 3 (C, D) patients with microsatellite unstable (A, C) and stable (B, D) tumors. Cut-off values for both genes were the most significant within the 25% and 75% interquartile ranges.

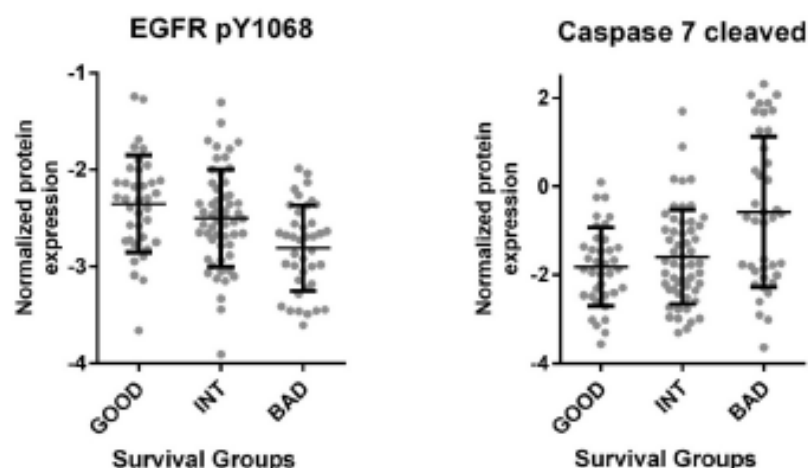


Figure 4. TCGA based proteome analysis of colon cancer tumor tissue reveals increased EGFR phosphorylation and decreased Caspase 7 cleavage in the good prognosis group. RNA seq. and proteome data for 132 colon cancer primary tumor tissues downloaded from "cancerzenomix.nih.gov" via the TCGA data portal classified according to the SU signature (bad survivors: 40, good survivors: 37, intermediate survivors: 55) revealed increased EGFR 1068 phosphorylation among patients with better prognosis ($p < 0.001$, 1-way Anova) (left). The same analysis showed increased Caspase 7 cleavage in patients within the worse survival group ($p < 0.0001$) (right). Caspase 7 cleavage was directly correlated with ULBP2 and inversely with SEMA5A expression ($p < 0.001$ for both series by Pearson's r). The mean and standard deviation for each group are indicated.

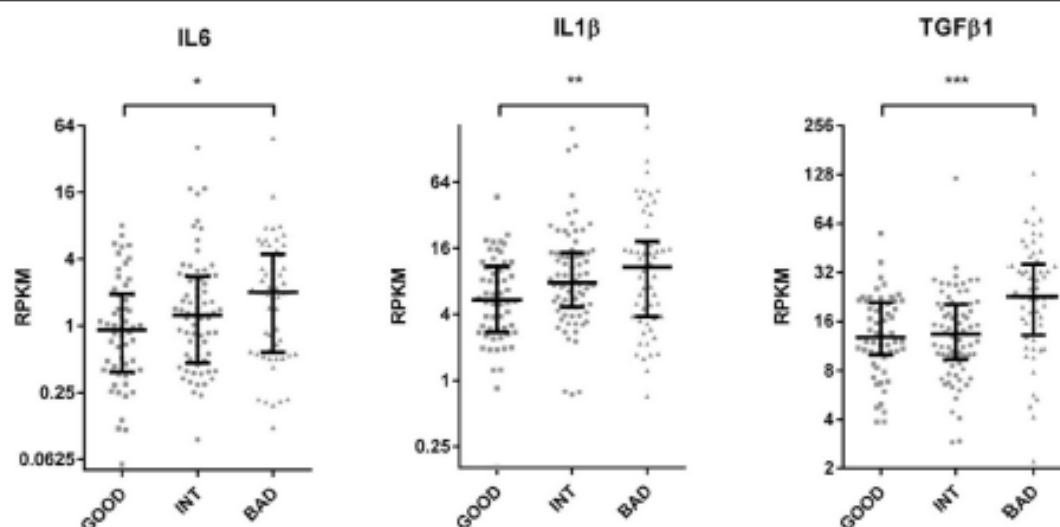


Figure 5. TCGA based RNA seq. analysis of colon cancer tumor tissue reveals increased inflammatory cytokine gene expression in the bad prognosis group. TCGA colon tumor samples were stratified into GIB groups using median expression values as cut-offs for ULBP2 and SEMA5A. RPKM values of IL6, IL1 β , TGF β 1 are plotted for "good", "intermediate" and "bad" groups. T-test p values between "good" and "bad" groups. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.0001$. RPKM: reads per kilobase per million mapped reads. The median and inter-quartile ranges for each group are indicated.

SU-GIB as a potential predictor of drug responsiveness

As prognostic classifiers are especially valuable if they can also predict response to therapy we aimed to identify drugs that would be suitable for the treatment of especially the bad prognosis group as identified by SU-GIB. For this purpose, we correlated IC50 values of all drugs tested for colon cancer cell lines in the CGP database [5] with SEMA5A and

ULBP2 expression, and identified 13 drugs whose IC50 values showed significant direct and inverse correlations with SEMA5A and ULBP2 expression, respectively (Table S11). Among these drugs, the most significant correlation with both genes in the expected direction was for the PI3K/mTOR inhibitor NVP-BEZ235 (Figure 6), suggesting that the SU-GIB stratification can possibly be utilized as a companion diagnostic test of sensitivity for a variety of drugs.

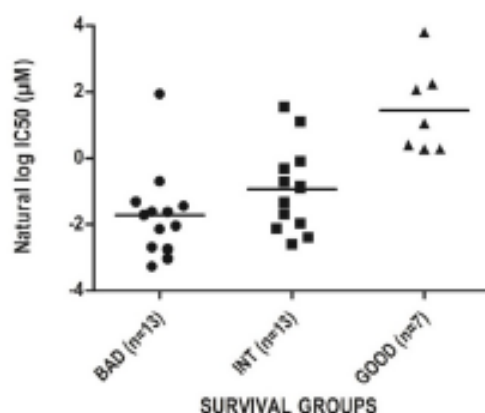


Figure 6. Chemosensitivity profiles of colon cancer cell lines corresponding to SU-G, -I and -B phenotypes. *In silico* analysis of colon cancer cell lines as classified by the SU signature into G, I or B phenotypes for NVP-BEZ235 cytotoxicity based on CGP data. Cut-off values used to generate a SU-GIB stratification were either the SSAT generated (SEMA5A) or the median expression values (ULBP2).

Discussion

Despite the availability of a large number of molecular tests capable of determining prognosis of colon cancer patients, very few have progressed beyond diagnostic and initial validation studies. An extensive report by the US Department of Health and Human Services which reviewed prognostic tests for cancer highlighted critical aspects that would have to be considered in the design of future discovery and validation studies, to help speed their introduction into routine clinical practice [1]. As the report emphasized the use of different patient cohorts, we designed this study to include multiple cohorts for both *in silico*, as well as *ex vivo* validation. The same report stresses the importance of performing MVAs that not only include all confounding factors but also utilize OS and/or CSS as end measures, in addition to DFS. We therefore utilized OS/CSS in our discovery method, subsequently validating our gene signature for DFS as well. Our results suggest that the SU-GIB stratification can be developed as a clinical prognostic test for colon cancer, especially since the strength of the test is comparable to that of an *in silico* adapted version of Oncotype DX colon. Obviously extensive validation studies are required to define the best patient group that would benefit from the SU-GIB test, as has been done previously for others [13, 14]. This would need to be followed by studies that would assess if treatment decisions can be altered by the SU-GIB test, which is considered a major criterion for clinical usability [15]. The *in silico* predictions for MSI status and Oncotype scores performed in this study are approximations at best, and actual test results of

these would need to be included in validation studies.

To our knowledge, this is the first report to associate SEMA5A and ULBP2 expression with clinical outcome in colon cancer. Semaphorins were first identified as axon guidance molecules involved in regulation of neural connections [16]. Type 5 semaphorins are transmembrane proteins present in both invertebrates and vertebrates [17]. In cancer, their upregulation has been related to both good and bad prognosis [18]. For example *Sema5a* transfection to pancreatic cancer cell lines resulted in higher cell invasion *in vitro*, and injection of such cells to nude mice enhanced tumorigenicity and tumor growth *in vivo* [19]. Similarly, SEMA5A has been shown to promote *in vitro* migration and invasion of gastric cancer cell lines [20]. On the other hand, SEMA5A has been reported to impede glioma cell motility [21], and has been associated with better prognosis in women with non-small cell lung cancer [22]. According to the Human Protein Atlas, SEMA5A expression occurs mostly in epithelial cells (www.proteinatlas.org), possibly indicating that this is an epithelial cell marker for colon cancer. On the other hand, analysis of ULBP2 expression in the same database reveals that it occurs more often in fibroblasts, compared to surrounding cells in various organs, and therefore in cells that are more mesenchymal in nature. ULBP2, a NKG2D ligand, is highly expressed in transformed cells [23] and is thought to provide a ligand for NK cells, facilitating tumor cell death [24-26]. The reason up-regulation of ULBP2 is related to a worse outcome could be because secretion or cleavage of this protein is related to reduced NK mediated tumor lysis [27]. In this line, an increase in expression of the tumor-associated or the secreted form of ULBP2 has been associated with an unfavorable outcome for several cancers including melanoma, non-small cell lung carcinoma and ovarian carcinoma [28-30]. Hence, the known molecular characteristics of these two genes help understand how they might affect prognosis in colon cancer. Our findings show that patients with a better outcome have increased phosphorylation of EGFR at Tyrosine 1068. The autophosphorylation of this residue is necessary for Grb2-EGFR binding and is strongly inhibited by Gefitinib [31, 32]. It is somewhat surprising then, that a PI3K/mTOR inhibitor is effective preferentially on tumors from patients with a worse prognosis, when both PI3K and mTOR are activated downstream to EGFR. However, our analysis of the GSE39582 database, which contains mutation data, shows that tumors from only 20% of SU-G patients contain Kras or Braf mutations, while tumors from B patients are mutated at a rate of 60%. This suggests that most patients with EGFR Tyr 1068 phosphorylation might

not harbor Kras/Braf mutations; and that the PI3K-mTOR inhibitor BEZ235 might be more effective on tumors with Kras/Braf mutations, rather than those with EGFR autophosphorylation. Several reports indicate an inverse association of EGFR/Kras/MAPK and PI3K activity and caspase 7 activation by its cleavage [33, 34], similar to our findings. Caspase 7 activation is involved in inflammation [35] and enhances NFkB activity [11]. It is therefore, not surprising that we find several inflammatory cytokines upregulated among SU-B tumors. Inflammation, in turn, is closely associated with epithelial to mesenchymal transition [36]. Our results show that SU-B tumors are significantly more mesenchymal in their phenotype compared to SU-G tumors, as assessed by Vimentin and E-cadherin expression, two genes which are considered reliable markers of mesenchymal and epithelial cells, respectively [37]. In summary, our data shows that the SU-GIB stratification not only defines distinct prognostic sub-groups but also functionally distinct colon cancer sub-types. Further validation of this signature, especially in prospectively studied cohorts, and further characterization of SU-GIB stratified tumors could reveal various novel options for treatment and diagnosis of colon cancer.

Supplementary Material

Supplementary figures and tables.

<http://www.jcancer.org/v08p1113s1.pdf>

Abbreviations

OS, overall survival; DSS, disease-specific survival; DFS, disease-free survival; SSAT, semi-supervised survival analysis tool; RMA, robust multi-array average; TCGA, The Cancer Genome Atlas; SU-GIB, SEMA5A/ULBP2 expression based good-intermediate-bad prognosis classification; MSI, microsatellite instable; MSS, microsatellite stable; MVA, multivariate analysis; R, Bioconductor.

Acknowledgments

We thank Drs. Koray D. Kaya, Sukru Atakan and Alper Poyraz for helpful guidance and discussions.

Funding support

TUBITAK 1125304 to AOG; TUBITAK BİDEB to SD, Mİ, BED, EO and SD(2); NIH/NCI Cancer Center Support Grant P30 CA008748 to MG.

Competing Interests

The authors have declared that no competing interest exists.

References

- Meleth S, Reeder-Hayes K, Ashick M, Clark R, Funkhouser W, Wines R, et al. Technology Assessment of Molecular Pathology Testing for the Estimation of Prognosis for Common Cancers. Technology Assessment of Molecular Pathology Testing for the Estimation of Prognosis for Common Cancers. Rockville (MD): Agency for Healthcare Research and Quality of the U.S. Department of Health and Human Services; 2014.
- Smith JJ, Deane NG, Wu F, Merchant NR, Zhang B, Jiang A, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*. 2010; 138: 958-68.
- Martins L, de Reynies A, Duval A, Selvaraj S, Gaub MP, Vescovo L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS medicine*. 2013; 10: e1001453.
- Barretina J, Caporaso G, Shimony N, Vekatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012; 483: 603-7.
- Garnett MJ, Edelman HJ, Heideman SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. 2012; 483: 570-5.
- O'Connell MJ, Lavery I, Yothers G, Paik S, Clark-Langone KM, Lopatin M, 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. *J Clin Oncol*. 2010; 28: 3937-44.
- Tian S, Roepman P, Popovici V, Michaut M, Majewski I, Salazar R, et al. A robust genomic signature for the detection of colorectal cancer patients with microsatellite instability phenotype and high mutation frequency. *J Pathol*. 2012; 228: 586-95.
- Li Z, Tian T, Lv F, Chang Y, Wang X, Zhang L, et al. Scl-1 promotes proliferation of pancreatic cancer cells via upregulation of cyclin D1 expression. *PLoS ONE*. 2013; 8: e59203.
- Lajmi E, Leconte L, Pedron-Mazoyer S, Ropert S, Racul W, Lavalette S, et al. Netrin-4 inhibits angiogenesis via binding to neogenin and recruitment of UNC5B. *Proc Natl Acad Sci U S A*. 2008; 105: 12491-6.
- Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*. 2001; 25: 402-8.
- Ehrer S, Petrilli V, Kawaner I, Minotti R, Castillo R, Santoro R, et al. Inflammasome-activated caspase 7 cleaves PARP1 to enhance the expression of a subset of NF-kappaB target genes. *Mol Cell*. 2012; 46: 200-11.
- Wang H, Wang HB, Zhou BH, Li CL, Zhang F, Wang XF, et al. Epithelial-mesenchymal transition (EMT) induced by TNF-alpha requires AKT/GSK-3beta-mediated stabilization of snail in colorectal cancer. *PLoS ONE*. 2013; 8: e56664.
- Yothers G, O'Connell MJ, Lee M, Lopatin M, Clark-Langone KM, Millward C, et al. Validation of the 12-gene colon cancer recurrence score in NSABP C-07 as a predictor of recurrence in patients with stage II and III colon cancer treated with fluorouracil and leucovorin (FU/LV) and FU/LV plus oxaliplatin. *J Clin Oncol*. 2013; 31: 4812-9.
- You YN, Rustin RB, Sullivan JD. Oncotype DX(R) colon cancer assay for prediction of recurrence risk in patients with stage II and III colon cancer: A review of the evidence. *Surgical oncology*. 2015; 24: 61-6.
- Srinivasan G, Renfro LA, Behrens RJ, Lopatin M, Chao C, Socot GS, et al. Prospective multicenter study of the impact of oncotype DX colon cancer assay results on treatment recommendations in stage II colon cancer patients. *The oncologist*. 2014; 19: 402-7.
- He Z, Wang KC, Koprivica V, Ming G, Song HJ. Knowing how to navigate: mechanisms of semaphorin signaling in the nervous system. *Sci STKE*. 2002; 2002: re1.
- Bahri SM, Chia W, Yang X. Characterization and mutant analysis of the *Drosophila* *sema 5c* gene. *Dev Dyn*. 2001; 221: 323-30.
- Tarnagone L. Emerging role of semaphorins as major regulatory signals and potential therapeutic targets in cancer. *Cancer Cell*. 2012; 22: 145-52.
- Sadanandam A, Varney ML, Singh S, Ashour AE, Moriaux N, Deb S, et al. High gene expression of semaphorin 5A in pancreatic cancer is associated with tumor growth, invasion and metastasis. *International journal of cancer Journal international du cancer*. 2010; 127: 1373-83.
- Pan G, Zhu Z, Huang J, Yang C, Yang Y, Wang Y, et al. Semaphorin 5A promotes gastric cancer invasion/metastasis via urokinase-type plasminogen activator/phosphoinositide 3-kinase/protein kinase B. *Dis Dis Sci*. 2013; 58: 2197-204.
- Li X, Law JW, Lee AY. Semaphorin 5A and plectin-B3 regulate human glioma cell motility and morphology through Rac1 and the actin cytoskeleton. *Oncogene*. 2012; 31: 595-610.
- Lu TP, Tsai MH, Lee JM, Hsu CP, Chen PC, Lin CW, et al. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev*. 2010; 19: 2590-7.
- Groh V, Ritzelhart R, Secret H, Bauer S, Grabstein KH, Spies T. Broad tumor-associated expression and recognition by tumor-derived gamma delta T cells of MICA and MICB. *Proc Natl Acad Sci U S A*. 1999; 96: 6679-84.
- Cosman D, Mullberg J, Sutherland CL, Chin W, Armitage R, Fanslow W, et al. ULBP6, novel MHC class I-related molecules, bind to CMV glycoprotein UL16

- and stimulate NK cytotoxicity through the NKG2D receptor. *Immunity*. 2001; 14: 125-33.
25. Ohashi M, Eagle RA, Townsend J. Post-translational modification of the NKG2D ligand RAET1G leads to cell surface expression of a glycosylphosphatidylinositol-linked isoform. *J Biol Chem*. 2010; 285: 16408-15.
 26. Hayakawa Y, Kelly JM, Westwood JA, Darcy PK, Dieffenbach A, Raulet D, et al. Cutting edge: tumor rejection mediated by NKG2D receptor-ligand interaction is dependent upon perforin. *J Immunol*. 2002; 169: 5377-81.
 27. Waldhauer I, Steinle A. Proteolytic release of soluble UL16-binding protein 2 from tumor cells. *Cancer Res*. 2006; 66: 2520-6.
 28. Yamaguchi K, Chikumi H, Shimizu A, Takata M, Kinoshita N, Hashimoto K, et al. Diagnostic and prognostic impact of serum-soluble UL16-binding protein 2 in lung cancer patients. *Cancer science*. 2012; 103: 1405-13.
 29. Nuckel H, Switala M, Sellmann L, Horn PA, Dürig J, Dührsen U, et al. The prognostic significance of soluble NKG2D ligands in B-cell chronic lymphocytic leukemia. *Leukemia*. 2010; 24: 1152-9.
 30. Rothe A, Jachimowicz ED, Boeckmann S, Madlener M, Kessler J, Reiners KS, et al. The bispecific immunoligand ULBP2-αCEA redirects natural killer cells to tumor cells and reveals potent anti-tumor activity against colon carcinoma. *International journal of cancer Journal international du cancer*. 2014; 134: 2829-40.
 31. Batzer AG, Rottin D, Urra JM, Skolnik EY, Schlessinger J. Hierarchy of binding sites for Grb2 and Shc on the epidermal growth factor receptor. *Mol Cell Biol*. 1994; 14: 5192-201.
 32. Prucharson J, McCormick F, Elisei DW, Tabet O. EGFR inhibition evokes innate drug resistance in lung cancer cells by preventing Akt activity and thus inactivating Bcl-1 function. *Proc Natl Acad Sci U S A*. 2015; 112: E3855-63.
 33. Chen G, Kronenberg P, Teugels E, Umelo IA, De Groot J. Targeting the epidermal growth factor receptor in non-small cell lung cancer cells: the effect of combining RNA interference with tyrosine kinase inhibitors or cetuximab. *BMC medicine*. 2012; 10: 28.
 34. Bai H, Li H, Li W, Gu T, Yang J, Cao D, et al. The PI3K/AKT/mTOR pathway is a potential predictor of distinct invasive and migratory capacities in human ovarian cancer cell lines. *Oncotarget*. 2015; 6: 25520-32.
 35. Lamloeff M, Kanneganti TD. Caspase-7: a protease involved in apoptosis and inflammation. *The international journal of biochemistry & cell biology*. 2010; 42: 21-4.
 36. Lopez-Novoa JM, Nieto MA. Inflammation and EMT: an alliance towards organ fibrosis and cancer progression. *EMBO molecular medicine*. 2009; 1: 305-14.
 37. Lee JM, Dedhar S, Kalluri R, Thompson EW. The epithelial-mesenchymal transition: new insights in signaling, development, and disease. *J Cell Biol*. 2006; 172: 973-81.