DRUG SELECTION FOR MALIGNANT MELANOMA USING BIOMARKERS GENERATED BY WEIGHTED GENE CO-EXPRESSION NETWORK ANALYSIS (WGCNA)

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

SEMİH ALPSOY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
BIOINFORMATICS

JANUARY 2017

Approval of the thesis:

DRUG SELECTION FOR MALIGNANT MELANOMA USING BIOMARKERS GENERATED BY WEIGHTED GENE CO-EXPRESSION NETWORK ANALYSIS (WGCNA)

Submitted by **Semih Alpsoy** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics Program, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Director, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics, METU** _____

Assist. Prof. Dr. Aybar Can Acar
Supervisor, **Health Informatics, METU** _____

Assoc. Prof. Dr. Ali Osmay Güre
Co-supervisor, **Molecular Biology and Genetics Dept., Bilkent University** _____

Examining Committee Members:

Assoc. Prof. Dr. Tolga Can
Computer Engineering, METU _____

Assist. Prof. Dr. Aybar Can Acar
Supervisor, Health Informatics, METU _____

Assoc. Prof. Dr. Ali Osmay Güre
Co-supervisor, Molecular Biology and Genetics, Bilkent University _____

Assist. Prof. Dr. Ercüment Çiçek
Computer Engineering, Bilkent University _____

Assist. Prof. Dr. Nurcan Tunçbağ
Health Informatics, METU _____

Date: 06.01.2017

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Semih Alpsoy

Signature : _____

ABSTRACT

DRUG SELECTION FOR MALIGNANT MELANOMA USING BIOMARKERS GENERATED BY WEIGHTED GENE CO-EXPRESSION NETWORK ANALYSIS (WGCNA)

ALPSOY, SEMİH

M.S., Bioinformatics Program

Supervisor: Assist. Prof. Dr. Aybar Can Acar

Co-supervisor: Assoc. Prof. Dr. Ali Osmay Güre

January 2017, 163 pages

Chemotherapy is one of the widely applied treatment choices for cancer patients; however, it might not be effective in the majority of patients due to the inability of foreseeing which patients respond to which chemotherapeutic agents. In order to ascertain appropriate chemotherapy for patients, thus, drug biomarkers predicting the response of the patients to chemotherapy should be discovered and translated into clinical practice to decide the ideal chemotherapeutic agents in a patient-centric manner. In this way, it might be possible to tackle cancer disease more effectively, extend the life expectancy of the patients, and economize health expenditures substantially. In addition, discovering drug biomarkers might pave the way for drug target identification, drug discovery process, and eludicating drug mechanism of actions. Because of all these reasons, a systems biology based network approach known as Weighted Gene Co-Expression Network Analysis (WGCNA) is utilized in this study to discover candidate biomarkers for anti-cancer drugs profiled in two large pharmacogenomics studies, the Cancer Cell Line Encyclopedia (CCLE) and the Cancer Genome Project (CGP). In the study, malignant melanoma is selected as a model disease, and only the common anti-cancer drugs between the two pharmacogenomics studies screened against human malignant melanoma cell lines are considered. Both gene expression and drug sensitivity data available in the studies are integrated to identify candidate biomarkers for these common anti-cancer drugs. Next, support vector machine regression (SVR) machine learning algorithm is employed to assess the predictive ability of the identified candidate biomarkers both individually and in combinations. For that purpose, the CCLE expression data of the candidate biomarkers and the CCLE drug sensitivity data are trained in the first step. Predictive ability of these candidate biomarkers is tested in an independent CGP dataset later on. Thereby, in-silico validation of several candidate

biomarkers could be accomplished. In conclusion, this thesis shows that the WGCNA methodology is a powerful approach for identifying gene expression-based candidate drug biomarkers for malignant melanoma. The thesis also shows that proper combinations of the candidate biomarkers generated by the WGCNA methodology improve anti-cancer drug sensitivity prediction significantly, and only a few gene combinations are sufficient to predict anti-cancer drug sensitivity powerfully.

Keywords: Malignant Melanoma, Weighted Gene Co-Expression Network Analysis, Biomarker Discovery, Drug Selection, Personalized Medicine

ÖZ

AĞIRLANDIRILMIŞ GEN KO-EKSPRESYON AĞ ANALİZİNİNDEN (WGCNA) ELDE EDİLEN BİYOBELİRTEÇLERİN KULLANILMASIYLA MALİGN MELANOMDA İLAÇ SEÇİMİ

ALPSOY, SEMİH

Yüksek Lisans, Biyoenformatik Programı

Danışman: Yar. Doç. Dr. Aybar Can Acar

Eş Danışman: Doç. Dr. Ali Osmay Güre

Ocak 2017, 163 sayfa

Kemoterapi, kanser hastalarında en yaygın kullanılan tedavi seçeneklerinden biri olmasına karşın hastaların hangi kemoterapötik ajanlara yanıt vereceği öngörülemediğinden hastaların önemli bir kısmında etkili olamayabilmektedir. Bu nedenle, hastalara uygun kemoterapiyi belirlemek için hastaların kemoterapiye yanıtını tahmin edebilen ilaç biyobelirteçleri keşfedilmeli ve hastaya özgü en ideal kemoterapötik ajanlara karar verilmesi adına biyobelirteçler klinik uygulamaya dönüştürülmelidir. Bu sayede, kanser hastalığıyla daha etkili mücadele etmek, hastaların yaşam süresini uzatmak ve sağlık harcamalarından önemli derecede tasarruf sağlamak mümkün olabilir. Dahası, ilaç biyobelirteçlerinin keşfedilmesi ilaç hedeflerinin bulunmasına, ilaç keşfi işlemine ve etki mekanizmasının anlaşılmasına ön ayak olabilir. Tüm bu nedenlerden ötürü, bu çalışmada Ağırlandırılmış Gen Ko-Ekspresyon Ağ Analizi (WGCNA) olarak bilinen ağ tabanlı sistem biyolojisi yaklaşımı iki büyük çaplı farmakogenomik çalışma olan Kanser Hücre Hattı Ansiklopedisi (CCLE) ve Kanser Genom Projesi (CGP)' nde profillenmiş ilaçların biyobelirteçlerinin keşfedilmesi amacıyla kullanılmaktadır. Çalışmada, malign melanom bir model hastalık olarak seçilmekte ve sadece iki farmakogenomik çalışmada insan malign melanom hücre hatlarına uygulanmış ortak anti-kanser ilaçları göz önünde bulundurulmaktadır. Bu ortak anti-kanser ilaçların potansiyel biyobelirteçlerinin belirlenmesi için iki farmakogenomik çalışmadaki gen ekspresyon ve ilaç hassasiyet verileri entegre edilmektedir. Sonrasında, destek vektör makinesi regresyon (SVR) makine öğrenme algoritması aday biyobelirteçlerin tek başına ve kombinasyonlar halinde tahmin edebilme gücünü incelemek amacıyla çalıştırılmaktadır. Bu amaçla, ilk aşamada aday biyobelirteçlerin CCLE çalışmasındaki gen ekspresyon ve ilaç hassasiyet verileri eğitilmektedir. Daha sonra, bu biyobelirteçlerin tahmin gücü bağımsız CGP verisetinde test edilmektedir. Böylece, bazı biyobelirteç adaylarının *in-siliko*

doğrulaması yapılmaktadır. Sonuç olarak, bu tez malignant melanoma için gen ekspresyon tabanlı ilaç biyobelirteç adaylarını tespit etmesi bakımından WGCNA metodolojisinin kuvvetli bir yaklaşım olduğunu göstermektedir. Tez ayrıca WGCNA metodu ile elde edilen uygun biyobelirteç kombinasyonlarının önemli bir miktarda anti-kanser ilaç hassasiyet tahminini iyileştirdiğini ve sadece birkaç gen kombinasyonunun güçlü bir şekilde anti-kanser ilaç hassasiyetini tahmin edebildiğini göstermektedir.

Anahtar Kelimeler: Malign Melanom, Ağırlandırılmış Gen Ko-Ekspresyon Ağ Analizi, Biyobelirteç Keşfi, İlaç Seçimi, Kişiselleştirilmiş Tıp

To my family and love of my life,

ACKNOWLEDGEMENTS

Firstly, I would like to thank my thesis advisor Assist. Prof. Dr. Aybar Can Acar of the Graduate School of Informatics at METU. The door to Prof. Acar office was always open whenever I ran into a trouble spot or had a question about my research or writing. His guidance, encouragement, intellectual support, and comments throughout my study are invaluable in the emergence of this thesis.

I would like to express my gratitude to my co-advisor Assoc. Prof. Dr. Ali Osmay Güre of the Molecular Biology and Genetics Department at Bilkent University for his continuous support and motivation in carrying out my study. His stupendous guidance, along with his intellectual support, plays a leading role in coming alive of this thesis.

Besides my advisor and co-advisor, I would like to thank the rest of my thesis committee: Assoc. Prof. Dr. Tolga Can, Assist. Prof. Dr. Ercüment Çiçek, and Assist. Prof. Dr. Nurcan Tunçbağ, not only for their insightful comments and encouragement but also for incenting me to widen my research from various perspectives.

I would like to extend my gratitudes to my dear friends for all their emotional supports, constant motivations, and great suggestions throughout my study. I would particularly like to thank Oğuz Emre Özdemir, Remzi Çelebi, and Volkan Orhan for their companionship, friendship, and care.

I greatly acknowledge Ministry of Science, Industry, and Technology of Turkey to fund this thesis project with the Technopreneurship Support Program (Grant ID: 0980.TGSD.2015).

I am deeply grateful to my parents Pembegül and Burhan Alpsoy, my brother Eray Alpsoy for their encouragement, endless help, confidence, understanding, and supporting me spiritually throughout my life. I feel lucky to have all of them in my life.

Last but not the least, I would like to express my deepest gratitude to the love of my life Ecem Özmeriç for her efforts, emotional support, care, friendship, patience, and extreme encouragements.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ vi

ACKNOWLEDGEMENTS ix

TABLE OF CONTENTS xi

LIST OF FIGURES xv

LIST OF TABLES xxi

LIST OF ABBREVIATIONS xxv

CHAPTERS

1. INTRODUCTION 1

1.1. Motivation 1

1.2. Scope and goal 2

1.3. Contribution 3

1.4. Outline of thesis 4

2. BACKGROUND AND RELATED WORKS 5

2.1. Cancer 5

2.2. Malignant melanoma 8

2.2.1. Epidemiology 8

2.2.2. Histological subtypes 8

2.2.3. Signs and symptoms 9

2.2.4. Causes and risk factors 10

2.2.5. Treatment 12
2.3. Weighted gene co-expression network analysis 14
2.3.1. Description 14
2.3.2. Analysis goals 14
2.3.3. Methodology 15
2.4. Support vector machine 18
2.4.1. Description 18
2.4.2. Theory behind the SVM 18
2.4.2.1. The linear case 19
2.4.2.2. The non-linear case 21
2.4.3. Parameter optimization 22
2.5. Related works 23
2.5.1. Previous works applied the WGCNA methodology 23
2.5.2. Previous works applied the SVM for drug response prediction 25
3. METHODOLOGY 27
3.1. Overview 27
3.2. Data sources 27
3.2.1. Cancer Cell Line Encyclopedia 27
3.2.2. Cancer Genome Project 28
3.3. Data pre-processing 29
3.3.1. Pre-processing of gene expression data 29
3.3.1. Pre-processing of drug sensitivity data 31
3.4. Identification of gene expression-based candidate drug biomarkers 33
3.5. In-silico validation of identified candidate drug biomarkers 35
4. RESULTS 39
4.1. Quality control and data pre-processing 39

4.2. WGCNA 40
4.2.1. Construction of gene co-expression network 40
4.2.2. Identification of network modules 42
4.2.3. Relating modules to drug sensitivity 47
4.2.4. Identification of candidate drug biomarkers 48
4.3. Predictive ability of candidate biomarkers 49
4.3.1. Predictive ability of single candidate biomarkers 50
4.3.1.1. Anti-cancer drug sensitivity prediction using the CGP gene expression data of all the MM cell lines and the CGP drug sensitivity data of the nine common drugs 53
4.3.1.1.1. IC50 prediction results 53
4.3.1.1.2. Activity Area prediction results 58
4.3.1.2. Anti-cancer drug sensitivity prediction using only the CGP gene expression data of the shared MM cell lines between the CCLE and the CGP studies 59
4.3.1.2.1. IC50 prediction results 61
4.3.1.2.2. Activity Area prediction results 63
4.3.2. Predictive ability of combined candidate biomarkers 64
4.3.2.1. IC50 prediction results 65
4.3.2.2. Activity Area prediction results 68
5. DISCUSSION 73
5.1. The WGCNA methodology is powerful in identification of candidate biomarkers 73
5.2. Combinations of candidate biomarkers improve drug sensitivity prediction 74
5.3. There are limitations for accurate assessment of predictors 75
5.3.1. Drug sensitivity data is highly inconsistent between the studies 75
5.3.2. Small sample size precludes reliable estimation of predictive power 76
5.3.3. Anti-cancer drugs profiled in the two studies do not have remarkable cytotoxic activity against the MM 77

6. CONCLUSION AND FUTURE STUDIES 79

6.1. Conclusion 79

6.2. Future work 81

REFERENCES 83

APPENDICES

A. THE PLOTS GENERATED BY DATA PRE-PROCESSING 89

B. HEATMAP PLOTS 95

C. CORRELATION SCORES 103

D. THE TREND OF PREDICTIVE POWER WITH VARYING NUMBER OF COMBINATIONS 123

E. THE TREND OF CORRELATION SCORES WITH VARYING NUMBER OF COMBINATIONS 153

LIST OF FIGURES

Figure 2.1: Barplot of the new cancer cases and deaths worldwide in 2012 as to the regions grouped into either less developed or more developed.....6

Figure 2.2: Piechart of the most common cancers worldwide in 2012.....6

Figure 2.3: Piechart of identifiable and/or potentially preventable factors estimated to cause cancer.....7

Figure 2.4: The ABCDE guideline used in clinics for early detection of melanoma.....10

Figure 2.5: Ultraviolet light is a modulator in malignant melanoma development.....12

Figure 2.6: The workflow of the WGCNA to identify centrally located intramodular hub genes which are the representatives of the modules obtained by hierarchical clustering.....17

Figure 2.7: Visual representation of classification task by the SVM.....19

Figure 2.8: Visual representation of regression task by the SVR.....20

Figure 3.1: Count of the genes profiled in the CCLE and the CGP studies.....29

Figure 3.2: Count of the malignant melanoma cell lines examined in the CCLE and the CGP studies.....30

Figure 3.3: Count of the anti-cancer drugs screened against the malignant melanoma cell lines.....31

Figure 3.4: The workflow depicting identification of gene expression-based candidate drug biomarkers by utilizing the WGCNA method.....34

Figure 3.5: The workflow which illustrates the assessment of the predictive ability of the candidate drug biomarkers both individually and in combinations.....36

Figure 3.6: The workflow which illustrates the assessment of the predictive ability of the candidate drug biomarkers both individually and in combinations after the CGP drug sensitivity data is excluded.....37

Figure 4.1: Scale independence and mean connectivity plots....41

Figure 4.2: Frequency distribution and scale free plots....42

Figure 4.3: Network plot generated by the most varying 8000 genes....43

Figure 4.4: Network statistics obtained form the network constructed by the most varying 8000 genes....43

Figure 4.5: Hierarchical clustering dendrogram of the most varying 8000 genes....44

Figure 4.6: Cluster dendrogram of module eigengenes before and after merging similar modules....44

Figure 4.7: Hierarchical clustering dendrogram of the most varying 8000 genes before and after merging the similar modules....45

Figure 4.8: Barplot of the gene counts in the modules before and after merging the similar modules....46

Figure 4.9: Barplot of the gene ontology terms for significantly enriched modules....47

Figure 4.10: Count of identified candidate biomarkers per modules for each of the nine anti-cancer drugs....49

Figure 4.11: The scatterplots which illustrate the poor correlation of IC50 and Activity Area values....60

Figure 4.12: The RMSE plots which depict the trend of prediction error when the censored IC50 values are included in the CCLE drug sensitivity data....67

Figure 4.13: The RMSE plots which depict the trend of prediction error when both the censored and extrapolated IC50 values are excluded from drug sensitivity data....69

Figure 4.14: The RMSE plots which depict the trend of prediction error when the CCLE Activity Area values are used for prediction....71

Figure A. 1: Images of the CCLE microarray samples visualized for quality control....91

Figure A. 2: Images of the CGP microarray samples visualized for quality control....92

Figure A. 3: The density distribution plots of the CCLE microarray samples before and after RMA normalization....93

Figure A. 4: The density distribution plots of the CGP microarray samples before and after RMA normalization..... 93

Figure A. 5: Boxplots of the CCLE microarray samples before and after RMA normalization..... 94

Figure A. 6: Boxplots of the CGP microarray samples before and after RMA normalization…..... 94

Figure A. 7: Boxplot of the gene expression levels of all the malignant melanoma cell lines after setting gene expression levels to the same level..... 95

Figure A. 8: Boxplot of the gene expression levels of the common malignant melanoma cell after setting gene expression levels to the same level..... 95

Figure B. 1: The heatmap plot of the saddlebrown module.....97

Figure B. 2: The heatmap plot of the darkslateblue module.....98

Figure B. 3: The heatmap plot of the darkolivegreen module..... 98

Figure B. 4: The heatmap plot of the darkmagenta module.....99

Figure B. 5: The heatmap plot of the green module..... 99

Figure B. 6: The heatmap plot of the purple module.....100

Figure B. 7: The heatmap plot of the darkturquoise module..... 100

Figure B. 8: The heatmap plot of the skyblue module.....101

Figure B. 9: The heatmap plot of the magenta module.....101

Figure B. 10: The heatmap plot of the darkgreen module..... 102

Figure B. 11: The heatmap plot of the darkorange module..... 102

Figure B. 12: The heatmap plot of the brown module.....103

Figure B. 13: The heatmap plot of the blue module..... 103

Figure B. 14: The heatmap plot of the white module..... 104

Figure B. 15: The heatmap plot of the black module..... 104

Figure D. 1: The scatterplots which depict the predictive performance for AZD0530 when the censored IC50 values are included in the CCLE drug sensitivity data.....131

Figure D. 2: The scatterplots which depict the predictive performance for AZD6244 when the censored IC50 values are included in the CCLE drug sensitivity data.....132

Figure D. 3: The scatterplots which depict the predictive performance for Erlotinib when the censored IC50 values are included in the CCLE drug sensitivity data.....133

Figure D. 4: The scatterplots which depict the predictive performance for Lapatinib when the censored IC50 values are included in the CCLE drug sensitivity data.....134

Figure D. 5: The scatterplots which depict the predictive performance for PD0325901 when the censored IC50 values are included in the CCLE drug sensitivity data.....135

Figure D. 6: The scatterplots which depict the predictive performance for PF2341066 when the censored IC50 values are included in the CCLE drug sensitivity data.....136

Figure D. 7: The scatterplots which depict the predictive performance for PLX4720 when the censored IC50 values are included in the CCLE drug sensitivity data.....137

Figure D. 8: The scatterplots which depict the predictive performance for Sorafenib when the censored IC50 values are included in the CCLE drug sensitivity data.....138

Figure D. 9: The scatterplots which depict the predictive performance for TAE684 when the censored IC50 values are included in the CCLE drug sensitivity data.....139

Figure D. 10: The scatterplots which depict the predictive performance for AZD6244 when both censored and extrapolated IC50 values are excluded from the drug sensitivity data.....140

Figure D. 11: The scatterplots which depict the predictive performance for PD0325901 when both censored and extrapolated IC50 values are excluded from the drug sensitivity data.....141

Figure D. 12: The scatterplots which depict the predictive performance for TAE684 when both the censored and extrapolated IC50 values are excluded from the drug sensitivity data..... 142

Figure D. 13: The scatterplots which depict the predictive performance for AZD0530 when the CCLE Activity Area values are used for prediction.....143

Figure D. 14: The scatterplots which depict the predictive performance for AZD6244 when the CCLE Activity Area values are used for prediction.....144

Figure D. 15: The scatterplots which depict the predictive performance for Erlotinib when the CCLE Activity Area values are used for prediction.....145

Figure D. 16: The scatterplots which depict the predictive performance for Lapatinib when the CCLE Activity Area values are used for prediction.....146

Figure D. 17: The scatterplots which depict the predictive performance for PD0325901 when the CCLE Activity Area values are used for prediction..... 147

Figure D. 18: The scatterplots which depict the predictive performance for PF2341066 when the CCLE Activity Area values are used for prediction.....148

Figure D. 19: The scatterplots which depict the predictive performance for PLX4720 when the CCLE Activity Area values are used for prediction.....149

Figure D. 20: The scatterplots which depict the predictive performance for Sorafenib when the CCLE Activity Area values are used for prediction.....150

Figure D. 21: The scatterplots which depict the predictive performance for TAE684 when the CCLE Activity Area values are used for prediction.....151

Figure E. 1: The barplots which demonstrate the correlation scores for AZD0530 when the censored IC50 values are included in the CCLE drug sensitivity data.....153

Figure E. 2: The barplots which demonstrate the correlation scores for AZD6244 when the censored IC50 values are included in the CCLE drug sensitivity data.....154

Figure E. 3: The barplots which demonstrate the correlation scores for Erlotinib when the censored IC50 values are included in the CCLE drug sensitivity data.....154

Figure E. 4: The barplots which demonstrate the correlation scores for Lapatinib when the censored IC50 values are included in the CCLE drug sensitivity data.....155

Figure E. 5: The barplots which demonstrate the correlation scores for PD0325901 when the censored IC50 values are included in the CCLE drug sensitivity data.....155

Figure E. 6: The barplots which demonstrate the correlation scores for PF2341066 when the censored IC50 values are included in the CCLE drug sensitivity data.....156

Figure E. 7: The barplots which demonstrate the correlation scores for PLX4720 when the censored IC50 values are included in the CCLE drug sensitivity data.....156

Figure E. 8: The barplots which demonstrate the correlation scores for Sorafenib when the censored IC50 values are included in the CCLE drug sensitivity data.....157

Figure E. 9: The barplots which demonstrate the correlation scores for TAE684 when the censored IC50 values are included in the CCLE drug sensitivity data.....157

Figure E. 10: The barplots which demonstrate the correlation scores for AZD6244 when both the censored and extrapolated IC50 values are excluded from drug sensitivity data..158

Figure E. 11: The barplots which demonstrate the correlation scores for PD0325901 when both the censored and extrapolated IC50 values are excluded from drug sensitivity data..158

Figure E. 12: The barplots which demonstrate the correlation scores for TAE684 when both the censored and extrapolated IC50 values are excluded from drug sensitivity data.....159

Figure E. 13: The barplots which demonstrate the correlation scores for AZD0530 when the CCLE Activity Area values are used for prediction..... 159

Figure E. 14: The barplots which demonstrate the correlation scores for AZD6244 when the CCLE Activity Area values are used for prediction..... 160

Figure E. 15: The barplots which demonstrate the correlation scores for Erlotinib when the CCLE Activity Area values are used for prediction..... 160

Figure E. 16: The barplots which demonstrate the correlation scores for Lapatinib when the CCLE Activity Area values are used for prediction..... 161

Figure E. 17: The barplots which demonstrate the correlation scores for PD0325901 when the CCLE Activity Area values are used for prediction.....161

Figure E. 18: The barplots which demonstrate the correlation scores for PF2341066 when the CCLE Activity Area values are used for prediction.....162

Figure E. 19: The barplots which demonstrate the correlation scores for PLX4720 when the CCLE Activity Area values are used for prediction..... 162

Figure E. 20: The barplots which demonstrate the correlation scores for Sorafenib when the CCLE Activity Area values are used for prediction..... 163

Figure E. 21: The barplots which demonstrate the correlation scores for TAE684 when the CCLE Activity Area values are used for prediction..... 163

LIST OF TABLES

Table 3.1: Names of anti-cancer drugs shared between the CCLE and the CGP studies.....31

Table 3.2: Count of malignant melanoma cell lines investigated for sensitivity profiling…32

Table 4.1: Scale free fitting index R2 and connectivity values for different β choices.....40

Table 4.2: Gene ontology terms of the significantly enriched modules and their p-values...46

Table 4.3: The list of anti-cancer drugs for which the WGCNA could identify candidate biomarkers..... 47

Table 4.4: The RMSE values determined for the best performing single candidate biomarkers in the CCLE IC50 prediction.....50

Table 4.5: The RMSE values determined for the best performing single candidate biomarkers in the CCLE IC50 prediction after removing the censored IC50 values.....51

Table 4.6: The RMSE values determined for the best performing single candidate biomarkers in the CCLE Activity Area prediction..... 52

Table 4.7: Count of all the malignant melanoma cell lines profiled in the CCLE study.....54

Table 4.8: The RMSE values determined for the best performing single candidate biomarkers when the extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs in the CCLE study..... 54

Table 4.9: Maximum screening concentration of drugs along with count of malignant melanoma cell lines screened against these drugs in the CGP study..... 55

Table 4.10: The RMSE values determined for the best performing single candidate biomarkers when the extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs in the CGP study..... 56

Table 4.11: Count of the malignant melanoma cell lines remaining after all the extrapolated and censored IC50 values are removed from drug sensitivity data.....57

Table 4.12: The RMSE values determined for the best performing single candidate biomarkers when both the extrapolated and censored IC50 values in the drug sensitivity data are excluded..... 58

Table 4.13: The RMSE values determined for the best performing single candidate biomarkers in the CGP Activity Area prediction..... 58

Table 4.14: The RMSE values determined for the best performing single candidate biomarkers in the CCLE IC50 prediction.....62

Table 4.15: The RMSE values determined for the best performing single candidate biomarkers in the CCLE IC50 prediction after excluding the censored IC50 values.....63

Table 4.16: The RMSE values determined for the best performing single candidate biomarkers in the CCLE Activity Area prediction..... 63

Table 4.17: List of the best performing candidate biomarker combinations for the nine drugs when IC50 is used as the drug sensitivity measure..... 66

Table 4.18: List of the best performing candidate biomarker combinations for the three drugs when IC50 is used as the sensitivity measure.....68

Table 4.19: List of the best performing biomarker combinations for the nine drugs when Activity Area is used as the drug sensitivity measure..... 69

Table C. 1: The modules eigengene values of which are significantly correlated to the CCLE IC50 values..... 103

Table C. 2: List of hub genes identified as candidate biomarkers for all the nine anti-cancer drugs 104

Table C. 3: The correlation scores determined for the best performing single candidate biomarkers when the CCLE IC50 values are used for prediction.....114

Table C. 4: The correlation scores determined for the best performing single candidate biomarkers after the censored IC50 values are excluded from the CCLE drug sensitivity data.....114

Table C. 5: The correlation scores determined for the best performing single candidate biomarkers when the CCLE Activity Area values are used for prediction.....114

Table C. 6: The correlation scores determined for the best performing single candidate biomarkers after the CGP IC50 values are censored to the maximum screening concentration of the drugs in the CCLE study.....115

Table C. 7: The correlation scores determined for the best performing single candidate biomarkers after the CGP IC50 values are censored to the maximum screening concentration of the drugs in the CGP study..... 115

Table C. 8: The correlation scores determined for the best performing single candidate biomarkers after all the extrapolated and censored IC50 values are excluded from drug sensitivity data..... 116

Table C. 9: The correlation scores determined for the best performing single candidate biomarkers when Activity Area values are used for prediction.....116

Table C. 10: The correlation scores determined for the best performing single candidate biomarkers when the extrapolated IC50 values are included in the drug sensitivity data..116

Table C. 11: The correlation scores determined for the best performing single candidate biomarkers when the extrapolated IC50 values are removed from the drug sensitivity data.....117

Table C. 12: The correlation scores determined for the best performing single candidate biomarkers when Activity Area values are used for prediction.....117

Table C. 13: The correlation scores determined for the best performing combined candidate biomarkers when the CGP gene expression data is used for IC50 prediction.....118

Table C. 14: The correlation scores determined for the identified best performing combined candidate biomarkers when the CGP gene expression data is used for IC50 prediction after removing the censored IC50 values..... 119

Table C. 15: The correlation scores determined for the identified best performing combined candidate biomarkers when the CGP gene expression data is used for Activity Area prediction..... 120

Table D. 1: Predictive power of the best performing single/combined candidate biomarkers when the CGP expression data is used for IC50 prediction.....123

Table D. 2: Predictive power of the best performing candidate biomarkers when the CGP expression data is used for IC50 prediction after removing the censored IC50 values from the CCLE drug sensitivity data..... 124

Table D. 3: Predictive power of the best performing candidate biomarkers when the CGP expression data is used for Activity Area prediction..... 125

Table D. 4: The RMSE values determined for the best performing combined candidate biomarkers when the censored IC50 values are included in the CCLE drug sensitivity data..... 127

Table D. 5: The RMSE values of the best performing combined candidate biomarkers when the censored IC50 values are excluded from the CCLE drug sensitivity data.....128

Table D. 6: The RMSE values of the best performing combined candidate biomarkers when the CCLE Activity Area values are used for prediction.....129

LIST OF ABBREVIATIONS

ABCDE	Asymmetry border color diameter evolving
ABCF1	ATP binding cassette subfamily F member 1
AD	Alzheimer' s disease
ASPM	Abnormal spindle microtubule assembly
ATF4	Activating transcription factor 4
AUC	Area under curve
BRAF	B-Raf proto-oncogene serine/threonine-protein kinase
CASC3	Cancer susceptibility candidate 3
CCLE	Cancer cell line encyclopedia
CDKN2A	Cyclin dependent kinase inhibitor 2A
CDK5	Cyclin dependent kinase 5
CGP	Cancer genome project
CHAC1	Cation transport regulator-like protein 1
CT	Computed tomography
CTDSP2	Small C-terminal domain phosphatase 2
DAVID	Database for Annotation, Visualization and Integrated Discovery
DNA	Deoxyribonucleic acid
DEP	Diesel exhaust particles
EC50	Half maximal effective concentration
EGFR	Epidermal growth factor receptor
EGFRvIII	Epidermal growth factor receptor variant III
ERBB2	Erb-B2 receptor tyrosine kinase 2
ERBB3	Erb-B2 receptor tyrosine kinase 3
EYA1	Eyes absent homolog 1
GO	Gene ontology
GSK	GlaxoSmithKline
HAND	HIV - association neurocognitive disorder

HIPK2	Homeodomain interacting protein kinase 2
HIV	Human immunodeficiency virus
HMEC	Human microvascular endothelial cell
HO-1	Heme oxygenase 1
IC50	Half maximal inhibitory concentration
IFN-g	Interferon gamma
JAK1	Janus kinase 1
LDOC1	Leucine zipper down-regulated in cancer 1
LECT1	Leukocyte cell derived chemotaxin 1
LOOCV	Leave one out cross validation
MAP3K14	Mitogen-activated protein kinase kinase kinase 14
MC1R	Melanocortin 1 receptor
MEK	Mitogen-activated protein kinase kinase 1
MHC	Major histocompatibility complex
MIT	Massachusetts Institute of Technology
MM	Malignant melanoma
NF-kB	Nuclear factor kappa B
NRAS	Neuroblastoma RAS viral oncogene homolog
OX-PAPC	1-palmitoyl-2-arachidonyl-*sn*-glycero-3-phosphorylcholine
PGAM2	Phosphoglycerate mutase 2
PGF	Placental growth factor
PM	Perfect Match
PSEN1	Presenilin 1
QTL2	Quantitative trait loci 2
QTL5	Quantitative trait loci 5
QTL10	Quantitative trait loci 10
QTL19	Quantitative trait loci 19
RAF	Raf proto-oncogene serine/threonine protein kinase
RBF	Radial basis function
RMA	Robust multi-array average
RMSE	Root mean square error
RNF125	Ring finger protein 125
RT-PCR	Reverse transcription polymerase chain reaction

SASH1	SAM And SH3 domain containing 1
siRNA	Small interfering ribonucleic acid
SMP	Small molecule perturbagen
SVM	Support vector machine
SVR	Support vector regression
TNPO3	Transportin 3
TOM	Topological overlap matrix
UPR	Unfolded protein response
UV	Ultraviolet radiation
WGCNA	Weighted gene co-expression network analysis
XBP1	X-box binding protein 1
XP	Xeroderma pigmentosum
YWHAZ	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase

CHAPTER 1

INTRODUCTION

1.1. Motivation

Tailoring of treatment as to the genetic background of each patient has been regarded invaluable to assure effective therapeutic interventions for disease progression in complex diseases having a genetic basis such as neurodegeneration and cancer. Personalized medicine (or precision medicine), emerged as a demand to tailor patient-specific treatment, became an outstanding field to explain the underlying molecular mechanism of these diseases as it might identify the most effective therapeutic options for patients according to their genetic backgrounds, lifestyles, and environments.

The popularity of personalized medicine has increased due to striking advancements in science and technology, which enabled the invention of high throughput medical devices profiling genome fast and accurate. These devices resulted in the accumulation of massive amount of data to be utilized in clinics; however, medical decisions could not solely been made according to the output of results. There was a need for tools, tests, or devices that could interpret the data to give a clinicially valueable decision. For this purpose, various computational tools and tests have been developed. In addition, powerful devices having ability to sequence of human genome fastly and accurately have been invented.

One of the major topics of the personalized medicine is pharmacogenomics, which is the study of transcriptomics variations as to drug response. Scientists interested in pharmacogenomics primarily aim to discover drug biomarkers so as to predict drug effectiveness. As such biomarkers give information about the mechanism of action, safety, and efficacy of a drug, they are becaming an essential part of drug development. Conventional drug development approaches have not been able to satisfy the drug industry owing to the fact that most of the drugs entering phase trials fail in efficacy, so drugs are hardly approved. So considering huge amount of expenditures and increasing costs to develop drugs, they have turned their interest into drug biomarker discovery. Drug biomarkers are also invaluable to understand heterogenous response of patients to chemotherapy. Clinicians mostly do not know whether a drug is effective in one patient is also effective in another. A “one-dose-fits-all” approach is mostly followed, but every patient responds the drug treatment with different doses. Thus, eliminating “trial-and-error”

approach in clinics is quite crucial to achieve effective treatment outcomes as it might minimize toxicities of drugs and adverse drug reactions when the drugs are combined in therapy. Thereby, those patients who do not respond the chemotherapy can be guided to alternative treatments that are more suitable for their genetic backgrounds.

Several computational approaches have been developed to identify drug biomarkers. However, most of them could not achieve to discover reliable biomarkers. This ordinarily stems from the underperformance of conventional reductionist methods. Conventional approaches have a drawback of focusing on a single scale that hinders complete understanding of the system. Thereby, the lack of systems based approaches mostly elicits underperformance in prediction. In addition, conventional approaches are incompetent in identifying the association between genomics and pharmacological entities. The general trend is to fit a model assuming that association of selected genomic features with pharmacological data is linear. However, it is commonly observed the association is non-linear. When feature selection approaches are taken into account to be the driving tool in performance evaluation of models, and they are mostly ineffective, conventional approaches usually crash in biomarker discovery. However, systems approaches behaving functional units in the model as the systems are excellent alternatives to compensate insufficiency of conventional methods. In spite of focusing on single scale, systems approaches focus on multi-scales enabling more robust predictions. Therefore, systems-based approaches in place of reductionist approaches are gaining popularity in biomarker research; nevertheless, there is still need for highly competent and outperforming systems approaches that could be successfully applied in biomarker studies.

Weighted Gene Co-Expression Network Analysis (WGCNA), a systems biology based co-expression network approach, is a widely used powerful approach for biomarker discovery. The WGCNA has the potential to identify therapeutic targets that can be regarded as the biomarkers. Constructing a model that inputs such candidate biomarkers obtained from the WGCNA may reveal their predictive power for drug sensitivity. For this purpose, utilizing highly efficient machine learning algorithms such as SVM and random forest could be applied to test the performances of candidate biomarkers coming from the WGCNA. Thereby, *in-silico* validation of candidate biomarkers might be accomplished accurately and powerfully. Subsequent to *in-silico* validation, these candidate biomarkers could be validated by *in-vitro* and *in-vivo* pre-clinical studies, so that they might be translated in clinics to decide the ideal drugs for each patient. In this respect, the WGCNA, along with machine learning approaches, holds a great promise to discover biomarkers. And so, they might be applied to determine the most effective chemotherapy for the patients as to their genetic backgrounds.

1.2. Scope and goal

The primary goal of this study is to show that proper combinations of the candidate biomarkers determined by the WGCNA method are more powerful in drug sensitivity prediction than single candidate biomarkers. In this respect, the study aims to demonstrate proper biomarker combinations might be highly effective in drug sensitivity prediction even

though single candidate biomarkers have poor predictive power. So it would be possible to show that combinations of candidate biomarkers are more likely to be validated in clinical studies. In addition, this study aims to show that combinations of only a few candidate biomarkers are sufficient to predict anti-cancer drug sensitivity powerfully.

The first step of analysis must construct a co-expression network by the WGCNA methodology after gene expression data of the 38 malignant melanoma (MM) cell lines gathered from the Cancer Cell Line Encyclopedia (CCLE) is normalized, and the most varying 8000 genes is selected for the analysis to ensure technical variability is less than biological variability. The second step should give a list of genes (candidate drug biomarkers) which have been identified in the literature as associated with the MM. The association indicates the WGCNA methodology is effective in identifying the MM related genes or already approved MM biomarkers, so that the rest of the genes is likely to be regarded as the candidate biomarkers. The last step should demonstrate that gene features (candidate drug biomarkers) obtained by the WGCNA analysis have high potential to predict anti-cancer drug sensitivity when models trained by the CCLE gene expression and drug sensitivity data are shown to be validated in an independent Cancer Genome Project (CGP) study.

The specified purposes above utilize two different large pharmacogenomics studies, the CCLE and the CGP. The CCLE is a data source which includes expression profiles of more than thousand human cancer cell lines. There are 61 MM cell lines expression profiles of which are profiled in the CCLE study. It also includes sensitivity profiles of the 24 anti-cancer drugs screened against 479 cell lines. Apart from the CCLE, the CGP is another data source which contains both gene expression and drug sensitivity data of the 139 anticancer drugs screened against 53 cancer cell lines. There are 42 MM cell lines screened by these anti-cancer drugs in the CGP study. In contrast to the CCLE study, each of the 139 anti-cancer drugs was screened against varying numbers of MM cell lines. Both of the studies share 29 MM cell lines and 15 anti-cancer drugs, so it is possible to validate trained models in the CGP test data using expression profiles of the shared MM cell lines and sensitivity profiles of the shared anti-cancer drugs between the studies.

1.3. Contribution

The major two contributions of this thesis are to show that *in-vitro* anti-cancer drug sensitivity can be predicted by basal gene expression profiles of candidate biomarkers identified by the WGCNA systems biology approach, and proper combinations of identified candidate biomarkers predict anti-cancer drug sensitivity more powerfully than single candidate biomarkers. Another contribution is to reveal that the models developed in this study can make accurate predictions for anti-cancer drug sensitivity with only a few features although the majority of computational models requires tens or hundreds of features to construct high performance models. It is also pointed the two pharmacogenomics studies, the CCLE and the CGP, have inconsistent drug sensitivity data that precludes model efficacy in predicting anti-cancer drug sensitivity. However, it is demonstrated that predictive ability of models is powerful when gene expression data of common MM cell lines between the

CCLE and the CGP studies is utilized with only the CCLE drug sensitivity data. This improved prediction is shown to be resulted from high consistency between gene expression profiles of the cell lines between the studies even though different platforms are used for measuring gene expression levels. As a final word, it is the first time demonstrated in this thesis that support vector machine regression (SVR) machine learning algorithm could powerfully be utilized to validate models trained by the CCLE data in an independent CGP data.

1.4. Outline of thesis

This thesis is composed of 6 major chapters, which includes background and related works, methodology, results, discussion, and conclusion and future studies. In *Chapter 2*, background information about basics of cancer, biology of the MM, workflow and principles of the WGCNA methodology, basics of support vector machine algorithm, and data repositories are given. In *Chapter 3*, the data repositories at which expression and drug sensitivity data of the MM cell lines are available and methods to analyze them in order to predict drug sensitivity of various anti-cancer drugs are described. The detailed explanation of the WGCNA methodology such as used similarity and adjacency matrices, soft thresholding approach, test of scale freeness, topological overlap matrix based dissimilarity measure inputted in hierarchical clustering is also explained and presented via visuals and flowcharts explicitly. In *Chapter 4*, the results of the WGCNA identifying candidate biomarkers to predict anti-cancer drug response of the drugs in MM and their predictive performances both individually and in combinations by the SVR method are presented. In *Chapter 5*, the success of the WGCNA methodology in identifiying candidate biomarkers for anti-cancer drug sensitivity, the assessment of predictive power of candidate biomarker combinations with the SVR method, and limitations of the study are discussed. Possible improvements contibuting to identify more powerful candidate biomarkers are also discussed in this chapter. In *Chapter 6*, finally, conclusions are expressed, and future studies that could be conducted for achieving to identify more powerful candidate biomarkers having potential to predict sensitivity of anti-cancer drugs with low therapeutic index are mentioned.

CHAPTER 2

BACKGROUND AND RELATED WORKS

2.1. Cancer

Cancer can be defined as a group of diseases that emerges due to the development of uncontrolled cell division and spreading of the uncontrolled division to the other parts of the body. As abnormal cell division continues, cancer cells outnumber the normal cells. This results in the development of possible signs and symptoms such as fatigue, unintended weight changes, difficulty in swallowing, hoarseness, unexpected muscle pains, bleeding or bruising, and unexplained fevers (Mayo Clinic, 2015). It is one of the most commonly observed diseases that leads to morbidity and mortality worldwide. According to National Cancer Institute (2015), tobacco is the leading risk factor for cancer that brings about 20% of global deaths. In addition, one-third of the cancer deaths have resulted from high body mass index, low fruit and vegetable intake, lack of physical activity, tobacco, and alcohol use. The GLOBOCAN project held by the International Agency for Research on Cancer (2014) shows that new cases diagnosed for cancer is approximately 14.1 million, and cancer puts the death 8.2 million people worldwide. 57% of the total new cases is diagnosed in less developed regions, and 5% of the global total dies in this less developed regions. **Figure 2.1** presents the data of the GLOBOCAN project showing the number of cancer cases and deaths worldwide for both less developed regions and more developed regions. The GLOBOCAN project also reports that lung cancer is the most deadly cancer accounting for 1.6 million deaths. Liver cancer and stomach cancer follow the lung cancer as a cause of death with 745.000 and 723.000 deaths, respectively (See **Figure 2.2**).

Genetic changes via mutations to the DNA within cells as a result of the interaction with physical, chemical, and biological carcinogens are the primary causes of cancer. **Figure 2.3** shows some of these most severe environmental factors accompanying mutations that may lead to carcinogenesis. Genes inside the DNA give instructions to the cell to perform regular cellular activities such as division and growth. However, errors made in the instructions are likely to disrupt the regulation, so that normal cells may become cancerous. Mayo Clinic (2015) declares that gene mutations can instruct the normal cells in 3 ways: rapid growth allowance, failure in stopping the uncontrolled cell growth, and mistakes made when repairing DNA errors. These are the most common types of mutations observed in cancer cells although several other types are also known, but they are less frequent. As age

progresses, the mutations inherited directly from parents, and the mutations acquired throughout the life accumulate in cells cause to cancer.

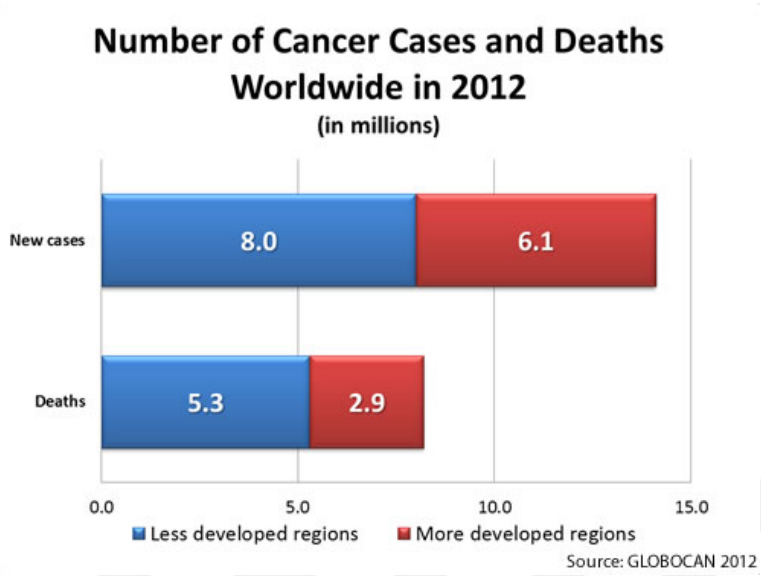

Figure 2.1: Barplot of the new cancer cases and deaths worldwide in 2012 as to the regions grouped into either less developed or more developed (International Agency for Research on Cancer, 2014)

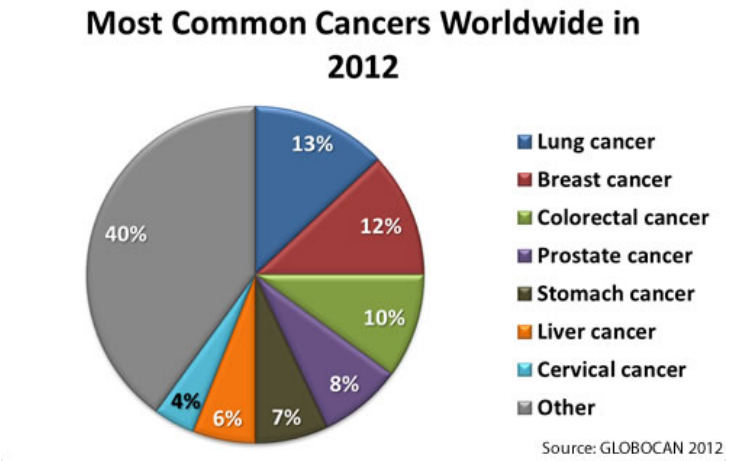

Figure 2.2: Piechart of the most common cancers worldwide in 2012 (International Agency for Research on Cancer, 2014)

The mortality of cancer can be lessened provided that patients are detected early and the right treatment starts immediately. Early diagnosis is pretty favorable if there is no available screening method to detect cancer. Unless these patients are diagnosed in early stages, there

might be no potent treatment option to cure it since in later stages it can metastasize and invade the other parts of the body in such a way that no treatment effort could improve the opportunity of recovery. Screening, which is the second method of detecting cancer early, is powerful in identifying the individuals who are likely to develop a specific type of cancer or the patients who have already developed a specific type of cancer when the individuals do not show any signs or symptoms. However, it may not be affordable or accessible for the majority of the population.

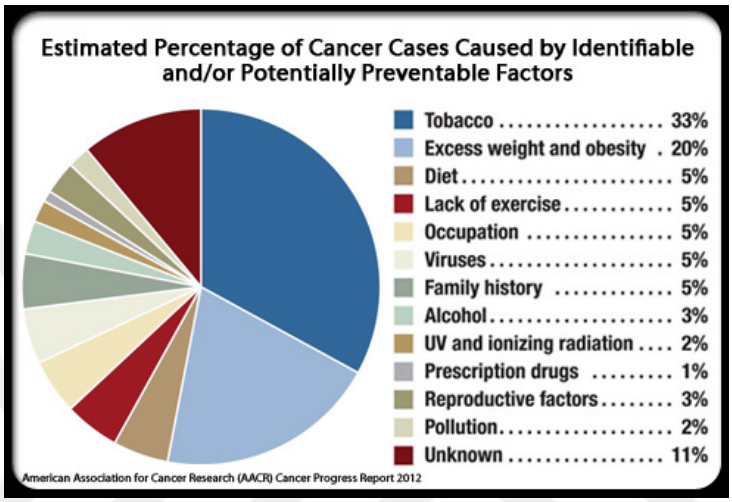

Figure 2.3: Piechart of identifiable and/or potentially preventable factors estimated to cause cancer (American Association for Cancer Research, 2012)

The most common treatment options used in cancer are surgery, chemotherapy, radiation therapy, targeted therapy, immunotherapy, and stem cell transplantation (National Cancer Institute, 2015). Surgery is generally the most effective option in the treatment of patients who are diagnosed in the early stages of cancer since cancer has not invaded other parts of the body. Chemotherapy is the medication that the patients are prescribed to take anti-cancer drugs to erase cancer cells. Though high cytotoxicity of anti-cancer drugs frightens the vast majority of the patients, it is commonly accepted that drug treatment kills rapidly dividing cancer cells in the expense of killing normal cells such as cells in the bone marrow, digestive tract, and hair follicles. Radiation therapy destroys cancer cells by high-energy particles or waves. It is one of the most appealed treatment options in cancer. Targeted therapy intervenes in blocking the growth and invasion of cancer with molecular targets identified as having roles in the development of cancer. Thus, it aims to destroy only the cancer cells without harming the normal cells. Immunotherapy is the use of body's immune response to treat cancer by augmenting the attack to the cancer cells as stimulating immune system further and giving immune system components such as artificial proteins functioning in the immune system. Stem cell transplantation, lastly, is the treatment strategy to compensate the destroyed cells in the bone marrow by replacing them with stems cells that have the ability to produce healthy cells.

2.2. Malignant melanoma

Malignant melanoma (MM) is a type of skin cancer that emerges due to malignancy of pigment-producing melanocytes in the skin, iris, and rectum. It is pointed to be the most dangerous skin cancer. The characteristic genetic alterations leading to the development of the MM has been identified; however, it still accounts for the majority of deaths that occur as a result of skin cancers. The most common type is cutaneous melanoma, which develops in the skin and is responsible for 75% of the deaths resulted from skin cancers although there are several rare non-skin melanomas (Schadendorf et al., 2015).

2.2.1. Epidemiology

The MM mainly develops in white populations because of intense ultraviolet radiation (UV) penetrating the fair skin easily, while low incidence rated acral and mucosal melanomas develop in pigmented populations from Africa and Asia. Different populations exhibit varying degrees of incidence rates for the MM worldwide, but Australia and New Zeland are the two countries that have been reported to have the highest incidence rates approaching 60 cases per 100.000 inhabitants per year. The incidence rate is nearly 20 cases per 100.000 inhabitants per year in Europe and is nearly 30 cases per 100.000 inhabitants per year in the United States. However, the incidence rate is 1 case per 100.000 inhabitants per year in dark-skinned populations of Africa and Asia. Individuals whose ages are between 40 and 60 constitute clinically the highest risk group for the MM; nevertheless, it can occur adults and older peoples greater than 80. Furthermore, it is one of the most common type of cancers that can be observed among adults aged between 20 and 29, and the median age of diagnosis is 57 (Schadendorf et al., 2015).

American Cancer Society (2015) estimates that about 76,380 new melanoma cases consisting of 46,870 men and 29,510 women will be diagnosed in the United States in 2016. Unfortunately, it is expected that approximately 10,130 people consisting of 6750 men and 3380 women will die at the end of the year. Schadendorf et al. (2015) show the incidence rate of the MM increased 17-fold in men and 9-fold in women from 1950 to 2007. They also show it increased with similar folds in Australia, Central Europe, and Scandinavia. The rate is still increasing due to the sun exposure, sunburns, and rise of longevity in these areas.

2.2.2. Histological subtypes

Histological classification of the MM has been an intense topic to identify patient groups that could gain benefit from obtaining appropriate therapy choices. After rigorous studies have been conducted to elucidate the subtypes of melanoma, clinicians could be able to classify it into four broad subtypes (Smoller, 2006):

a) Superficial spreading melanoma
b) Nodular melanoma
c) Lentigo maligna and lentigo maligna melanoma

d) Acral lentiginous melanoma

Superficial spreading melanoma: Neoplasms displaying this histological pattern constitute nearly 75% of all melanomas (Smoller, 2006). Outward and flat growth on the surface of the skin occurs in early phase. This phase is called as the radial growth phase that may last years; however, melanoma alters the growth direction to the inward at the end of early phase, so that it enters the vertical growth phase in which a bump-shaped appearance is observed above the skin.

Nodular melanoma: It is the second most common type of melanomas following superficial spreading melanoma. Neoplasms of this histological pattern emerge mostly in middle-aged adults. It shares several histological features with superficial spreading melanomas, but sharp circumscription is typical unlike superficial spreading melanomas show poor circumscription. In addition, it does not have any radial growth phase. After neoplasms invade the epidermis, they are likely to enter a vertical growth phase which leads to aggressive downward growth (Smoller, 2006).

Lentigo maligna and lentigo maligna melanoma: Lentigo maligna is observed in elderly people whose heads and necks expose to sun damage. There is no available incidence rates for the subtype, yet it is evident that the rate increases dramatically. Lentigo maligna melanoma, which is the least common subtype of melanoma, emerges from lentigo melanoma by invading the dermis and has a long lasting radial growth phase (Tung and Vidimos, 2010).

Acral lentiginous melanoma: It is the rearest subtype of melanoma observed on acral surfaces. The areas of the skin without hair such as palms, soles, and nails are the potent sites for development. It may spread more quickly than previously mentioned melanomas (Tung and Vidimos, 2010).

2.2.3. Signs and symptoms

The MM is extremely heterogenous, so signs and symptoms vary from one patient to another. However, it is generally observed that new spots appear on the skin, or size, shape, or color of an existing mole changes. The “ABCDE” rule is pretty useful in assessing the signs. Even though it is not valid for all cases as nodular melanoma, the guideline works perfectly for the majority of melanomas (Melanoma Action Coalition, n.d.). “ABCDE” is an abbreviation of the words listed: asymmetry, border irregularity, color, diameter, and evolution (See **Figure 2.4**).

a) Asymmetry: Irregular shaped moles are typical.
b) Border irregularity: The edges of some melanomas are irregular, blurred, rough, or notched.
c) Color: Changes in shade or distribution of mole color are frequent.
d) Diameter: The width is greater than six millimeters.
e) Evolution: Asymmetry, border irregularity, color, and diameter changes over time.

Symptoms of the MM may include;

- Persistent sore
- New growth
- Itchiness, tenderness, or pain
- Spread of pigmentation and swelling to the outside the border of spot
- Change in the surface of a mole

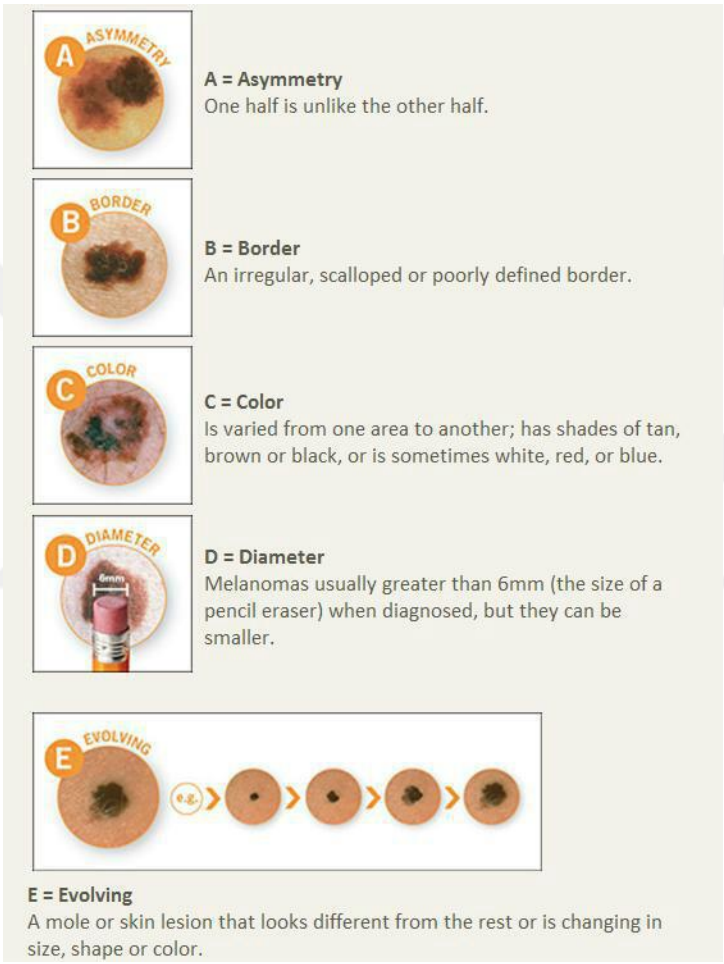

Figure 2.4: The ABCDE guideline used in clinics for early detection of melanoma (Melanoma Action Coalition, n.d.)

2.2.4. Causes and risk factors

The MM develops as a result of complex interplay of genetic and environmental factors. It is shown that continuous exposure to the UV originating from sunlight can induce melanocyte tumorigenesis. Apart from the UV irradiation, mutation or deletion of CDKN2A, a tumor suppressor gene in cell cycle regulation, is a genetic factor in the MM

development. Moreover, the BRAF mutation activating the RAF/MEK pathway is crucial in both the MM development and metastasis (Gruber et al., 2008). Genes which encode proteins related to pigmentation, DNA repair, cell growth, and differentiation or detoxification of metabolites also contribute. For example, MC1R, a pigmentation gene, is known to increase the susceptibility of individuals who carry the CDKN2A mutations (Hayward, 2003).

It is significant to detect individuals who have the predisposition to the MM or risk factors triggering the MM development. Thereby, individuals can avoid the chance of developing the disease. Nevertheless, having a risk factor does not suggest the disease always appear as it is known that some individuals may have no risk factors, but they might get the disease. American Cancer Society (2016) establishes risk factors as follows;

a) The UV light exposure
b) Moles
c) Fair skin, freckling, and light hair
d) Family history
e) Personal history
f) Weakened immune system
g) Older age
h) Male gender
i) Xeroderma pigmentosum

The UV light exposure: The UV radiation is the main factor in damage to skin due to the disruption of regulation in genes that control skin cell growth. Especially, intermittent exposure to the sun may give rise to sunburns that are likely to contribute to the risk of developing the MM. It triggers the appearance of many melanocytes in the skin since cells express genes interacting with a protein named interferon-gamma (IFN-g). These changes do not occur in non-melanoma cells, suggesting that IFN-g is essential in the MM development after the UV exposure (NIH, 2011). **Figure 2.5** depicts how exposure the sun may lead to the MM development.

Moles: A mole or nevus is a kind of benign pigmented tumor that can appear innate or a certain time elapsed. Moles normally do not give any harm to the body, but a person who has more than one mole is likely to develop the MM (American Cancer Society, 2016).

Fair skin, freckling, and light hair: Whites are more prone to develop the MM than blacks. Whites with red or blond hair, blue or green eyes, or fair skin that freckles or burns easily are at greater risk (American Cancer Society, 2016).

Family history: 10% of the MM patients has the family history. Both genetic and environmental factors account for the familial background. Mutations run in a family, and lifestyle of frequent sun exposure has a direct effect on the appearance of the disease (American Cancer Society, 2016).

Personal history: Individuals who suffered from the MM before have a higher risk of developing it. Furthermore, those who have basal or squamous skin cell cancers are at greater risk group (American Cancer Society, 2016).

Weakened immune system: Immune system is useful in fighting with agents that may result in skin cancers. Individuals whose immune systems are deficient, thus, are more prone to the MM (American Cancer Society, 2016).

Older age: Individuals who are older than 30 are more likely to develop the MM than younger individuals. However, individuals who have family history develops the disease in younger ages (American Cancer Society, 2016).

Male gender: In the United States, males older than 45 are at a greater risk than women older than 45, whereas males younger than 45 are at a lower risk than women younger than 45 (American Cancer Society, 2016).

Xeroderma pigmentosum: Xeroderma pigmentosum (XP) is a condition that skin cells might not repair damage occurred to the DNA. Thus, the XP individuals are in a risk group especially in younger ages (American Cancer Society, 2016).

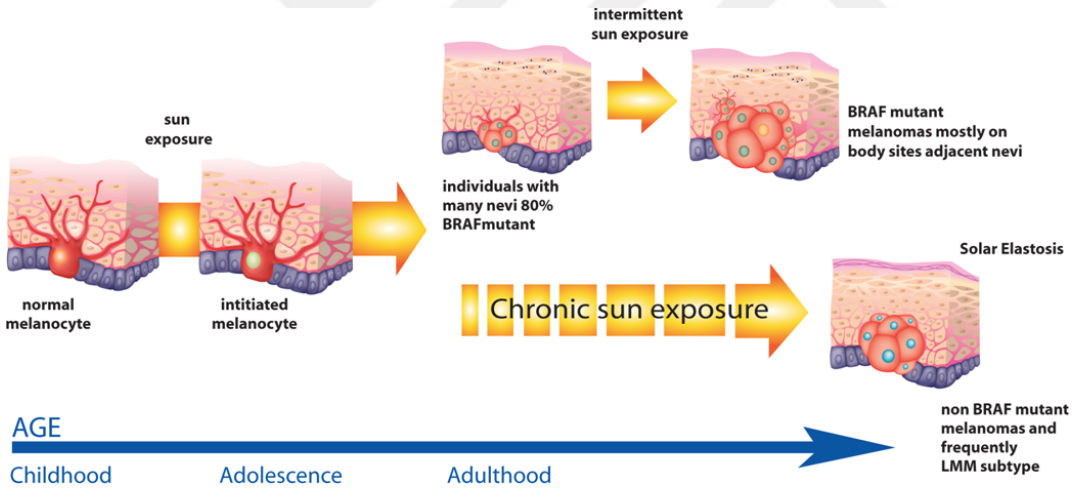

Figure 2.5: Ultraviolet light is a modulator in malignant melanoma development (Walker and Hacker, 2011)

2.2.5. Treatment

The type of treatment varies from patient to patient since the MM is a heterogenous disease, so the most appropriate therapy is decided according to stage and location of the disease. Early stage melanomas are not severe. It is not necessary to take any further treatment after they are surgically removed. Surgical operation is the only treatment in early stages. However, the MM can spread to the other parts of the body in later stages (American Cancer

Society, 2016). In addition to surgical operation, there are several treatment options that can be applied:

- Surgery to remove affected lymph nodes
- Chemotherapy
- Radiation therapy
- Immunotherapy
- Targeted therapy

Surgery to remove affected lymph nodes: Surgical operations may be necessary provided that the MM spread to the lymph nodes. However, when the MM spreads to the distant parts of the body such as organs, surgery is not a feasible option (Mayo Clinic, 2016).

Chemotherapy: Anti-cancer drugs are administered into either mouth or vein. After intake, drugs travel throughout bloodstream to attack cancer cells. This therapy option is recommended for advanced MM patients. However, chemotherapy is not an effective treatment option in the MM. It is used only to relive symptoms or extend survival of patients. American Cancer Society (2015) lists dacarbazine, temozolomide, paclitaxel, carmustine, cisplatin, carboplatin, and vinblastine that might be useful in treatment. Depending on the overall health of the patient, some of the drugs may be given in combination to enhance therapeutic efficacy. Nevertheless, combination strategy may lead to serious side effects such as hair loss, nausea, vomitting, loss of appetite, and increased risk of infection.

Radiation therapy: High-powered energy beams such as X-rays or particles are used to eradicate the MM in radiation therapy after surgical operation to remove lymph nodes is accomplished. The procedure is applied to relieve symptoms of the MM, so that it may not spread to the other parts of the body (Mayo Clinic, 2016).

Immunotherapy: It is a type of biological therapy that uses substances naturally made by body to help immune system for finding and destroying the MM cells. There are six main categories for immunotherapy approaches in the MM currently: checkpoint inhibitors, oncolytic virus therapies, cancer vaccines, adoptive T cell therapy, monoclonal antibodies, and cytokines. The typical agents for the MM administered in clinics are ipilimumab, pembrolizumab, and nivolumab (Cancer Research UK, 2016).

Targeted therapy: It is an approach to develop agents which are effective in destroying cancer cells but not normal cells. These agents interfere with specific molecules in specific pathways that are responsible for growth and spread of tumor to prevent associated side effects more effectively than chemotherapy or radiation therapy since they directly attack to the targets. Melanoma Research Foundation (n.d.) proposes there are five broad approved targeted therapy options;

a) Vemurafenib
b) Trametinib
c) Dabrafenib
d) Vemurafenib + cobimetinib

e) Dabrafenib + trametinib

2.3. Weighted gene co-expression network analysis (WGCNA)

2.3.1. Description

Weighted gene co-expression network analysis (WGCNA), also known as the weighted correlation network analysis, is a data mining method that constructs a gene co-expression network using correlation patterns among genes (Langfelder and Horvath, 2008). It is regarded to be a powerful network approach in analyzing high-dimensional biological data. It can be used for data reduction, clustering, feature selection, data integration, and data exploration. Network approach is intuitive to most biologists, and software implementation is simple, user-friendly, and comprehensive; thus, the WGCNA has gained a great popularity since its release. Although it can be applied to various types of biological problems, statisticians have widely used it in genomic applications mostly to identify candidate biomarkers or therapeutic targets.

2.3.2. Analysis goals

The WGCNA methodology is established majorly based on correlation networks, which are constructed on the basis of correlations between variables (Langfelder and Horvath, 2008). Correlation network methodology allows statisticians to use network language to extract pairwise relationships between network nodes. Thereof, many biological analyses required to identify key drivers in biological networks can be accomplished. For this purpose, the WGCNA method can be used for the following list of analyses:

- Distinct clusters (modules) of interconnected nodes can be identified.
- Higly connected hub genes, representative genes of the specific modules in which it is located centrally, can be identified to focus on a few biologically interesting modules instead of focusing thousand of genes. Thanks to this data reduction, multiple testing problems can be alleviated.
- Identification of significant modules can be achieved to relate these modules to external data.
- Annotation of network modules by defining a measure for module membership can be created to show which modules are more closely related to the identified modules.
- Network neighborhood can be defined to identify highly connected nodes to a given set of nodes in order to find interacting nodes that might be interesting.
- Screening nodes as to node significance or network topological property such as high connectivity can be achieved.
- Differential network analysis that contrasts one network to another network can be used to identify changing network parameters such as shape, size, and pattern among networks.

- Consensus module analysis can be used to obtain shared modules among many networks. These modules can be regarded building blocks of networks that are essential for the specific biological process.

2.3.3. Methodology

Defining a gene co-expression similarity measure is the first step in network construction. Pairwise correlation of genes is used to get this similarity measure. It is denoted by a pair of genes i and j such that;

$$s_{ij} = |\text{cor}(x_i, x_j)| \quad \textbf{(2.1)}$$

where s_{ij} is the measure of similarity, x_i and x_j are the measures of expression of genes i and j across multiple samples. This measure is called as unsigned co-expression similarity. Most of the co-expression network approaches use the unsigned measure; however, it raises serious problems since absolute value ignores relevant biological information such as activation or repression (Langfelder and Horvath, 2008).

The similarity measure is necessary to define the network, but it gives no information about how strongly genes are connected to each other. Thus, an adjacency matrix $A = [a_{ij}]$ is defined to quantify connectedness after similarity measure is transformed to connection strength. There are two possible ways to transform similarity measure into connection strength:

a) Hard thresholding
b) Soft thresholding

Hard thresholding

The similarity measure is transformed into network adjacency such that adjacency $a_{ij} = 1$ if $s_{ij} \geq t$ and 0 otherwise. Here, t is a threshold constant that can take any values between 0 and 1. Gene connections take discrete values when hard thresholding is used. It, however, may result in loss of information as to threshold choice. For example, if t is equal to 0.9, then the values below than 0.9 are encoded as non-connected, whereas the values higher than or equal to 0.9 are encoded as connected. In this case, 0.89 is a high similarity value, but hard thresholding classifies it as non-connected. This is a serious problem for interpretation of connectedness. Since there is no weight between nodes, the network constructed by hard thresholding is called as the unweighted network (Langfelder and Horvath, 2008).

Soft thresholding

Instead of using hard thresholding approach, the WGCNA has an alternative approach named soft thresholding. It preserves continuous values in order not to lose any information of gene connectedness. In this case, there is a weight between nodes, so the network constructed by soft thresholding is called as the weighted network (Langfelder and Horvath, 2008). The WGCNA uses following power function to assess connection strength;

$$a_{ij} = (s_{ij})^{\beta} \quad \textbf{(2.2)}$$

where β is the soft thresholding parameter. $\beta = 6$ is the default value for unsigned networks. Scale free criterion, which assumes that degree distribution of a network decays by a power law, is optionally applied to find the best β values for thresholding (Langfelder and Horvath, 2008). Scale free networks consist some nodes that have many more connections called as hubs compared to most of the nodes in the network. Most biological networks are considered to be scale free since it is known that only a few nodes are hubs, and the rest is much less connected (Barabasi and Bonabeau, 2003). Therefore, provided that approximate scale free topology is reached, the smallest value of β can be chosen to construct a weighted network.

The advantage of power function is that it relates co-expression similarity to weighted network adjacency linearly on a logarithmic scale as $\log(a_{ij}) = \beta \log(s_{ij})$. This equation simply suggests higher β values transforms high similarities into high adjacencies. Nevertheless, it approximates 0 as similarity decreases.

The next step in the analysis is to identify distinct modules, which are composed of highly interconnected genes. One needs to define a network proximity measure in order to produce the modules. The WGCNA uses the topological overlap matrix (TOM) as the proximity measure though there are several measures available since it is a great approach to find biologically meaningful modules (Langfelder and Horvath, 2008). The proximity of a pair of genes in the modules takes values between 0 and 1. The values which are closer to 1 is said to be highly interconnected. The TOM measure includes both the adjacency of genes and the connection strengths these genes share with other genes. Since it considers shared neighbours, network proximity measure is highly robust when compared to available methods (Langfelder and Horvath, 2008). This measure is an input for module detection methods to produce modules. However, the default method is hierarchical clustering. When the clustering analysis is complete, the output of modules is summarized with a value known as module eigengene that represents the genes inside of a given module (Langfelder and Horvath, 2008). It is a weighted average of the standardized module gene expression data that makes it the first principal component of the standardized expression profiles (Langfelder and Horvath, 2008). Accordingly, every module has a unique module eigengene value. This results in a reduction of hundreds or thousands of expression profiles of genes to single module eigengene value. Thus, instead of dealing with a number of genes, it is possible to use the eigengene value when relating the module to external data. However, before relating it to external data, one should check for the functionality of the modules. It is

wise to apply gene ontology information to test whether the modules are biologically meaningful as some modules may reflect noise in place of a true signal.

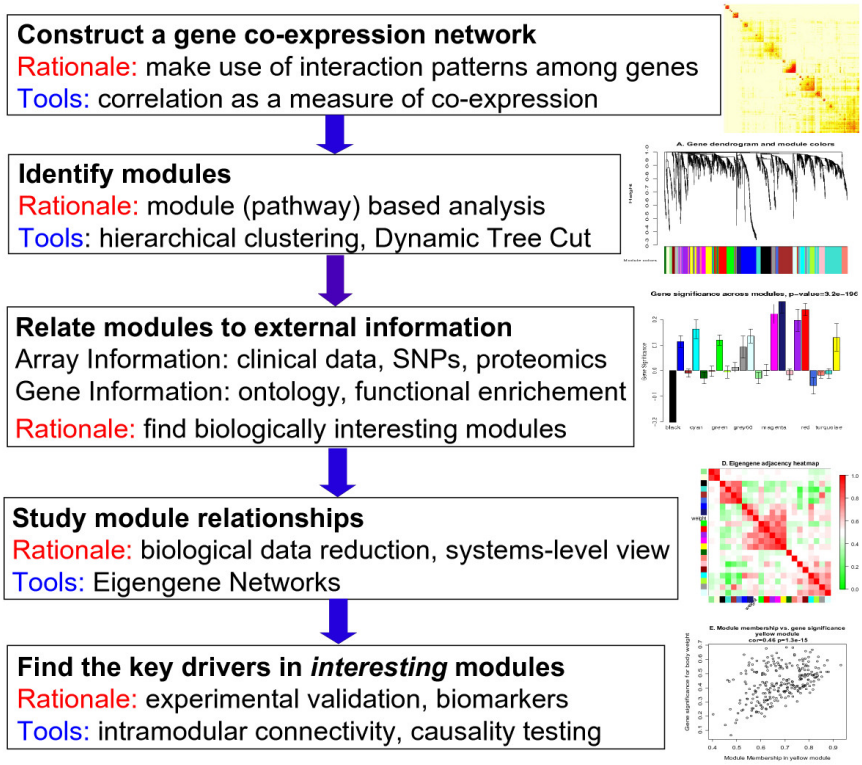

Figure 2.6: The workflow of the WGCNA to identify centrally located intramodular hub genes which are the representatives of the modules obtained by hierarchical clustering (Langfelder & Horvath, 2008)

The final step is to relate functional modules to the biological trait of interest. It can be achieved by correlating the module eigengene value of modules to external data. This gives rise to eigengene significance measure that can be used in the selection of modules for further analysis. Moreover, one can construct eigengene modules to deduct the relationships of modules with each other. The ones which show greater similarity in their eigengene values can be merged to obtain more functional modules, so that they can be subsequently related to external data. After selection of modules, it is essential to identify key drivers in each module. The WGCNA has two connectivity measures, module membership and intramodular connectivity for this task (Langfelder and Horvath, 2008). Module membership correlates module eigengene value with individual expression of genes inside a given module, whereas intramodular connectivity is the value of the sum of adjacencies with respect to module genes although they are equivalent in practice (Langfelder and Horvath, 2008). The genes connectivity values of which are greatest in the modules are regarded as the intramodular hub genes that explain the expression profile of modules greatly. This gives an opportunity to represent each module with an only single gene which collects the highest

variance. **Figure 2.6** shows the workflow of the WGCNA in identifying such intramodular hub genes.

2.4. Support vector machine

2.4.1. Description

Support vector machine (SVM) is a machine learning algorithm that uses supervised learning approach to perform classification and regression analysis (Burges, 1998). It classifies samples in two groups by finding a hyperplane that can separate them as wide as possible in classification case. The vectors which define the hyperplane is called as support vectors (Fletcher, 2009). In addition to classification, a function that has the ability to predict the value of interest can be generated for regression tasks. As it occurs in classification, regression analysis requires a loss function penalizing the values outside of an interval from actual values (Lanlan et al., 2015). Provided that the best parameters are chosen for optimization task in model construction, it may have an outstanding predictive ability in classification and regression problems. However, it needs data have already been labeled as it can not perform unsupervised clustering. Therefore, an extension of SVM, support vector clustering, was developed to improve the algorithm (Ben-Hur et al., 2001). The major applications of the SVM used in real life include time series forecasting, handwriting recognition, text categorization, bankruptcy prediction, face identification and recognition, and biological and medical aid (Gaspar et al., 2012).

2.4.2. Theory behind the SVM

The idea behind the SVM is to find a hyperplane which separates two classes by maximizing the distance between them in order to build a classifier that divides data into training and test sets. Values in the training data labeled for each class are used to produce a model that can predict the labeled values of the test data provided that only the test data features are applied. For instance, **Figure 2.7** visualizes that some positively and negatively labeled samples could be divided into two groups by maximum-margin hyperplanes after optimal hyperplane is determined.

The SVM could also be applied for regression tasks by defining a loss function that tolerates errors within a certain distance from the actual value. This error tolerating region is called as the epsilon intensive zone, and variables outside of the zone determine the cost of errors on the training points. **Figure 2.8** depicts a hypothetical case for linear regression task that could be performed by Support Vector Regression (SVR), which is a name given for regression task.

2.4.2.1. The linear case

The SVM is used to find a hyperplane that separates training data into two classes. However, it is not always possible to find a hyperplane. In this case, a hinge loss function is introduced to separate samples into two classes. In this respect, linear SVM could perform classification task in two possible ways:

a) Linearly separable case (Hard margin)
b) Linearly non-separable case (Soft margin)

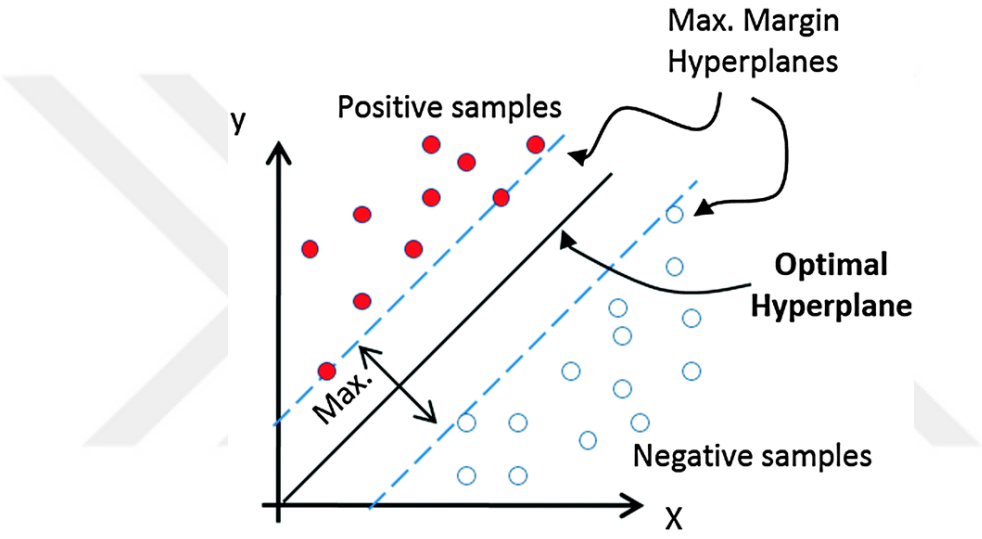

Figure 2.7: Visual representation of classification task by the SVM (Fernandes-Lozano et al., 2014)

The linearly separable case

Given that N points exist in the training data $\{(x_1, y_1),(x_2, y_2),(x_3, y_3),\ldots,(x_N, y_N)\}$, where $x_i \in R^d$ and $y_i \in \{-1, +1\}$, a hyperplane classifier can be built. It should be noted the data satisfies the following constraints;

$$w \cdot x_i + b \geq 1, \quad y_i = 1 \quad \& \quad w \cdot x_i + b \leq -1, \quad y_i = -1 \tag{2.3}$$

where w is normal to the hyperplane, |b| / ||w|| is the distance from the hyperplane to the origin, and ||w|| is the Euclidean norm of vector w. The following equation can be obtained after these two constraints are combined together

$$y_i (w \cdot x_i + b) \geq 1, \quad \forall i \tag{2.4}$$

Two hyperplanes, H_1 and H_2, can be determined for points lying on **Equation 2.3**. The margin m is extracted as the distance between H_1 and H_2.

$$m = \frac{|1 - b|}{\|w\|} - \frac{|-1 - b|}{\|w\|} = \frac{2}{\|w\|} \quad \textbf{(2.5)}$$

From the **Equation 2.5**, the hyperplane that separates two classes maximally can be obtained after the following optimization problem;

$$\text{Minimize}_{(w,b)} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i (w \,.\, x_i + b) \geq 1, \quad \forall\, i \quad \textbf{(2.6)}$$

The minimization formula can be derived after Lagrange multiplier $\alpha_i \geq 0$ is introduced for each constraints in **Equation 2.6** such that;

$$\text{Maximize } L(b, w, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} a_i y_i (w \,.\, x_i - b) + \sum_{i=1}^{N} \alpha_i \quad \textbf{(2.7)}$$

Equation 2.7 is maximized with respect to w and b. In this step, derivatives of L (b, w, α) with respect to all the α disappear.

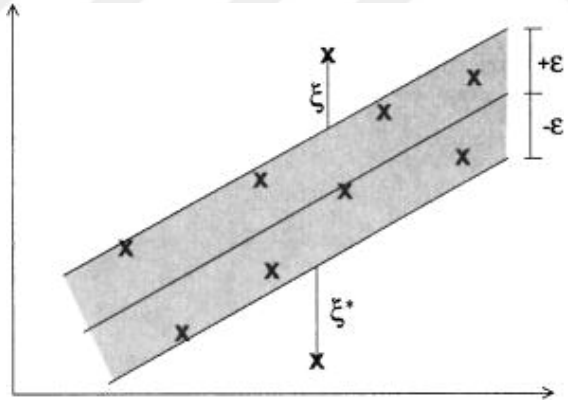

Figure 2.8: Visual representation of regression task by the SVR (Clarke et al., 2003)

The linearly non-separable case

In the case of perfectly separable margin, it is not possible that any observation lies on the boundary of the margin. In this scenario, classification function can linearly separate the classes safely due to the maximization of the margin. This is known as hard margin. However, classes may not be separated perfectly in some cases. In this case, the SVM formulation can not find a solution since classification error precludes the existence of

hyperplanes in this non-perfectly separable case. What is more, the classifier may overfit when it looks for a hyperplane that separates the margin perfectly although the data is linearly separable. The presence of outliers, for example, may be the cause of mentioned overfitting. Thus, a soft margin approach was developed for the SVM. A slack variable ε_i is introduced in **Equation 2.3** and **Equation 2.4** to allow classification errors at the expense of a cost proportional to the value of ε_i . Thus, the new constraints with slack variables become;

$$\forall i \quad \begin{cases} w \cdot x_i + b \geq +1 - \epsilon_i & y_i = +1 \\ w \cdot x_i + b \leq -1 - \epsilon_i & y_i = -1 \\ \epsilon_i \geq 0 \end{cases} \tag{2.8}$$

Equation 2.8 permits some instances to be placed into another class. This, however, reduces the effect of outliers. Large value of ε_i is not desirable since it may lead to trivial or sub-optimal solutions. Therefore, misclassification error has to be adjusted introducing the slack variable to **Equation 2.6**.

$$\text{Minimize}_{(w,\epsilon)} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \epsilon_i \tag{2.9}$$

Equation 2.9 is subject to the constraints in **Equation 2.8**. Here, C is the cost coefficient used for misclassification penalty. So, the solution of **Equation 2.9** can be written as;

$$\frac{\partial L_p}{\partial w} = 0 \quad \Leftrightarrow \quad w = \sum_i \alpha_i y_i x_i \tag{2.10}$$

where $\alpha_i > 0$ are the support vectors of the SVM solution. Now, maximization formula can be obtained and solved as the function is maximized with respect to α and minimized with respect to b and w.

$$\text{Maximize}_{\alpha} L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i . x_j) \quad \text{subject to} \quad \forall i \begin{cases} \sum_i \alpha_i y_i = 0 \\ C \leq \alpha_i \leq 0 \end{cases} \tag{2.11}$$

2.4.2.2. The non-linear case

Linear SVM has linear decision boundaries; however, the SVMs can be extended to more general decision boundaries. Nevertheless, the kernel trick can be used to have non-linear boundaries thanks to dot product of vectors in **Equation 2.12**. Given that a kernel function exists as a dot product in feature space such that,

$$K(x_i, x_j) = f(x_i) . f(x_j) \tag{2.12}$$

The dot product in **Equation 2.13** can be replaced with new kernel function K. The final equation becomes

$$\text{Maximize}_{\alpha}\, L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K\left(x_i . x_j\right) \quad \text{subject to} \quad \forall i \begin{cases} \sum_i \alpha_i y_i = 0 \\ C \le \alpha_i \le 0 \end{cases} \qquad \textbf{(2.13)}$$

Non-linear SVMs transform the training data into a higher dimensional feature space to obtain a hyperplane with maximum margin that separates the classes. Kernel functions are applied to find such a hyperplane. The most widely used kernel functions are;

a) Polynomial kernel (homogenous): $K\left(x_i, x_j\right) = \left(x_i . x_j\right)^d$

b) Polynomial kernel (inhomogenous): $K\left(x_i, x_j\right) = \left(x_i . x_j + 1\right)^d$

c) RBF kernel: $K\left(x_i, x_j\right) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$

d) Sigmoid kernel: $K\left(x_i, x_j\right) = \tanh\left(\gamma\left(x_i, x_j\right) - \phi\right)$

2.4.3. Parameter optimization

Finding a hyperplane maximizing the distance between two classes is an optimization problem. A training procedure using variables to separate distinct classes is, thus, required to extract the optimum parameters in model construction, so that the SVM can learn the rule of function to make accurate decisions about which group it will classify the new sample. An improvement in performance can be basically accomplished with tuning the slack variables penalty weight (*C*) to identify the best tradeoff while generalizing model by allowing misclassification errors (Gaspar et al., 2012). Misclassification errors are penalized with large *C* values; therefore, the hyperplane obtained by the soft margin SVM strongly avoids misclassification errors while sacrificing generalization (Chen et al., 2004). As *C* approaches infinity, a hard margin SVM behavior appears. However, low *C* values slightly penalize the misclassification errors, giving rise to failure in accurate separation. Since it is independent of kernel choice, minimization of the value is easily adjusted. The Linear kernel only requires *C* should be tuned appropriately; however, additional parameters should also be tuned when using other kernels. The most common choice is to use radial basis function (RBF) kernel, which includes sigma and epsilon parameters to be tuned accordingly with the C parameter. In addition to the RBF, there are several different kernel choices that are highly popular such as polynomial and sigmoid kernels. These kernels are optimized in a similar manner, but the RBF has an advantage due to its flexibility in fitting the data when compared to other kernel methods (Gaspar et al., 2012).

The SVM is powerful in separating the classes by classification or predicting the actual values by regression when an appropriate choice of kernel is chosen and the SVM

parameters are fine tuned. However, data itself is also important as the SVM algorithm strictly depends on data to plot each training example in a high dimensional space. Although the kernel is responsible for the translation of high dimensional space, the separation is mostly related to the available feature set. Features are extremely important in determining the separability of data into two classes since correlation of each future to its class affects the process (Chang and Lin, 2013). What is more, some features may contribute negatively to the classification process. Thus, it is essential to select the best features for the SVM to be able to generalize the model. That explains why numerous optimization strategies give importance to feature selection process (Czarnecki et al., 2015).

In optimization task, the efficacy of the SVM is generally dependent on four strategies:

a) Correct kernel choice
b) Optimal kernel parameters
c) Misclassification penalty choice
d) Feature selection

The strategies mentioned can improve performance significantly. Huang and Wang (2006), for instance, pointed in their study that when C and gamma parameters are considered together in optimization with the RBF kernel choice and Grid search are used for feature selection, the performance of model increases superficially. However, a vast majority of the efforts focuses on specific kernel choices and optimization of parameters, while ignoring the comparison of different kernels as parameters are optimized. The study of Huang and Wang is a supportive evidence that simultaneous feature selection and parameter optimization indeed improve the performance of the classifier. Thereof, different kernel choices with optimized parameters combined with appropriate feature selection methods should be assessed together to improve the performance of classifier further.

2.5. Related works

2.5.1. Previous works applied the WGCNA methodology

Xue et al. (2013) studied the transcriptome changes from oocyte to morula in human and mouse embryos. The WGCNA identified functional modules for each developmental stage. In addition, they observed that nearly all of the human stage-specific modules were conserved, but developmental specificity and timing differed between these two species. They also determined several key genes that might be important in mammalian pre-implantation development.

Levine et al. (2013) investigated the genes responsible for dysregulation in individuals with the HIV-association neurocognitive disorder (HAND). They used gene expression data derived from brains of the HIV^+ patients to illuminate the pathogenicity of the HAND. The

WGCNA approach found TNPO3 gene associated with neurocognitive impairment. Moreover, they used these data, along with gene expression data derived from Alzheimer's Disease (AD) patients, to identify shared pathways for pathogenesis. Cytoplasm, energy, mitochondrion, tricarboxylic acid cycle, transit peptide, and synaptic vesicle pathways were suppressed in both diseases, whereas cell differentiation, activator, repeat, cell communication, regulation of transcription, and phosphorylation pathways were increased. They also observed that CTDSP2, CASC3, PGF, SASH1, HIPK2 genes were upregulated in both groups.

De Jong et al. (2012) applied the WGCNA method to identify genetic factors of schizophrenia. They used transcriptome data obtained from schizophrenia patients and healthy controls. As a result of their analysis, they identified modules including brain-expressed genes. The central gene in this module, ABCF1, was regulated by the MHC complex; thus, they concluded the MHC complex might give rise to schizophrenia.

Miller et al. (2008) explored the molecular targets for Alzheimer's disease (AD) with the WGCNA approach. They identified distinct functional modules most of which were related to disease progression. They also investigated the impact of gene expression changes to the progression of the AD and normal aging to compare module conservation. Two modules, the module related to mitochondrial processes such as energy metabolism and the module related to synaptic plasticity, emerged to conserve between two conditions. They identified CDK5, YWHAZ and PSEN1 genes were central in both the AD and aging, suggesting they have roles in the progression of the disease.

Gong et al. (2007) inquired the effect of ambient air pollution to cardiovascular mortality and morbidity. They explored the diesel exhaust particles (DEP) and oxidized 1-palmitoyl-2-arachidonyl-sn-glycero-3-phosphorylcholine (OX-PAPC) on genome-wide gene expression by using human microvascular endothelial cells (HMEC). The WGCNA approach indicated both the DEP extract and ox-PAPC co-regulated many genes. They enriched the identified modules and observed the modules were relevant to vascular inflammation. Their in-vivo experimentation study with hypercholesterolemic mice also demonstrated these particles resulted in upregulation of HO-1, XBP1, and ATF4 genes in the module related to the liver.

Horvath et al. (2006) studied to identify new molecular targets for glioblastoma that may give rise to targeted therapy options. They used two independent datasets of clinical tumor samples for the WGCNA methodology and identified a module downstream of the mutant epidermal growth factor receptor, EGFRvIII. They indicated Erlotinib, which is an epidermal growth factor tyrosine kinase inhibitor, could inhibit this receptor. In addition, they identified ASPM gene in the module to be a candidate molecular target in glioblastoma after they inhibited the gene to demonstrate that ASPM is essential in tumor cell and neural stem cell proliferation.

Ghazalpour et al. (2006) applied the WGCNA methodology to identify genetic regulatory loci associated with mouse weight. For this purpose, they used liver gene expression data of female mice and genetic marker data from an F2 mouse intercross. After identifying several

modules were strongly related to weight, they investigated the genes which have a strong correlation with body weight in these modules and demonstrated QTL2, QTL5, QTL10, and QTL19 are key loci which coordinately regulate the modules.

Gargalovic et al. (2006) used gene expression profiling of endothelial cells to construct a co-expression network with the WGCNA approach. They observed that the genes in some identified modules were significantly enriched for known pathways. Moreover, they validated genes in the modules enriched for an unfolded protein response (UPR) by the siRNA and the UPR inducer tunicamycin. They predicted a gene of unknown function (CHAC1) present in the module and is a target for the UPR transcriptional activator ATF4.

Oldham et al. (2006) investigated the molecular bases of brain organization between human and chimpanzee brains using gene expression measure for network construction. They produced distinct modules correspond to the different brain regions and analyzed the conservation of the modules between the species. They identified the module associated with cerebral cortex was weakly conserved than the module associated with subcortical brain regions. Furthermore, they identified LDOC1, EYA1, LECT1, and PGAM2 genes had significantly changed in human - chimpanzee evolution.

2.5.2. Previous works applied the SVM for drug response prediction

Dong et al. (2015) constructed an in-silico model that could predict anti-cancer drug response from gene expression and drug sensitivity data available in the Cancer Cell Line Encyclopedia (CCLE). They constructed the model with Support Vector Machine (SVM) and a recursive feature selection tool. After checking the robustness of the model with cross-validation, they tested the predictive ability in an independent Cancer Genome Project (CGP) dataset. The performance of their model was great for most of the drugs profiled in the CCLE. However, when they used the CGP data as test data, only three drugs out of eleven drugs shared between the CCLE and the CGP achieved a satisfactory performance. They suggested genomic features were powerful in anti-cancer drug response prediction and concluded their model could be effective in personalized medicine due to its high predictive ability in drug response prediction for certain drugs.

Hejase and Chan (2015) applied non-linear SVM to predict the drug response of breast cancer cell lines. They integrated proteomic, gene expression, RNA-seq, DNA methylation, and DNA copy number variation data to increase the predictive ability of their model. Although they also used different machine learning algorithms other than the SVM, it appeared top three performing ensemble approaches were the SVM family of supervised learning methods with weighted probabilistic c-index scores 0.562, 0.554, and 0.549.

Jang et al. (2014) assessed different modeling approaches to compare model efficacy in drug sensitivity prediction. They used gene expression, copy number, and mutation data in the CCLE and the CGP studies to construct their models. They observed the SVM was one of

the most powerful algorithms for drug sensitivity prediction. They also noted that the efficiency of the SVM increased when the data were integrated.

Kovalev et al. (2013) examined the predictive ability of machine learning methods such as the SVM, Naive Bayesian, Logistic Regression and Linear Discriminant Analysis in tuberculosis drug response using X-ray and CT images of tuberculosis patients. They observed the highest classification accuracy of drug response was 75% and the SVM was the top performing method.

Ruderfer et al. (2009) tried to predict small-molecule perturbagen (SMP) response from gene expression data of yeasts measured in an SMP-free medium. They used the SVM for classification task that divided the yeast populations into sensitive or resistant to SMP. In this way, they could identify drug response over 70% accuracy.

CHAPTER 3

METHODOLOGY

3.1. Overview

In this chapter, two large pharmacogenomics studies containing molecular and drug sensitivity data of human cancer cell lines are introduced. The pre-processing steps of gene expression and drug sensitivity data of malignant melanoma (MM) cell lines subsetted from these studies is explained, and the workflow of the WGCNA methodology used for identification of gene expression-based candidate drug biomarkers is presented. In addition, in-silico validation process of identified biomarkers is explained in detail, and the results of the validation are illustrated with visuals and tables.

3.2. Data sources

In the study, two large pharmacogenomics studies, the Cancer Cell Line Encyclopedia and the Cancer Genome Project, are used as the data sources.

3.2.1. Cancer Cell Line Encyclopedia

The CCLE is a pharmacogenomics study conducted to understand the genetic characterization of various human cancer cell lines. The Broad Institute of MIT & Harvard collaborated with the Novartis Institutes for Biomedical Research and its Genomics Institute of the Novartis Research Foundation to complete this project. It includes gene expression, chromosomal copy number variation, and massively parallel sequencing data of the 947 human cancer cell lines. In addition, pharmacological profiles of the 24 anti-cancer drugs screened against the 479 cell lines are available in the CCLE study. Human cancer cell lines were obtained from 36 different tumor types and characterized by several genomic technology platforms. Targeted massively parallel sequencing was used to determine the mutational status of more than 1,600 genes. Mass spectrometric genotyping identified 392 recurrent mutations affecting 33 known cancer genes. High-density single nucleotide polymorphism arrays (Affymetrix SNP.0) were used for DNA copy number measurement.

Affymetrix U133 plus 2.0 arrays quantified the mRNA expression levels for each of the cell lines (Barretina et al., 2012).

There are 61 MM cell lines for which gene expression was profiled in the CCLE study. However, not all the cell lines were screened against the 24 anti-cancer drugs. Thus, the sensitivity profile of each drug was constructed with the varying number of cell lines screened. At this point, it should be noted that none of the drugs profiled in the CCLE study has been currently administered in the treatment of the MM in clinics, but it is assumed that they may have a significant cytotoxic activity that has not been identified yet.

The CCLE study reports the sensitivity data with 4 different parameters for all drugs: IC50, EC50, Amax, and Activity Area.

IC50: The concentration of a drug required for 50% inhibition of desired activity in *in-vitro* studies is defined as IC50. It is also known as the half-maximum inhibitory concentration.

EC50: The concentration at which a drug produces 50% of the maximal possible effect in in-vitro studies is defined as EC50. It is also known as the half maximal effective concentration.

Amax: The maximal effect level of a drug is defined as Amax.

Activity Area: The area above the dose-response curve is defined as Activity Area.

3.2.2. Cancer Genome Project

The CGP is an effort, similar to the CCLE, that aims to identify molecular causes of cancer and discover therapeutic biomarkers which could influence significantly the design, cost, and success of anti-cancer drug development. The project was funded by the Wellcome Trust Sanger Institute and the National Institute of Health. It includes human cancer cell lines which were subjected to exome sequencing of the 64 commonly mutated cancer genes, genome-wide analysis of copy number gain and loss using Affymetrix SNP.0 microarrays, and expression profiling of the 14,500 genes using Affymetrix HT-U133A microarrays. In addition, it contains pharmacological profiles of the 139 anti-cancer drugs (Garnett et al, 2012).

The CGP study includes expression profile of 42 MM cell lines, which is less than the CCLE study profiling 61 MM cell lines. However, the 139 anti-cancer drugs were not screened against all the cell lines. None of these drugs has been approved for the MM treatment as it in the case of the anti-cancer drugs profiled in the CCLE study.

The CGP study reports the sensitivity data with 2 different parameters, IC50 and AUC, for all the drugs. IC50 is the common drug parameter measure used in both the CCLE and the

CGP studies, while AUC parameter, which is the area under the dose-response curve, is only available in the CGP study.

3.3. Data pre-processing

Both the gene expression and drug sensitivity data available in the CCLE and the CGP studies are pre-processed before implementing the WGCNA methodology to identify candidate drug biomarkers.

3.3.1. Pre-processing of gene expression data

The CCLE and the CGP studies contain basal gene expression profiles of the MM cell lines analyzed by the DNA microarray technology. However, the expression profiles were obtained by two different platforms even though both studies used Affymetrix technology. The CCLE study used Affymetrix Human Genome U133 Plus 2.0 Array, while the CGP study used Affymetrix Human Genome U133A Array (Haibe-Kains et al., 2013). Thereby, the two data contain different number of genes. The CCLE data profiled expression of 19,178 genes; on the other hand, the CGP data profiled expression of 12,153 genes. 12,153 genes are common between the two studies. **Figure 3.1** shows count of genes expressions of which were profiled in a venn diagram.

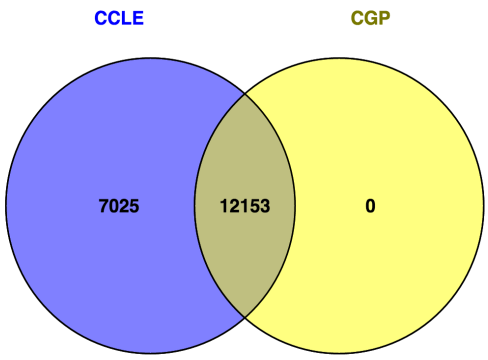

Figure 3.1: Count of the genes profiled in the CCLE and the CGP studies (Oliveros, J.C., 2007 - 2015)

Apart from different platform choice, the MM cell lines used in the studies mostly differ. The CCLE and the CGP studies used 61 and 42 MM cell lines, respectively. Nevertheless, only 29 of the cell lines are shared between the studies. **Figure 3.2** shows count of the MM cell lines investigated in the studies.

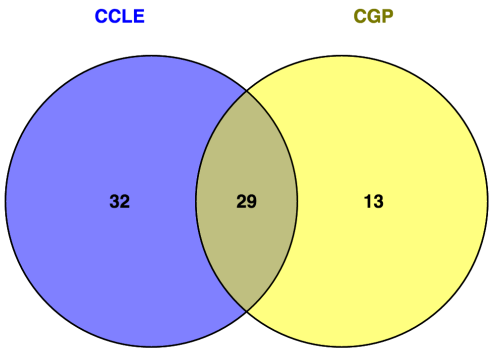

Figure 3.2: Count of the malignant melanoma cell lines examined in the CCLE and the CGP studies (Oliveros, J.C., 2007 - 2015)

Raw gene expression data (Affymetrix .CEL files) of the 61 MM cell lines investigated in the CCLE study is downloaded from Gene Expression Omnibus (GEO) website with GSE36133 accession code. However, only the data of the 38 MM cell line samples in the study is selected for statistical analyses since the fifteen common drugs between the CCLE and the CGP studies were screened against only these cell lines. In addition, the raw gene expression data (Affymetrix .CEL files) of the 42 MM cell lines investigated in the CGP study is downloaded from ArrayExpress website with E-MTAB-783 accession code. As in the case of the CCLE study, the CGP study did not screen the drugs against all the malignant melanoma cell lines. So only the screened malignant melanoma cell lines are selected for the analyses.

R programming language is used to perform all pre-processing steps of the gene expression data. 'affyPLM' package is used to inspect images of the arrays for quality control (Bolstad et al., 2005). Using the same package, density distribution plots are produced as diagnostic plots to show whether arrays required normalization. 'ggplot2' package is used to plot boxplots of samples to identify expression levels of the arrays before normalization (Wickham, 2009). After carefully inspecting images of the arrays for quality control and producing density distribution plots along with boxplots of the arrays, RMA normalization is performed to calculate a single expression value for a transcript from a set of values, such that it removes background chip effects and normalizes intensity values. 'affy' package is used for normalization procedure (Gautier et al., 2004). Lastly, 'sva' package is applied to combine and homogenize cell lines between the CCLE and the CGP studies. A uniform model is generated to remove batch effects and unwanted variation in gene expression data by using the package (Leek et al., 2016). Thereby, gene expression levels of the arrays in the CCLE and the CGP studies, which use different platforms for gene expression profiling, are set on the same scale. A boxplot of all the arrays in the two studies is also produced to show that noise is properly removed.

3.3.1. Pre-processing of drug sensitivity data

The CCLE study identified pharmacological profiles of the 24 anti-cancer drugs against the 61 MM cell lines, while the CGP study identified pharmacological profiles of the 139 anti-cancer drugs against the 42 MM cell lines. 15 anti-cancer drugs are common between the two studies (Barretina et al., 2012; Garnett et al., 2012). Count of the drugs used in the studies is visualized as a venn diagram in **Figure 3.3**. Names of the shared drugs is also given in **Table 3.1**.

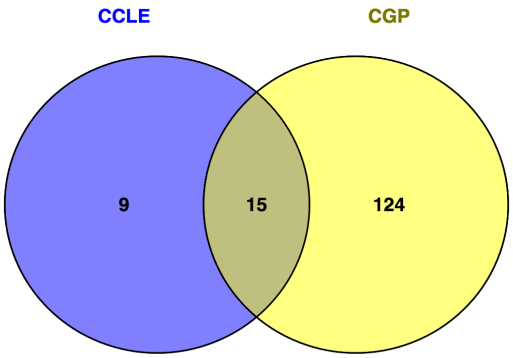

Figure 3.3: Count of the anti-cancer drugs screened against the malignant melanoma cell lines (Oliveros, J.C., 2007 - 2015)

Table 3.1: Names of anti-cancer drugs shared between the CCLE and the CGP studies

Drug Name	**Target(s)**
17-AAG	HSP90
AZD0530	SRC, ABL1
AZD6244	MEK1/2
Erlotinib	EGFR
Lapatinib	EGFR, ERBB2
Nilotinib	ABL
Nutlin-3	MDM2
TAE684	ALK
Paclitaxel	Microtubules
PD0325901	MEK1/2
PD0332991	CDK4/6

Table 3.1 (Continued)

Drug Name	Target(s)
PF2341066	cMET, ALK
PHA665752	MET
PLX4720	BRAF
Sorafenib	PDGFRA, PDGFRB, KDR, KIT, FLT3

The drug sensitivity data of the common drugs profiled in the CCLE study is downloaded from the CCLE (http://www.broadinstitute.org/ccle/) website. In addition, the drug sensitivity data of the common drugs used in the CGP study is downloaded from the CGP (http://www.cancerrxgene.org/downloads/) website. Subsequent analyses are restricted to the common drugs between the two studies. However, these drugs were not screened against all the malignant melanoma cell lines in both of the studies. Count of malignant melanoma cell lines investigated in the both studies for sensitivity profiling is tabulated in **Table 3.2**.

Table 3.2: Count of malignant melanoma cell lines investigated for sensitivity profiling

Drug Name	Count (CCLE)	Count (CGP)
17-AAG	40	41
AZD0530	40	18
AZD6244	40	36
Erlotinib	40	16
Lapatinib	40	16
Nilotinib	33	41
Nutlin-3	40	41
Paclitaxel	40	18
PD0325901	40	39
PD0332991	40	35
PF2341066	40	18
PHA665752	40	18
PLX4720	40	41
Sorafenib	40	18
TAE684	40	18

3.4. Identification of gene expression-based candidate drug biomarkers

Before implementing the WGCNA method, the most varying 8000 genes in the CCLE study are selected among 19,178 genes expressions of which were profiled by Affymetrix microarray technology. In addition, the 38 MM cell lines screened by all the common drugs between the CCLE and the CGP studies except Nilotinib are selected to apply the WGCNA method. Nilotinib is intentionally disregarded because it was screened against less MM cell lines. Only the probesets which have the greatest variation for the genes across the 38 samples are determined and picked for the analysis since multiple probesets were used for a particular gene in the microarray experiments. The final expression dataset, thereby, consists the 8000 most varying genes across the 38 MM cell lines.

R programming language is used to apply the WGCNA methodology to identify candidate biomarkers. 'WGCNA' package is employed to construct a co-expression network, cluster genes into network modules, relate modules to drug sensitivity data, and identify centrally located intramodular hub genes regarded as candidate biomarkers (Langfelder and Horvath, 2008).

In the first step of the identification of candidate biomarkers, the pair-wise correlation patterns among the 8000 genes across the 38 MM cell lines are described to construct a co-expression network. After the co-expression network is constructed, highly co-expressed genes are clustered into network modules by hierarchical clustering. A network plot is generated by Cytoscape to investigate the statistics of the constructed network such as node and edge count, clustering coefficient, network density, and average number of neighbours. Clusters are produced by a dynamic tree cut approach, which clusters the modules according to the shape of the clusters. The obtained modules are summarized with eigengene values, which can be considered as the weighted average of the expression profiles of the genes inside the modules. Next, module enrichment analysis is performed by the Database for Annotation, Visualization and Integrated Discovery (DAVID) tool to check whether hierarchical clustering produces cohesive and functional modules (Huang et al., 2008). Afterwards, an eigengene network, which is the co-expression network between module eigengenes, is constructed to merge the modules for which eigengene values are at least 70% similar. In this way, highly co-expressed genes are grouped into one big module. Then, eigengene values of the merged modules are re-calculated, and module enrichment analysis is repeated to identify functional modules. Only significantly enriched modules are selected for the rest of the analysis. Thereafter, drug sensitivity data (IC50) is integrated to the expression data by correlating sensitivity profiles of the drugs to the eigengene values of the significantly enriched modules. The modules which have significant correlations (P-value < 0.05) are selected to identify intramodular hub genes. As the last step, intramodular connectivity values greater than 0.70 are determined within the significantly enriched and correlated modules. Individual expression of these hub genes is correlated to sensitivity profiles of each drug, and the hub genes having significant correlations (P-value < 0.05) are regarded as the candidate drug biomarkers. The workflow for identifying gene expression based candidate drug biomarkers by the WGCNA methodology is given in **Figure 3.4**.

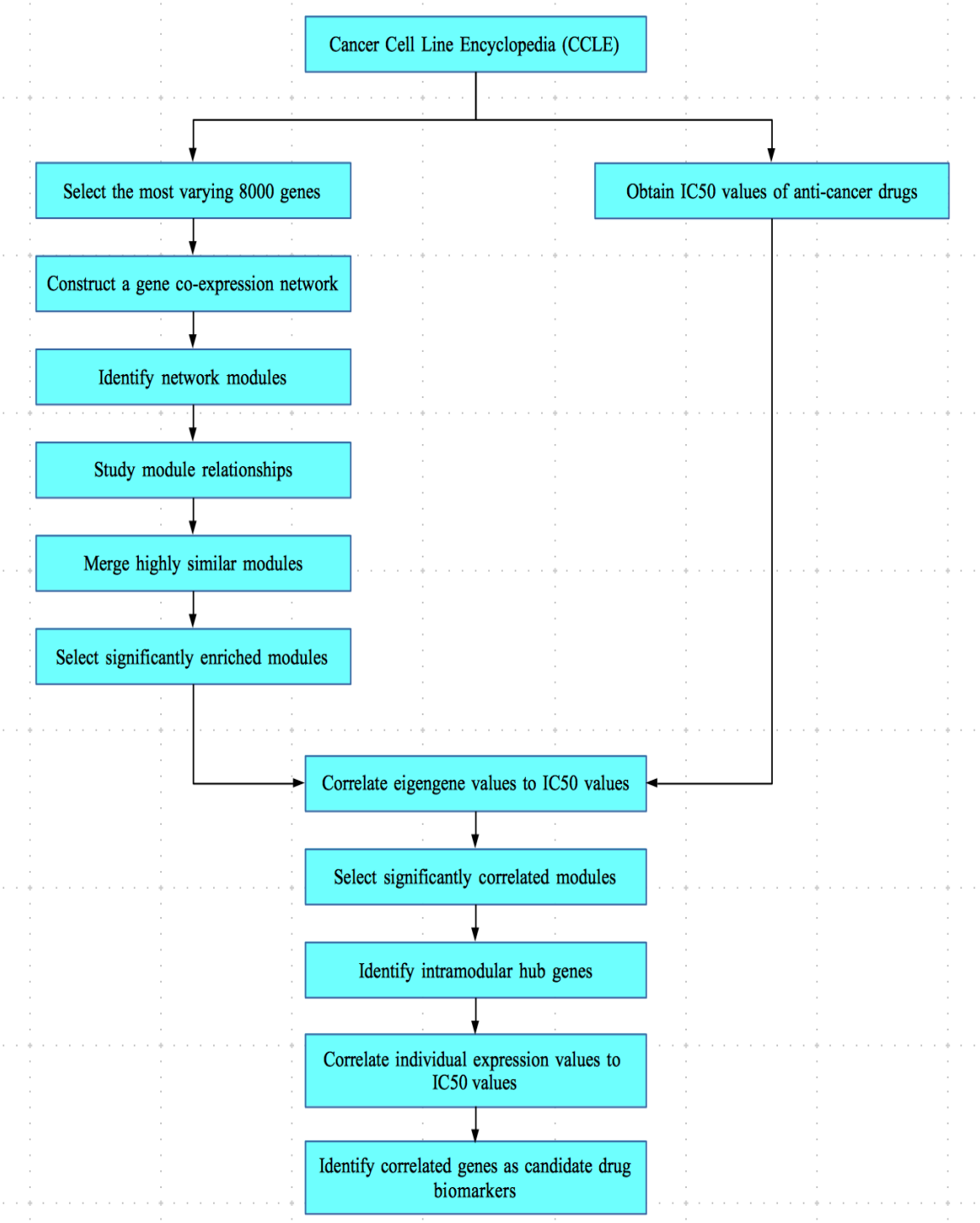

Figure 3.4: The workflow depicting identification of gene expression-based candidate drug biomarkers by utilizing the WGCNA method

3.5. In-silico validation of identified candidate drug biomarkers

The identified candidate drug biomarkers are investigated to determine how well they could predict anti-cancer drug sensitivity. For this purpose, the expression data of the MM cell lines and sensitivity profiles of anti-cancer drugs shared between the CCLE and the CGP studies are used to construct in-silico models by a machine learning method, support vector regression (SVR). However, only the candidate biomarkers (hub genes) having the highest correlation scores (top 5) within each pharmacologically significant module identified for each anti-cancer drug are selected for statistical modeling. In the first step of the model construction, the gene expression data of the selected candidate biomarkers and drug sensitivity data of the cell lines (IC50) in the CCLE study are trained both individually and in combinations. Leave-one-out cross validation (LOOCV) technique is performed as model validation technique to assess how the predictions could be generalized to independent data sets. The best model having the optimal parameters for the SVR is obtained by grid search in the range of cost: $\{2^1, 2^2, ….., 2^9\}$ and epsilon: {0.01, 0.02, 0.03,……, 0.2}. Next, the predicted sensitivity values are compared to the actual values in the CCLE by Pearson and Spearman correlations, and the predictive ability of the models is assessed by Root Mean Square Error (RMSE) measure, which measures the difference between the values predicted by a model and value actually observed. After comparison, the predictive ability of the candidate biomarkers both individually and in combinations is evaluated in the CGP data by the same SVR models generated by the CCLE data. In this step, only the probesets shared between the CCLE (the most varying 8000 genes) and the CGP expression data are selected since these two studies used different microarray platforms. If there is no shared probeset between the two expression data, then only the shared genes are selected from these expression data without considering probeset information. On the other hand, non-shared genes are discarded as the predictive ability of these genes could not be assessed in the CGP data. Thus, only the genes shared between the studies are regarded as candidate biomarkers in the study. For testing the predictive ability, both the expression profiles of the candidate biomarkers and sensitivity profiles of the anti-cancer drugs (IC50 and AUC) in the CGP study are used to predict anti-cancer drug sensitivity after logarithmic IC50 values are transformed to actual IC50 values by taking inverse logarithm and AUC values are transformed to Activity Area values by dividing AUC values to the number of drug concentrations as Haibe-Kains et al. (2013) performed. Furthermore, IC50 values are processed in three different ways, and anti-cancer drug sensitivity is predicted for the following three different cases:

1. The CGP IC50 values exceeding the fixed maximum screening concentration (8 μM) in the CCLE study are censored to 8 μM.

2. The CGP IC50 values of each drug exceeding the maximum screening concentration of each drug in the CGP study, which differs across the drugs, are censored to the maximum screening concentration of the relevant drug.

3. The IC50 values equal or greater than the maximum screening concentration of each drug in the CCLE and the CGP studies are excluded from drug sensitivity data.

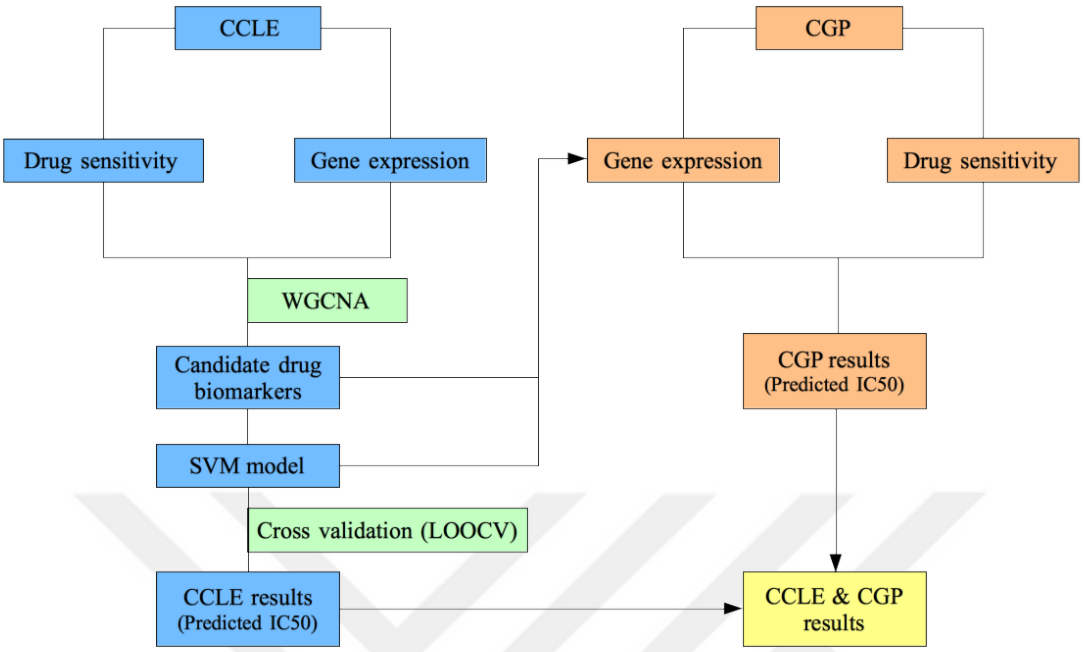

Figure 3.5: The workflow which illustrates the assessment of the predictive ability of the candidate drug biomarkers both individually and in combinations

Activity Area value does not require further processing since it can be measured in any screening concentrations. So Activity Area values are used for anti-cancer drug sensitivity prediction after the CGP AUC values are transformed to Activity Area values.

After predictions are performed, the predicted CGP sensitivity values are compared to the actual CGP values similar to the comparison of the CCLE predictions. RMSE scores are used for comparing the predictive power of the models generated with the data in the two studies. The RMSE scores are transformed to percent error in prediction by dividing the subtraction of the actual and predicted drug sensitivity values to the concentration range and multiplying the resultant with 100. In the final part, the predictions obtained by the CCLE and the CGP data are compared, and the best single candidate biomarkers and biomarker combinations are determined according to the RMSE scores. **Figure 3.5** illustrates all the steps of identification of the best candidate biomarkers both individually and in combinations.

Although anti-cancer drug sensitivity prediction was performed for the common drugs between the CCLE and the CGP studies, predictive ability of the candidate biomarkers is poor due to the inconsistency in sensitivity profiles of the drugs in the two studies (Haibe-Kains et al., 2013). So instead of using the sensitivity data of the two studies, only the CCLE sensitivity data is chosen to assess the predictive ability of the candidate biomarkers. In this way, it is aimed to show that the candidate biomarkers could predict anti-cancer drug sensitivity powerfully provided that drug sensitivity data were consistent. In addition, only

the shared cell lines between the two studies are selected to assess the performance of the candidate biomarkers for sensitivity prediction. IC50 values are also processed in three different ways as in the case of IC50 prediction performed before excluding the CGP drug sensitivity data. **Figure 3.6** depicts the final workflow in which only the CCLE drug sensitivity data is used as the drug sensitivity data.

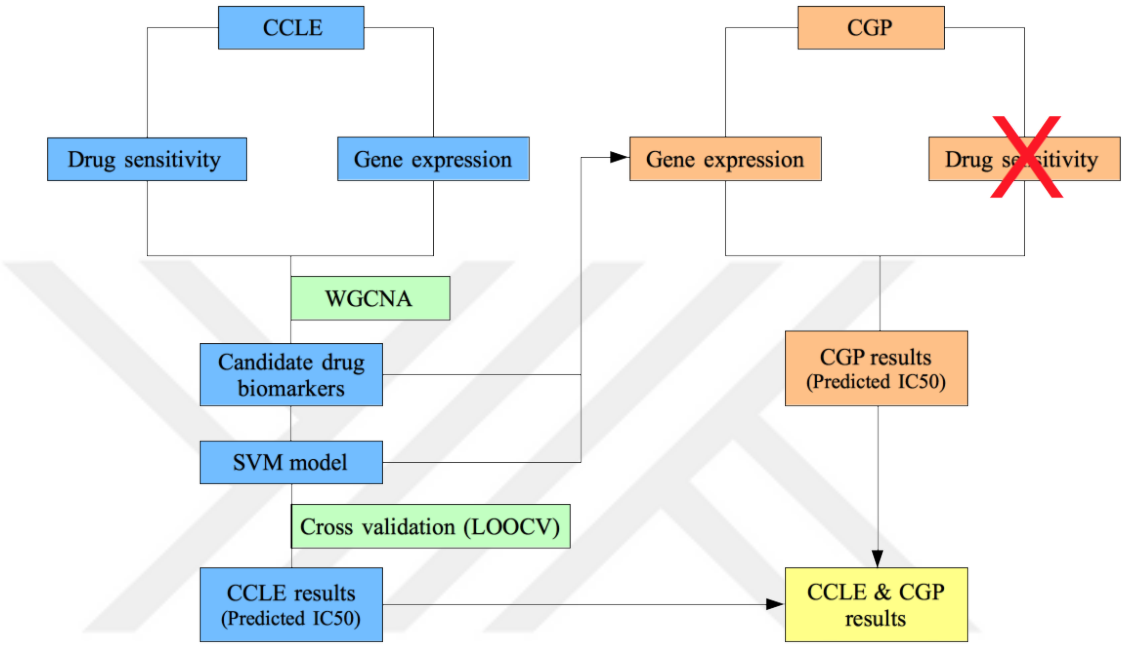

Figure 3.6: The workflow which illustrates the assessment of the predictive ability of the candidate drug biomarkers both individually and in combinations after the CGP drug sensitivity data is excluded

CHAPTER 4

RESULTS

4.1. Quality control and data pre-processing

Two large pharmacogenomics studies, the CCLE and the CGP, are used as the data sources in this study. Both of these studies include microarray gene expression profiles of the MM cell lines and sensitivity profiles of various anti-cancer drugs applied to these cell lines. It is assumed that using gene expression and drug sensitivity data available in the two studies could be significant to identify gene expression-based candidate drug biomarkers. However, one should control the quality of microarray samples to start data analysis. Thus, the quality of images is controlled before data analysis. **Figure A. 1** shows that the quality of the CCLE microarray sample images is fine. Similarly, **Figure A. 2** shows that the quality of the CGP microarray sample images is also fine. So there is no need to remove any array samples.

Once image analysis is complete, RMA normalization is performed to remove unwanted variation arising from technical artifacts. To control how well the normalization is performed, density distribution plots and boxplots of microarray samples are created before and after RMA normalization. Density distribution plots before RMA normalization produce varying density curves for each sample, whereas density distribution plots after RMA normalization produce a consensus curve for all samples (See **Figure A. 3 & Figure A. 4**). These plots imply that normalization is performed to the two microarray data successfully.

In addition, boxplots before and after RMA normalization are plotted to show that normalization sets the average expression values of the microarray samples to a common scale (See **Figure A. 5 & Figure A. 6**). Thereby, boxplots, along with density distribution plots, confirm that RMA normalization efficiently adjusts the microarray expression data for effects which arise from technical variations rather than biological variations.

The gene expression levels of the samples in the CCLE and the CGP expression data are also set to the same scale to compare the predictive ability of the identified candidate biomarkers. For this purpose, two boxplots are generated to identify how well the samples are set to the same scale as to their expression levels. **Figure A. 7** visualizes the gene expression levels of all the MM cell lines investigated in both the CCLE and the CGP studies, while **Figure A. 8** visualizes the gene expression levels of the common MM cell

lines investigated between the CCLE and the CGP studies. It is observed that the gene expression levels are leveraged in both cases. This shows that the technical variability (batch effects) is achieved to be removed.

4.2. WGCNA

4.2.1. Construction of gene co-expression network

The gene expression profile of the CCLE MM cell lines is used to construct a gene co-expression network. Although 19,178 genes are profiled, here only the most varying 8000 genes are selected for network construction. In this way, the genes which have significant biological variation are determined, and technical variation resulted from experimental protocols & procedures is minimized. In addition, construction of a network restricted to 8000 genes is computationally more efficient.

After selection of the most varying 8000 genes, the measure of similarity between the gene expression profiles is identified. This similarity measure is important since it is a measure of concordance between gene expression profiles across the experiments. Values of the similarity measure lies between -1 and 1. However, connection measures of genes can not get negative values, so the similarity matrix is transformed into an adjacency matrix by soft thresholding. Optimal β value for soft thresholding is determined by the scale-free topology criterion. A linear regression model fitting index R^2 is used to observe whether the network satisfies the scale-free topology. In the study, R^2 and connectivity values for different threshold β choices are determined (See **Table 4.1**).

β = 6 is chosen for soft thresholding since $R^2 > 0.8$ and mean connectivity is high enough. **Figure 4.1** plots scale independence and mean connectivity. It can be seen from the plot that R^2 is greater than 0.80 when β = 6, and mean connectivity value is sufficiently high (greater than 1) at this β value. Different β values ensuring the above conditions could also be selected since soft thresholding approach is highly robust for different threshold choices. However, as β values increase, mean connectivity decreases towards 0. So β values ensuring mean connectivity values sufficiently high should be preferred.

Table 4.1: Scale free fitting index R^2 and connectivity values for different β choices

Power	R^2	Mean Connectivity	Median Connectivity	Maximum Connectivity
1	7.56×10^{-5}	1431.92	1417.42	2242.69
2	0.24	400.16	381.22	906.13
3	0.60	141.88	127.16	433.81

Table 4.1 (Continued)

Power	R^2	Mean Connectivity	Median Connectivity	Maximum Connectivity
4	0.79	59.11	48.50	250.12
5	0.86	27.81	20.49	163.83
6	0.91	14.42	9.36	114.56
7	0.95	8.11	4.56	83.90
8	0.97	4.88	2.36	63.73
9	0.98	3.11	1.28	52.39
10	0.98	2.09	0.72	46.73
12	0.96	1.08	0.25	38.78
14	0.98	0.64	0.09	33.41
16	0.97	0.42	0.04	29.47
18	0.94	0.30	0.01	26.43
20	0.95	0.22	0.01	24.00

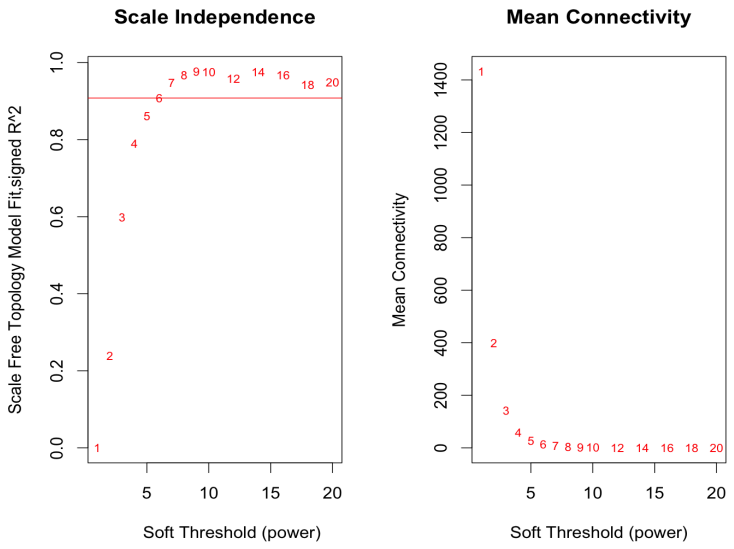

Figure 4.1: Scale independence and mean connectivity plots

In order to check whether the network is scale-free, a scale-free plot is generated for $\beta = 6$ (See **Figure 4.2**). Scale-free fitting index R^2 value is sufficiently high, so the network is pointed to be scale-free. This implies very few genes have high connectivities, while most of

the genes have low connectivities. Thus, it is shown that there must be existing hub genes in the network.

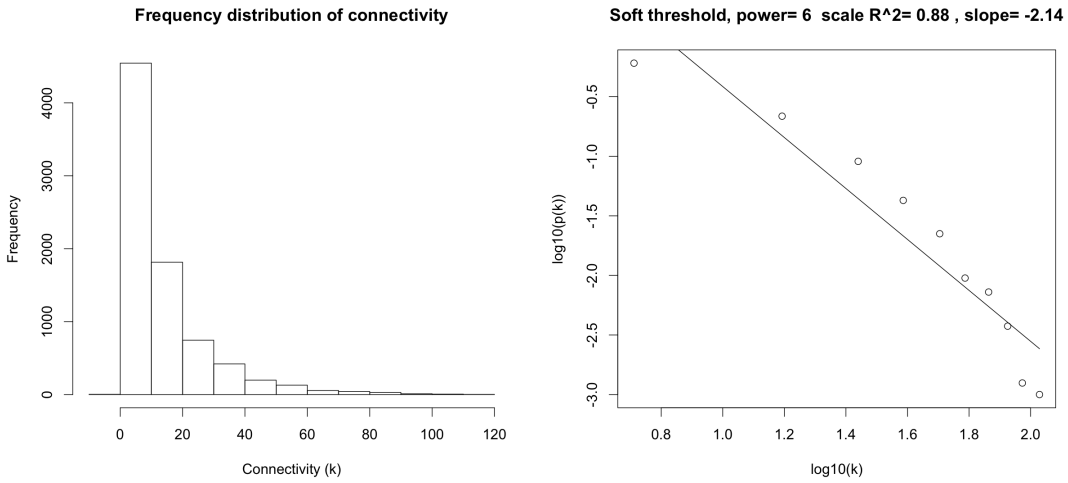

Figure 4.2: Frequency distribution and scale free plots

A network plot for the most varying 8000 genes is also generated (See **Figure 4.3**). However, only adjacencies between genes higher than 0.10 are selected for visualizing the network since Cytoscape could not visualized the 34,699,000 edges obtained in network construction when all adjacencies between genes are considered. It is not surprising that there are so many edges in the network because soft thresholding approach does not allow to lose weak connections as opposed to hard thresholding approach. Choosing 0.10 as an adjacency threshold produces a network including 5609 nodes (genes) and 67,374 edges (connections) between the genes. Detailed network statistics are also shown in **Figure 4.4**.

4.2.2. Identification of network modules

Once the network is constructed, subsets of nodes which are highly connected to each other are detected. In network terminology, these subsets of nodes are called as modules. As opposed to traditional dissimilarity measures such as Euclidean distance and Manhattan distance, topological overlap dissimilarity measure is used as an input to hierarchical clustering for module detection. Furthermore, modules are produced by dynamic tree cut approach which adaptively cuts branches of the dendrogram depending on their shapes. In this way, more coherent modules are obtained compared to the modules which are produced by a traditional dissimilarity measure and constant height cutoff value. In this study, 49 distinct modules are obtained after hierarchical clustering. **Figure 4.5** visualizes the hierarchical clustering dendrogram plot illustrating the identified modules depicted with different colors.

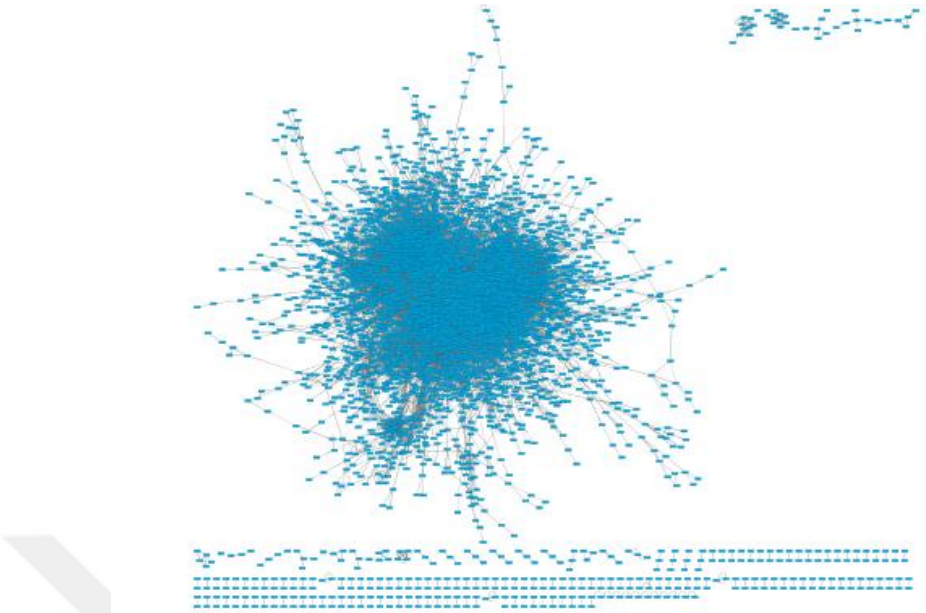

Figure 4.3: Network plot generated by the most varying 8000 genes

Each of the produced modules is represented by a module eigengene value. The module eigengene value could be regarded as the weighted average of the expression profiles of the genes inside a given module. In other words, module eigengenes are the first principal components of the modules that explain the greatest variation among genes in the modules. By using these module eigengenes, an eigengene network, which is the co-expression network of module eigengenes, is constructed to identify how well the modules are related to each other. **Figure 4.6** shows the clustering dendrogram of the module eigengenes that explains how modules are related to each other. However, the modules which have at least 70% eigengene value similarity to the other modules are merged together in order to produce more coherent and functional modules. **Figure 4.6** also shows the clustering dendrogram of the module eigengenes after similar modules are merged.

Clustering coefficient :	**0.311**	Number of nodes :	**5609**
Connected components :	**174**	Network density :	**0.004**
Network diameter :	**18**	Network heterogeneity :	**1.830**
Network radius :	**1**	Isolated nodes :	**11**
Network centralization :	**0.059**	Number of self-loops :	**52**
Shortest paths :	**27191734 (86%)**	Multi-edge node pairs :	**206**
Characteristic path length :	**4.902**	Analysis time (sec) :	**2312.399**
Avg. number of neighbors :	**23.912**		

Figure 4.4: Network statistics obtained form the network constructed by the most varying 8000 genes

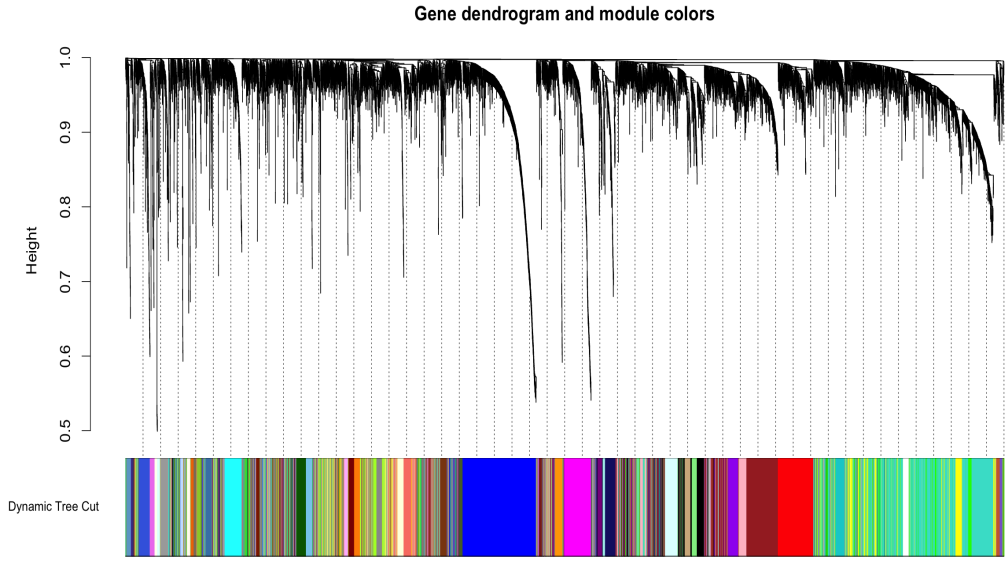

Figure 4.5: Hierarchical clustering dendrogram of the most varying 8000 genes

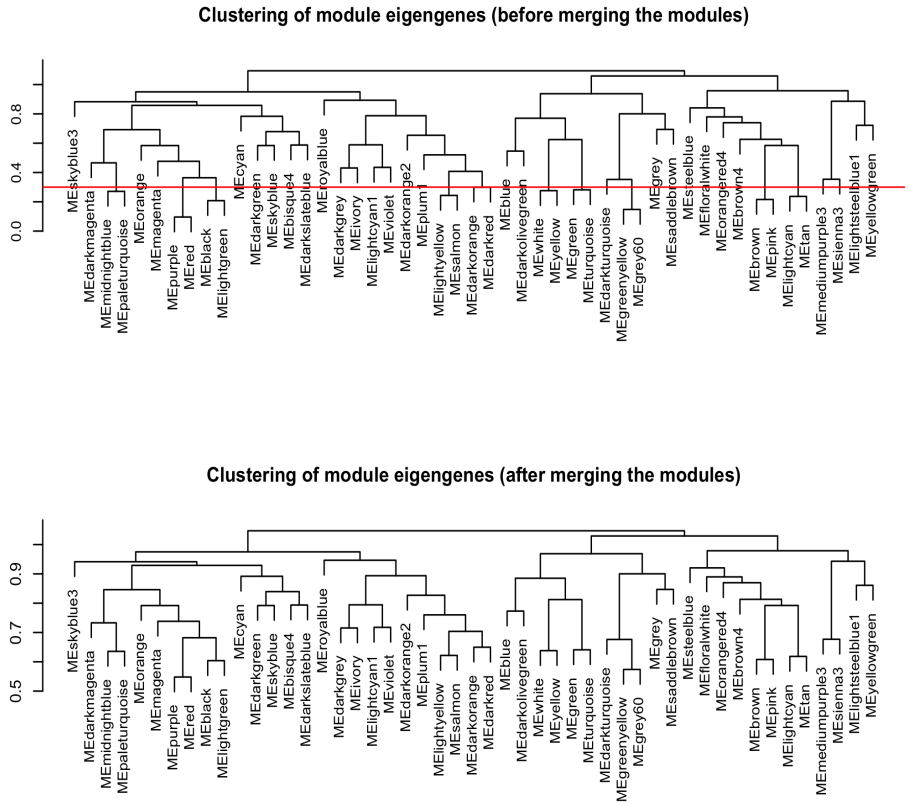

Figure 4.6: Cluster dendrogram of module eigengenes before and after merging similar modules

When similar modules are merged, the number of modules decreases from 49 to 37. Hierarchical clustering dendrogram visualizing the modules before and after merging is given in **Figure 4.7**. In addition, count of the genes in each module before and after merging similar modules is given in **Figure 4.8**.

In order to determine whether these modules are distinct and functional, module enrichment analysis is performed by the Database for Annotation, Visualization and Integrated Discovery (DAVID) tool (Huang et al., 2008). The 15 modules among the total 37 modules are significantly enriched for a known biological process, so these modules are considered to be functional and pharmacologically significant. The significantly enriched modules along with gene ontology terms and p-values are tabulated in **Table 4.2**. Furthermore, barplot of the gene ontology terms as to their significance is given in **Figure 4.9**. For the rest of the analyses, only these significantly enriched modules are selected for identifying gene expression-based candidate drug biomarkers.

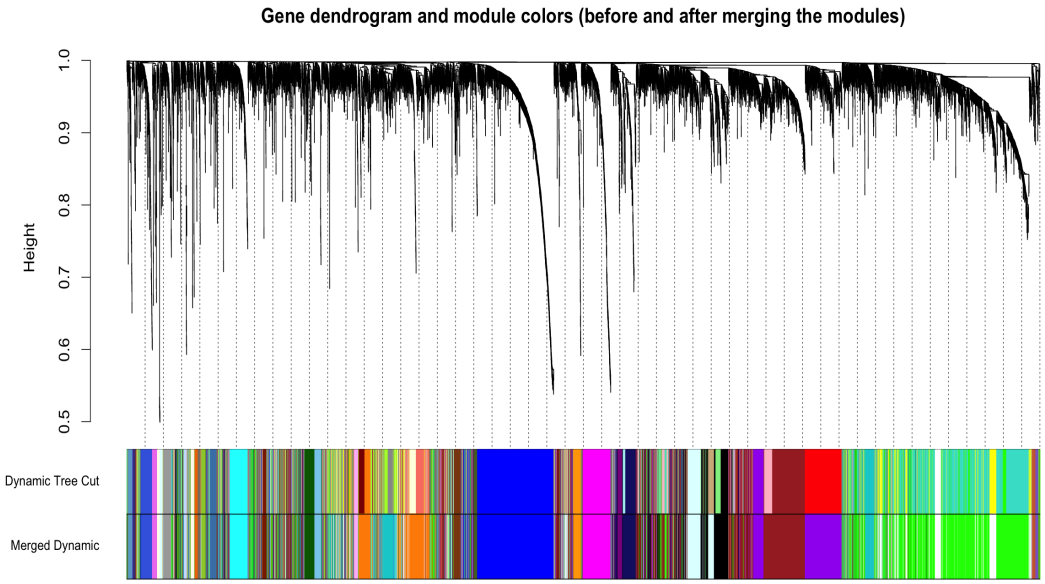

Figure 4.7: Hierarchical clustering dendrogram of the most varying 8000 genes before and after merging the similar modules

Heatmap plots for the significantly enriched modules are also created. In the plots, rows are the genes in the modules, and columns are the CCLE malignant melanoma cell lines. **Figures B. 1 – B. 15** visualize the heatmap plots of each significantly enriched module. The modules which include highly co-expressed genes should show characteristic band structures. Indeed, the heatmap plots of the significantly enriched modules exhibit these band structures, so it is clear that these modules contain highly co-expressed genes.

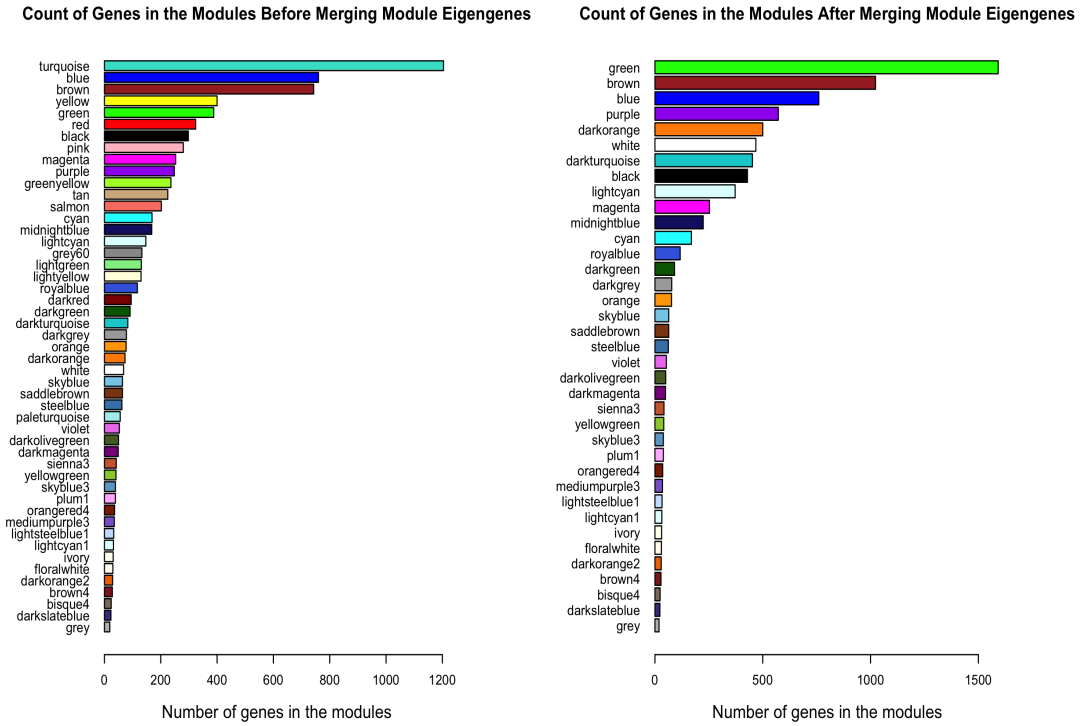

Figure 4.8: Barplot of the gene counts in the modules before and after merging the similar modules

Table 4.2: Gene ontology terms of the significantly enriched modules and their p-values

Modules	**GO Term**	**P-value**	**P-value (Benjamini)**
Saddlebrown	Regulation of transcription	1.9×10^{-35}	3.4×10^{-33}
Darkslateblue	Nucleosome assembly	1.4×10^{-19}	9.0×10^{-18}
Darkolivegreen	Calcium dependent cell adhesion	8.4×10^{-16}	1.7×10^{-13}
Darkmagenta	Inflammatory response	6.4×10^{-13}	2.7×10^{-10}
Green	Intracellular transport	1.5×10^{-12}	5.2×10^{-9}
Purple	Blood vessel development	4.3×10^{-8}	1.1×10^{-4}
Darkturquoise	DNA replication	1.6×10^{-7}	2.9×10^{-4}
Skyblue	Antigen processing and presentation	3.0×10^{-7}	1.1×10^{-4}
Magenta	Tube development	3.1×10^{-7}	4.5×10^{-4}
Darkgreen	Response to virus	4.4×10^{-7}	3.8×10^{-4}
Darkorange	Cell adhesion	9.2×10^{-7}	9.0×10^{-4}

Table 4.2 (Continued)

Modules	GO Term	P-value	P-value (Benjamini)
Brown	Regulation of cell morphogenesis	2.1×10^{-6}	6.4×10^{-3}
Blue	Cell adhesion	3.5×10^{-6}	8.2×10^{-3}
White	Cell cycle phase	5.8×10^{-6}	1.1×10^{-2}
Lightcyan	Glycoprotein biosynthesis pathway	4.8×10^{-5}	4.6×10^{-2}
Black	Regulation of cell migration	2.3×10^{-5}	4.9×10^{-2}

4.2.3. Relating modules to drug sensitivity

After identifying distinct and functional modules by hierarchical clustering, the logical next step is to relate them to external information. In this study, modules are related to drug sensitivity data of the anti-cancer drugs shared between the CCLE and the CGP studies. As a sensitivity measure, IC50 value is selected for determining related modules. In order to relate the modules, the module eigengene values of each functional modules are correlated to the CCLE IC50 values of the anti-cancer drugs, and significantly correlated modules are determined for each of the fifteen anti-cancer drugs. The CCLE data but not the CGP data is used in this step since drugs are applied to wider concentrations than the drugs profiled in the CGP. In addition, the CCLE study includes more MM cell lines screened by these drugs, and the CCLE IC50 values are not extrapolated when screening concentrations do not reach an IC50 value.

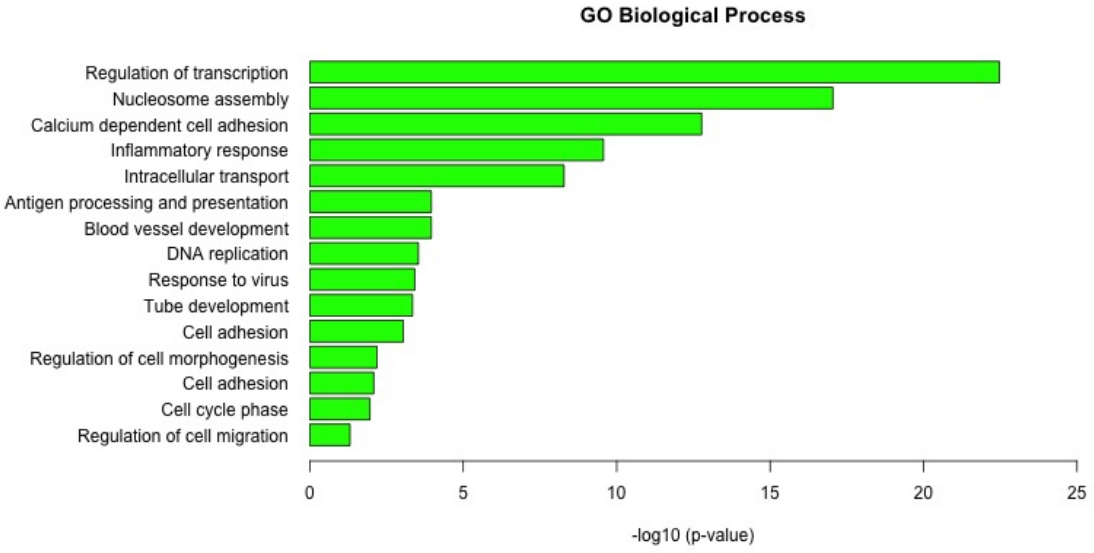

Figure 4.9: Barplot of the gene ontology terms for significantly enriched modules

Correlation analysis determines significant modules for the nine anti-cancer drugs among the fifteen anti-cancer drugs. Thus, only these drugs are selected for identifying candidate biomarkers. List of these anti-cancer drugs, along with their targets, class, and organizations, are tabulated in **Table 4.3**. In addition, Pearson correlation score and p-values of the significantly correlated modules to the IC50 values of the nine drugs are given in **Table C. 1**. The drugs which do not show significant correlations are regarded to have poor cytotoxic activity in the MM cell lines, and it is assumed that there might not be MM specific gene expression-based biomarkers for these ineffective drugs.

Table 4.3: List of anti-cancer drugs for which the WGCNA could identify candidate biomarkers

Drug Name	Target(s)	Class	Organization
AZD0530	Src, Abl/Bcr-Abl	Kinase inhibitor	AstraZeneca
AZD6244	MEK	Kinase inhibitor	AstraZeneca
Erlotinib	EGFR	Kinase inhibitor	Genentech
Lapatinib	EGFR, HER2	Kinase inhibitor	GlaxoSmithKline
PD0325901	MEK	Kinase inhibitor	Pfizer
PF2341066	c-MET/ALK	Kinase inhibitor	Pfizer
PLX4720	RAF	Kinase inhibitor	Plexxikon
Sorafenib	Raf Kinase B/C	Kinase inhibitor	Bayer
TAE684	ALK	Kinase inhibitor	Novartis

4.2.4. Identification of candidate drug biomarkers

Identification of the pharmacologically significant modules for each of the nine anti-cancer drugs is significant since there might be key gene drivers in these modules having the ability to predict anti-cancer drug sensitivity. To find these key drivers, attention is focused on the hub genes which have the greatest connectivity to the other genes in a given module. Now that the hub genes are the central players in the modules, they may be regarded as representative genes of the modules they are located. Thus, only using a few representative genes for each module is sufficient to explain the biological activity of the modules. In this way, one can reduce the number of genes significantly and do not need to consider hundreds or thousands of genes in the analysis. This greatly alleviates multiple comparison problems. From this aspect, the approach is considered to be highly robust. In this analysis, the genes which have at least 0.70 intramodular connectivity values are regarded to be hub genes. In addition, only the hub genes whose individual gene expression values show significant correlations to the IC50 values (P-value < 0.05) are selected as candidate biomarkers. All of the identified hub genes for each nine anti-cancer drug are given in **Table C. 2**. **Figure 4.10**

also visualizes the distribution of the identified candidate biomarkers for the nine drugs as to the modules they belong.

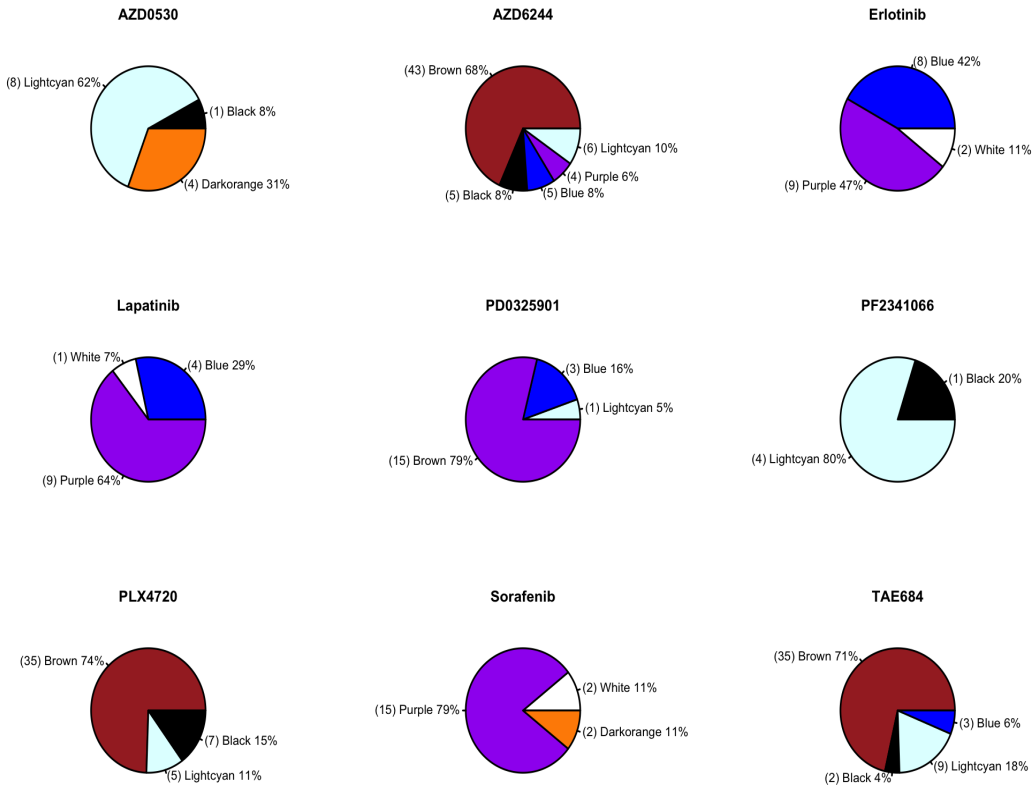

Figure 4.10: Count of identified candidate biomarkers per modules for each of the nine anti-cancer drugs

4.3. Predictive ability of candidate biomarkers

The WGCNA method identifies gene expression-based candidate biomarkers for the nine anti-cancer drugs among the fifteen anti-cancer drugs shared between the CCLE and the CGP studies. However, predictive ability of these candidate biomarkers for drug sensitivity can not be determined by the WGCNA method. Thus, a machine learning algorithm, the SVR, is applied to determine how well the candidate biomarkers predict drug sensitivity. In the study, predictive ability is investigated both individually and in combinations for drug sensitivity. The reason why the SVR is preferred is that it is effective and superior in extracting the non-linear relations between gene expression and drug sensitivity. In addition, it can perform powerful predictions in small sample sizes.

4.3.1. Predictive ability of single candidate biomarkers

Predictive ability of the single candidate biomarkers is assessed by using gene expression and drug sensitivity data. In the study, IC50 and Activity Area are used as the sensitivity measures since both the CCLE and the CGP studies report drug sensitivities with these two common measures. In addition, at most three genes having the greatest intramodular connectivity and correlation scores among each pharmacologically significant module for the nine drugs are selected for assessing the predictive ability of the candidate biomarkers for drug sensitivity since there are tens of intramodular hub genes in some of the modules.

In the first step, the CCLE gene expression data of the single candidate biomarkers is trained along with the CCLE IC50 values of the nine drugs in a non-linear SVR model. Next, the CCLE IC50 values are predicted for each drug by the generated model. The RMSE values determined for the best performing single candidate biomarkers among the top three single candidate biomarkers selected for each drug are inspected and tabulated in **Table 4.4**. These top performing candidate biomarkers are regarded as having the highest predictive ability for IC50 prediction (See **Table D. 1**). Then, the predicted CCLE IC50 values are correlated to the actual CCLE IC50 values. The results show that only Pearson correlation score of RNF125 gene determined for PD0325901 is significant, and only Spearman correlation score of NAV3 gene determined for Lapatinib is significant. **Table C. 3** also tabulates the correlation scores of all the top performing single candidate biomarkers for the nine drugs as the CCLE data is used to predict IC50.

Table 4.4: The RMSE values determined for the best performing single candidate biomarkers in the CCLE IC50 prediction

Gene Name	Module Name	Drug Name	RMSE (CCLE)	Error in Prediction (%)
TSPAN13	Darkorange	AZD0530	1.41	17.63
SLC23A2	Brown	AZD6244	3.33	41.63
C8orf4	Blue	Erlotinib	1.18	14.75
PAOX	Blue	Lapatinib	1.54	19.25
RNF125	Lightcyan	PD0325901	2.67	33.38
CLCN7	Lightcyan	PF2341066	1.25	15.63
SAMM50	Brown	PLX4720	3.09	38.63
MPRIP	Purple	Sorafenib	0.59	7.38
MFSD12	Lightcyan	TAE684	2.07	25.88

In the following step, IC50 prediction is performed by excluding the censored CCLE IC50 values from the CCLE drug sensitivity data. However, only AZD6244, PD0325901, and

PLX4720 remain to have sufficient IC50 values, so only these drugs are considered in prediction. **Table 4.5** tabulates the top performing single candidate biomarkers for these three drugs. This table clearly indicates error in prediction lowers for AZD6244, PD0325901, and TAE684. This shows that excluding censored CCLE IC50 values from the CCLE drug sensitivity data improves the predictive ability of the single candidate biomarkers determined for the three drugs.

Table 4.5: The RMSE values determined for the best performing single candidate biomarkers in the CCLE IC50 prediction after removing the censored IC50 values

Gene Name	Module Name	Drug Name	RMSE (CCLE)	Error in Prediction (%)
DAAM1	Lightcyan	AZD6244	0.75	14.88
RNF125	Lightcyan	PD0325901	1.50	21.99
NTF3	Black	PLX4720	1.58	21.76

The actual CCLE IC50 values are also correlated to the predicted CCLE IC50 values in order to grasp which single candidate biomarkers are good at prediction (See **Table C. 4**). The results show that both Pearson and Spearman correlation scores of DAAM1 gene determined for AZD6244 is significant and only Pearson correlation score of NTF3 gene determined for PLX4720 is significant. However, neither Pearson nor Spearman correlation of RNF125 gene determined for PD0325901 is significant.

After IC50 prediction, Activity Area is predicted for the shared nine drugs between the CCLE and the CGP studies. Activity Area has an advantage over IC50. It generally gives more accurate measurements than IC50 because Activity Area can always be measured irrespective of drug screening concentration. In IC50 measurement, however, drug screening concentration is significant. Indeed, the CCLE study screened all the drugs till 8 micro molars (μM), and IC50 could not be generated in the given concentration interval for a portion of the drugs. So the IC50 values are censored to 8 μM for the drugs which do not reach an IC50 value till the maximum concentration value. These censored values affect negatively the performance of the models in prediction as the models confuse in predicting IC50 values due to the repeating censored values. Thereby, Activity Area is expected to give more reliable results than IC50.

Similar to IC50 prediction, Activity Area prediction is performed for the drugs after the single candidate biomarkers are trained by the CCLE data. The RMSE values of the best performing single candidate biomarkers among the top three candidate single biomarkers selected for each drug are inspected and tabulated in **Table 4.6**. This table shows that Activity Area prediction error is lower than IC50 prediction error. So it can be considered that Activity Area is more powerful in assessing predictive ability of the single candidate biomarkers than IC50.

The actual CCLE Activity Area values are also correlated to the predicted CCLE Activity Area values. The top performing candidate biomarker of each drug is selected to investigate

correlation scores. **Table C. 5** tabulates the correlation scores as the CCLE Activity Area values are used to predict drug sensitivity. The results show that both Pearson and Spearman correlations of relevant single candidate biomarkers for the drugs except PF2341066 are significant. However, only Pearson correlation score of CLCN7 is significant for PF2341066. These results strongly imply that Activity Area prediction is performed more powerfully than IC50 prediction. So Activity Area could be regarded as a better indicator of drug sensitivity in this case.

Table 4.6: The RMSE values determined for the best performing single candidate biomarkers in the CCLE Activity Area prediction

Gene Name	Module Name	Drug Name	RMSE (CCLE)	Error in Prediction (%)
FLJ42627	Lightcyan	AZD0530	0.054	5.43
PAOX	Blue	AZD6244	0.12	11.72
LRP5	Blue	Erlotinib	0.029	2.89
NAV3	Purple	Lapatinib	0.050	5.02
PAOX	Blue	PD0325901	0.16	15.67
CLCN7	Lightcyan	PF2341066	0.036	3.60
APOD	Lightcyan	PLX4720	0.098	9.82
ETHE1	Purple	Sorafenib	0.038	3.82
PROS1	Brown	TAE684	0.070	7.05

Predictive ability of the single candidate biomarkers is assessed in an independent CGP study after training with the CCLE data. Before the assessment, the common genes between the most varying 8000 genes selected from the CCLE gene expression data and the genes profiled in the CGP study are determined as to the probeset ID of the genes. So only these common genes are selected to assess the predictive ability of the single candidate biomarkers. In some cases, however, genes are selected as to the gene names without inspecting probeset IDs when there is no shared probeset between the two gene expression data. In addition, the gene expression levels of the samples profiled in the CCLE and the CGP studies are set to common scale because different microarray platforms are used to measure gene expression profiles of the MM cell lines. After these adjustments, the CGP drug sensitivity is predicted in two different ways;

1. Both the CGP gene expression data of all the MM cell lines and the CGP drug sensitivity data of the nine common drugs are considered in drug sensitivity prediction.

2. Only the CGP gene expression data of the MM cell lines shared between the CCLE and the CGP studies are considered in drug sensitivity prediction.

4.3.1.1. Anti-cancer drug sensitivity prediction using the CGP gene expression data of all the MM cell lines and the CGP drug sensitivity data of the nine common drugs

Predictive ability of the best performing single candidate biomarkers determined by the CCLE data is tested in the CGP data by using the gene expression profiles of all the MM cell lines and the sensitivity profiles of the nine common drugs in the CGP study. For this purpose, the CGP IC50 and Activity Area values are separately used in drug sensitivity prediction.

4.3.1.1.1. IC50 prediction results

The CGP IC50 values are used in drug sensitivity prediction after processing the values in three different ways:

1. The extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs (8 μM) profiled in the CCLE study.
2. The extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs (differs among drugs) profiled in the CGP study.
3. The extrapolated CGP IC50 values are excluded from the CGP drug sensitivity data.

Censoring the extrapolated CGP IC50 values to the maximum screening concentration of the drugs profiled in the CCLE study

The extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs (8 μM) profiled in the CCLE study. It is observed that most of the cell lines screened against the nine drugs could not reach their actual IC50 values, so they are extrapolated. For example, none of the cell lines screened against AZD0530, Erlotinib, and PF2341066 has IC50 values lower than 8 μM, and some of the cell lines screened against AZD6244, Lapatinib, PD0325901, PLX4720, Sorafenib, and TAE684 have IC50 values lower than 8 μM. Only the IC50 values of PD0325901 are completely lower than 8 μM. **Table 4.7** tabulates the count of cell lines having IC50 values lower than 8 μM for each drug in the CGP study.

Since there are only three drugs (AZD6244, PD0325901, PLX4720) having sufficiently large number of IC50 values, they are selected to assess the predictive ability of the single candidate biomarkers. Otherwise, the CGP IC50 values of the rest of the drugs would be transformed to majorly repeating numbers that might impede the assessment of the predictive power. For assessment, the RMSE values determined for the best performing single candidate biomarkers of the three drugs determined by the CCLE data are inspected

and tabulated in **Table 4.8**. When the error in prediction obtained from the CGP data is compared to that of the CCLE data, the predictive ability of the single candidate biomarkers seems to be amply lower for the CGP case. This indicates that censoring the CGP IC50 values to the maximum screening concentration of the drugs profiled in the CCLE study does not improve the predictions.

Table 4.7: Count of all the malignant melanoma cell lines profiled in the CCLE study

Drug Name	Count of Cell Lines	Count of Cell Lines (IC50 values lower than 8 µM)
AZD0530	14	-
AZD6244	31	16
Erlotinib	12	-
Lapatinib	12	-
PD0325901	34	34
PF2341066	14	-
PLX4720	36	18
Sorafenib	13	1
TAE684	14	5

Table C. 6 also shows that neither Pearson nor Spearman correlation of any three drugs is significant.

Table 4.8: The RMSE values determined for the best performing single candidate biomarkers when the extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs in the CCLE study

Gene Name	Module Name	Drug Name	RMSE (CGP)	Error in Prediction (%)
DAAM1	Lightcyan	AZD6244	4.40	55.00
PAOX	Blue	PD0325901	0.94	11.75
BAMBI	Lightcyan	PLX4720	3.44	43.00

Predictive ability of the single candidate biomarkers is poor as expected since the IC50 values are highly inconsistent between the CCLE and the CGP studies (Haibe-Kains et al., 2013). Indeed, as the CCLE IC50 values are correlated to the CGP IC50 values after censoring and common malignant melanoma cell lines in the two studies are considered, correlation scores and R^2 are determined to be poor (See **Figure 4.11 a**). This inconsistency between the two studies impairs the assessment of the predictive performance significantly.

Censoring the extrapolated CGP IC50 values to the maximum screening concentration of the drugs profiled in the CGP study

In the first approach, the extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs (8 μM) profiled in the CCLE study, so that sensitivity profiles of the drugs in the studies could be comparable in similar scale. However, the inconsistency in drug sensitivity data between the studies in this scale impedes adversely the predictive performance of the models. Thus, it is assumed that censoring the extrapolated CGP IC50 values of each drug to the maximum screening concentrations of the relevant drugs profiled in the CGP study. In this way, the predictive performances might be determined more reliably. For this reason, all the extrapolated CGP IC50 values of each drug are censored to the maximum concentrations of the relevant drugs as shown in **Table 4.9**.

Table 4.9: Maximum screening concentration of drugs along with count of malignant melanoma cell lines screened against these drugs in the CGP study

Drug Name	Maximum Screening Concentration (μM)	Count of Cell Lines	Count of Cell Lines (IC50 values lower than max. screening conc.)
AZD0530	2	14	-
AZD6244	4	31	17
Erlotinib	2	12	-
Lapatinib	2	12	-
PD0325901	0.25	34	31
PF2341066	2	14	-
PLX4720	10	36	22
Sorafenib	4	13	3
TAE684	2	14	3

All the malignant melanoma cell lines screened against AZD0530, Erlotinib, Lapatinib, and PF2341066 have IC50 values greater than the maximum screening concentrations of each corresponding drug. In addition, the IC50 values of the cell lines screened against Sorafenib and TAE684 are mostly larger than the maximum screening concentration of the drugs. So these drugs are not considered in IC50 prediction. Thereby, only the IC50 values of AZD6244, PD0325901, and PLX4720 are predicted.

Since there are only three anti-cancer drugs (AZD6244, PD0325901, PLX4720) having sufficiently large numbers of measured IC50 values after selecting IC50 values lower than the maximum screening concentration of the drugs profiled in the CGP study, these drugs are selected to assess the predictive ability of the single candidate biomarkers.

The RMSE values determined for the best performing single candidate biomarkers of the three drugs are inspected and tabulated in **Table 4.10**. The table shows that predictive ability of the single candidate biomarkers is extremely poor when the CGP IC50 values of the drugs are censored to the maximum screening concentrations of these drugs profiled in the CGP study. This suggests that all the extrapolated or censored IC50 values should be removed from the CGP drug sensitivity data in order to perform reliable IC50 predictions.

Table 4.10: The RMSE values determined for the best performing single candidate biomarkers when the extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs in the CGP study

Gene Name	Module Name	Drug Name	RMSE (CGP)	Error in Prediction (%)
PAOX	Blue	AZD6244	1.86	23.25
APOE	Brown	PD0325901	0.44	176.00
BAMBI	Lightcyan	PLX4720	4.57	45.70

Table C. 7 shows neither Pearson nor Spearman correlation of any three drugs is again significant.

Predictive ability of the single candidate biomarkers is poor as in the case of censoring the IC50 values to the maximum screening concentration of the drugs profiled in the CCLE study since the IC50 values are highly inconsistent between the studies. When the CCLE IC50 values are correlated to the CGP IC50 values after censoring and common malignant melanoma cell lines between the two studies are considered, correlation scores and R^2 are determined to be poor (See **Figure 4.11 b**).

Removing the extrapolated CGP IC50 values from the CGP drug sensitivity data

Instead of using extrapolated or censored values in IC50 prediction, it is assumed that using only measured experimental IC50 values of the MM cell lines might improve the predictive performance of the models. Since there is a reasonable amount of cell lines having IC50 values lower than the maximum screening concentrations of AZD6244, PD0325901, and PLX4720, IC50 prediction is performed only for these three drugs after the extrapolated or the censored IC50 values are removed from both the CCLE and the CGP drug sensitivity data. **Table 4.11** tabulates the count of malignant melanoma cell lines left after removing the extrapolated or the censored IC50 values. Next, IC50 prediction is performed after training the selected CCLE IC50 values, along with the CCLE expression data of the single candidate biomarkers, and testing the selected CGP IC50 values along with the CGP expression data of the single candidate biomarkers.

Table 4.11: Count of the malignant melanoma cell lines remaining after all the extrapolated and censored IC50 values are removed from drug sensitivity data

Drug Name	Count of Cell Lines (CCLE)	Count of Cell Lines (CGP)
AZD0530	5	-
AZD6244	26	17
Erlotinib	4	-
Lapatinib	8	-
PD0325901	33	31
PF2341066	14	-
PLX4720	24	22
Sorafenib	10	3
TAE684	22	3

The RMSE values of the best performing single candidate biomarkers of the three drugs when all the extrapolated and the censored IC50 values are removed from the drug sensitivity data are inspected and tabulated in **Table 4.12**. The table shows that predictive ability of the single candidate biomarkers is still extremely poor. This implies that excluding non-measured IC50 values from the drug sensitivity data is not effective since the inconsistency between the two drug sensitivity data again precludes accurate assessment of predictive power.

Table C. 8 shows that neither Pearson nor Spearman correlation of any three drugs is significant. The CCLE and the CGP IC50 values of the common malignant melanoma cell lines between the studies are also correlated. The results show that only Spearman correlation of PLX4720 shows a significant increase among the three drugs (See **Figure 4.11 c**). However, there is still inconsistency in IC50 values between the studies. This shows that it is not possible to perform powerful predictions when IC50 is used as drug sensitivity measure.

4.3.1.1.2. Activitiy area prediction results

Now that the CCLE and the CGP IC50 values are highly inconsistent, IC50 is regarded to be not a reliable estimator to assess the predictive ability of the single candidate biomarkers. So Activity Area is decided to be used as an alternative to IC50 in drug sensitivity prediction.

Table 4.12: The RMSE values determined for the best performing single candidate biomarkers when both the extrapolated and censored IC50 values in the drug sensitivity data are excluded

Gene Name	Module Name	Drug Name	RMSE (CGP)	Error in Prediction (%)
PAOX	Blue	AZD6244	0.55	28.95
SCUBE2	Brown	PD0325901	0.27	22.88
BAMBI	Lightcyan	PLX4720	0.98	16.39

As opposed to IC50, Activity Area is neither censored nor extrapolated in the CCLE and the CGP studies. Since measuring the area of a dose-response curve is possible in all screening concentrations, Activity Area is obtained for all the drugs profiled in the studies. So it has an advantage over IC50 in this manner. In addition, the absence of censored or extrapolated values give the opportunity to perform more efficient predictions as predictive ability of a model is more powerful when non-repeating and measured values exist in the data.

After the CGP Activity Area values (AUC is transformed to Activity Area) of the drugs are tested along with the CGP gene expression data of the single candidate biomarkers with the same model obtained by the CCLE data, the CGP Activity Area values are predicted for all the nine drugs profiled both in the CCLE and the CGP studies. The RMSE values determined for the best performing single candidate biomarkers of all the nine drugs are inspected and tabulated in **Table 4.13**. The table shows that predictive ability of the single candidate biomarkers is extremely poor as in the case of IC50 prediction. This again shows how inconsistency between the two drug sensitivity data ruins accurate assessment of predictive performance (See **Figure 4.11 d**).

Table 4.13: The RMSE values determined for the best performing single candidate biomarkers in the CGP Activity Area prediction

Gene Name	Module Name	Drug Name	RMSE (CGP)	Error in Prediction (%)
FLJ42627	Lightcyan	AZD0530	0.90	90.11
PAOX	Blue	AZD6244	0.54	54.00
LRP5	Blue	Erlotinib	0.93	93.17
NAV3	Purple	Lapatinib	0.98	98.02
PAOX	Blue	PD0325901	0.28	28.42
CLCN7	Lightcyan	PF2341066	0.92	92.12
APOD	Lightcyan	PLX4720	0.67	67.65
ETHE1	Purple	Sorafenib	0.88	88.41

Table 4.13 (Continued)

Gene Name	Module Name	Drug Name	RMSE (CGP)	Error in Prediction (%)
LRP5	Blue	Erlotinib	0.93	93.17
PROS1	Brown	TAE684	0.83	82.78

Correlation scores of any genes determined for the relevant drugs are, however, not significant (**Table C. 9**).

4.3.1.2. Anti-cancer drug sensitivity prediction using only the CGP gene expression data of the shared MM cell lines between the CCLE and the CGP studies

Predictive ability of the single candidate biomarkers is assessed by constructing models that train the CCLE gene expression data of the MM cell lines, along with the CCLE drug sensitivity data of the nine drugs. Then, these models are tested by using the CGP gene expression data of the MM cell lines, along with the CGP sensitivity data of the same drugs. However, there could not be identified any single candidate biomarkers having a potential to predict drug sensitivity of any drugs. This poor performance of the models is as a result of inconsistency in both IC50 and Activity Area values between the CCLE and the CGP studies. Thus, the CGP drug sensitivity data is no longer considered in drug sensitivity prediction, while the CCLE drug sensitivity data is held for predictions. The reason why the CCLE drug sensitivity data is held is that it includes more reliable sensitivity data due to both its screening drugs against cancer cell lines in wider concentration ranges and higher sensitivity of the cytotoxicity assay it used for measuring sensitivity values (Haverty et al., 2016). Thereby, it is decided to use only the CGP gene expression data, which is highly concordant to the CCLE gene expression data, in drug sensitivity prediction. In addition, instead of using all the MM cell lines, only the common MM cell lines between the studies are considered in drug sensitivity prediction. In this way, it is expected that predictive ability of the single candidate biomarkers might be assessed more accurately, reliably, and powerfully. In this scenario, the CCLE IC50 values are used in two different ways after the shared malignant melanoma cell lines between the studies are selected:

1. The censored CCLE IC50 values are included in the CCLE drug sensitivity data
2. The censored CCLE IC50 values are removed from the CCLE drug sensitivity data

On the other hand, the CCLE Activity Area values are used with no modification since the CCLE Activity Area values are neither censored nor extrapolated.

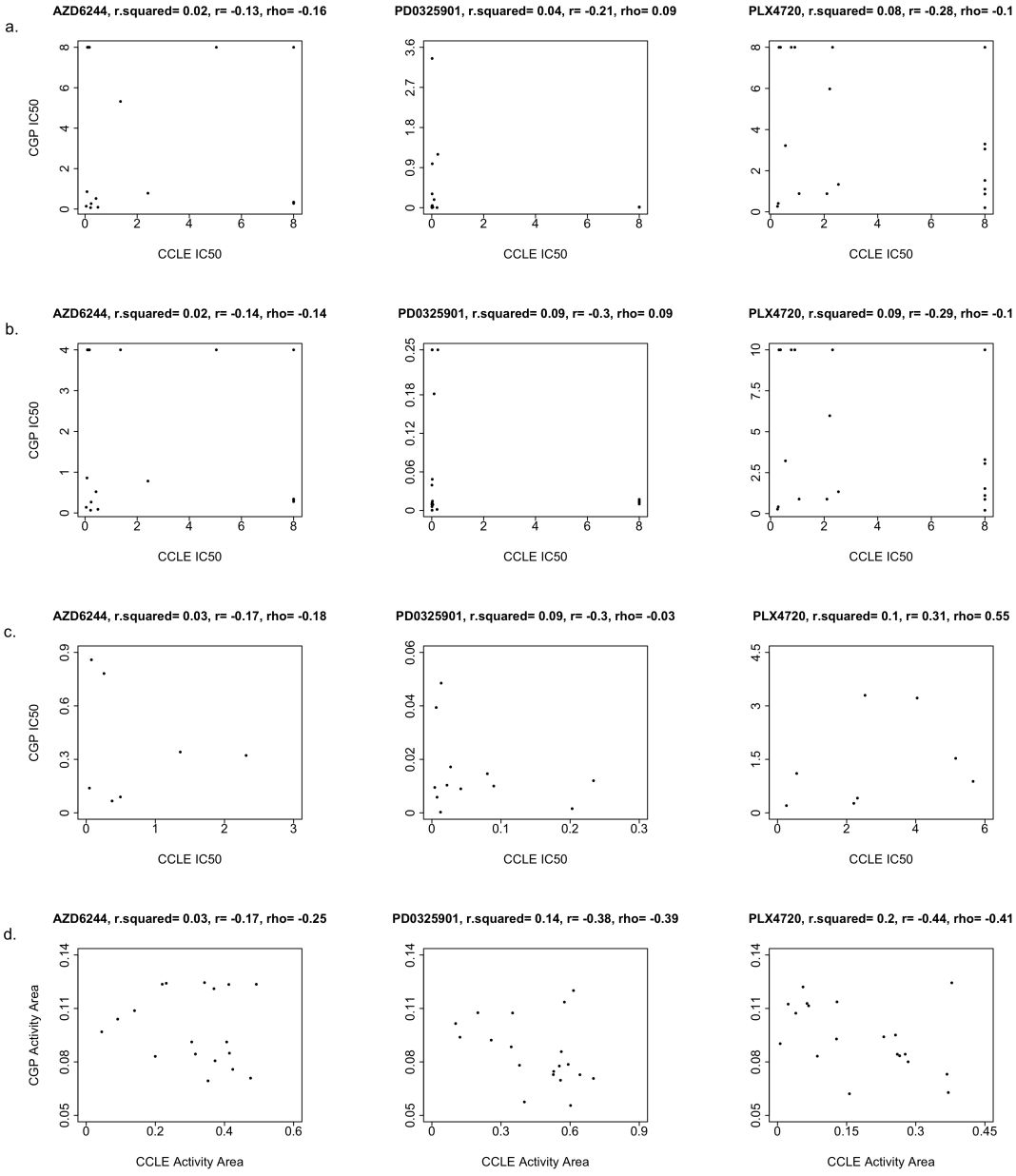

Figure 4.11: The scatterplots which illustrate the poor correlation of IC50 and Activity Area values. (a) Correlation scores for IC50 values after censoring the CGP IC50 values to the maximum screening concentration of the drugs profiled in the CCLE study. (b) Correlation scores for IC50 values after censoring the CGP IC50 values to the maximum screening concentration of the relevant drugs in the CGP study. (c) Correlation scores for IC50 values after removing all the extrapolated or censored IC50 values from the drug sensitivity data. (d) Correlation scores for Activity Area values.

4.3.1.2.1. IC50 prediction results

The CCLE IC50 is predicted by including or removing the censored IC50 values in order to show that how the censored IC50 values affect the predictive power. It is expected that more powerful predictions might be performed after removing the censored IC50 values since repeating values might lead to inaccurate predictions. On the other hand, removing censoreddata decreases the sample size which might lower the efficiency of the models in predicting drug sensitivity. However, small sample size is assumed to be more favorable in IC50 prediction than having censored values in the drug sensitivity data.

Including the censored IC50 values in the CCLE drug sensitivity data

Excluding the CGP drug sensitivity data in IC50 prediction certainly contribute to a more reliable assessment of predictive performance. However, it is still uncertain whether using only the CCLE drug sensitivity data improves the performance noteworthily. So the CCLE IC50 values are used with no modification in the first step.

The RMSE values determined for the best performing single candidate biomarkers of all the nine drugs obtained when only the CCLE drug sensitivity data is used as drug sensitivity data are inspected and tabulated in **Table 4.14**. The table shows that the RMSE values determined by using only the CCLE data are quite similar to the RMSE values determined by feeding the CGP gene expression data. Similarity in predictive performance suggests that excluding the CGP drug sensitivity data from predictions is an effective approach for assessment of predictive power. The similarity also suggests that gene expression data of the common MM cell lines in independent CGP study could be used powerfully to test the predictive ability of the candidate biomarkers. However, the predictions are still poor for the majority of the drugs since the censored values are not removed from the drug sensitivity data. So removing the censored CCLE IC50 values is assumed to improve predictions further and give more reliable estimate of the predictive power..

Table C. 10 shows that both Pearson and Spearman correlation scores of genes determined for AZD6244, Lapatinib, PD0325901, PLX4720, and TAE684 are significant. On the other hand, only Pearson correlation score of BAMBI gene determined for AZD0530 is significant.

Excluding the censored CCLE IC50 values from the CCLE drug sensitivity data

The CCLE drug sensitivity data contains many censored IC50 values for the drugs screened against the MM cell lines. This is not surprising since none of the drugs profiled in the CCLE study has been shown to be effective in the MM treatment. So they mostly do not inhibit the growth of the MM cell lines sufficiently. For this reason, the majority of the MM cell lines could not reach their IC50 point in the screened concentration interval. IC50 values of these MM cell lines are censored to the maximum screening concentration in the

CCLE study, but the IC50 values are not accurate. These inaccurate IC50 values lead to poor performance of the models in IC50 prediction. Therefore, all the censored IC50 values are removed from the CCLE drug sensitivity data.

Table 4.14: The RMSE values determined for the best performing single candidate biomarkers in the CCLE IC50 prediction

Gene Name	Module Name	Drug Name	RMSE (CGP)	Error in Prediction (%)
BAMBI	Lightcyan	AZD0530	1.19	14.88
RAB38	Brown	AZD6244	2.47	30.88
FAM172A	White	Erlotinib	1.38	17.25
NAV3	Purple	Lapatinib	1.80	22.50
RNF125	Lightcyan	PD0325901	2.57	32.13
BAMBI	Lightcyan	PF2341066	1.27	15.88
BAMBI	Lightcyan	PLX4720	3.03	37.88
HIVEP3	Darkorange	Sorafenib	0.77	9.63
PROS1	Brown	TAE684	1.69	21.13

After excluding the censored data, there leaves 14 MM cell lines for only three drugs, AZD6244, PD0325901, and TAE684. The rest of the drugs has either totally or mostly censored IC50 values. Thus, IC50 predictions are performed for only these drugs. The RMSE values determined for the best performing single candidate biomarkers of these three drugs are tabulated in **Table 4.15**. The table shows that excluding the censored data, indeed, improves the predictive power. The RMSE values of the three drugs are lower than the RMSE values determined for the case on which the censored CCLE IC50 values are included in predictions. This shows that censored values should be removed for accurate assessment of predictive performance.

Table C. 11 shows Pearson correlation score of only the genes determined for PD0325901 and TAE684 is significant, while only Spearman correlation score of the gene determined for AZD6244 is significant.

Although predictive ability of the single candidate biomarkers is high when the CCLE data is trained, it is poor for most of the single candidate biomarkers when the CGP gene expression data is used to test the model. Since the gene expression data of the common MM cell lines are concordant between the studies, the poor performance is likely resulted from overfitting. The reason why the CCLE data overfits can be explained by small sample size. In model selection step, the LOOCV is used to tune hyperparameters, and the best model is selected. However, because of small sample size, the outputs of the LOOCV are highly correlated with each other. The mean of these highly correlated quantities has higher

variance, so the test error estimate resulting from the LOOCV tends to have higher variance. This explains why the predicted CCLE IC50 values may not correlate significantly to the actual CCLE IC50 values.

Table 4.15: The RMSE values determined for the best performing single candidate biomarkers in the CCLE IC50 prediction after excluding the censored IC50 values

Gene Name	Module Name	Drug Name	RMSE (CGP)	Error in Prediction (%)
BAMBI	Lightcyan	AZD6244	0.78	32.37
APOE	Brown	PD0325901	0.25	25.00
PROS1	Brown	TAE684	1.23	17.01

4.3.1.2.2. Activity area prediction results

Using the actual CCLE IC50 values in drug sensitivity prediction could not assess the predictive ability of the most of the single candidate biomarkers in the independent CGP data. So the CCLE Activity Area values of the drugs are chosen for assessment of predictive performance. The RMSE values determined for the best performing single candidate biomarkers of all the nine drugs are tabulated in **Table 4.16**. The table shows that the RMSE values of all the genes determined for the relevant drugs except PD0325901 are much lower than the RMSE values determined when the CGP Activity Area values are used as drug sensitivity data. This implies that the removal of the inconsistent CGP drug sensitivity data improves the predictive power and allows their more accurate assessment.

Table 4.16: The RMSE values determined for the best performing single candidate biomarkers in the CCLE Activity Area prediction

Gene Name	Module Name	Drug Name	RMSE (CGP)	Error in Prediction (%)
BAMBI	Lightcyan	AZD0530	0.29	29.12
MPRIP	Purple	AZD6244	0.22	21.77
NAV3	Purple	Erlotinib	0.27	27.48
PAOX	Blue	Lapatinib	0.24	24.00
RNF125	Blue	PD0325901	0.27	27.17
TFPI2	Black	PF2341066	0.28	29.57
BAMBI	Lightcyan	PLX4720	0.32	32.00
MPRIP	Purple	Sorafenib	0.21	21.03
BAMBI	Lightcyan	TAE684	0.24	23.72

Table C. 12 shows that both Pearson and Spearman correlation scores of the genes determined for AZD6244 and Sorafenib are significant. The table also show that only Pearson correlation score of the genes determined for AZD0530 and PD0325901 are significant

4.3.2. Predictive ability of combined candidate biomarkers

Single candidate biomarkers are used in predicting IC50 and Activity Area values of the anti-cancer drugs profiled between the CCLE and the CGP studies. However, predictive ability of the single candidate biomarkers is determined to be not powerful in both of these two studies, i.e, a trained model in the CCLE mostly fails in independent CGP test data. At this step, combining the proper single candidate biomarkers is assumed to increase the predictive ability of the models. In this way, combined candidate biomarkers could predict drug sensitivity values both in the CCLE and the CGP studies more powerfully. However, there appears several important considerations at this point such as which combinations should be used, how many combinations should be generated, and which methods should be applied in identifying effective combinations. In the thesis, highly connected intramodular hub genes selected from different pharmacologically significant modules produced by the WGCNA methodology are assumed to improve the predictive performance. Thereby, using expression profiles of only a few genes might be sufficient to boost the predictive power and hence predictions significantly. However, at most five candidate biomarker combinations are generated to identify how many combinations are sufficient and which combinations boost the predictions significantly since there is no drug having more than five different pharmacologically significant modules. Furthermore, in contrast to the assessment of the predictive power by using the CGP drug sensitivity data, only the CCLE drug sensitivity data is used for assessment since drug sensitivity data is highly inconsistent between the two studies. Lastly, only the common MM cell lines between the studies are selected for predictions in order to validate the results obtained from one study to another study.

After selecting the CCLE drug sensitivity data as drug data and expression profiles of the common MM cell lines between the studies as gene expression data, IC50 prediction is performed in two different ways:

1. The censored CCLE IC50 values are included in the CCLE drug sensitivity data
2. The censored CCLE IC50 values are removed from the CCLE drug sensitivity data

Activity Area prediction is also performed; however, the CCLE Activity Area values are not processed as in the case of IC50 values since Activity Area values can be generated for any screening concentrations.

Predictive performance of the combinations is assessed by the RMSE scores. In this way, the best performing combinations are determined for each drug. In addition, the actual drug sensitivity values of the best performing combinations are correlated to the predicted values, so Pearson and Spearman correlation scores are determined for the combinations.

4.3.2.1. IC50 prediction results

IC50 prediction is first performed by training the CCLE gene expression and drug sensitivity data. In this way, combinations of the candidate biomarkers having significant predictive power are determined. Next, predictive ability of these combined candidate biomarkers is tested by the CGP gene expression data. Thereby, the most effective candidate biomarker combinations for each drug are identified.

Including the censored IC50 values in the CCLE drug sensitivity data

The WGCNA method identifies tens of intramodular hub genes (candidate biomarkers) for several pharmacologically significant modules. For this reason, only the top performing three intramodular hub genes having the highest correlation scores and intramodular connectivity values are selected as candidate biomarkers. IC50 prediction is performed after these candidate biomarkers are selected. However, the censored IC50 values are included in the CCLE drug sensitivity data in order to determine how including the censored data affect the predictive performance.

The best performing combinations are determined according to their RMSE scores in **Table D. 4**. This table shows that the best performing combinations for AZD0530, Erlotinib, Lapatinib, PF2341066, and Sorafenib are binary, for PD0325901 and PLX4720 are triple, for AZD6244 and TAE684 are quadruple. It seems increasing combinations do not improve predictive power further for the five drugs. This is not surprising since sensitivity data of these five drugs mostly contain censored values that hinder powerful predictions. On the other hand, increasing combinations improve predictive power for the rest of the drugs since they mostly kill MM cell lines, and sensitivity data of the four drugs contain less censored values. However, predictive power for AZD6244 does not increase after quadruple combination even though there are five different modules determined for the drug.

Correlation scores of the predicted IC50 values to the actual IC50 values suggest that Pearson correlation scores of all the biomarker combinations determined for the relevant drugs are significant. On the other hand, Spearman correlation scores of binary combination determined for Sorafenib and triple combinations determined for Erlotinib, Lapatinib, PD0325901, and Sorafenib are not significant. (See **Table C. 13**).

After the best performing combinations are determined for all the nine drugs, they are tabulated along with the best performing single candidate biomarkers (See **Table D. 1**). In this way, the trend of predictive power obtained by varying number of combinations is demonstrated. In addition, the scatterplots which illustrate the behavior of prediction errors, correlation scores, and R^2 values determined for each combination are generated for each drug (See **Figures D. 1 - D. 9**). The barplots which display the trend of Pearson and Spearman correlation scores with varying number of correlations are also generated for all the drugs (See **Figures E. 1 - E. 9**).

The best performing single candidate biomarkers/biomarker combinations identified from all possible combinations for each of the nine drugs are tabulated in **Table 4.17**. This table tpoints that biomarker combinations outperform single candidate biomarkers in predictive power and at most quadruple combination is sufficient to obtain the highest predictive power. So increasing number of combinations might not improve the predictive power onwards quadruple combination; on the contrary, it might decrease the predictive power. RMSE plots are also generated to show the trend of prediction error for the possible number of gene combinations at this point (See **Figure 4.12**).

Table 4.17: List of the best performing candidate biomarker combinations for the nine drugs when IC50 is used as the drug sensitivity measure

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
BAMBI - MAP3K14	AZD0530	0.20	2.50	0.33	4.13
DAAM1 - ERBB2 - MPRIP - RAB38	AZD6244	1.24	15.50	1.46	18.25
ITGA5 - LRP5	Erlotinib	0.74	9.25	0.86	10.75
ITGA5 - LRP5	Lapatinib	1.19	14.88	1.12	14.00
APOE - ERBB2 - RNF125	PD0325901	1.95	24.38	2.37	29.63
BAMBI - TFPI2	PF2341066	0.84	10.50	0.80	10.00
APOE - DAAM1 - MAP3K14	PLX4720	2.32	29.00	2.52	31.50
ETHE1 - FAM172A - HIVEP3	Sorafenib	0.061	0.76	0.56	7.00

Excluding the censored IC50 values from the CCLE drug sensitivity data

The censored CCLE IC50 values in the CCLE drug sensitivity data are removed to perform more reliable predictions since repeating IC50 values impede the performance of models adversely. However, only three drugs (AZD6244, PD0325901, and TAE684) remain to have sufficient amount of IC50 values for predictions. So predictive performance of the combined candidate biomarkers is assessed only for these three drugs.

Combinations for the three drugs are generated and their predictive performances are tabulated in **Table D. 5**. This table shows that the best performing combinations are triple for PD0325901 and TAE684 and quadruple for AZD6244. However, statistical model overfits when predicting IC50 values of PD0325901 and TAE684 due to low sample size after removing the censored values. So predictions for these two drugs are not reliable. Only

the sensitivity prediction determined for AZD6244 is reliable in this scenario, and increasing biomarker combination count improves predictive power till quinary combination as in the case of prediction performed when censored IC50 values are included in the CCLE drug sensitivity data.

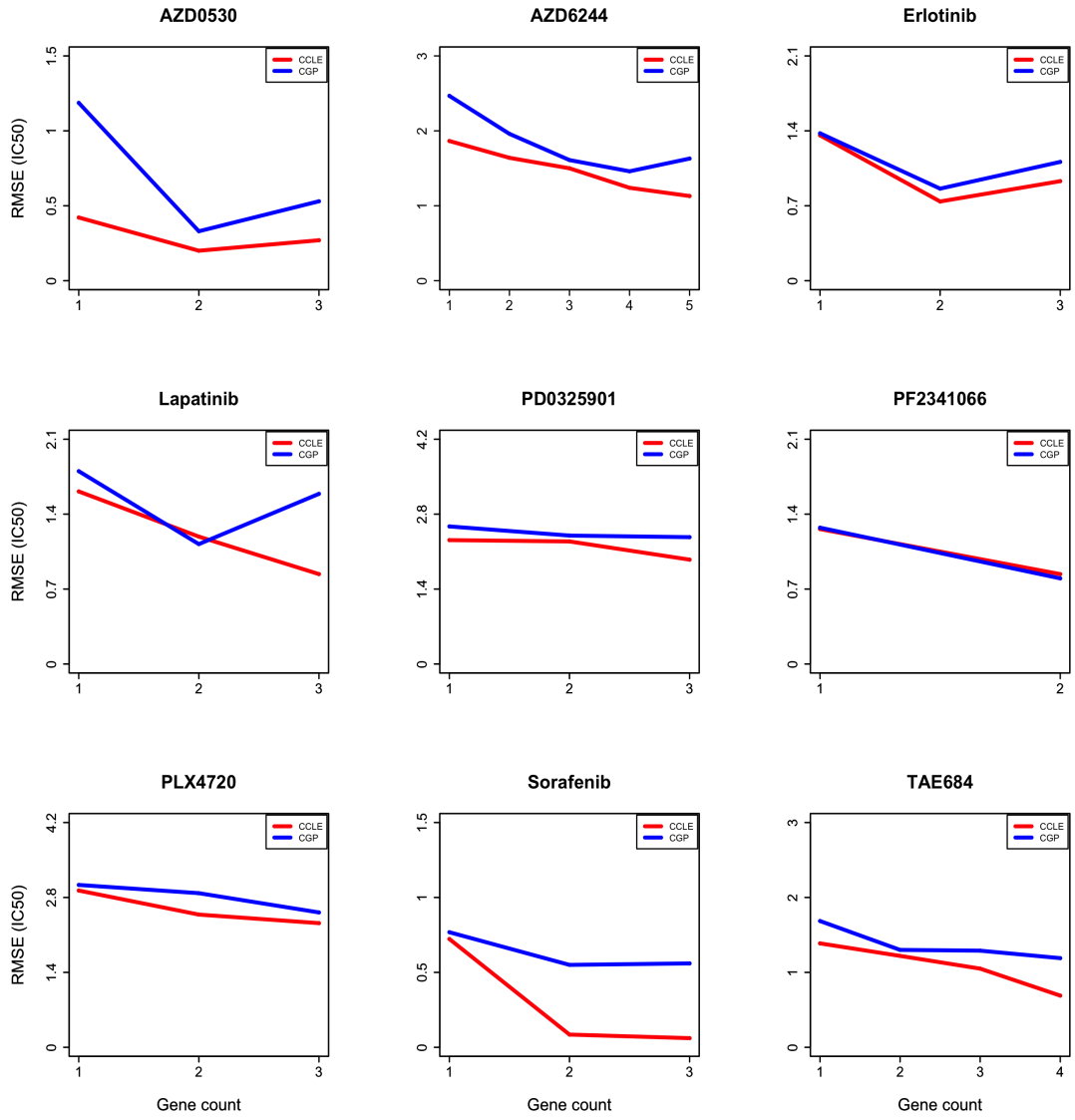

Figure 4.12: The RMSE plots which depict the trend of prediction error when the censored IC50 values are included in the CCLE drug sensitivity data

Correlation scores determined for the three drugs shows that Pearson correlation scores of all biomarker combinations of the relevant drugs are significant. Nevertheless, Spearman correlation scors of binary combination determined for PD0325901 and triple combinations determined for PD0325901 and TAE684 are not significant (See **Table C. 14**). The best performing single candidate biomarkers and biomarker combinations determined for the

three drugs are tabulated in **Table D. 2**. In addition, scatterplots for predive power and barplots for correlations score are generated (See **Figures D. 10 - D. 12** and **Figures E. 10 - E. 12**).

All the best performing biomarker combinations are tabulated in **Table 4.18.** This table demonstrates that biomarker combinations outperform single candidate biomarkers in predictive power, and at most quadruple combination is sufficient to obtain the highest predictive power. This means that increasing number of combinations might not improve the predictive power onwards quadruple combination. RMSE plots are also generated to show the trend of prediction error for the possible number of gene combinations (See **Figure 4.13**).

Table 4.18: List of the best performing candidate biomarker combinations for the three drugs when IC50 is used as the sensitivity measure

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
DAAM1 - ERBB2 - MPRIP - SLC23A2	AZD6244	0.40	16.60	0.38	15.77
APOE - ERBB2 - RNF125	PD0325901	6.8×10^{-5}	0.0068	0.16	16.00
BAMBI - C8orf4 - MAP3K14 - PROS1	TAE684	0.56	7.75	0.88	12.17

4.3.2.2. Activity Area prediction results

After IC50 predictions, Activity Area predictions are performed for the combined candidate biomarkers. As in the case of IC50 prediction, binary, triple, quadruple, and quinary combinations are generated for Activity Area prediction.

The predictive performances of the best performing combinations determined for each drug are given in **Table D. 6**. This table shows that both the CCLE and the CGP RMSE values of the binary combinations generated for all the drugs are lower than that of the single candidate biomarkers. The best performing combinations determined for AZD0530, Erlotinib, Lapatinib, PD0325901, PF2341066, and Sorafenib are binary, for PLX4720 is triple, for TAE684 is quadruple, and for AZD6244 is quinary. Using Activity Area data shows the expected increasing trend in predictive power for combinations generated by selecting one biomarker candidate from each pharmacologically significant module. This behaviour could not clearly be observed with IC50 values because of repeating censored values. In this respect, assessment of predictive power with Activity Area values is more powerful.

Correlation scores determined for biomarker combinations are tabulated in **Table C. 15**. This table shows Pearson correlation scores of all the combinations except triple combinations of Erlotinib and Sorafenib are significant. On the other hand, the table shows Spearman correlation scores of the combinations except binary combinations of Erlotinib and TAE684 and triple combinations of Erlotinib, Lapatinib, and Sorafenib are significant. Predictive performances of the candidadate biomarkers, scatterplots for predictive power, and barplots for correlation scores are also given in **Table D. 3**, **Figures D. 13 – D. 21**, and **Figures E. 13 – E. 21**, respectively.

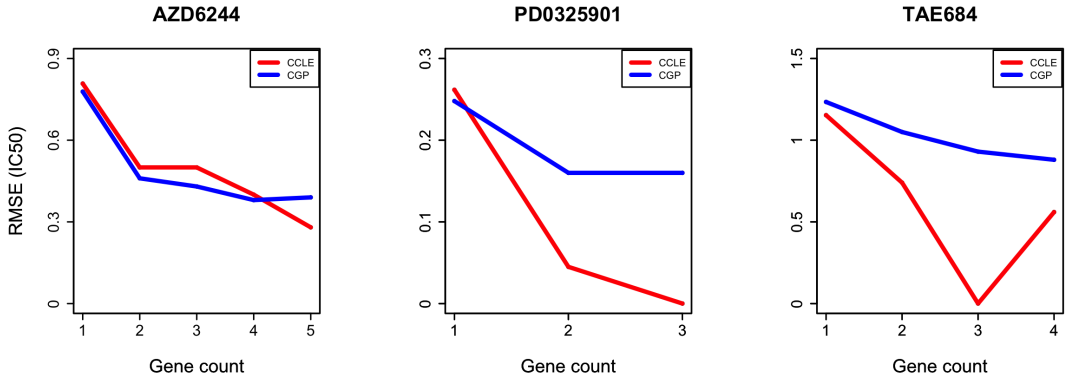

Figure 4.13: The RMSE plots which depict the trend of prediction error when both the censored and extrapolated IC50 values are excluded from drug sensitivity data

The best performing single candidate biomarkers or biomarker combinations identified from all possible combinations for each of the three drugs are tabulated in **Table 4.19**. This table demonstrates that biomarker combinations outperform single candidate biomarkers in predictive power as in the case of IC50 prediction. RMSE plots are also generated to show the trend of prediction error for the possible number of gene combinations (See **Figure 4.14**).

Table 4.19: List of the best performing biomarker combinations for the nine drugs when Activity Area is used as the drug sensitivity measure

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
BAMBI - MAP3K14	AZD0530	0.18	17.78	0.30	30.26
BAMBI - MAP3K14 - MPRIP - PROS1 - RAB38	AZD6244	0.048	4.75	0.13	13.08
LRP5 - NAV3	Erlotinib	0.22	21.73	0.26	26.28

Table 4.19 (Continued)

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
LRP5 - PAOX	Lapatinib	0.16	15.66	0.21	20.82
ERBB2 - RNF125	PD0325901	0.22	21.57	0.23	22.96
BAMBI - TFPI2	PF2341066	0.21	20.69	0.25	25.18
BAMBI - MAP3K14 - PROS1	PLX4720	0.20	19.56	0.24	24.43
HIVEP3 - MPRIP	Sorafenib	0.14	14.27	0.20	20.00
MAP3K14 - MFSD12 - PAOX - PROS1	TAE684	0.13	13.38	0.15	15.13

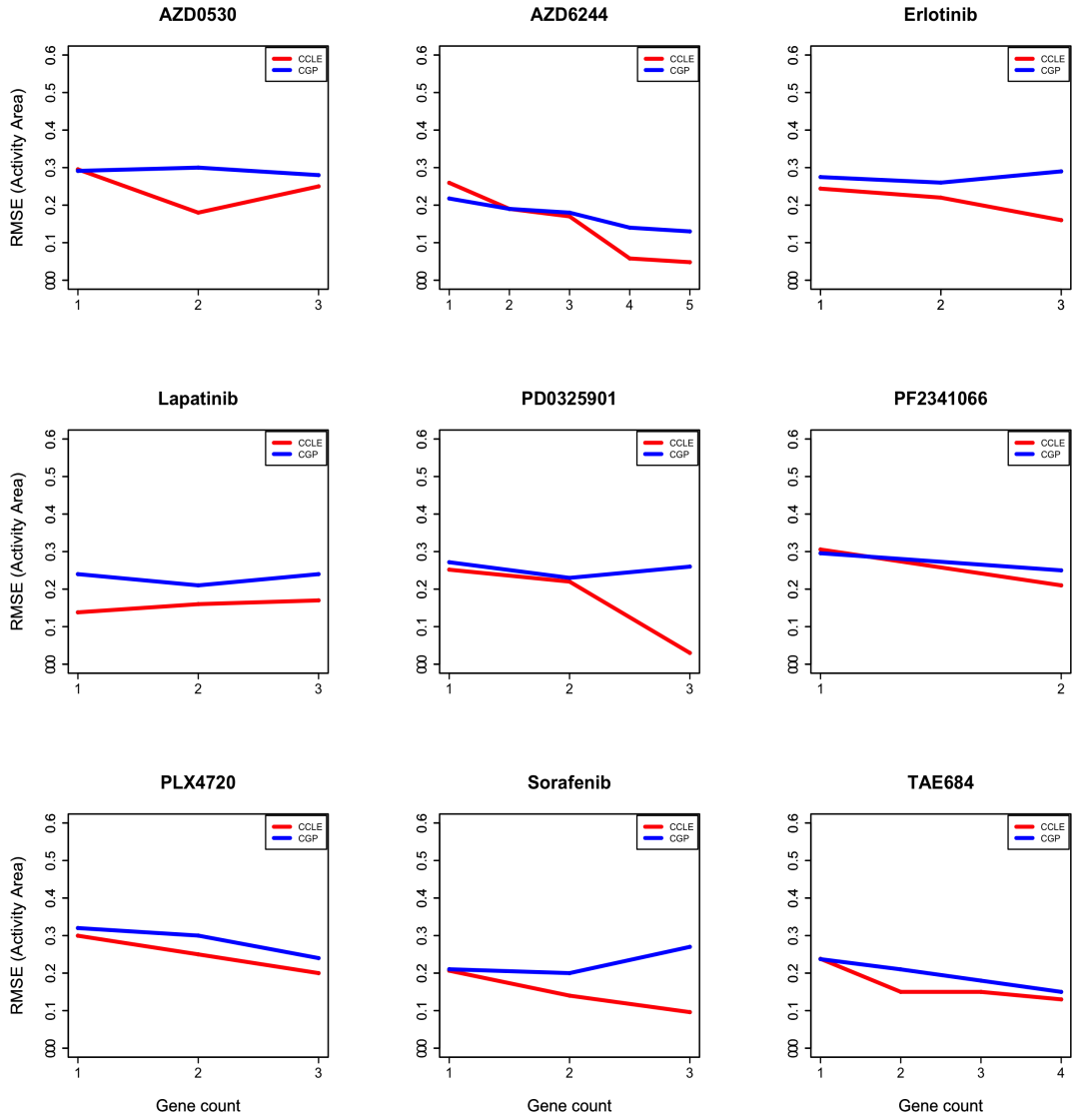

Figure 4.14: The RMSE plots which depict the trend of prediction error when the CCLE Activity Area values are used for prediction

CHAPTER 5

DISCUSSION

5.1. The WGCNA methodology is powerful in identification of candidate biomarkers

The WGCNA methodology identifies centrally located intramodular hub genes inside pharmacologically significant modules determined for the nine anti-cancer drugs among the fifteen anti-cancer drugs profiled both in the CCLE and the CGP studies. These hub genes are assumed to be candidate drug biomarkers for malignant melanoma (MM) since disease producing agents are known to attack the central players in a network. If that is the case, the hub genes identified by the WGCNA should have functions related to the development and progression of the MM. Indeed, the majority of the genes have been shown to be responsible for both the development and progression of the MM. For instance, MAP3K14 identified for AZD0530, AZD6244, PLX4720, and TAE684 is an important gene in the MM development. Its depletion lowers the expression of genes leading to tumor growth and pro-survival factors. Thu et al. (2012) show that MAP3K14 regulates β-catenin and NF-κβ regulated transcription for development and growth of the MM. it might be a candidate biomarker or therapeutic target for the MM. RNF125 identified for AZD0530, AZD6244, PD0325901, PLX4720, and TAE684 is a key gene that develops resistance to BRAF inhibitors. Kim et al. (2015) demonstrate that downregulation of RNF125 upregulates receptor tyrosine kinases via JAK1 degulation. They demonstrate that when JAK1 and EGFR signaling are blocked, RNF125 expression lowers, so BRAF resistance could be avoided in melanoma. ERBB2 identified for AZD6244 and PD0325901 is a therapeutic target for BRAF/NRAS cutaneous melanomas when it forms a complex with ERBB3 (Capparelli et al., 2015). Even though most of the identified hub genes are related to the MM disease, the rest of the hub genes have not been shown to be associated with the MM yet. This does not show that they are not significant in development or progression of the disease. These genes might be involved in tumor development and progression, but they might not have been investigated yet. In addition, they seem to be irrelevant to the MM disease because of not being considered with other genes having high associations to the MM development or progression. So they might be novel therapeutic targets or biomarkers for the MM, and experimental works should be performed for both demonstrating their roles in the MM development or progression and validating them in independent studies.

5.2. Combinations of candidate biomarkers improve drug sensitivity prediction

There are numerous studies conducted to identify powerful and reliable genomic predictors for determining anti-cancer drug sensitivity; however, most of the studies fail to identify such predictors. The major reason for this failure obviously emerges from the efforts aiming to identify single predictors. However, single predictors could not be effective in elucidation of drug activity since the vast majority of the drugs attack more than one targets. So identification of these multiple targets is extremely important for predicting drug sensitivity effectively. In this study, indeed, it is shown that combinations of candidate drug biomarkers (predictors) improve anti-cancer drug sensitivity substantially. When all the anti-cancer drugs for which single candidate biomarkers are identified are considered, it can be seen that proper combinations generated by selecting only one hub gene from each pharmacologically significant modules obtained by the WGCNA methodology always have higher predictive power than their single counterparts. Binary combinations always improve anti-cancer drug sensitivity prediction tremendously, but increasing combinations onwards binary combinations improves predictive power slightly in most cases. Interestingly, increasing combination counts does not improve the predictive power further after quadruple and quinary combinations.

The predictive power of candidate biomarkers both individually and in combinations is investigated for several different scenarios due to inconsistency in drug sensitivity data between the CCLE and the CGP studies. In this way, not only the predictive power of combinations is assessed but the effect of inconsistency in drug sensitivity prediction and possible efforts to overcome the inconsistency are inspected as well. In the first step, the expression profiles of the candidate biomarkers of all the MM cell lines and pharmacological profiles (IC50 and Activity Area) of the anti-cancer drugs screened against these cell lines in the CCLE study are trained with a non-linear SVR machine learning algorithm. Then, the predictive power of the candidate biomarkers is tested in independent CGP data. However, the assessment could not be performed because of inconsistency in drug sensitivity data between the two studies. So only common malignant MM cell lines are selected, and the CGP drug sensitivity data is removed from analyses. These adjustments lead to a more reliable assessment of predictive power. In order to determine the best performing biomarker combinations after the adjustments, anti-cancer drug sensitivity is predicted. For the drug sensitivity measure, both IC50 and Activity Area are selected as the drug sensitivity measures. When IC50 is selected as the drug sensitivity measure, the best performing combined candidate biomarkers always have higher predictive power than the best performing single candidate biomarkers. As IC50 values are included in the CCLE drug sensitivity data, the best performing candidate biomarker combinations are binary combination (BAMBI - MAP3K14) for AZD0530, quadruple combination (DAAM1 - ERBB2 - MPRIP - RAB38) for AZD6244, binary combination (ITGA5 - LRP5) for Erlotinib, binary combination (ITGA5 - LRP5) for Lapatinib, triple combination (APOE - ERBB2 - RNF125) for PD0325901, binary combination (BAMBI - TFPI2) for PF2341066, triple combination (APOE - DAAM1 - MAP3K14) for PLX4720, triple combination (ETHE1 - FAM172A - HIVEP3) for Sorafenib, and quadruple combination (BAMBI - C8orf4 - MAP3K14 - PROS1) for TAE684. However, as IC50 values are excluded from the CCLE drug sensitivity data, the best performing candidate biomarker combinations are quadruple combination (DAAM1 - ERBB2 - MPRIP - SLC23A2) for AZD6244, triple combination (APOE - ERBB2 - RNF125) for PD0325901, and quadruple combination

(BAMBI - C8orf4 - MAP3K14 - PROS1) for TAE684. So, IC50 removal does not change the best performing combinations of PD0325901 and TAE684, while it replaces RAB38 with SLC23A2 in quadruple combination generated for AZD6244. Similar to IC50 predictions, the candidate biomarker combinations always have higher predictive powers than the single candidate biomarkers when Activity Area is selected as the drug sensitivity measure. The best performing candidate biomarker combinations are binary combination (BAMBI - MAP3K14) for AZD0530, quinary combination (BAMBI - MAP3K14 - MPRIP - PROS1 - RAB38) for AZD6244, binary combination (LRP5 - NAV3) for Erlotinib, binary combination (LRP5 - PAOX) for Lapatinib, binary combination (ERBB2 - RNF125) for PD0325901, binary combination (BAMBI - TFPI2) for PF2341066, triple combination (BAMBI - MAP3K14 - PROS1) for PLX4720, binary combination (HIVEP3 - MPRIP) for Sorafenib, and quadruple combination (MAP3K14 - MFSD12 - PAOX - PROS1) for TAE684. Only combination generated for AZD0530 is same when IC50 and Activity Area predictions are compared. This shows that drug sensitivity measure is an important part of the assessment of predictive power. Activity Area is a more reliable measure since it can be measured in any screening concentration. Nevertheless, IC50 could not be measured when screening concentration is not sufficient for 50% inhibition. Unmeasured IC50 values are either censored to the maximum screening concentration as the CCLE reports or extrapolated as the CGP reports. Both of these approaches are not accurate, so the values are not reliable.

5.3. There are limitations for accurate assessment of predictors

Drug sensitivity prediction is a challenge in most cases as drug sensitivity is inconsistent between the studies, sample sizes are small for reliable estimation of predictive power, and drugs screened against the samples usually do not show a significant cytotoxic activity. In this study, all the three mentioned drawbacks challenge reliable and powerful estimation of predictive power. Even so, several strategies such as removing inconsistent drug sensitivity data, excluding non-measured values, and using the SVR machine learning algorithm that employs powerfully in small sample sizes are performed to assess the predictive power. However, there is still a need for further studies aiming to overcome these limitations. In this way, it would be possible to identify reliable drug biomarkers or therapeutic targets for the MM disease.

5.3.1. Drug sensitivity data is highly inconsistent between the studies

The CCLE and the CGP studies report common drug sensitivity measures, IC50 and Activity Area. However, these drug sensitivity measures are highly inconsistent between the studies when the shared MM cell lines and anti-cancer drugs between the studies are compared. There are five possible reasons for this inconsistency. First, the two studies use different pharmacological assays to measure drug sensitivity. In order to identify how different assay choice affects the consistency of drug sensitivity, Haibe-Kains et al. (2013) investigated the GlaxoSmithKline (GSK) data at which the pharmacological assay (Cell Titer-Glo Luminescent Cell Viability Assay kit from Promega) is same with the CCLE used.

As filtering common cell lines and drugs inspected in the CCLE, the CGP, and the GSK data, they identified that the GSK IC50 values were more consistent with the CCLE IC50 values though overall consistency was still quite poor. Nonetheless, they state that which study gives a more realiable sensitivity measure can not be known solely by comparisons. Second, different experimental protocols followed in the studies impact the accuracy of drug sensitivity measures. Standardization of experimental protocols might improve the efficiency of drug sensitivity measurement; however, Garnett et al. (2013) show that even though the same experimental protocols were followed with the same cell lines for an anti-cancer drug, Camptothecin, the correlation scores between the measurements increased ($r < 0.60$), but the correlation was still not so high as expected. Phenotypic differences among cell lines and different passage counts might explain the discrepancy in drug sensitivity measure, yet there is no study conducted for the effects of phenotype and passage count on the accuracy of drug sensitivity measure. Third, the two studies measured drug sensitivity with different concentration ranges. The CGP study mostly screened the drugs in a narrower concentration range than the CCLE study. If same screening concentration ranges had been determined in the studies, more reliable and powerful predictions would have been performed. Fourth, the CGP study extrapolates the IC50 values of the cell lines which do not reach 50% inhibition in given concentration range. Extrapolated values are non-measured values, so they do not show the actual drug sensitivity measure. Fifth, genetic heterogeneity of the cell lines might have an impact on displaying different drug responses. This might explain that how anti-cancer drug response could not be measured effectively even though the same experimental protocols were followed with same screening concentration ranges. Thereof, biological factors should also be considered for obtaining accurate drug sensitivity measures. In conclusion, both technical and biological considerations should be taken into consideration when drug sensitivity is measured. In this way, consistency in drug sensitivity could be achieved.

5.3.2. Small sample size precludes reliable estimation of predictive power

Small sample size is an important limitation of the study. The CCLE and the CGP studies investigated 62 and 42 MM cell lines respectively for expression profiling. However, anti-cancer drugs profiled in these two studies were not screened against all the MM cell lines, so the count of the MM cell lines used for drug sensitivity measure is even less than the mentioned numbers. It is hard to predict drug sensitivity with such a low amount of cell lines via machine learning approaches since they require large sample sizes for learning the behavior of data. Otherwise, statistical model might overfit the data, so predictive power would not be assessed powerfully and reliably. Moreover, testing a model is more meaningful in large sample sizes. Nevertheless, a machine learning algortihm having superior performance in low sample sizes could be employed for compansating all these limitations. For this purpose, the SVR machine learning algorithm was chosen for predictions. In contrast to the SVR, most of the other machine learning algortihms are unfortunately poor in predictive performance in low sample sizes. From this aspect, using only the SVR for drug sensitivity prediction is a limitation since other algorithms were not applied for predictive power. Therefore, it is essential that pharmacogenomics studies should screen large amount of cell lines for powerful anti-cancer drug sensitivity prediction.

5.3.3. Anti-cancer drugs profiled in the two studies do not have remarkable cytotoxic activity against the MM

The common anti-cancer drugs profiled between the two studies have not been shown to have a significant cytotoxic activity against the MM. For this reason, the concentration range of these drugs required for killing the MM cell lines is mostly insufficient. Larger screening concentrations would give an idea about their cytotoxic activity; however, the cost of the studies would be extremely higher, and it would be meaningless to observe a cytotoxic activity in a screening concentration that human cells can not tolerate in chemotherapy. Thus, IC50 could not be obtained for a significant proportion of the MM cell lines. Activity Area, on the other hand, could be obtained for any screening concentration, but it does not show anything when anti-cancer drugs do not have a significant cytotoxic activity. If pharmacogenomics studies screened anti-cancer drugs having a high cytotoxic activity for the MM, it would be easier to assess the predictive power of the identified candidate biomarkers for drug sensitivity more powerfully. Thereof, anti-cancer drugs having the high cytotoxic activity for the MM should be included in future studies in order to identify more effective predictors (biomarkers) that could determine the ideal doses required for each of the MM cell lines according to their genetic background.

CHAPTER 6

CONCLUSION AND FUTURE STUDIES

6.1. Conclusion

Personalized treatment of cancer according to genetic background of each patient is an important research topic in this era. It is widely assumed that huge amount of patient data produced by the high throughput sequencing technology will contribute to identify genomic predictors that might pave the way for choosing the right anti-cancer drugs in chemotherapy for each patient in a patient-centric manner. For this purpose, numerious studies conducted to identify such predictors, but only a small fraction of them could be successful due to the lack of high quality computational studies in the field. Reductionist approaches were mostly followed in the past, but they were poor in understanding the complexity of the biological question of interest. So instead of following reductionist approaches, scientists are turning their attention to systems approaches, which enable them to investigate the biological question as a system. Systems approaches are gaining popularity in personalized medicine since it has been shown many times that they are highly effective in discovering the biological complexity. So it is considered that they might be major tools in identification of biomarkers or therapeutic targets in cancer.

In this study, the WGCNA systems biology based network approach is used to identify gene expression-based candidate drug biomarkers for the MM by using expression profiles of the MM cell lines and sensitivity profiles of the anti-cancer drugs screened against these cell lines in the CCLE study. In the first step, the expression profiles of the 38 MM cell lines having large amount of drug sensitivity data obtained for the 24 anti-cancer drugs profiled in the CCLE study are selected for constructing a gene co-expression network as to the correlation patterns of genes across the MM cell line samples. After network construction, module detection is performed by hiearchical clustering by a dynamic tree cut algorithm, which clusters highly co-expressed genes into distinct network modules. The modules which are at least 70% similar are merged into a single module since higly similar modules are likely to account for similar biological activitiy. Next, module eigengene value, which is the average expression profiles of the genes inside each module, is determined for the merged modules. Then, module enrichment analysis with the DAVID tool is performed to observe whether the modules are functional, and the functional modules are selected for further analyses. The eigengene values of these functional modules are correlated to the sensitivity profiles (IC50) of the drugs available in the CCLE study, and only highly correlated (P-

value < 0.05) modules are filtered (Pharmacologically significant modules). After pharmacologically significant modules are determined, the centrally located intramodular hub genes having the greatest connectivity to other genes (Intramodular connectivity > 0.70) are identified for each pharmacologically significant module. At this step, individual expression profiles of these hub genes are correlated to the IC50 values of the MM cell lines in order to identify significantly correlated hub genes inside each of the pharmacologically significant modules. These hub genes are regarded to be candidate drug biomarkers or therapeutic targets for the MM. If the WGCNA methodology is applied powerfully for identifying such candidate biomarkers, they should have functions related to the MM development or progression. Indeed, the majority of the identified candidate biomarkers have already been associated with the MM disease, so it is shown that the WGCNA methodology works to identify MM specific candidate biomarkers powerfully.

Identification of single candidate biomarkers for the MM is an important task, but their predictive power for drug sensitivity is also crucial. So predictive ability of these candidate biomarkers is assessed by the SVR machine learning algorithm both individually and in combinations. Combinations of these candidate biomarkers are generated since it is assumed that proper combinations might improve anti-cancer drug sensitivity prediction substantially. Proper combinations are generated by selecting only one intramodular hub gene from each pharmacologically significant module as selecting only one representative gene from each module having a distinct biological process is assumed to be sufficient for anti-cancer drug sensitivity prediction. However, at most 5 gene combinations are generated since no drug has more than 5 pharmacologically significant modules. The predictive power of the candidate biomarkers is first assessed by the CCLE data with the SVR machine learning algorithm both individually and in combinations. LOOCV is used as model validation technique, and drug sensitivity measures (IC50 and Activity Area) are predicted. Here, RMSE is used as a model performance metric. The less the RMSE value of the models, the more accurate the predictions are performed. In this way, the best performing candidate biomarkers and biomarker combinations for each drug are determined by the CCLE data. After prediction of drug sensitivity measures, the actual CCLE values are correlated to the predicted CCLE values. Thereby, it is checked whether the predicted values for the best performing candidate biomarkers or biomarker combinations are significantly correlated to the actual values. The predictive power is also tested in an independent CGP data with the same model obtained by the CCLE data. Nevertheless, predictive power is low when the CGP data is fed into the model since high inconsistency in drug sensitivity data between the two studies leads to failure of accurate assessment. So the CGP IC50 values are processed in three different ways, but the CGP Activity Area values do not need any modification:

1. The extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs (8 μM) profiled in the CCLE study.

2. The extrapolated CGP IC50 values are censored to the maximum screening concentration of the drugs (differs among drugs) profiled in the the CGP study.

3. The extrapolated CGP IC50 values are excluded from the CGP drug sensitivity data.

Even though all the three processing steps are performed, none of the approaches is effective in the assessment of predictive power. Therefore, the CGP drug sensitivity data is removed

from predictions. Only the CCLE drug sensitivity data is used since drugs are screened against broader screening concentration ranges and IC50 values are not extrapolated. In addition, only the common MM cell lines between the studies are selected as it is assumed that highly concordant gene expression data between the studies would give clues for accurate assessment of predictive power if the drug sensitivity data were assumed to be consistent. However, the CCLE IC50 values are processed in two different ways, while the CCLE Activity Area values remain to be same:

1. The censored CCLE IC50 values are included in the CCLE drug sensitivity data.
2. The censored CCLE IC50 values are removed from the CCLE drug sensitivity data.

After the processing step, the predictive power of the best performing single candidate biomarkers or biomarker combinations are determined according to both IC50 and Activity Area values (See **Table 4.21, Table 4.26, & Table 4.31**). The results show that the best-performing candidate biomarkers determined by including the IC50 values in the CCLE data are same with that of excluding the IC50 values from the CCLE data except for AZD6244. However, only the best performing candidate biomarkers determined by the CCLE Activity Area values of AZD0530 is same with that of determined by the CCLE IC50 values. This is likely have resulted from discrepancies in reported drug sensitivity measurement parameters (IC50 and Activity Area). In addition, the results show that combinations always have high predictive power than single candidate biomarkers. The results also show that only a few genes are sufficient to predict anti-cancer drug sensitivity powerfully. So predictive power reduces suddenly onwards a few gene combinations.

We select the MM as a model disease in this study to show that we have developed a strategy that identifies gene expression-based candidate drug biomarkers or drug targets for several diseases such as cancer, obesity, and neurological disorders. We believe that after the mentioned limitations are overcome, the identified candidate biomarkers obtained by the developed strategy have potential to be validated by the RT-PCR technique in both in-vitro and in-vivo clinical studies, so that they would be utilized in clinics to determine the most effective drugs in chemotherapy for each of the patients according to their genetic backgrounds. And so, we believe in tackling diseases more effectively, extending the life expectancy of the patients, and economizing health expenditures substantially.

6.2. Future Work

Prediction of anti-cancer drug sensitivity is challenging due to the lack of available biological and clinical data, technical difficulties in accurate drug sensitivity measurement, and varying experimental procedures among current studies. For this reason, most of the studies could not accomplish to identify powerful predictors or biomarkers that may contribute to the personalized treatment of patients suffering from cancer disease. However, in this study, it is demonstrated that combinations of the candidate biomarkers identified by the WGCNA method might be effective in anti-cancer drug sensitivity prediction although there are several aforementioned obstacles in the identification of reliable predictors. Even so, several forthcoming studies should be held to identify reliable predictors and translate

them in clinics to decide the most appropriate chemotherapy for cancer patients according to their genetic profiles.

- Instead of using only the MM cell lines, several other types of cancer cell lines can also be used to detect whether the methodology applied in this study could be reproduced for different cancer types. It can also be applied to other diseases such as diabetes and alzheimer in order to identify biomarkers that might predict which patient responds to which particular drug prior to onset of therapy.

- In addition to gene expression data, other external data such mutation, copy number variation, and methylation can be integrated to drug sensitivity data. Thereby, the assessment of candidate biomarkers can be accomplished more accurately, reliably, and powerfully.

- Finally, predictive ability of candidate biomarkers identified for any cancer types can be assessed both individually and in combinations further in additional pharmacogenomics studies such as NCI60, gCSI, and GSK in order to accomplish *in-silico* validation of candidate biomarkers. What is more, candidate biomarkers can be validated by several clinical studies. In this way, both *in-vitro* and *in-vivo* validation of candidate biomarkers could be performed, and those validated might be used in clinics to determine the patients who might respond chemotherapy beforehand.

REFERENCES

American Association for Cancer Research (2012). AACR cancer progress report 2012. Retrieved August 03, 2016, from http://cancerprogressreport.org/2011/Documents/2011_AACR_CPR_Text_8-03-12.pdf

American Cancer Society (2016). What is melanoma skin cancer? Retrieved August 02, 2016, from http://www.cancer.org/cancer/skincancermelanoma/detailedguide/melanoma-skin-cancer-key-statistics

American Cancer Society (2016). Risk factors for melanoma skin cancer. Retrieved August 02, 2016, from http://www.cancer.org/cancer/skincancermelanoma/detailedguide/melanoma-skin-cancer-risk-factors

Barabasi, A. L. and Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288, 0-69

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 307-603. JOUR. Retrieved from http://dx.doi.org/10.1038/nature11003

Ben-Hur, A., Horn, D., Siegelman, H. T., & Vapnik, V. (2001). Support Vector Clustering. *Journal of Machine Learning Research,* 2 (2001), 125-137

Bolstad, B. M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R. A., & Speed, T. P. (2005). Quality Assessment of Affymetrix GeneChip Data in Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Gentleman R, Carey V, Huber W, Irizarry R, and Dudoit S. (Eds.), Springer, New York.

Burges, J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121-167. http://doi.org/10.1023/A:1009715923555

Cancer Research UK (2016). Biological therapy for malignant melanoma. Retrieved August 03, 2016, from http://www.cancerresearchuk.org/aboutcancer/type/melanoma/treatment/biological-therapy-for-melanoma

Capparelli, C., Rosenbaum, S., Berman-Booty, L. D., Salhi, A., Gaborit, N., Zhan, T., … Aplin, A. E. (2015). ErbB3/ErbB2 complexes as a therapeutic target in a subset of wild-type BRAF/NRAS cutaneous melanomas. Cancer Research, 75(17), 3554-3567. http://doi.org/10.1158/0008-5472.CAN-14-2959

Chang, C., & Lin, C. (2013). LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology* (TIST), 2, 1-39. http://doi.org/10.1145/1961189.1961199

Chen, D. R., Wu, Q., Ying, Y. M., & Zhou, D. X. (2004). Support vector machine soft margin classifiers: Error analysis. Journal of Machine Learning Research, 5(2004), 1143-1175.

Clarke, S., Griebsch, J., & Simpson, T. (2005). Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses. Journal of Mechanical Design, Vol. 127(6), 1077-1087. http://doi:10.1115/1.1897403

Czarnecki, W. M., Podlewska, S., & Bojarski, A. J. (2015). Robust optimization of SVM hyperparameters in the classification of bioactive Drugs. *Journal of Cheminformatics*, 1-15. http://doi.org/10.1186/s13321-015-0088-0

De Jong, S., Boks, M. P. M., Fuller, T. F., Strengman, E., Janson, E., de Kovel, C. G. F., Ori, A. P. S., Vi, N., Mulder, F., Blom, J. D., Glenthøj, B., Schubart, C. D., Cahn, W., Kahn., R. S., Horvath, S., & Ophoff, R. A. (2012). A gene co-expression network in whole blood of Schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PLoS ONE*, *7*(6), 1-10. http://doi.org/10.1371/journal.pone.0039498

Dong, Z., Zhang, N., Li, C., Wang, H., Fang, Y., Wang, J., & Zheng, X. (2015). Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. *BMC Cancer*, *15*(1), 489. http://doi.org/10.1186/s12885-015-1492-6

European Society of Medical Oncology (2015). Four distinct genomic subtypes of cutaneous melanoma. Retrieved August 02, 2016, from http://www.esmo.org/Oncology-News/Four-Distinct-Genomic-Subtypes-of%20Cutaneous-Melanoma

Fernandez-Lozano, C., Fernandez-Blanco, E., Dave, K., Pedreira, N., Gestal, M., Dorado, J., & Munteanu, C. R. (2014). Improving enzyme regulatory protein classification by means of SVM-RFE feature selection. *Molecular BioSystems*, *10*(5), 1063-1071. http://doi.org/10.1039/C3MB70489K

Fletcher, T. (2009). Support Vector Machines Explained. *Online. Http://sutikno. Blog. Undip. Ac. id/files/2011/11/SVM-Explained. pdf.[Accessed 06 06 2013]*, 1-19. http://doi.org/10.1002/9780470503065.app2

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., … Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, *483*(7391), 570-575. JOUR. Retrieved from http://dx.doi.org/10.1038/nature11005

Gargalovic, P. S., Imura, M., Zhang, B., Gharavi, N. M., Clark, M. J., Pagnon, J., Yang, W. P., He, A., Truong, A., Patel, S., Nelson, S. F., Horvath, S., Berliner, J. A., Kirchgessner T. G., & Lusis, A. J. (2006). Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proc Natl Acad Sci U S A*, *103*(34), 12741-12746. http://doi.org/10.1073/pnas.0605457103

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy: analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics,* 20,(3), 307-315

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E. E., Drake, T. A., Lusis, A. J., & Horvath, S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics*, *2*(8), 1182-1192. http://doi.org/10.1371/journal.pgen.0020130

Gong, K. W., Zhao, W., Li, N., Barajas, B., Kleinman, M., Sioutas, C., Horvath, S., Lusis, A. J, Nel, A., & Araujo, J. A. (2007). Air-pollutant chemicals and oxidized lipids exhibit genome-wide synergistic effects on endothelial cells. *Genome Biology*, *8*(7), R149. http://doi.org/10.1186/gb-2007-8-7-r149

Gruber, F., Kaštelan, M., & Brajac, I. (2008). Molecular and genetic mechanisms in melanoma. Collegium Antropologicum, 32 Suppl 2, 147-52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19138018

Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J. W. L., & Quackenbush, J. (2013). Inconsistency in large pharmacogenomic studies. *Nature*, *504*(7480), 389-393. Retrieved from http://dx.doi.org/10.1038/nature12831

Haverty, P. M., Lin, E., Tan, J., Yu, Y., Lam, B., Lianoglou, S., … Bourgon, R. (2016). Reproducible pharmacogenomic profiling of cancer cell line panels. Nature, 533(7603), 333-337. Retrieved from http://dx.doi.org/10.1038/nature17987

Hayward, N. K. (2003). Genetics of melanoma predisposition, 3053-3062. Retrieved from http://doi.org/10.1038/sj.onc.1206445

Hejase, H., & Chan, C. (2015). Improving Drug Sensitivity Prediction Using Different Types of Data. *CPT: Pharmacometrics & Systems Pharmacology*, *4*(2), 98-105. http://doi.org/10.1002/psp4.2

Horvath, S., Zhang, B., Carlson, M., Lu, K. V, Zhu, S., Felciano, R. M., Laurance, M. F., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, A. C., Liau, L. M., Wu, H., Geschwind, D. H., Febbo, P. G., Kornblum, H. I., Cloughesy, T. F., Nelson, S. F., & Mischel, P. S. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(46), 17402-17407. http://doi.org/0608396103/r10.1073/pnas.0608396103

Huang, C. L., & Wang, C. J. (2006). A GA-based feature selection and parameters optimizationfor support vector machines. *Expert Systems with Applications*, *31*(2), 231-240. http://doi.org/10.1016/j.eswa.2005.09.024

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*, *4*(1), 44-57. Retrieved from http://dx.doi.org/10.1038/nprot.2008.211

International Agency for Research on Cancer (2014). GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide in 2012. Retrieved August 02, 2016, from http://globocan.iarc.fr/Default.aspx

Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H., & Margolin, A. a. (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*,3-74. http://doi.org/10.1055/s-0029-1237430.Imprinting

Kim, H., Frederick, D. T., Levesque, M. P., Cooper, Z. A., Feng, Y., Krepler, C., … Ronai, Z. A. (2015). Downregulation of the Ubiquitin Ligase RNF125 Underlies Resistance of Melanoma Cells to BRAF Inhibitors via JAK1 Deregulation. *Cell Reports*, 11(9), 1458-1473. http://doi.org/10.1016/j.celrep.2015.04.049

Kovalev, V. A., Liauchuk, V. A., & Safonau, I. U. (2013). Examining the feasibility of predicting drug resistance of lung tuberculosis using image data, 122-125. Retrieved August 06, 2016, from http://www.elib.bsu.by/bitstream/123456789/52066/1/122-125.pdf

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559. http://doi.org/10.1186/1471-2105-9-559

Lanlan, Z., Juyang, L., Qilin, Z., & Yudong, W. (2015). Using Genetic Algorithm to Optimize Parameters of Support Vector Machine and Its Application in Material Fatigue Life Prediction, *8*(1), 21-26. http://doi.org/10.3968/6404

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E. & Storey, J. D. (2016). sva: Surrogate Variable Analysis. R package version 3.20.0

Levine, A. J., Miller, J. A., Shapshak, P., Gelman, B., Singer, E. J., Hinkin, C. H., Commins, D., Morgello, S., Grant, I., & Horvath, S. (2013). Systems analysis of human brain gene expression : mechanisms for HIV-associated neurocognitive impairment and common pathways with Alzheimer ' s disease. *BMC Medical Genomics*,(4), 1-20. http://doi:10.1186/1755-8794-6-4

Mayo Clinic (2015). Cancer. Retrieved August 02, 2016, from **http://www.mayoclinic.org/diseases-conditions/cancer/basics/definition/con-20032378**

Melanoma Research Foundation (n.d.). Melanoma treatment. Retrieved Retrieved August 03, 2016, from https://www.melanoma.org/understand-melanoma/melanoma-treatment

Miller, J. A., Oldham, M. C., & Geschwind, D. H. (2008). A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *J Neurosci*, *28*(6), 1410-1420. http://doi.org/10.1523/JNEUROSCI.4098-07.2008\r28/6/1410

National Cancer Institute (2015). About cancer. Retrieved August 02, 2016, from https://www.cancer.gov/about-cancer

National Institutes of Health (2011). How UV radiation triggers melanoma. Retrieved August 03, 2016, from https://www.nih.gov/news-events/nih-research-matters/how-uv-radiation-triggers-melanoma

Oldham, M. C., Horvath, S., & Geschwind, D. H. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 17973-8. http://doi.org/10.1073/pnas.0605938103

Oliveros, J.C. (2007 - 2015). Venny. An interactive tool for comparing lists with Venn's diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html

Ruderfer, D. M., Roberts, D. C., Schreilber, S. L., Perlstein, E. O., & Kruglyak, L. (2009). Using expression and genotype to predict drug response in yeast. *PLoS ONE*, *4*(9). http://doi.org/10.1371/journal.pone.0006907

Schadendorf, D., Fisher, D. E., Garbe, C., Gershenwald, J. E., Grob, J.-J., Halpern, A., Heryln, M., Marchetti, M. A., McArthur, G., Ribas, A., Roesch, A., & Hauschild, A. (2015). Melanoma. *Nature Reviews Disease Primers*, *1*, 15003. Retrieved from http://dx.http://doi.org/10.1038/nrdp.2015.3

Smoller, B. R. (2006). Histologic criteria for diagnosing primary cutaneous malignant melanoma. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *19 Suppl 2*, S34-S40. http://doi.org/10.1038/modpathol.3800508

Thu, Y. M., Su, Y., Yang, J., Splittgerber, R., Na, S., Boyd, A., … Richmond, A. (2012). NF-κB inducing kinase (NIK) modulates melanoma tumorigenesis by regulating expression of pro-survival factors through the β-catenin pathway. *Oncogene*, 31(20), 2580-2592. http://doi.org/10.1038/onc.2011.427

Tung, R. & Vidimos, A. (2010). Melanoma. Retrieved August 03, 2016, from http://www.clevelandclinicmeded.com/medicalpubs/diseasemanagement/dermatology/cutaneous-malignant-melanoma/

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer New York, 2009

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun Y. E., Liu, J., Horvath, S., & Fan, G. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, *500*(7464), 593-597. http://doi.org/10.1038/nature12364

APPENDIX A

THE PLOTS GENERATED BY DATA PRE-PROCESSING

Figure A. 1: Images of the CCLE microarray samples visualized for quality control

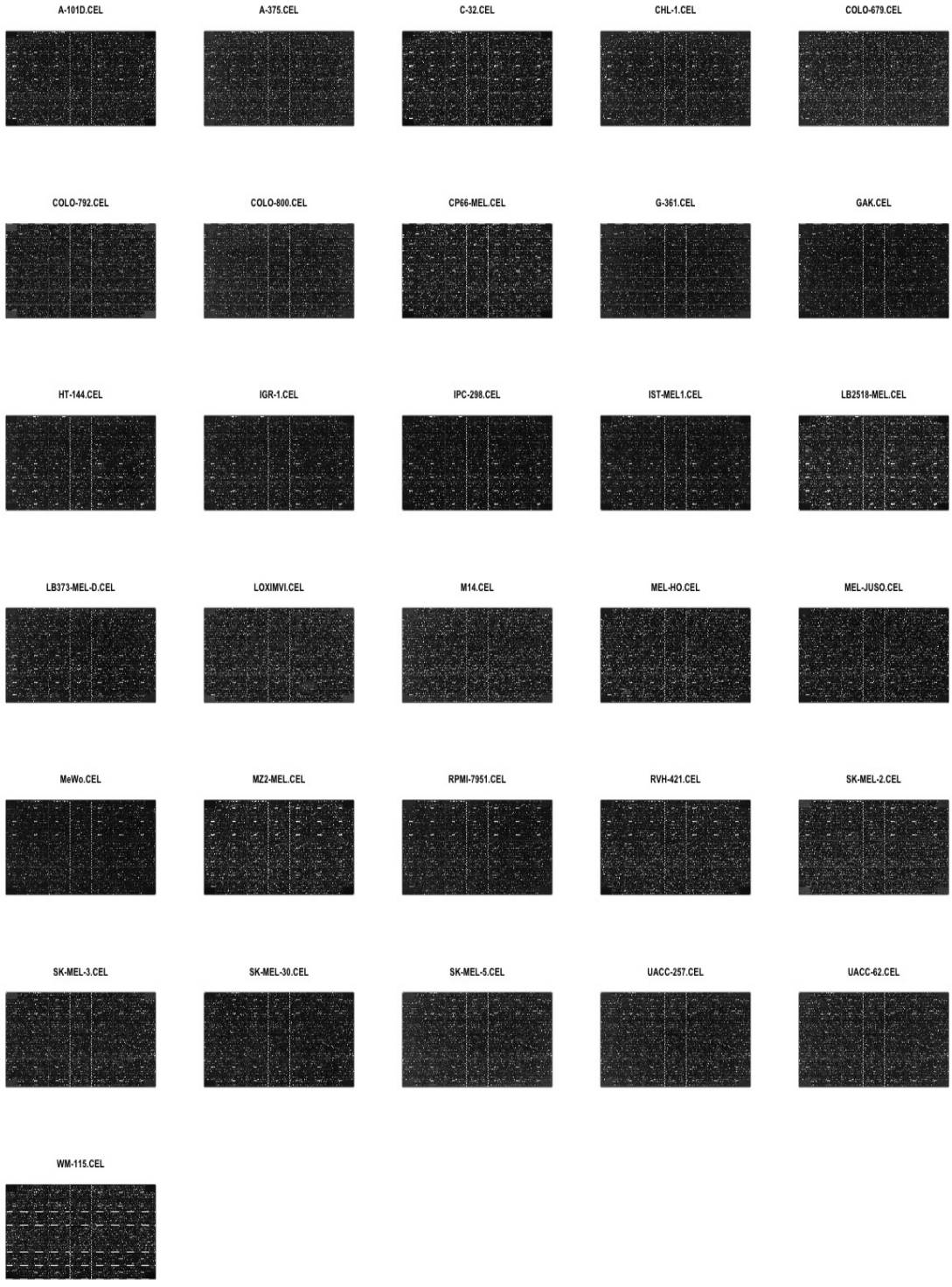

Figure A. 2: Images of the CGP microarray samples visualized for quality control

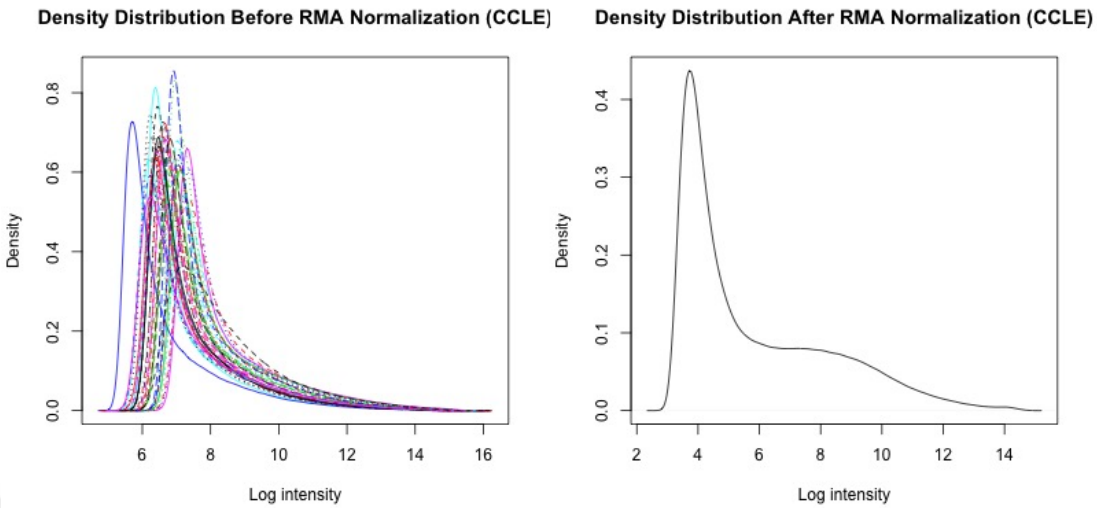

Figure A. 3: The density distribution plots which show the histogram of the perfect match intensities for the CCLE microarray samples before and after RMA normalization

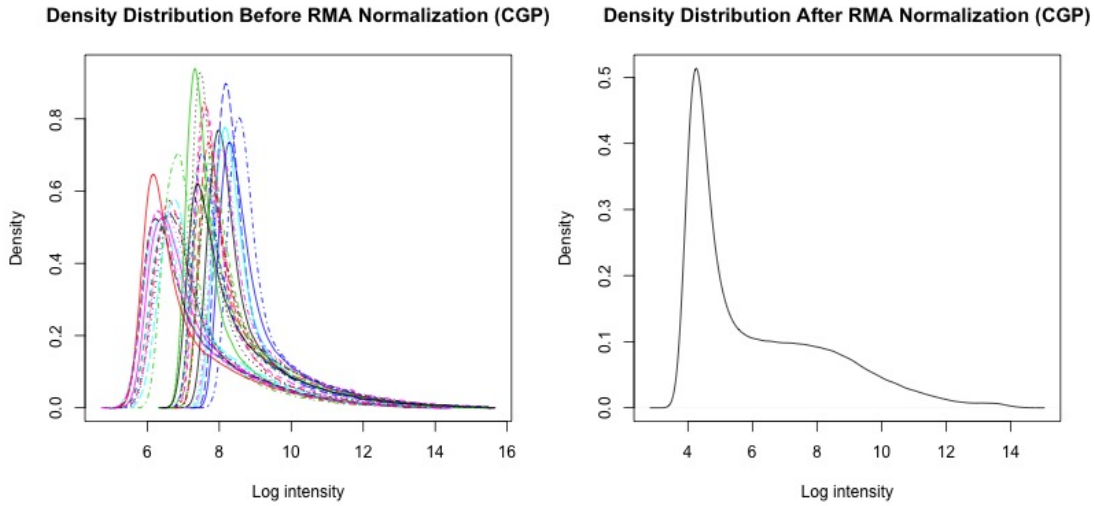

Figure A. 4: The density distribution plots which show the histogram of the perfect match intensities for the CGP microarray samples before and after RMA normalization

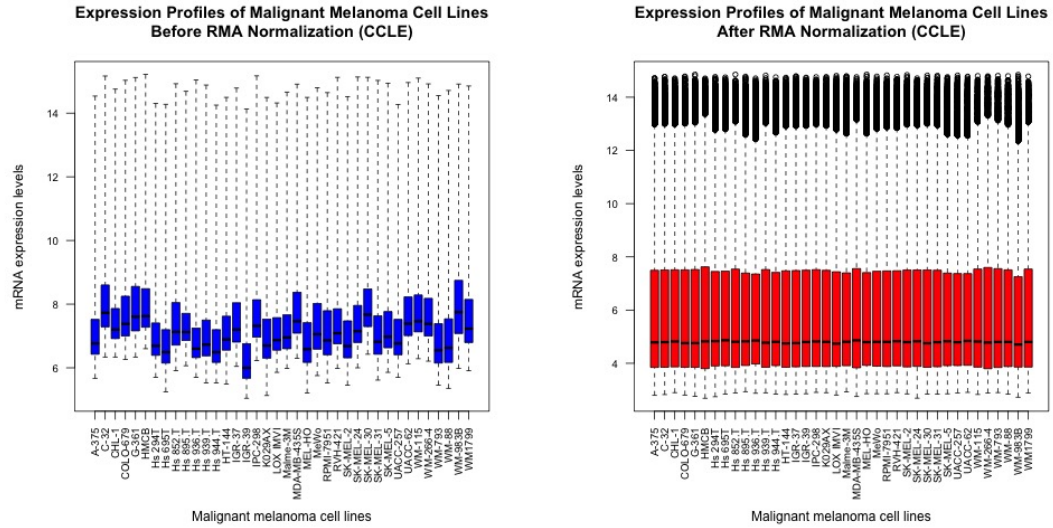

Figure A. 5: Boxplots of the CCLE microarray samples before and after RMA normalization

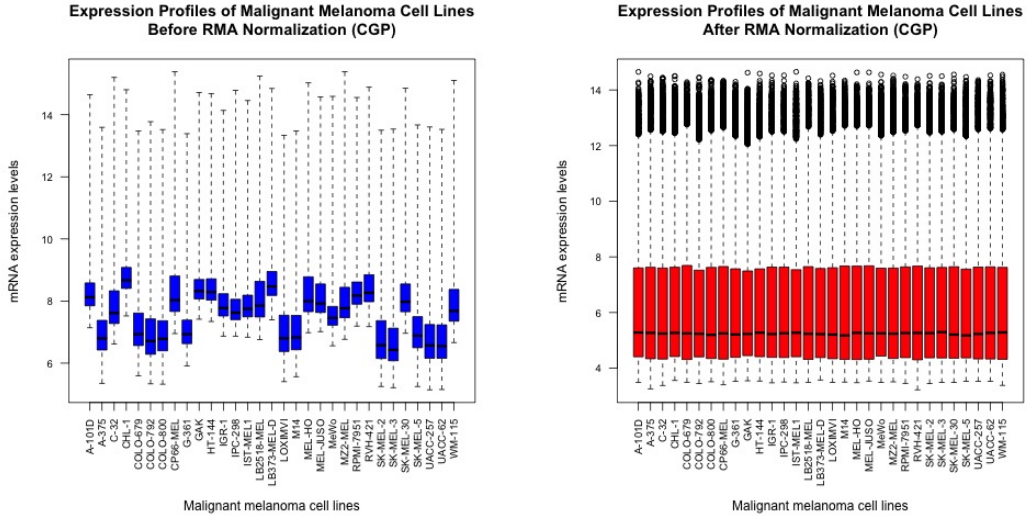

Figure A. 6: Boxplots of the CGP microarray samples before and after RMA normalization

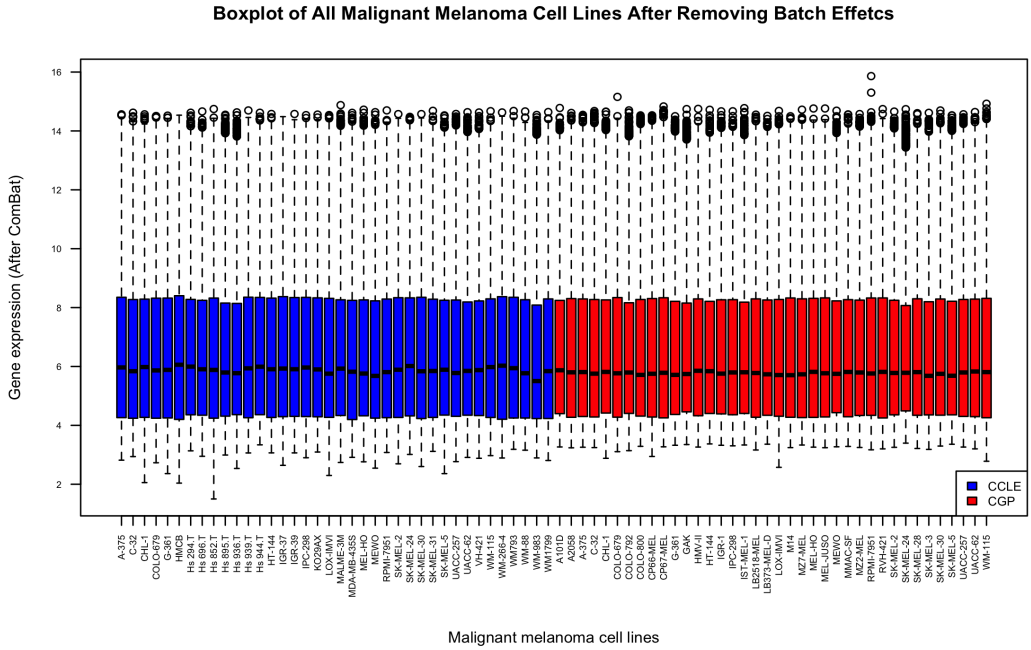

Figure A. 7: Boxplot of the gene expression levels of all the malignant melanoma cell lines investigated both in the CCLE and the CGP studies after setting gene expression levels of the cell lines to the same level to make the studies comparable

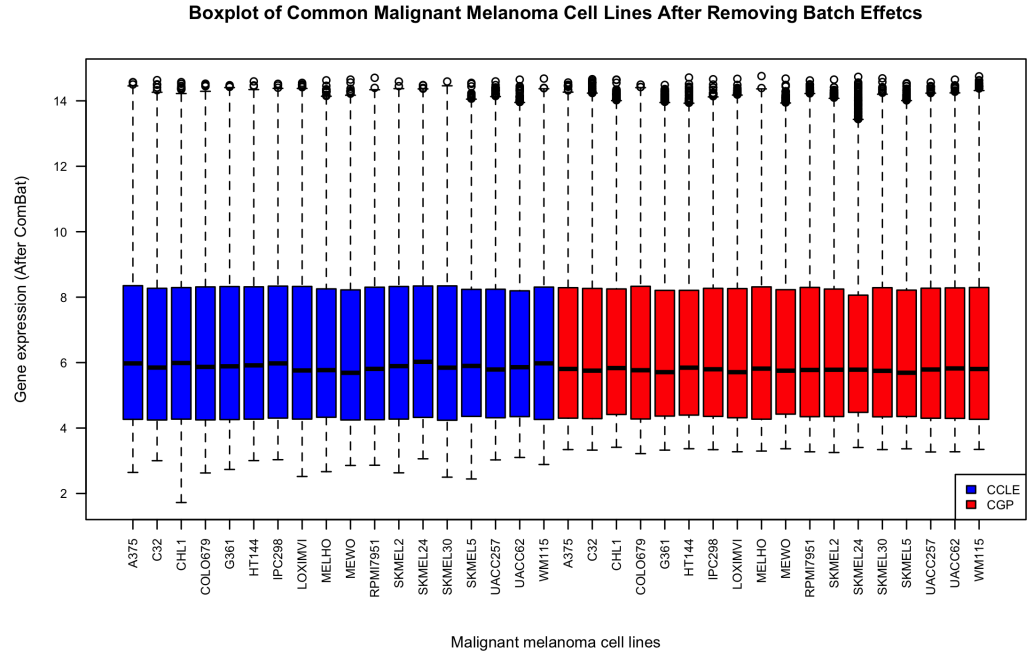

Figure A. 8: Boxplot of the gene expression levels of the common malignant melanoma cell lines investigated between the CCLE and the CGP studies after setting gene expression levels of the cell lines to the same level to make the studies comparable

APPENDIX B

HEATMAP PLOTS

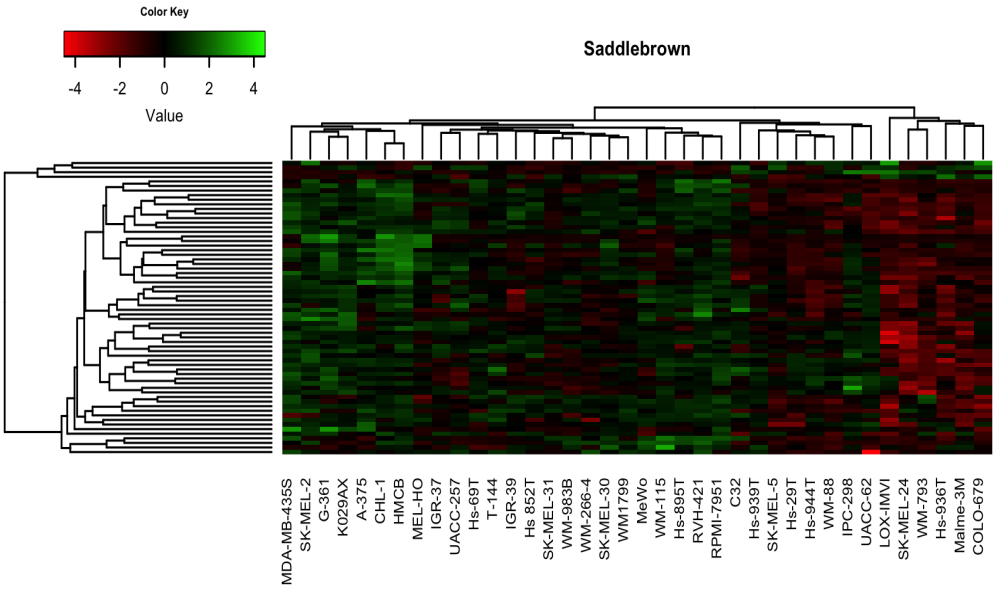

Figure B. 1: The heatmap plot of the saddlebrown module

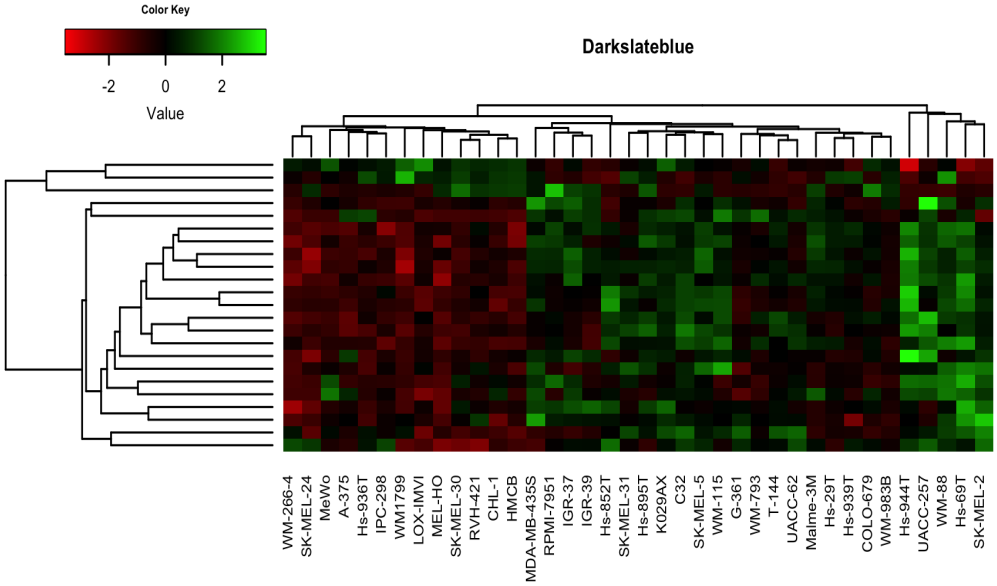

Figure B. 2: The heatmap plot of the darkslateblue module

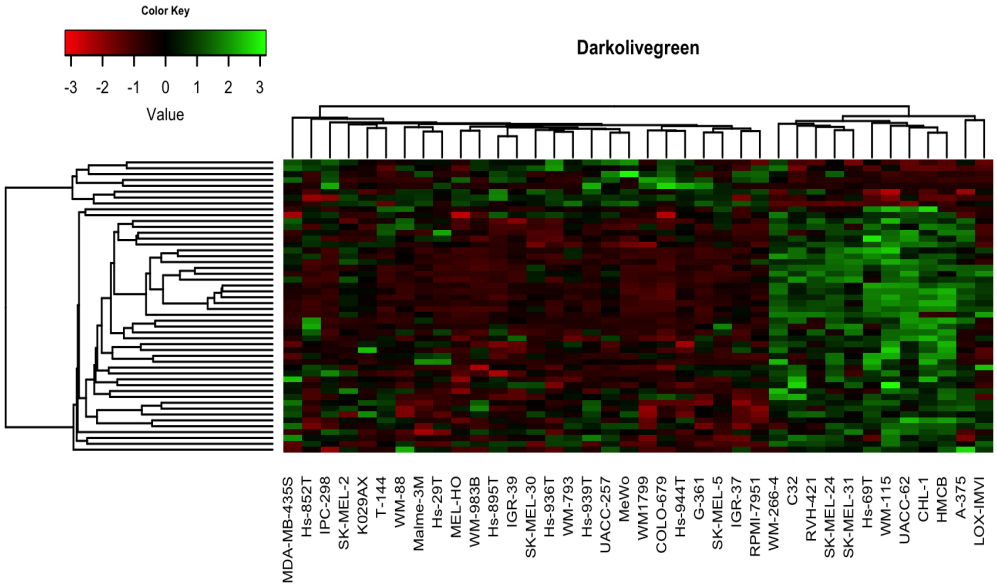

Figure B. 3: The heatmap plot of the darkolivegreen module

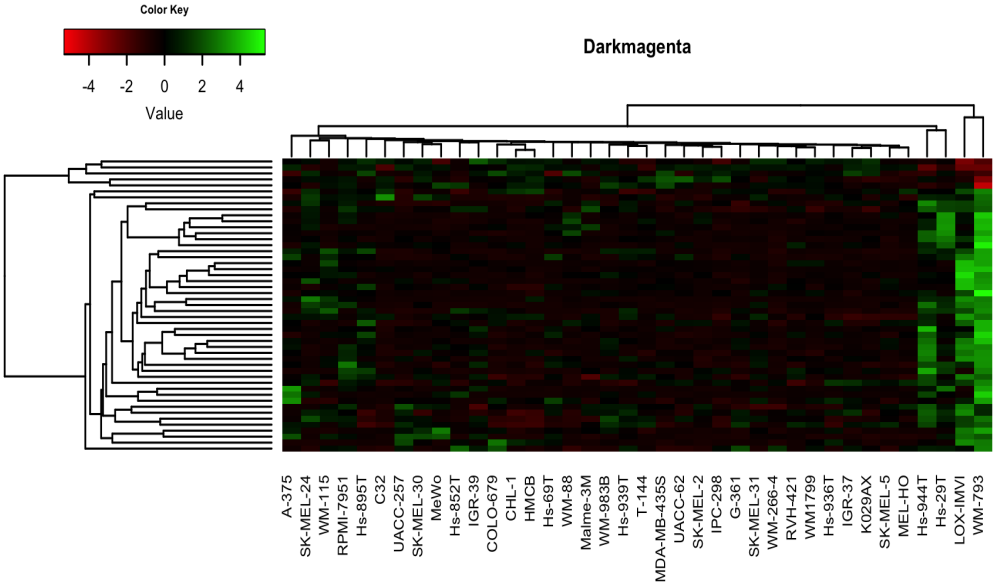

Figure B. 4: The heatmap plot of the darkmagenta module

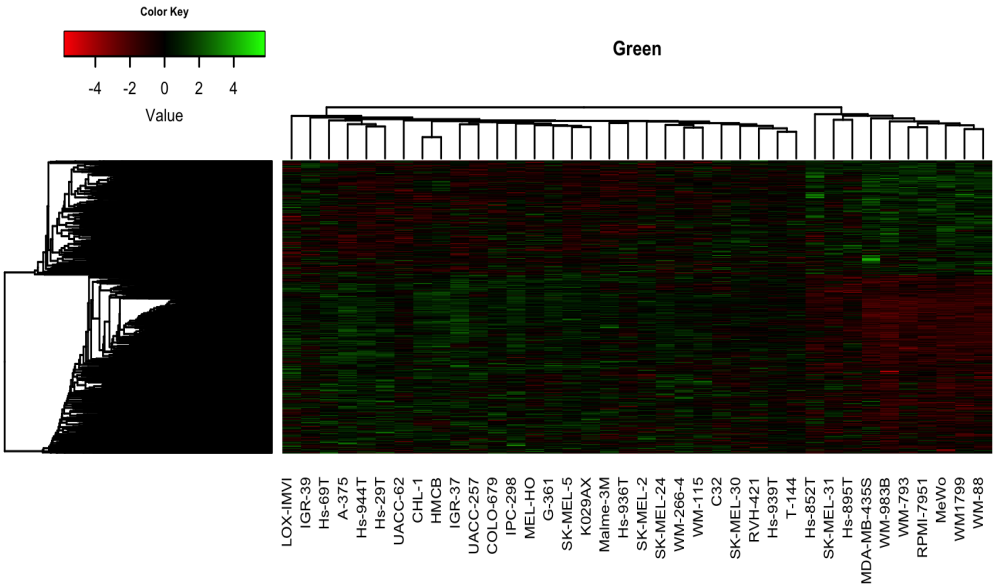

Figure B. 5: The heatmap plot of the green module

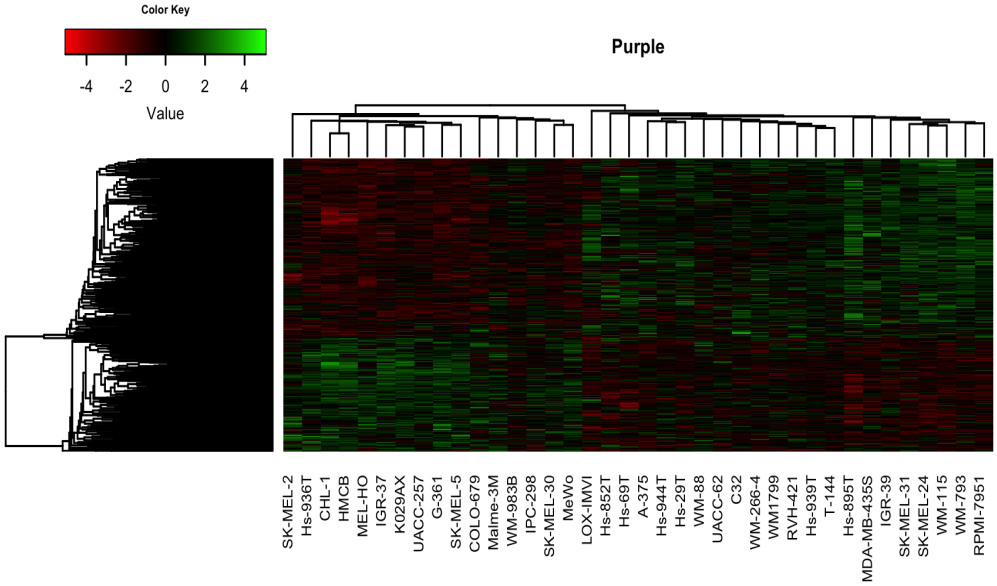

Figure B. 6: The heatmap plot of the purple module

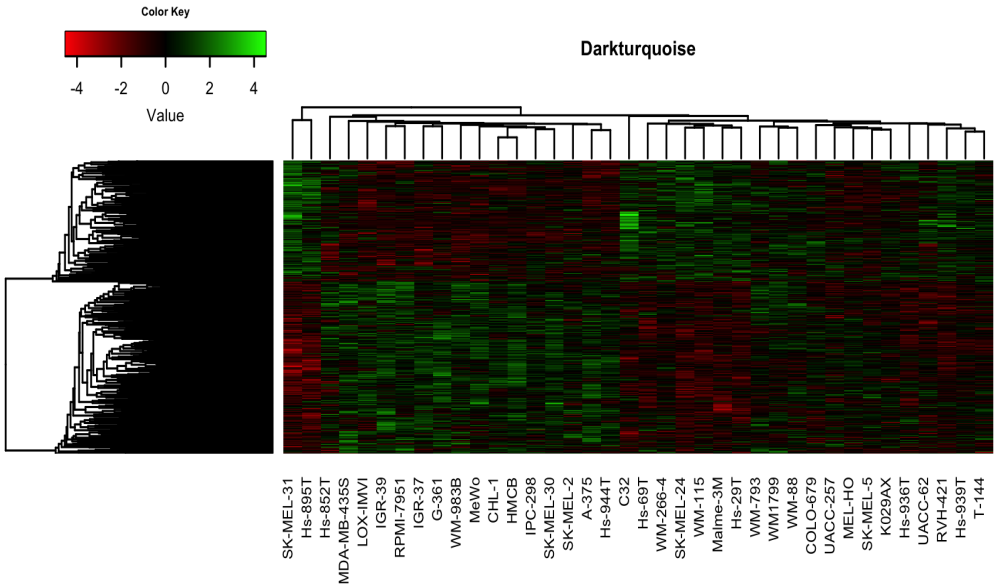

Figure B. 7: The heatmap plot of the darkturquoise module

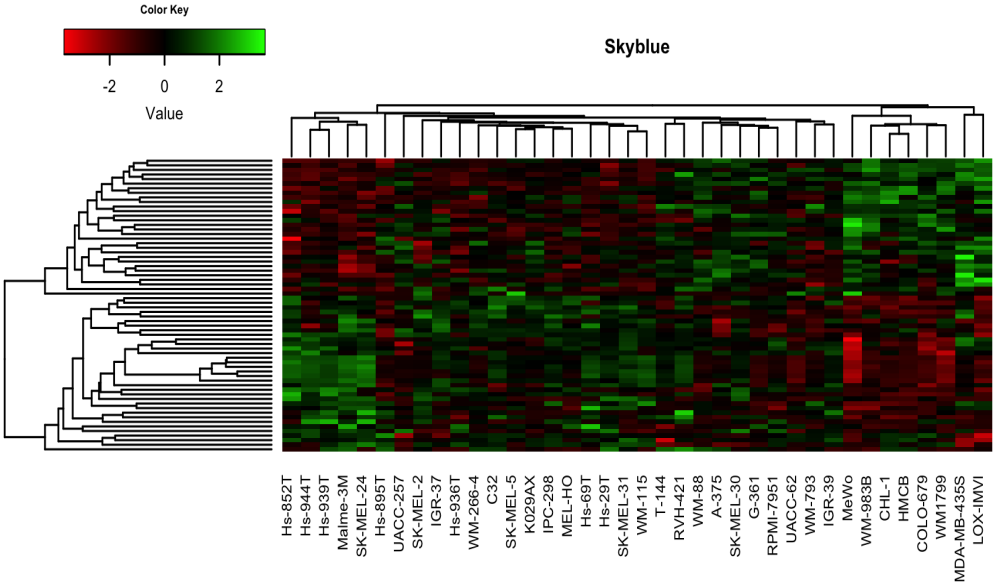

Figure B. 8: The heatmap plot of the skyblue module

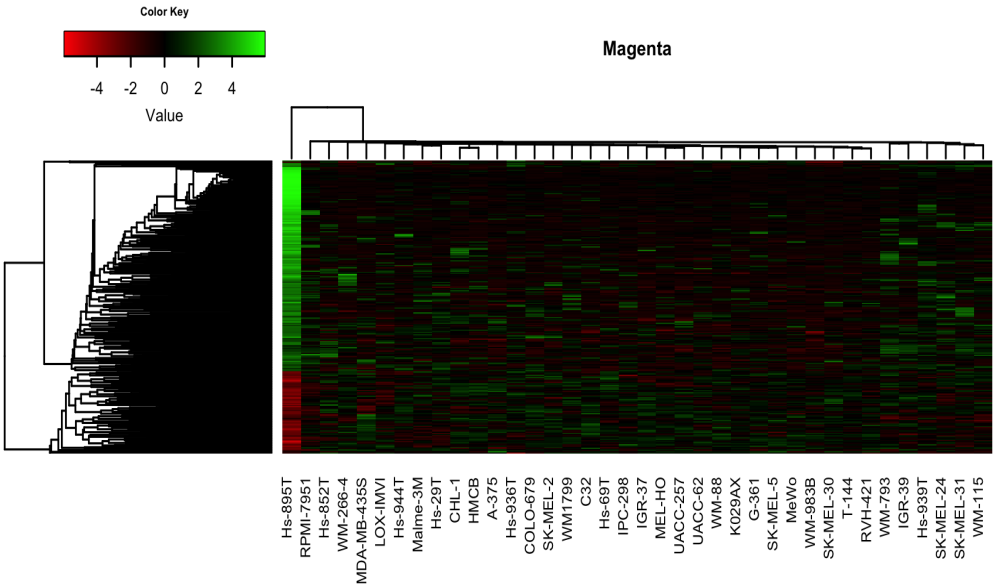

Figure B. 9: The heatmap plot of the magenta module

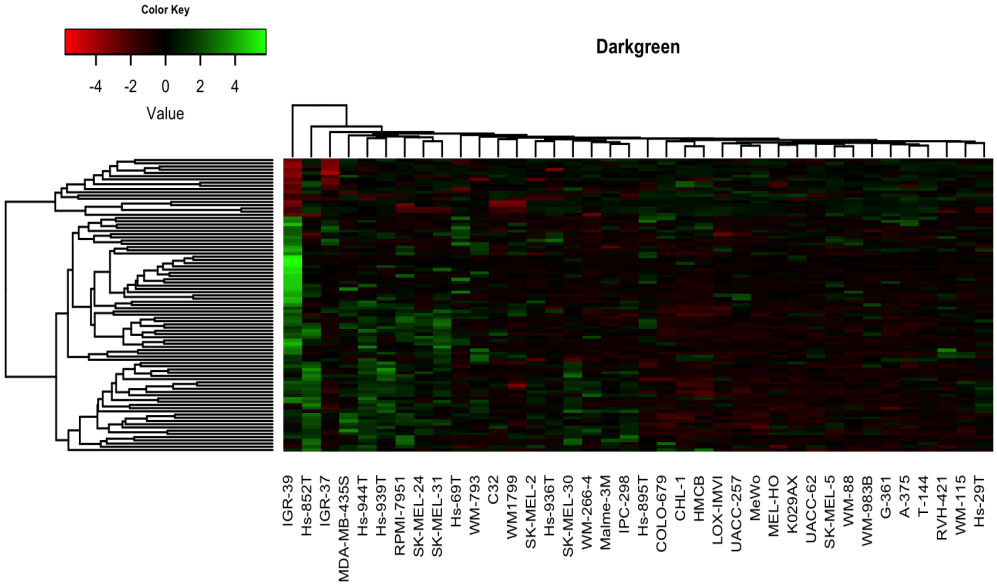

Figure B. 10: The heatmap plot of the darkgreen module

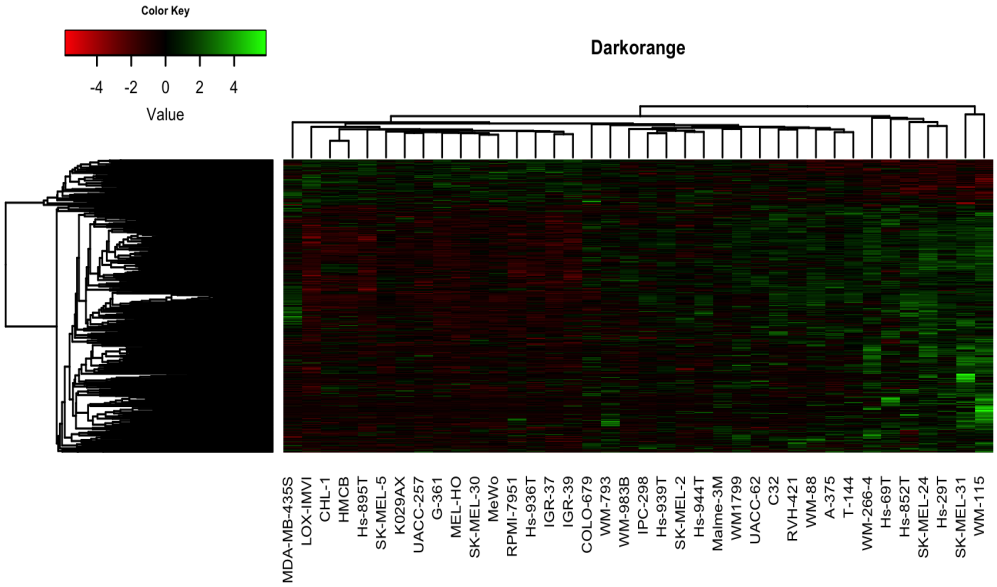

Figure B. 11: The heatmap plot of the darkorange module

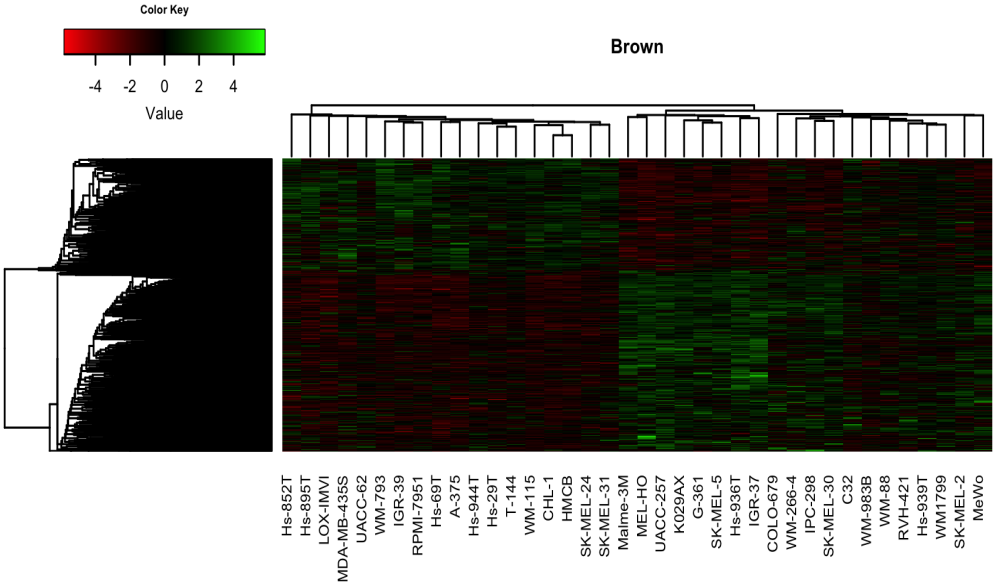

Figure B. 12: The heatmap plot of the brown module

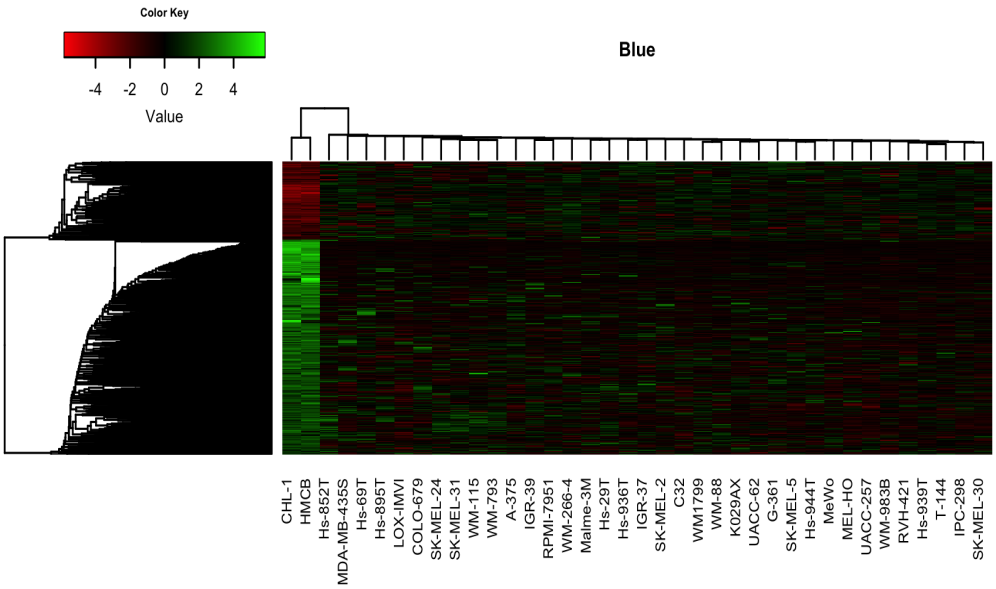

Figure B. 13: The heatmap plot of the blue module

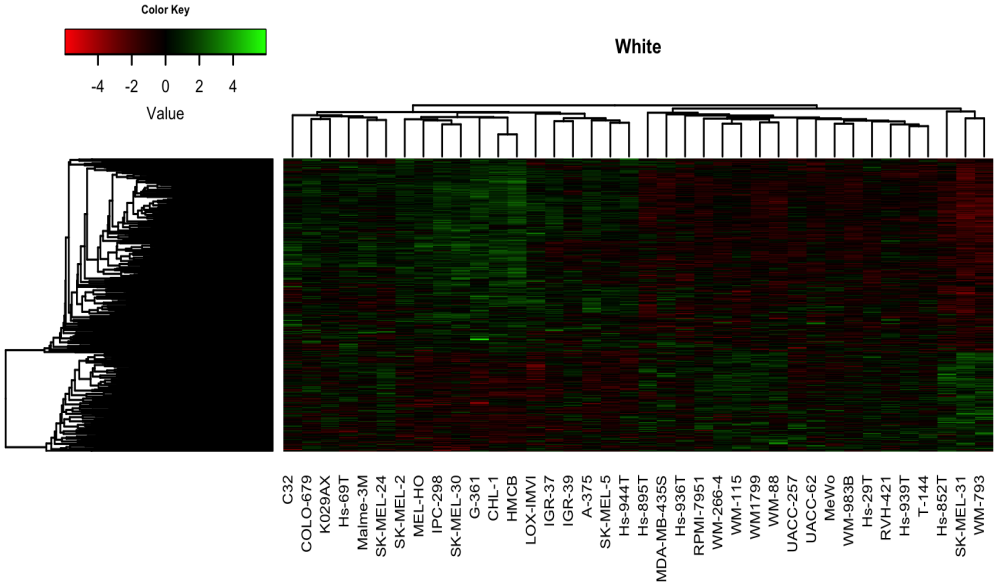

Figure B. 14: The heatmap plot of the white module

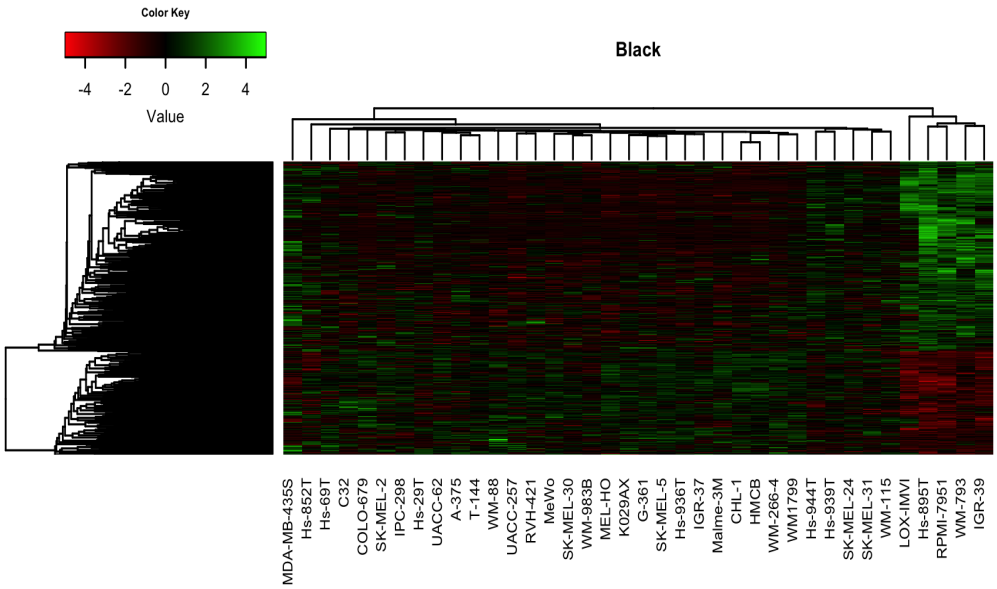

Figure B. 15: The heatmap plot of the black module

APPENDIX C

CORRELATION SCORES

Table C. 1: The modules eigengene values of which are significantly correlated to the CCLE IC50 values

Module Name	Drug Name	Pearson Correlation	P - value
Lightcyan	AZD0530	0.53	0.00064
Darkgreen	AZD0530	-0.42	0.0095
Darkturquoise	AZD0530	-0.40	0.014
Darkorange	AZD0530	0.34	0.031
Black	AZD0530	-0.34	0.033
Brown	AZD6244	-0.63	2.7×10^{-5}
Black	AZD6244	0.46	0.0036
Lightcyan	AZD6244	-0.45	0.0044
Blue	AZD6244	0.41	0.010
Purple	AZD6244	0.34	0.037
Magenta	AZD6244	0.33	0.038
Darkgreen	AZD6244	0.33	0.045
Blue	Erlotinib	-0.86	8.1×10^{-12}
White	Erlotinib	-0.54	0.00050
Purple	Erlotinib	0.49	0.0019
Saddlebrown	Erlotinib	-0.43	0.0075
Blue	Lapatinib	-0.69	1.5×10^{-6}
Purple	Lapatinib	0.43	0.0071
White	Lapatinib	-0.34	0.038

Table C. 1 (Continued)

Module Name	Drug Name	Pearson Correlation	P - value
Blue	PD0325901	0.44	0.0054
Lightcyan	PD0325901	-0.38	0.028
Brown	PD0325901	-0.35	0.030
Black	PF2341066	-0.50	0.0029
Lightcyan	PF2341066	0.47	0.0032
Magenta	PF2341066	-0.43	0.0064
Lightcyan	PLX4720	-0.49	0.0018
Brown	PLX4720	-0.48	0.0022
Black	PLX4720	0.48	0.0023
Darkmagenta	PLX4720	0.35	0.033
Darkgreen	PLX4720	0.34	0.038
White	Sorafenib	-0.51	0.0010
Purple	Sorafenib	0.50	0.0013
Darkturquoise	Sorafenib	-0.36	0.027
Darkorange	Sorafenib	0.34	0.036
Lightcyan	TAE684	0.59	9.1×10^{-5}
Brown	TAE684	0.50	0.0014
Black	TAE684	-0.39	0.015
Darkturquoise	TAE684	-0.39	0.016
Blue	TAE684	-0.33	0.043

Table C. 2: List of hub genes identified as candidate biomarkers for all the nine anti-cancer drugs

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
MAP3K14	AZD0530	Black	-0.51	0.0011	0.84
TRPM8	AZD0530	Darkorange	0.44	0.0058	0.73
ARHGAP15	AZD0530	Darkorange	0.43	0.0072	0.74
TSPAN13	AZD0530	Darkorange	0.37	0.024	0.80

Table C. 2 (Continued)

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
FHDC1	AZD0530	Darkorange	0.32	0.048	0.79
MRAP2	AZD0530	Darkturquoise	0.34	0.037	0.71
TRIM2	AZD0530	Lightcyan	0.73	1.7×10^{-7}	0.94
BAMBI	AZD0530	Lightcyan	0.45	0.0041	0.73
CHST6	AZD0530	Lightcyan	0.43	0.0068	0.74
FLJ42627	AZD0530	Lightcyan	0.40	0.014	0.83
RNF125	AZD0530	Lightcyan	0.34	0.035	0.90
DAAM1	AZD0530	Lightcyan	0.34	0.035	0.90
SLC23A2	AZD0530	Lightcyan	0.33	0.041	0.78
DPF3	AZD6244	Black	0.54	0.00045	-0.77
NTF3	AZD6244	Black	0.49	0.0020	-0.82
RGMB	AZD6244	Black	0.40	0.013	-0.71
MAP3K14	AZD6244	Black	0.39	0.015	0.84
MAN2A1	AZD6244	Black	0.34	0.038	0.72
ERBB2	AZD6244	Blue	0.53	0.00058	0.81
C8orf4	AZD6244	Blue	0.45	0.0046	-0.87
LOC153546	AZD6244	Blue	0.38	0.018	0.97
C1orf198	AZD6244	Blue	0.38	0.018	-0.71
PAOX	AZD6244	Blue	0.35	0.029	0.99
RAB38	AZD6244	Brown	-0.73	2.2×10^{-7}	0.91
APOE	AZD6244	Brown	-0.63	2.3×10^{-5}	0.88
C17orf58	AZD6244	Brown	-0.62	3.0×10^{-5}	0.70
TIMM50	AZD6244	Brown	-0.61	4.2×10^{-5}	0.90
D4S234E	AZD6244	Brown	-0.61	4.4×10^{-5}	0.93
LAMA1	AZD6244	Brown	-0.60	5.8×10^{-5}	0.80
PROS1	AZD6244	Brown	-0.60	6.9×10^{-5}	-0.75
TNFRSF14	AZD6244	Brown	-0.58	0.00015	0.77
GALNT3	AZD6244	Brown	-0.58	0.00015	0.92

Table C. 2 (Continued)

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
CPOX	AZD6244	Brown	-0.56	0.00023	0.74
ANKRD44	AZD6244	Brown	-0.56	0.00024	0.74
SETDB2	AZD6244	Brown	-0.56	0.00026	0.73
CHCHD6	AZD6244	Brown	-0.56	0.00030	-0.75
HELZ	AZD6244	Brown	-0.54	0.00048	-0.82
USP54	AZD6244	Brown	-0.54	0.00052	0.73
NPAT	AZD6244	Brown	-0.53	0.00058	0.84
SPRY1	AZD6244	Brown	-0.50	0.0013	0.81
GAPDHS	AZD6244	Brown	-0.50	0.0013	-0.76
SAMM50	AZD6244	Brown	-0.50	0.0016	0.79
SCUBE2	AZD6244	Brown	-0.49	0.0020	0.78
SLC18B1	AZD6244	Brown	-0.47	0.0027	0.70
TGFB1	AZD6244	Brown	0.45	0.0044	0.78
RAB17	AZD6244	Brown	-0.44	0.0058	0.73
ALDH3B2	AZD6244	Brown	-0.43	0.0065	-0.79
LOC100653010	AZD6244	Brown	-0.43	0.0067	-0.71
OMG	AZD6244	Brown	-0.42	0.0081	0.82
ST8SIA1	AZD6244	Brown	-0.42	0.0087	-0.80
P2RX4	AZD6244	Brown	-0.41	0.011	-0.71
C12orf66	AZD6244	Brown	-0.41	0.011	-0.80
RNMT	AZD6244	Brown	-0.41	0.011	0.76
TMEM87A	AZD6244	Brown	-0.38	0.020	0.72
GANC	AZD6244	Brown	-0.37	0.022	0.73
REPS1	AZD6244	Brown	-0.37	0.022	-0.71
CEACAM1	AZD6244	Brown	-0.37	0.024	0.82
TIMP2	AZD6244	Brown	-0.35	0.031	0.87
CTSH	AZD6244	Brown	-0.35	0.031	0.81
GMPR	AZD6244	Brown	-0.35	0.032	0.78

Table C. 2 (Continued)

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
WFDC1	AZD6244	Brown	-0.35	0.033	0.82
APOLD1	AZD6244	Brown	-0.35	0.034	-0.79
CTLA4	AZD6244	Brown	-0.34	0.035	0.72
BEST1	AZD6244	Brown	-0.33	0.042	0.84
CA14	AZD6244	Brown	-0.33	0.043	0.86
C15orf37	AZD6244	Brown	-0.33	0.045	0.73
RNF125	AZD6244	Lightcyan	-0.55	0.00038	0.90
DAAM1	AZD6244	Lightcyan	-0.38	0.017	0.90
CHST6	AZD6244	Lightcyan	-0.37	0.022	0.74
BAMBI	AZD6244	Lightcyan	-0.37	0.024	0.73
SLC23A2	AZD6244	Lightcyan	-0.34	0.038	0.78
APOD	AZD6244	Lightcyan	-0.32	0.047	0.77
MPRIP	AZD6244	Purple	0.51	0.0012	0.84
NGEF	AZD6244	Purple	0.47	0.0032	0.78
PTPN14	AZD6244	Purple	0.45	0.0045	0.84
IRS1	AZD6244	Purple	0.36	0.025	-0.73
THAP9	Erlotinib	Blue	-0.61	5.2×10^{-5}	-0.71
C8orf4	Erlotinib	Blue	-0.47	0.0028	-0.87
GAS2L1	Erlotinib	Blue	0.46	0.0033	0.77
LOC100505989	Erlotinib	Blue	-0.43	0.0064	-0.75
LRP5	Erlotinib	Blue	-0.43	0.0067	0.98
PAOX	Erlotinib	Blue	-0.43	0.0071	0.99
LRCH2	Erlotinib	Blue	-0.42	0.0089	0.77
C1orf198	Erlotinib	Blue	-0.38	0.020	-0.71
NAV3	Erlotinib	Purple	0.53	0.00057	0.71
NR3C1	Erlotinib	Purple	0.50	0.0014	0.87
ITGA5	Erlotinib	Purple	0.48	0.0025	0.71
WDFY2	Erlotinib	Purple	0.45	0.0047	0.71

Table C. 2 (Continued)

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
MAML2	Erlotinib	Purple	0.43	0.0064	0.76
LOC338758	Erlotinib	Purple	0.41	0.011	0.80
DHRS3	Erlotinib	Purple	0.40	0.013	0.75
THBS1	Erlotinib	Purple	0.39	0.016	-0.79
PITPNC1	Erlotinib	Purple	0.36	0.027	0.72
ARSB	Erlotinib	White	0.39	0.016	0.78
FAM172A	Erlotinib	White	0.35	0.033	0.78
THAP9	Lapatinib	Blue	-0.48	0.0023	-0.71
PAOX	Lapatinib	Blue	-0.40	0.013	0.99
LRP5	Lapatinib	Blue	-0.37	0.024	0.98
C1orf198	Lapatinib	Blue	-0.36	0.026	-0.71
PITPNC1	Lapatinib	Purple	0.43	0.0076	0.72
MAML2	Lapatinib	Purple	0.42	0.0080	0.76
ITGA5	Lapatinib	Purple	0.42	0.0082	0.71
NR3C1	Lapatinib	Purple	0.42	0.0091	0.87
WDFY2	Lapatinib	Purple	0.37	0.023	0.71
NAV3	Lapatinib	Purple	0.36	0.026	0.71
ETHE1	Lapatinib	Purple	0.36	0.027	0.80
LOC338758	Lapatinib	Purple	0.35	0.032	0.80
DHRS3	Lapatinib	Purple	0.34	0.037	0.75
FAM172A	Lapatinib	White	0.39	0.016	0.78
C8orf4	PD0325901	Blue	0.53	0.00064	-0.87
PAOX	PD0325901	Blue	0.37	0.023	0.99
ERBB2	PD0325901	Blue	0.34	0.035	0.81
NPAT	PD0325901	Brown	-0.48	0.0022	0.84
APOE	PD0325901	Brown	-0.42	0.0084	0.88
D4S234E	PD0325901	Brown	-0.40	0.013	0.93
LAMA1	PD0325901	Brown	-0.39	0.016	0.80

Table C. 2 (Continued)

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
SCUBE2	PD0325901	Brown	-0.38	0.017	0.78
TGFB1	PD0325901	Brown	0.38	0.018	0.78
PROS1	PD0325901	Brown	-0.36	0.025	-0.75
C15orf37	PD0325901	Brown	-0.36	0.026	0.73
TIMM50	PD0325901	Brown	-0.36	0.026	0.90
ALDH3B2	PD0325901	Brown	-0.36	0.026	-0.79
RNMT	PD0325901	Brown	-0.34	0.037	0.76
GALNT3	PD0325901	Brown	-0.33	0.042	0.92
TNFRSF14	PD0325901	Brown	-0.33	0.042	0.77
GAPDHS	PD0325901	Brown	-0.33	0.044	-0.76
ANKRD44	PD0325901	Brown	-0.33	0.046	0.74
RNF125	PD0325901	Lightcyan	-0.57	0.00018	0.9
SHC3	PF2341066	Black	-0.59	0.00010	0.86
TFPI2	PF2341066	Black	-0.33	0.043	0.94
SOX5	PF2341066	Lightcyan	0.47	0.0030	0.77
CHST6	PF2341066	Lightcyan	0.45	0.0047	0.74
BAMBI	PF2341066	Lightcyan	0.44	0.0057	0.73
CLCN7	PF2341066	Lightcyan	0.33	0.040	0.90
C7orf31	PF2341066	Magenta	0.44	0.0052	0.90
MAN2A1	PLX4720	Black	0.46	0.0038	0.72
SHC3	PLX4720	Black	0.43	0.0074	0.86
RGMB	PLX4720	Black	0.42	0.0095	-0.71
DPF3	PLX4720	Black	0.37	0.021	-0.77
NTF3	PLX4720	Black	0.35	0.030	-0.82
CYBRD1	PLX4720	Black	0.34	0.035	0.70
MAP3K14	PLX4720	Black	0.33	0.041	0.84
LAMA1	PLX4720	Brown	-0.57	0.00021	0.80
RAB38	PLX4720	Brown	-0.54	0.00041	0.91

Table C. 2 (Continued)

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
TIMM50	PLX4720	Brown	-0.54	0.00047	0.90
GAPDHS	PLX4720	Brown	-0.54	0.00049	-0.76
ANKRD44	PLX4720	Brown	-0.52	0.00073	0.74
SPRY1	PLX4720	Brown	-0.51	0.00098	0.81
LOC100653010	PLX4720	Brown	-0.51	0.0010	-0.71
D4S234E	PLX4720	Brown	-0.5	0.0013	0.93
GALNT3	PLX4720	Brown	-0.49	0.0017	0.92
PROS1	PLX4720	Brown	-0.49	0.0018	-0.75
SLC18B1	PLX4720	Brown	-0.48	0.0021	0.70
CHCHD6	PLX4720	Brown	-0.48	0.0021	-0.75
HELZ	PLX4720	Brown	-0.48	0.0022	-0.82
CPOX	PLX4720	Brown	-0.48	0.0022	0.74
APOE	PLX4720	Brown	-0.48	0.0023	0.88
ALDH3B2	PLX4720	Brown	-0.46	0.0036	-0.79
TMEM87A	PLX4720	Brown	-0.46	0.0038	0.72
C17orf58	PLX4720	Brown	-0.46	0.0039	0.70
RNMT	PLX4720	Brown	-0.43	0.0065	0.76
GANC	PLX4720	Brown	-0.43	0.0070	0.73
TGFB1	PLX4720	Brown	0.43	0.0077	0.78
TNFRSF14	PLX4720	Brown	-0.42	0.0079	0.77
SETDB2	PLX4720	Brown	-0.42	0.0093	0.73
ST8SIA1	PLX4720	Brown	-0.41	0.011	-0.8
SAMM50	PLX4720	Brown	-0.41	0.011	0.79
USP54	PLX4720	Brown	-0.38	0.018	0.73
C12orf66	PLX4720	Brown	-0.38	0.019	-0.80
CEACAM1	PLX4720	Brown	-0.37	0.023	0.82
NPAT	PLX4720	Brown	-0.37	0.023	0.84
SCUBE2	PLX4720	Brown	-0.37	0.024	0.78

Table C. 2 (Continued)

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
C15orf37	PLX4720	Brown	-0.37	0.024	0.73
CTSH	PLX4720	Brown	-0.36	0.025	0.81
OMG	PLX4720	Brown	-0.36	0.027	0.82
TIMP2	PLX4720	Brown	-0.35	0.030	0.87
RAB17	PLX4720	Brown	-0.35	0.030	0.73
RNF125	PLX4720	Lightcyan	-0.51	0.0011	0.90
BAMBI	PLX4720	Lightcyan	-0.39	0.014	0.73
APOD	PLX4720	Lightcyan	-0.39	0.016	0.77
DAAM1	PLX4720	Lightcyan	-0.38	0.018	0.90
CHST6	PLX4720	Lightcyan	-0.37	0.021	0.74
HIVEP3	Sorafenib	Darkorange	0.37	0.023	0.77
ARHGAP15	Sorafenib	Darkorange	0.33	0.044	0.74
MPRIP	Sorafenib	Purple	0.51	0.00094	0.84
MAML2	Sorafenib	Purple	0.50	0.0015	0.76
PTPN14	Sorafenib	Purple	0.46	0.0035	0.84
NR3C1	Sorafenib	Purple	0.45	0.0048	0.87
WDFY2	Sorafenib	Purple	0.42	0.0087	0.71
LOC338758	Sorafenib	Purple	0.42	0.0088	0.80
ETHE1	Sorafenib	Purple	0.41	0.010	0.80
STARD13	Sorafenib	Purple	0.38	0.019	0.73
DHRS3	Sorafenib	Purple	0.38	0.020	0.75
ARHGAP28	Sorafenib	Purple	-0.36	0.027	-0.71
ITGA5	Sorafenib	Purple	0.35	0.032	0.71
MTMR11	Sorafenib	Purple	0.35	0.032	0.86
PITPNC1	Sorafenib	Purple	0.34	0.036	0.72
NAV3	Sorafenib	Purple	0.34	0.036	0.71
HOXB5	Sorafenib	Purple	0.32	0.050	0.70
FAM172A	Sorafenib	White	0.59	0.00011	0.78

Table C. 2 (Continued)

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
ARSB	Sorafenib	White	0.34	0.037	0.78
MAP3K14	TAE684	Black	-0.37	0.022	0.84
RGMB	TAE684	Black	-0.32	0.047	-0.71
PAOX	TAE684	Blue	-0.36	0.025	0.99
LOC153546	TAE684	Blue	-0.35	0.032	0.97
C8orf4	TAE684	Blue	-0.35	0.034	-0.87
ST8SIA1	TAE684	Brown	0.61	5.3×10^{-5}	-0.80
ATP6V0A2	TAE684	Brown	0.57	0.00020	0.80
HEXA	TAE684	Brown	0.56	0.00025	-0.76
LAMA1	TAE684	Brown	0.55	0.00040	0.80
D4S234E	TAE684	Brown	0.53	0.00069	0.93
PROS1	TAE684	Brown	0.52	0.00086	-0.75
APOE	TAE684	Brown	0.52	0.00087	0.88
GALNT3	TAE684	Brown	0.51	0.00095	0.92
SCUBE2	TAE684	Brown	0.51	0.00097	0.78
TIMP2	TAE684	Brown	0.50	0.0015	0.87
ANKRD44	TAE684	Brown	0.49	0.0018	0.74
RAB17	TAE684	Brown	0.48	0.0022	0.73
CTSH	TAE684	Brown	0.47	0.0028	0.81
TNFRSF14	TAE684	Brown	0.46	0.0033	0.77
OMG	TAE684	Brown	0.46	0.0040	0.82
CEACAM1	TAE684	Brown	0.44	0.0059	0.82
P2RX4	TAE684	Brown	0.43	0.0078	-0.71
LDB3	TAE684	Brown	0.42	0.0088	0.78
C17orf58	TAE684	Brown	0.41	0.011	0.70
CELSR1	TAE684	Brown	0.40	0.012	-0.78
GAPDHS	TAE684	Brown	0.40	0.013	-0.76
SPRY1	TAE684	Brown	0.40	0.014	0.81

Table C. 2 (Continued)

Gene Name	Drug Name	Module Name	Pearson Correlation	P - value	Intramodular Connectivity
ACSBG1	TAE684	Brown	0.39	0.016	0.75
RAB38	TAE684	Brown	0.38	0.018	0.91
CPOX	TAE684	Brown	0.38	0.019	0.74
TIMM50	TAE684	Brown	0.38	0.020	0.90
NPAT	TAE684	Brown	0.37	0.021	0.84
LOC374443	TAE684	Brown	0.37	0.023	0.87
USP54	TAE684	Brown	0.35	0.032	0.73
LOC150568	TAE684	Brown	0.35	0.033	-0.75
SETDB2	TAE684	Brown	0.34	0.037	0.73
PLEKHO2	TAE684	Brown	0.34	0.040	-0.88
CA14	TAE684	Brown	0.33	0.041	0.86
ALDH3B2	TAE684	Brown	0.33	0.042	-0.79
C22orf25	TAE684	Brown	0.32	0.048	-0.78
MRAP2	TAE684	Darkturquoise	0.44	0.0056	0.71
MIPEPP3	TAE684	Lightcyan	0.56	0.00023	-0.73
BAMBI	TAE684	Lightcyan	0.56	0.00029	0.73
SLC23A2	TAE684	Lightcyan	0.45	0.0048	0.78
MFSD12	TAE684	Lightcyan	0.39	0.015	0.72
FLJ42627	TAE684	Lightcyan	0.39	0.015	0.83
CHST6	TAE684	Lightcyan	0.39	0.017	0.74
APOD	TAE684	Lightcyan	0.35	0.029	0.77
CLCN7	TAE684	Lightcyan	0.35	0.029	0.90
RNF125	TAE684	Lightcyan	0.35	0.030	0.90

Table C. 3: The correlation scores determined for the best performing single candidate biomarkers when the CCLE IC50 values are used for prediction

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
BAMBI	Lightcyan	AZD0530	0.14	0.39	0.24	0.14
RAB38	Brown	AZD6244	0.24	0.15	0.15	0.38
FAM172A	White	Erlotinib	0.18	0.29	0.32	0.052
NAV3	Purple	Lapatinib	0.13	0.43	0.32	0.048
RNF125	Lightcyan	PD0325901	0.44	0.0052	0.21	0.21
BAMBI	Lightcyan	PF2341066	0.018	0.92	0.070	0.68
BAMBI	Lightcyan	PLX4720	0.24	0.15	0.22	0.19
HIVEP3	Darkorange	Sorafenib	0.099	0.56	0.11	0.50
PROS1	Brown	TAE684	0.31	0.060	0.26	0.11

Table C. 4: The correlation scores determined for the best performing single candidate biomarkers after the censored IC50 values are excluded from the CCLE drug sensitivity data

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
DAAM1	Lightcyan	AZD6244	0.78	3.6×10^{-6}	0.55	0.0053
RNF125	Lightcyan	PD0325901	0.27	0.13	0.27	0.14
NTF3	Black	TAE684	0.67	0.00045	0.40	0.057

Table C. 5: The correlation scores determined for the best performing single candidate biomarkers when the CCLE Activity Area values are used for prediction

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
FLJ42627	Lightcyan	AZD0530	0.57	0.00019	0.47	0.0031
PAOX	Blue	AZD6244	0.62	0.000030	0.54	0.00041

Table C. 5 (Continued)

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
LRP5	Blue	Erlotinib	0.69	1.8×10^{-6}	0.54	0.00040
NAV3	Purple	Lapatinib	0.63	2.3×10^{-5}	0.52	0.00089
PAOX	Blue	PD0325901	0.62	3.6×10^{-5}	0.56	0.00031
CLCN7	Lightcyan	PF2341066	0.45	0.0049	0.32	0.052
APOD	Lightcyan	PLX4720	0.60	6.0×10^{-5}	0.59	0.00012
HOXB5	Purple	Sorafenib	0.52	0.00080	0.48	0.0024
CLCN7	Lightcyan	TAE684	0.53	0.00062	0.42	0.0090

Table C. 6: The correlation scores determined for the best performing single candidate biomarkers after the CGP IC50 values are censored to the maximum screening concentration of the drugs in the CCLE study

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
DAAM1	Lightcyan	AZD6244	0.058	0.77	-0.091	0.63
PAOX	Blue	PD0325901	0.21	0.24	0.29	0.092
BAMBI	Lightcyan	PLX4720	-0.085	0.62	-0.098	0.57

Table C. 7: The correlation scores determined for the best performing single candidate biomarkers after the CGP IC50 values are censored to the maximum screening concentration of the drugs in the CGP study

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
PAOX	Blue	AZD6244	0.19	0.30	0.20	0.29
APOE	Brown	PD0325901	0.11	0.55	0.20	0.25
BAMBI	Lightcyan	PLX4720	-0.075	0.66	-0.098	0.57

Table C. 8: The correlation scores determined for the best performing single candidate biomarkers after all the extrapolated and censored IC50 values are excluded from drug sensitivity data

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
PAOX	Blue	AZD6244	0.38	0.17	0.34	0.22
SCUBE2	Brown	PD0325901	-0.10	0.62	-0.13	0.54
BAMBI	Lightcyan	PLX4720	-0.026	0.92	-0.23	0.35

Table C. 9: The correlation scores determined for the best performing single candidate biomarkers when Activity Area values are used for prediction

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
FLJ42627	Lightcyan	AZD0530	0.12	0.68	-0.0022	0.99
PAOX	Blue	AZD6244	0.22	0.23	0.15	0.43
LRP5	Blue	Erlotinib	-0.35	0.27	-0.27	0.39
NAV3	Purple	Lapatinib	-0.32	0.31	-0.31	0.32
PAOX	Blue	PD0325901	0.17	0.35	0.22	0.21
CLCN7	Lightcyan	PF2341066	-0.17	0.56	0.015	0.96
APOD	Lightcyan	PLX4720	0.076	0.66	-0.042	0.81
ETHE1	Purple	Sorafenib	0.20	0.50	0.016	0.96
PROS1	Brown	TAE684	0.37	0.19	0.33	0.25

Table C. 10: The correlation scores determined for the best performing single candidate biomarkers when the extrapolated IC50 values are included in the drug sensitivity data

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
BAMBI	Lightcyan	AZD0530	0.87	2.3×10^{-6}	0.34	0.16
RAB38	Brown	AZD6244	0.74	0.00046	0.51	0.032

Table C. 10 (Continued)

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
FAM172A	White	Erlotinib	0.17	0.50	0.20	0.42
NAV3	Purple	Lapatinib	0.51	0.031	0.54	0.022
RNF125	Lightcyan	PD0325901	0.60	0.0085	0.41	0.094
BAMBI	Lightcyan	PF2341066	0.35	0.15	0.20	0.42
BAMBI	Lightcyan	PLX4720	0.59	0.011	0.53	0.025
HIVEP3	Darkorange	Sorafenib	0.19	0.46	0.24	0.33
PROS1	Brown	TAE684	0.64	0.0039	0.47	0.049

Table C. 11: The correlation scores determined for the best performing single candidate biomarkers when the extrapolated IC50 values are removed from the drug sensitivity data

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
BAMBI	Lightcyan	AZD6244	0.27	0.35	0.64	0.015
APOE	Brown	PD0325901	0.54	0.040	0.032	0.91
PROS1	Brown	TAE684	0.64	0.034	0.51	0.11

Table C. 12: The correlation scores determined for the best performing single candidate biomarkers when Activity Area values are used for prediction

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
BAMBI	Lightcyan	AZD0530	0.60	0.0088	0.31	0.21
MPRIP	Purple	AZD6244	0.69	0.0017	0.69	0.0019
NAV3	Purple	Erlotinib	0.16	0.52	0.18	0.47
PAOX	Blue	Lapatinib	0.41	0.087	0.22	0.38
RNF125	Blue	PD0325901	0.51	0.032	0.33	0.18

Table C. 12 (Continued)

Gene Name	Module Name	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
TFPI2	Black	PF2341066	0.38	0.12	0.44	0.070
BAMBI	Lightcyan	PLX4720	0.30	0.23	0.41	0.093
MPRIP	Purple	Sorafenib	0.68	0.0021	0.82	2.4×10^{-5}
BAMBI	Lightcyan	TAE684	0.31	0.21	0.28	0.26

Table C. 13: The correlation scores determined for the best performing combined candidate biomarkers when the CGP gene expression data is used for IC50 prediction

Gene Combination	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
BAMBI - MAP3K14	AZD0530	0.98	6.2×10^{-13}	0.65	0.0035
BAMBI - MAP3K14 - TSPAN13	AZD0530	0.97	3.1×10^{-11}	0.60	0.0086
BAMBI - MPRIP	AZD6244	0.81	5.6×10^{-5}	0.81	4.8×10^{-5}
ERBB2 - MPRIP - RAB38	AZD6244	0.86	4.1×10^{-6}	0.92	7.1×10^{-8}
DAAM1 - ERBB2 - MPRIP - RAB38	AZD6244	0.90	4.5×10^{-7}	0.86	5.0×10^{-6}
DAAM1 - ERBB2 - MPRIP - NTF3 - RAB38	AZD6244	0.87	3.3×10^{-6}	0.84	1.5×10^{-5}
ITGA5 - LRP5	Erlotinib	0.78	0.00014	0.57	0.013
FAM172A - ITGA5 - LRP5	Erlotinib	0.65	0.0037	0.34	0.16
ITGA5 - LRP5	Lapatinib	0.82	$2.7 \text{ X } 10^{-5}$	0.60	0.0091
FAM172A - ITGA5 - LRP5	Lapatinib	0.61	0.0066	0.27	0.29
C8orf4 - RNF125	PD0325901	0.65	0.0037	0.43	0.074
APOE - ERBB2 - RNF125	PD0325901	0.59	0.0093	0.27	0.28
BAMBI - TFPI2	PF2341066	0.82	$2.7 \text{ X } 10^{-5}$	0.65	0.0037

Table C. 13 (Continued)

Gene Combination	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
BAMBI - PROS1	PLX4720	0.55	0.017	0.56	0.017
APOE - DAAM1 - MAP3K14	PLX4720	0.70	0.0012	0.55	0.018
ETHE1 - FAM172A	Sorafenib	0.75	0.00035	0.36	0.14
ETHE1 - FAM172A - HIVEP3	Sorafenib	0.70	0.0012	0.24	0.34
MSFD12 - PROS1	TAE684	0.79	0.00010	0.68	0.0018
MAP3K14 - MFSD12 - PROS1	TAE684	0.79	8.7×10^{-5}	0.77	0.00017
BAMBI - C8orf4 - MAP3K14 - PROS1	TAE684	0.83	2.4×10^{-5}	0.71	0.00093

Table C. 14: The correlation scores determined for the identified best performing combined candidate biomarkers when the CGP gene expression data is used for IC50 prediction after removing the censored IC50 values

Gene Combination	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
MPRIP - SLC23A2	AZD6244	0.85	0.00013	0.57	0.035
BAMBI - ERBB2 - MPRIP	AZD6244	0.84	0.00017	0.90	5.1×10^{-4}
DAAM1 - ERBB2 - MPRIP - SLC23A2	AZD6244	0.90	1.2×10^{-5}	0.57	0.037
BAMBI - ERBB2 - MAP3K14 - MPRIP - SLC23A2	AZD6244	0.89	2.5×10^{-5}	0.70	0.0069
APOE - ERBB2	PD0325901	0.80	0.00033	0.26	0.35
APOE - ERBB2 - RNF125	PD0325901	0.81	0.00022	0.42	0.12
BAMBI - PROS1	TAE684	0.74	0.0088	0.52	0.11
BAMBI - MAP3K14 - PROS1	TAE684	0.88	0.00038	0.71	0.019

Table C. 14 (Continued)

Gene Combination	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
BAMBI - C8orf4 - MAP3K14 - PROS1	TAE684	0.82	0.0018	0.66	0.0031

Table C. 15: The correlation scores determined for the identified best performing combined candidate biomarkers when the CGP gene expression data is used for Activity Area prediction

Gene Combination	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
BAMBI - MAP3K14	AZD0530	0.56	0.015	0.49	0.038
BAMBI - MAP3K14 - TSPAN13	AZD0530	0.59	0.010	0.48	0.043
ERBB2 - MPRIP	AZD6244	0.75	0.00029	0.81	6.2×10^{-5}
BAMBI - MPRIP - PAOX	AZD6244	0.81	5.5×10^{-5}	0.80	0.00010
BAMBI - MAP3K14 - MPRIP - PROS1	AZD6244	0.88	1.5×10^{-6}	0.91	8.5×10^{-7}
BAMBI - MAP3K14 - MPRIP - PROS1 - RAB38	AZD6244	0.91	1.2×10^{-7}	0.95	2.7×10^{-6}
LRP5 - NAV3	Erlotinib	0.40	0.099	0.25	0.32
FAM172A - GAS2L1 - NAV3	Erlotinib	0.11	0.67	0.094	0.71
LRP5 - PAOX	Lapatinib	0.65	0.0034	0.44	0.067
FAM172A - LRP5 - NAV3	Lapatinib	0.53	0.024	0.11	0.65
ERBB2 - RNF125	PD0325901	0.66	0.0028	0.58	0.013
ERBB2 - NPAT - RNF125	PD0325901	0.60	0.0079	0.55	0.020
BAMBI - TFPI2	PF2341066	0.58	0.011	0.55	0.019
BAMBI - MAP3K14	PLX4720	0.49	0.038	0.53	0.027

Table C. 15 (Continued)

Gene Combination	Drug Name	Pearson Correlation	Pearson Correlation (p-value)	Spearman Correlation	Spearman Correlation (p-value)
BAMBI - MAP3K14 - PROS1	PLX4720	0.68	0.0018	0.58	0.014
HIVEP3 - MPRIP	Sorafenib	0.66	0.0027	0.53	0.025
FAM172A - HIVEP3 - MPRIP	Sorafenib	0.36	0.15	0.35	0.15
MFSD12 - PROS1	TAE684	0.54	0.020	0.36	0.14
MAP3K14 - MFSD12 - PROS1	TAE684	0.70	0.0012	0.55	0.020
MAP3K14 - MFSD12 - PAOX - PROS1	TAE684	0.80	7.7×10^{-5}	0.77	0.00028

APPENDIX D

THE TREND OF PREDICTIVE POWER WITH VARYING NUMBER OF COMBINATIONS

Table D. 1: Predictive power of the best performing single/combined candidate biomarkers when the CGP expression data is used for IC50 prediction

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
BAMBI	AZD0530	0.42	5.28	1.19	14.83
BAMBI - MAP3K14	AZD0530	0.20	2.50	0.33	4.13
BAMBI - MAP3K14 - TSPAN13	AZD0530	0.27	3.38	0.53	6.63
RAB38	AZD6244	1.86	23.25	2.47	30.88
BAMBI - MPRIP	AZD6244	1.64	20.50	1.96	24.50
ERBB2 - MPRIP - RAB38	AZD6244	1.50	18.75	1.61	20.13
DAAM1 - ERBB2 - MPRIP - RAB38	AZD6244	1.24	15.50	1.46	18.25
DAAM1 - ERBB2 - MPRIP - NTF3 - RAB38	AZD6244	1.13	14.13	1.63	20.38
FAM172A	Erlotinib	1.36	17.10	1.38	17.25
ITGA5 - LRP5	Erlotinib	0.74	9.25	0.86	10.75
FAM172A - ITGA5 - LRP5	Erlotinib	0.93	11.63	1.11	13.88
NAV3	Lapatinib	1.61	20.13	1.80	22.50
ITGA5 - LRP5	Lapatinib	1.19	14.88	1.12	14.00
FAM172A - ITGA5 - LRP5	Lapatinib	0.84	10.50	1.59	19.88
RNF125	PD0325901	2.32	28.94	2.57	32.13

Table D. 1 (Continued)

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
C8orf4 - RNF125	PD0325901	2.29	28.63	2.40	30.00
APOE - ERBB2 - RNF125	PD0325901	1.95	24.38	2.37	29.63
BAMBI	PF2341066	1.26	15.75	1.27	15.88
BAMBI - TFPI2	PF2341066	0.84	10.50	0.80	10.00
BAMBI	PLX4720	2.93	36.63	3.03	37.88
BAMBI - PROS1	PLX4720	2.48	31.00	2.88	36.00
APOE - DAAM1 - MAP3K14	PLX4720	2.32	29.00	2.52	31.50
HIVEP3	Sorafenib	0.72	9.00	0.77	9.63
ETHE1 - FAM172A	Sorafenib	0.085	1.06	0.55	6.88
ETHE1 - FAM172A - HIVEP3	Sorafenib	0.061	0.76	0.56	7.00
PROS1	TAE684	1.39	17.38	1.69	21.13
MSFD12 - PROS1	TAE684	1.22	15.25	1.30	16.25
MAP3K14 - MFSD12 - PROS1	TAE684	1.05	13.13	1.29	16.13
BAMBI - C8orf4 - MAP3K14 - PROS1	TAE684	0.69	8.63	1.19	14.88

Table D. 2: Predictive power of the best performing candidate biomarkers when the CGP expression data is used for IC50 prediction after removing the censored IC50 values from the CCLE drug sensitivity data

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
BAMBI	AZD6244	0.81	33.61	0.78	32.37
MPRIP - SLC23A2	AZD6244	0.50	20.71	0.46	19.09
BAMBI - ERBB2 - MPRIP	AZD6244	0.50	20.75	0.43	17.84
DAAM1 - ERBB2 - MPRIP - SLC23A2	AZD6244	0.40	16.60	0.38	15.77

Table D. 2 (Continued)

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
BAMBI - ERBB2 - MAP3K14 - MPRIP - SLC23A2	AZD6244	0.28	11.62	0.39	16.18
APOE	PD0325901	0.25	25.00	0.26	26.00
APOE - ERBB2	PD0325901	0.045	4.50	0.16	16.00
APOE - ERBB2 - RNF125	PD0325901	6.8×10^{-5}	0.0068	0.16	16.00
PROS1	TAE684	1.15	15.91	1.23	15.55
BAMBI - PROS1	TAE684	0.74	10.24	1.05	14.52
BAMBI - MAP3K14 - PROS1	TAE684	0.00044	0.0061	0.93	12.86
BAMBI - C8orf4 - MAP3K14 - PROS1	TAE684	0.56	7.75	0.88	12.17

Table D. 3: Predictive power of the best performing candidate biomarkers when the CGP expression data is used for Activity Area prediction

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
BAMBI	AZD0530	0.30	29.53	0.29	29.12
BAMBI - MAP3K14	AZD0530	0.18	17.78	0.30	30.26
BAMBI - MAP3K14 - TSPAN13	AZD0530	0.25	25.39	0.28	28.16
MPRIP	AZD6244	0.26	25.95	0.22	21.77
ERBB2 - MPRIP	AZD6244	0.19	19.32	0.19	19.01
BAMBI - MPRIP - PAOX	AZD6244	0.17	17.74	0.18	17.85
BAMBI - MAP3K14 - MPRIP - PROS1	AZD6244	0.058	5.76	0.14	14.47
BAMBI - MAP3K14 - MPRIP - PROS1 - RAB38	AZD6244	0.048	4.75	0.13	13.08
NAV3	Erlotinib	0.24	24.40	0.27	27.48

Table D. 3 (Continued)

Biomarker Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
LRP5 - NAV3	Erlotinib	0.22	21.73	0.26	26.28
FAM172A - GAS2L1 - NAV3	Erlotinib	0.16	15.60	0.29	29.29
PAOX	Lapatinib	0.14	13.82	0.24	24.00
LRP5 - PAOX	Lapatinib	0.16	15.66	0.21	20.82
FAM172A - LRP5 - NAV3	Lapatinib	0.17	16.79	0.24	23.82
RNF125	PD0325901	0.25	25.19	0.27	27.17
ERBB2 - RNF125	PD0325901	0.22	21.57	0.23	22.96
ERBB2 - NPAT - RNF125	PD0325901	0.030	2.99	0.26	26.26
TFPI2	PF2341066	0.31	30.56	0.28	29.57
BAMBI - TFPI2	PF2341066	0.21	20.69	0.25	25.18
BAMBI	PLX4720	0.30	29.97	0.32	32.00
BAMBI - MAP3K14	PLX4720	0.25	24.57	0.30	30.06
BAMBI - MAP3K14 - PROS1	PLX4720	0.20	19.56	0.24	24.43
MPRIP	Sorafenib	0.21	20.65	0.21	21.03
HIVEP3 - MPRIP	Sorafenib	0.14	14.27	0.20	20.00
FAM172A - HIVEP3 - MPRIP	Sorafenib	0.096	9.61	0.27	26.86
BAMBI	TAE684	0.24	23.80	0.24	23.72
MFSD12 - PROS1	TAE684	0.15	14.94	0.21	21.39
MAP3K14 - MFSD12 - PROS1	TAE684	0.15	14.67	0.18	18.02
MAP3K14 - MFSD12 - PAOX - PROS1	TAE684	0.13	13.38	0.15	15.13

Table D. 4: The RMSE values determined for the best performing combined candidate biomarkers when the censored IC50 values are included in the CCLE drug sensitivity data

Gene Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
BAMBI - MAP3K14 (lightcyan - black)	AZD0530	0.20	2.50	0.33	4.13
BAMBI - MAP3K14 - TSPAN13 (lightcyan - black - darkorange)	AZD0530	0.27	3.38	0.53	6.63
BAMBI - MPRIP (lightcyan - purple)	AZD6244	1.64	20.50	1.96	24.50
ERBB2 - MPRIP - RAB38 (blue - purple - brown)	AZD6244	1.50	18.75	1.61	20.13
DAAM1 - ERBB2 - MPRIP - RAB38 (lightcyan - blue - purple - brown)	AZD6244	1.24	15.50	1.46	18.25
BAMBI - IRS1 - MPRIP - NTF3 - PAOX (Lightcyan - brown - purple - black - blue)	AZD6244	0.00068	0.0085	2.61	32.63
ITGA5 - LRP5 (purple - blue)	Erlotinib	0.74	9.25	0.86	10.75
FAM172A - ITGA5 - LRP5 (white - purple - blue)	Erlotinib	0.93	11.63	1.11	13.88
ITGA5 - LRP5 (purple - blue)	Lapatinib	1.19	14.88	1.12	14.00
FAM172A - ITGA5 - LRP5 (white - purple - blue)	Lapatinib	0.84	10.50	1.59	19.88
C8orf4 - RNF125 (blue - lightcyan)	PD0325901	2.29	28.63	2.40	30.00
APOE - ERBB2 - RNF125 (brown - blue - lightcyan)	PD0325901	1.95	24.38	2.37	29.63
BAMBI - TFPI2 (lightcyan - black)	PF2341066	0.84	10.50	0.80	10.00
BAMBI - PROS1 (lightcyan - brown)	PLX4720	2.48	31.00	2.88	36.00

Table D. 4 (Continued)

Gene Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
APOE - DAAM1 - MAP3K14 (brown - lightcyan - black)	PLX4720	2.32	29.00	2.52	31.50
ETHE1 - FAM172A (purple - white)	Sorafenib	0.085	1.06	0.55	6.88
ETHE1 - FAM172A - HIVEP3 (purple - white - darkorange)	Sorafenib	0.061	0.76	0.56	7.00
MSFD12 - PROS1 (lightcyan - brown)	TAE684	1.22	15.25	1.30	16.25
MAP3K14 - MFSD12 - PROS1 (black - lightcyan - brown)	TAE684	1.05	13.13	1.29	16.13
BAMBI - C8orf4 - MAP3K14 - PROS1 (lightcyan - blue - black - brown)	TAE684	0.69	8.63	1.19	14.88

Table D. 5: The RMSE values of the best performing combined candidate biomarkers when the censored IC50 values are excluded from the CCLE drug sensitivity data

Gene Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
MPRIP - SLC23A2 (purple - brown)	AZD6244	0.50	20.71	0.46	19.09
BAMBI - ERBB2 - MPRIP (lightcyan - blue - purple)	AZD6244	0.50	20.75	0.43	17.84
DAAM1 - ERBB2 - MPRIP - SLC23A2	AZD6244	0.40	16.60	0.38	15.77
BAMBI - ERBB2 - MAP3K14 - MPRIP - SLC23A2 (lightcyan - blue - black - purple - brown)	AZD6244	0.28	11.62	0.39	16.18

Table D. 5 (Continued)

Gene Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
APOE - ERBB2 (brown - blue)	PD0325901	0.045	4.50	0.16	16.00
APOE - ERBB2 - RNF125 (brown - blue - lightcyan)	PD0325901	6.8×10^{-5}	0.0068	0.16	16.00
BAMBI - PROS1 (lightcyan - brown)	TAE684	0.74	10.24	1.05	14.52
BAMBI - MAP3K14 - PROS1 (lightcyan - black - brown)	TAE684	0.00044	0.0061	0.93	12.86
BAMBI - C8orf4 - MAP3K14 - PROS1 (lightcyan - blue - black - brown)	TAE684	0.56	7.75	0.88	12.17

Table D. 6: The RMSE values of the best performing combined candidate biomarkers when the CCLE Activity Area values are used for prediction

Gene Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
BAMBI - MAP3K14 (lightcyan - black)	AZD0530	0.18	17.78	0.30	30.26
BAMBI - MAP3K14 - TSPAN13 (lightcyan - black - darkorange)	AZD0530	0.25	25.39	0.28	28.16
ERBB2 - MPRIP (blue - purple)	AZD6244	0.19	19.32	0.19	19.01
BAMBI - MPRIP - PAOX (lightcyan - purple - blue)	AZD6244	0.17	17.74	0.18	17.85
BAMBI - MAP3K14 - MPRIP - PROS1 (lightcyan - black - purple - brown)	AZD6244	0.058	5.76	0.14	14.47

Table D. 6 (Continued)

Gene Combination	Drug Name	RMSE (CCLE)	Error in Prediction (CCLE, %)	RMSE (CGP)	Error in Prediction (CGP, %)
BAMBI - MAP3K14 - MPRIP - PROS1 - RAB38 (lightcyan - black - purple - blue - brown)	AZD6244	0.048	4.75	0.13	13.08
LRP5 - NAV3 (blue - purple)	Erlotinib	0.22	21.73	0.26	26.28
FAM172A - GAS2L1 - NAV3 (white - blue - purple)	Erlotinib	0.16	15.60	0.29	29.29
LRP5 - PAOX (blue - brown)	Lapatinib	0.16	15.66	0.21	20.82
FAM172A - LRP5 - NAV3 (white - blue - purple)	Lapatinib	0.17	16.79	0.24	23.82
ERBB2 - RNF125 (blue - lightcyan)	PD0325901	0.22	21.57	0.23	22.96
ERBB2 - NPAT - RNF125 (blue - brown - lightcyan)	PD0325901	0.030	2.99	0.26	26.26
BAMBI - TFPI2 (lightcyan - black)	PF2341066	0.21	20.69	0.25	25.18
BAMBI - MAP3K14 (lightcyan - black)	PLX4720	0.25	24.57	0.30	30.06
BAMBI - MAP3K14 - PROS1 (lightcyan - black - brown)	PLX4720	0.20	19.56	0.24	24.43
HIVEP3 - MPRIP (darkorange - purple)	Sorafenib	0.14	14.27	0.20	20.00
FAM172A - HIVEP3 - MPRIP (white - darkorange - purple)	Sorafenib	0.096	9.61	0.27	26.86
MFSD12 - PROS1 (lightcyan - brown)	TAE684	0.15	14.94	0.21	21.39
MAP3K14 - MFSD12 - PROS1 (black - lightcyan - brown)	TAE684	0.15	14.67	0.18	18.02
MAP3K14 - MFSD12 - PAOX - PROS1 (black - lightcyan - blue - brown)	TAE684	0.13	13.38	0.15	15.13

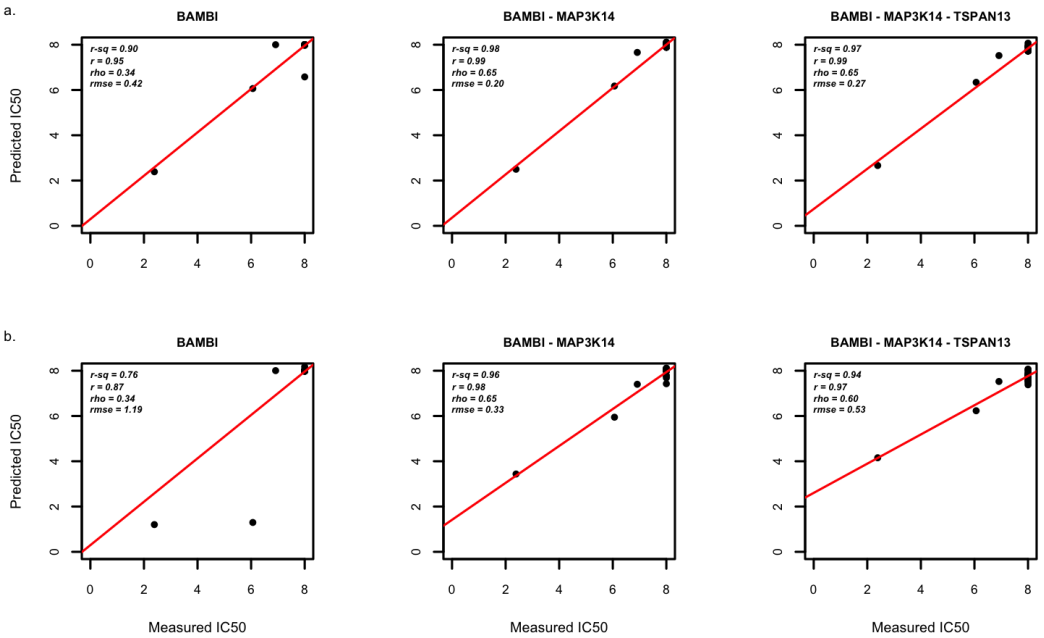

Figure D. 1: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for AZD0530 when the censored CCLE IC50 values are included in the CCLE drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

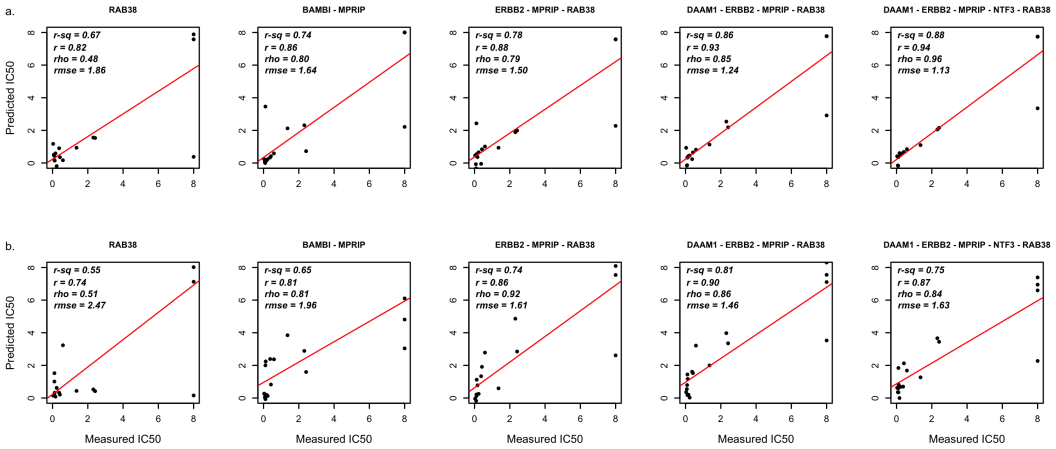

Figure D. 2: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for AZD6244 when the censored CCLE IC50 values are included in the CCLE drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

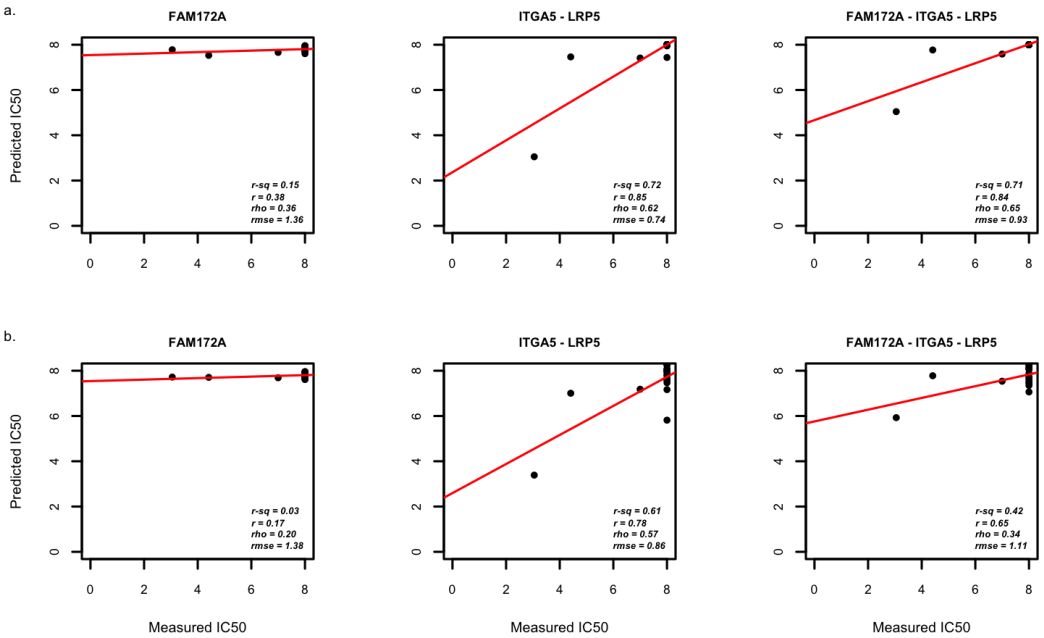

Figure D. 3: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for Erlotinib when the censored CCLE IC50 values are included in the CCLE drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

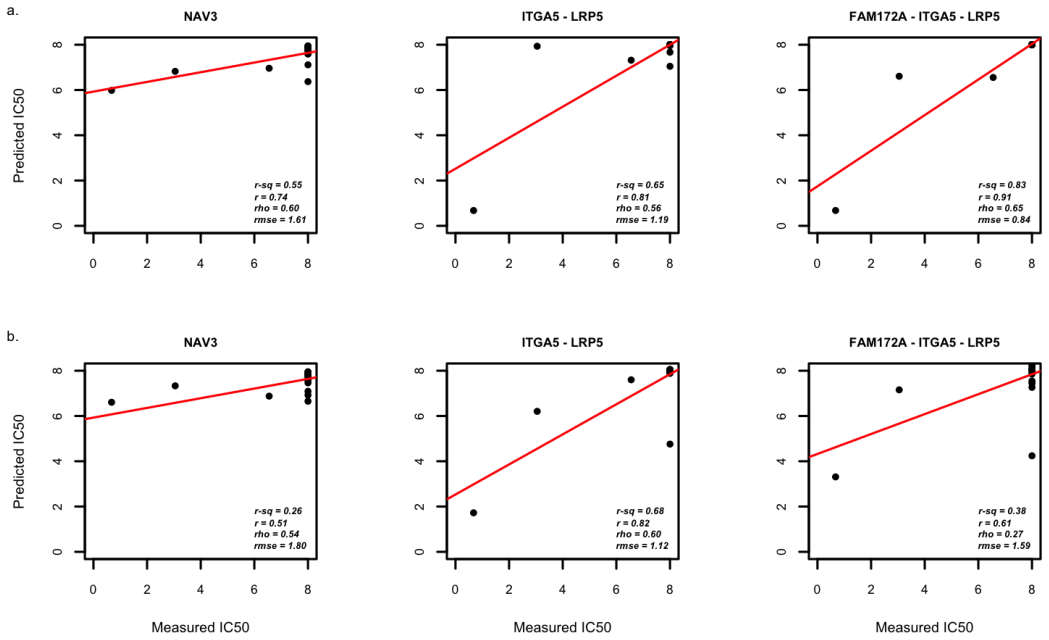

Figure D. 4: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for Lapatinib when the censored CCLE IC50 values are included in the CCLE drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

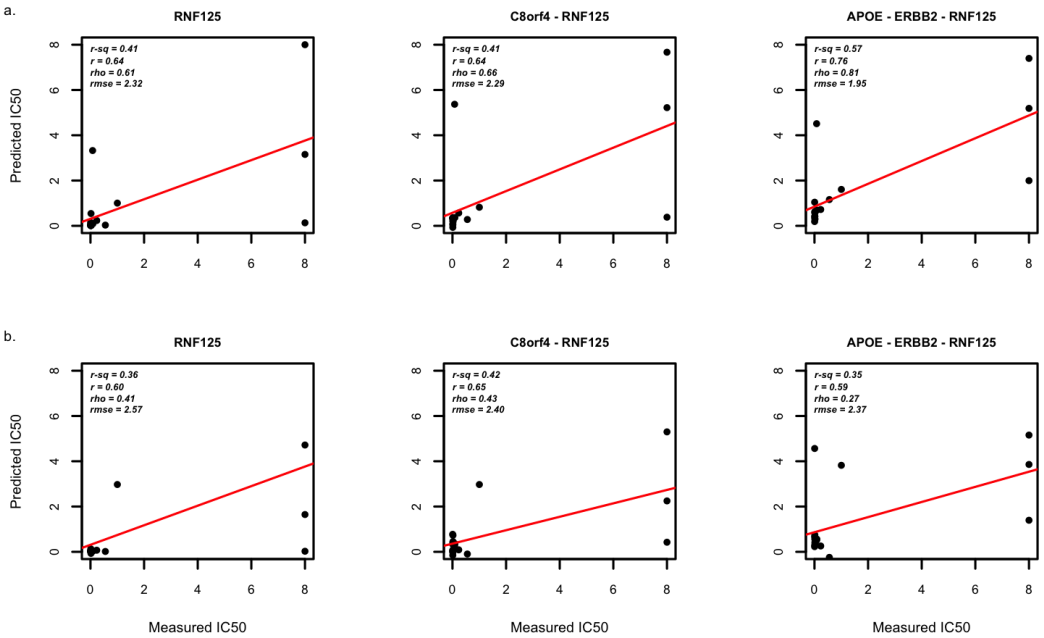

Figure D. 5: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for PD0325901 when the censored CCLE IC50 values are included in the CCLE drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

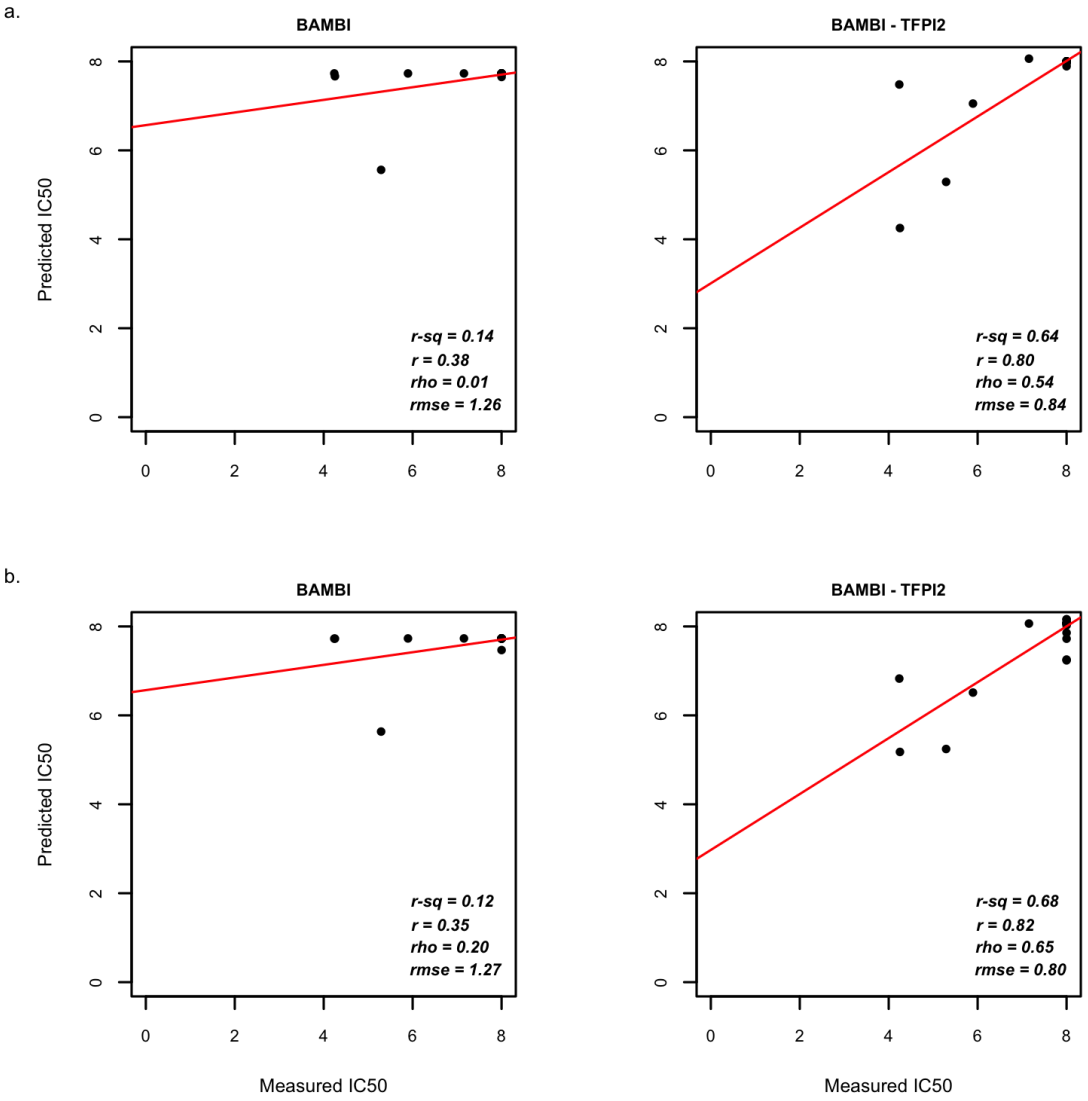

Figure D. 6: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for PF2341066 when the censored CCLE IC50 values are included in the CCLE drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

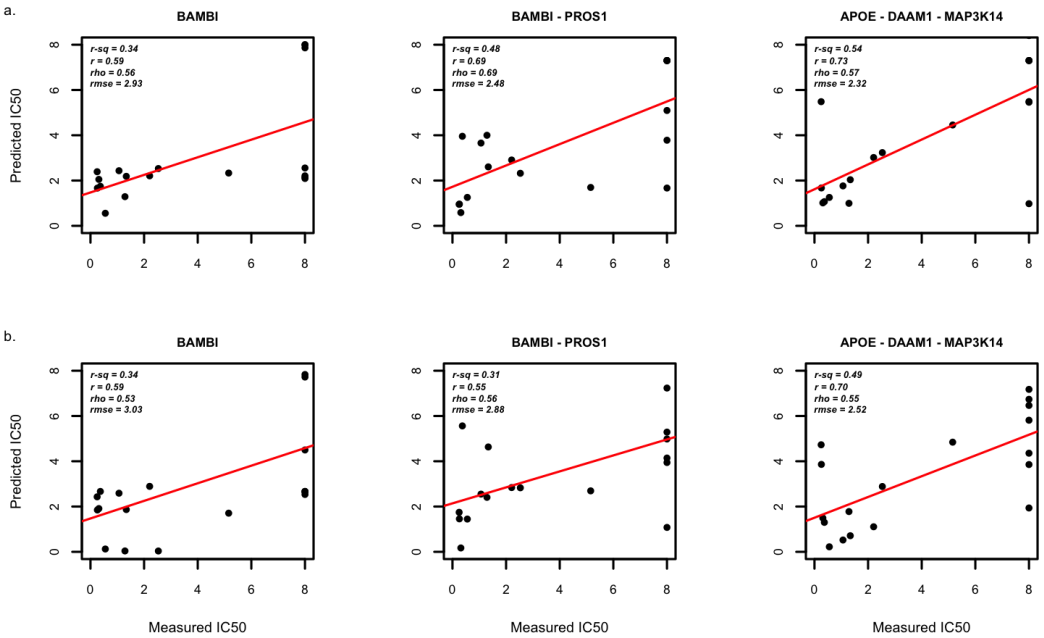

Figure D. 7: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for PLX4720 when the censored CCLE IC50 values are included in the CCLE drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

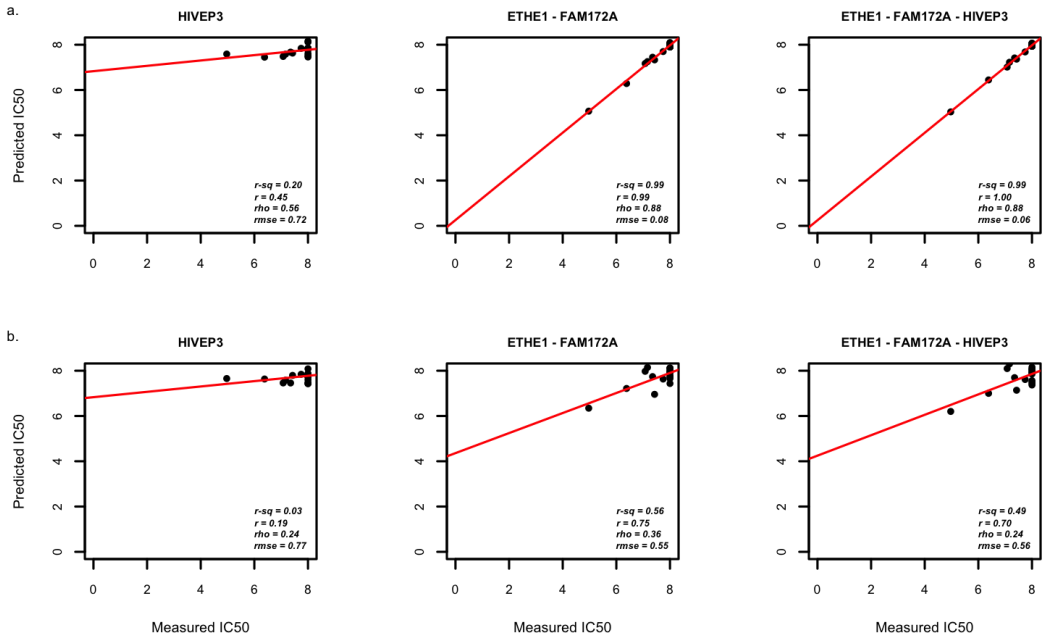

Figure D. 8: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for Sorafenib when the censored CCLE IC50 values are included in the CCLE drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

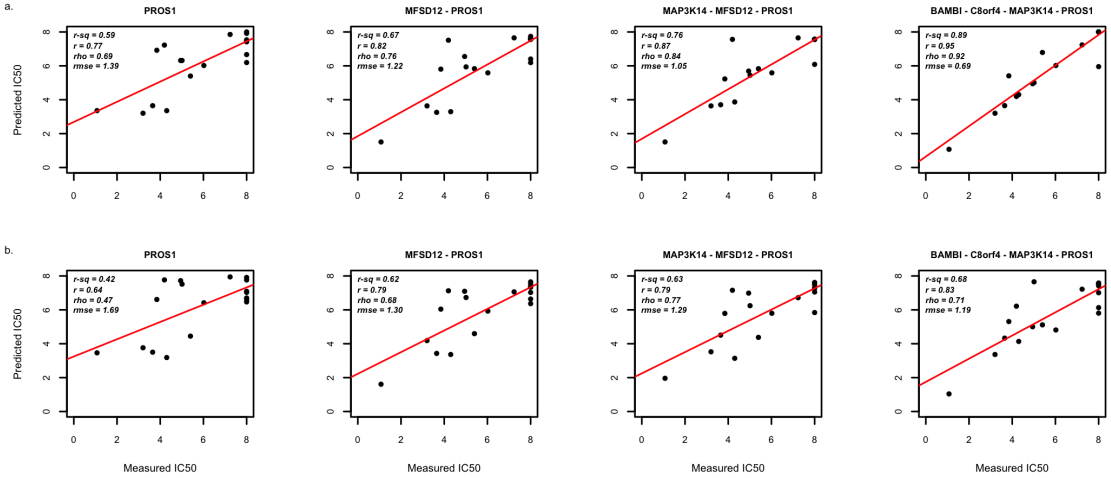

Figure D. 9: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for TAE684 when the censored CCLE IC50 values are included in the CCLE drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

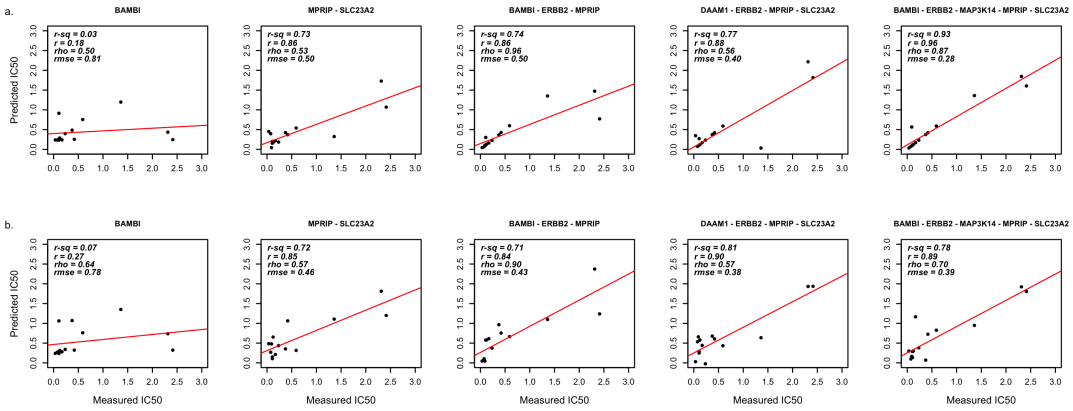

Figure D. 10: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for AZD6244 when both the censored CCLE IC50 values and the extrapolated CGP IC50 values are excluded from the CCLE and the CGP drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

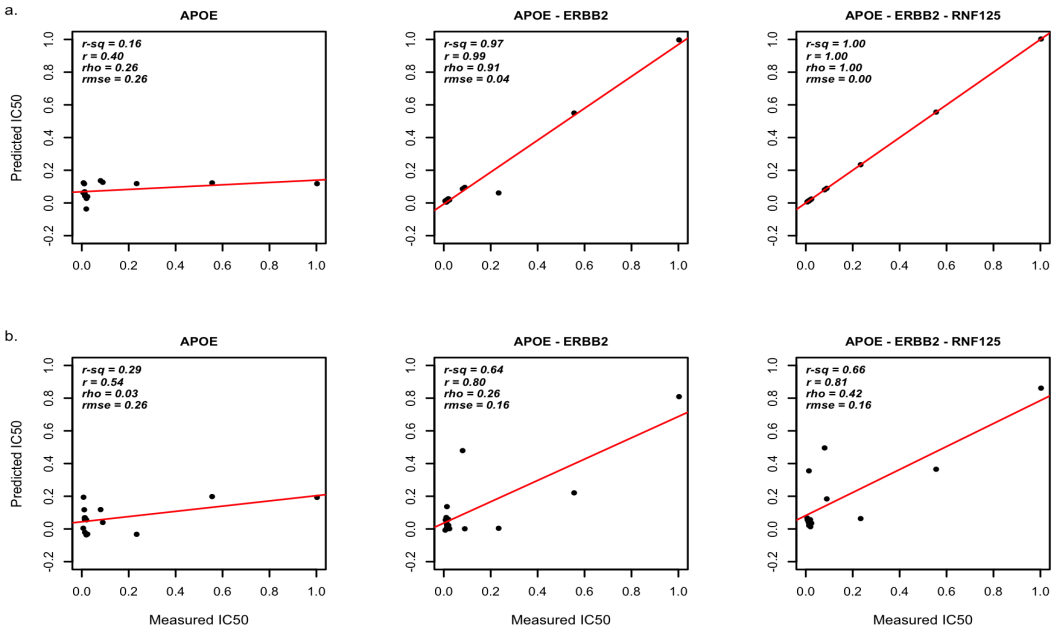

Figure D. 11: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for PD0325901 when both the censored CCLE IC50 values and the extrapolated CGP IC50 values are excluded from the CCLE and the CGP drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

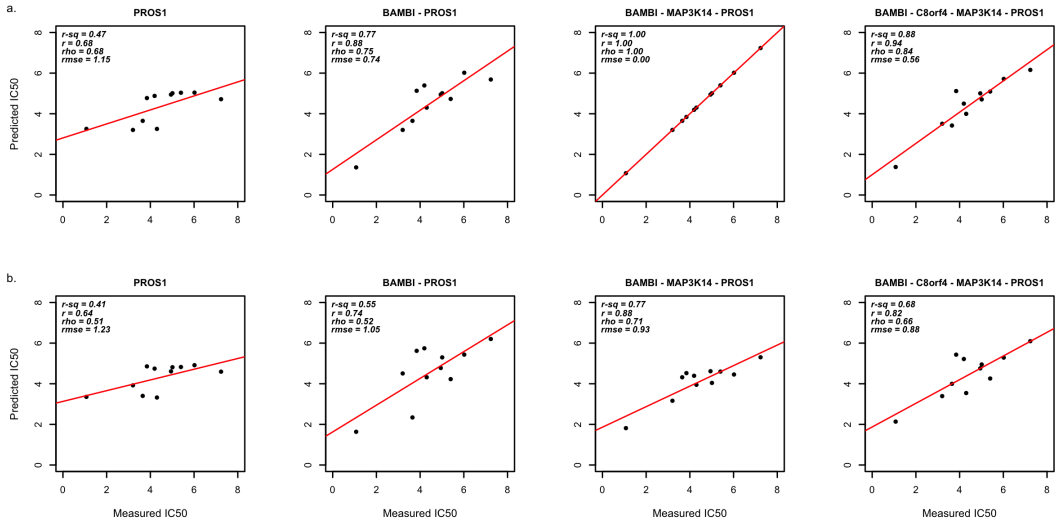

Figure D. 12: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for TAE684 when both the censored CCLE IC50 values and the extrapolated CGP IC50 values are excluded from the CCLE and the CGP drug sensitivity data. (a) IC50 prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

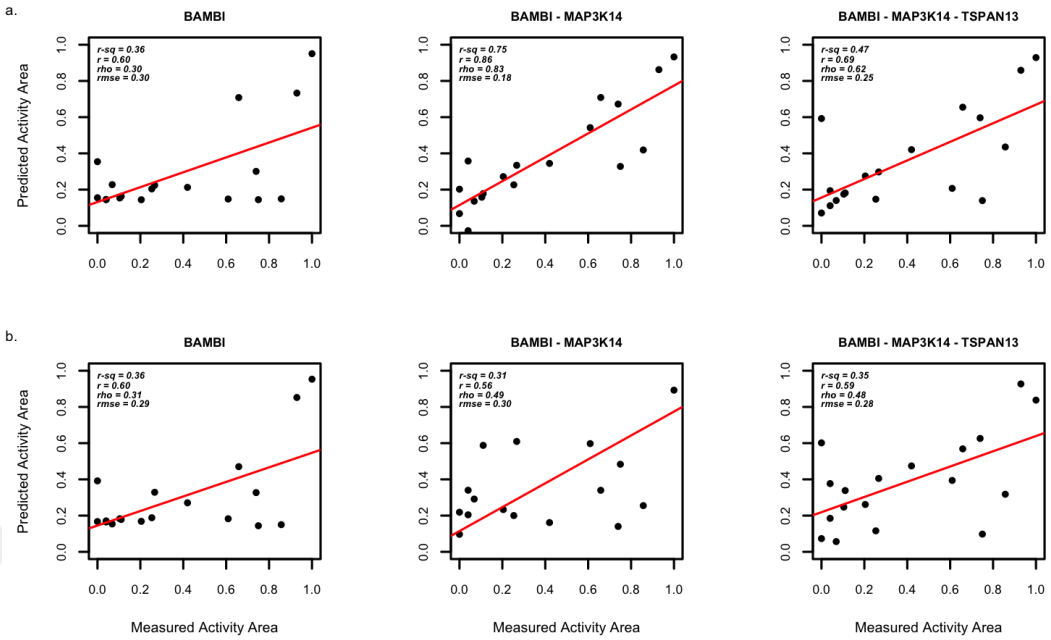

Figure D. 13: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for AZD0530 when the CCLE Activity Area values are used for prediction. (a) Activity Area prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

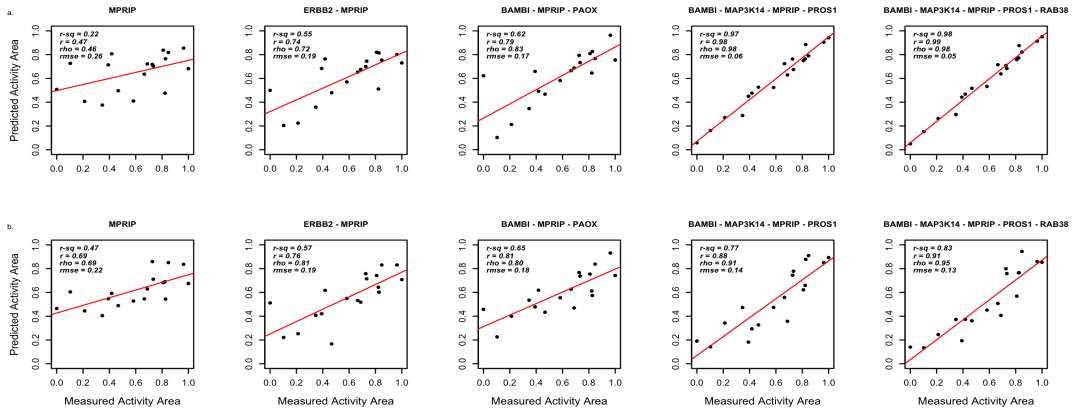

Figure D. 14: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for AZD6244 when the CCLE Activity Area values are used for prediction. (a) Activity Area prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

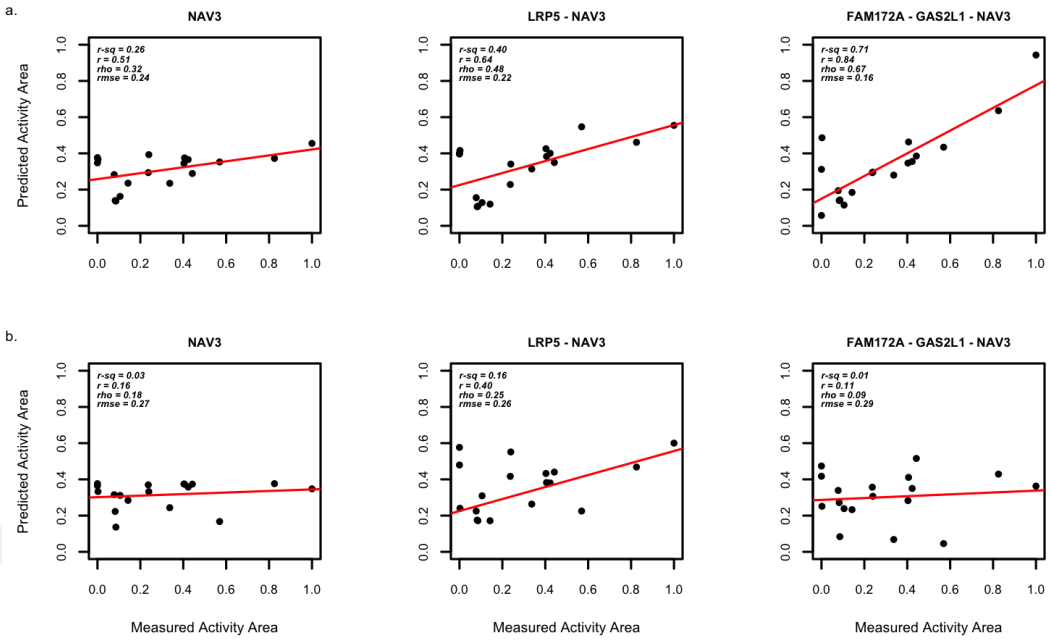

Figure D. 15: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for Erlotinib when the CCLE Activity Area values are used for prediction. (a) Activity Area prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

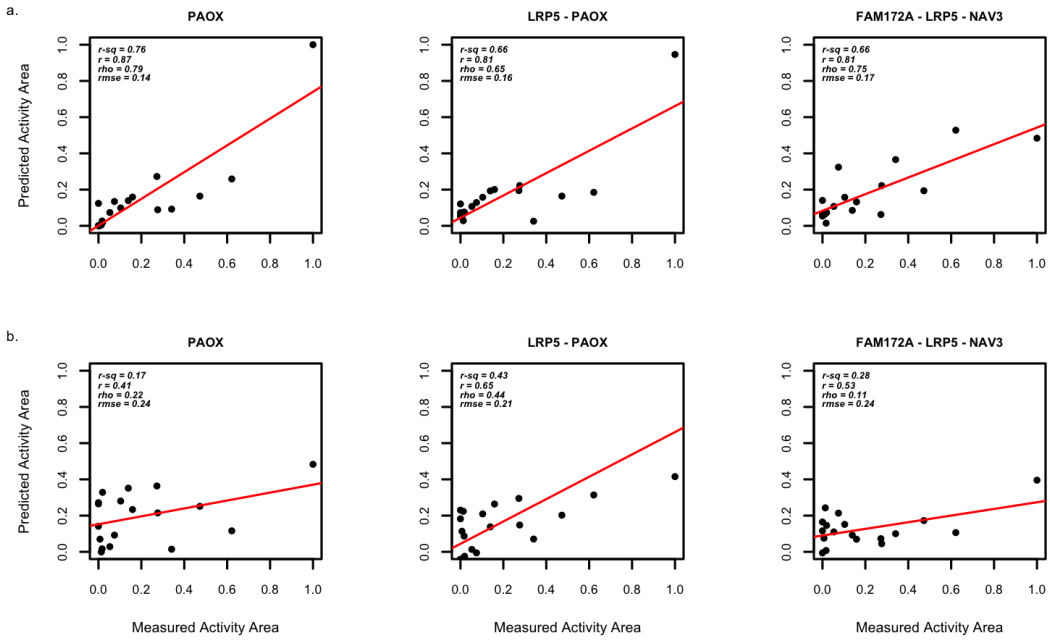

Figure D. 16: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for Lapatinib when the CCLE Activity Area values are used for prediction. (a) Activity Area prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

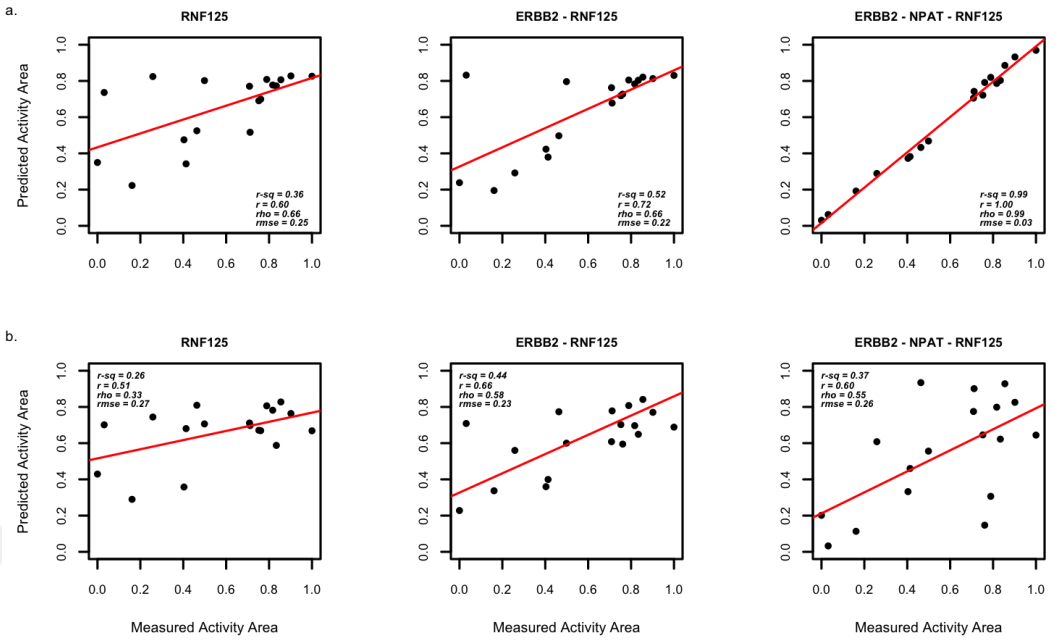

Figure D. 17: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for PD0325901 when the CCLE Activity Area values are used for prediction. (a) Activity Area prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

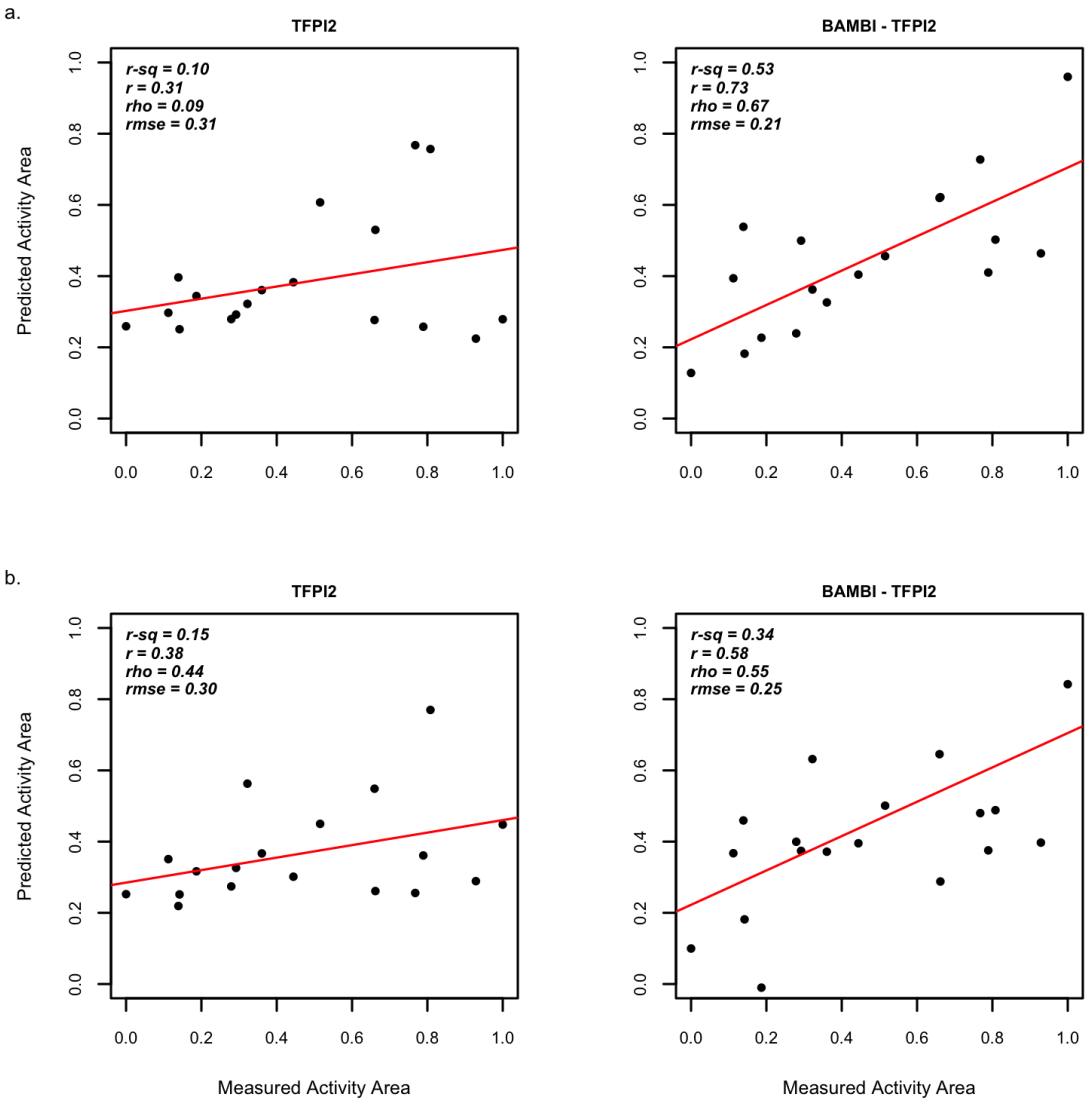

Figure D. 18: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for PF2341066 when the CCLE Activity Area values are used for prediction. (a) Activity Area prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

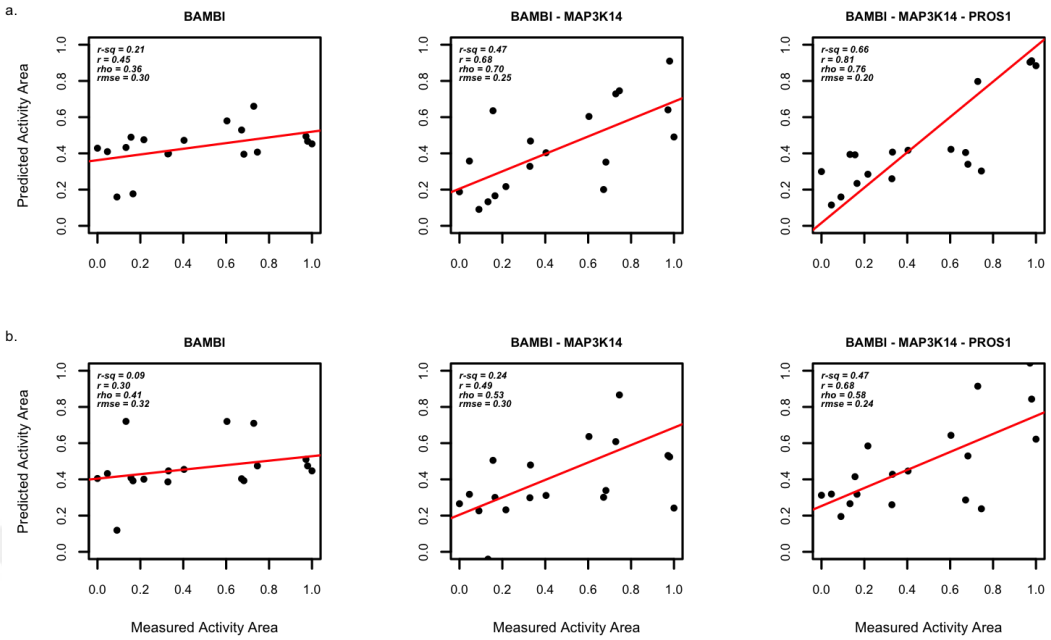

Figure D. 19: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for PLX4720 when the CCLE Activity Area values are used for prediction. (a) Activity Area prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

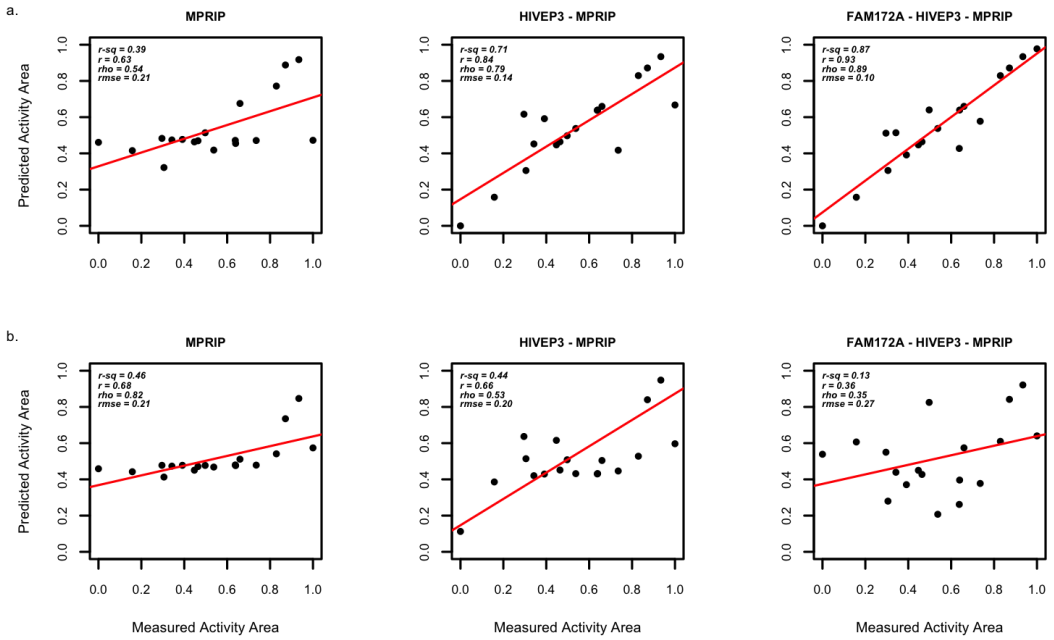

Figure D. 20: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for Sorafenib when the CCLE Activity Area values are used for prediction. (a) Activity Area prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

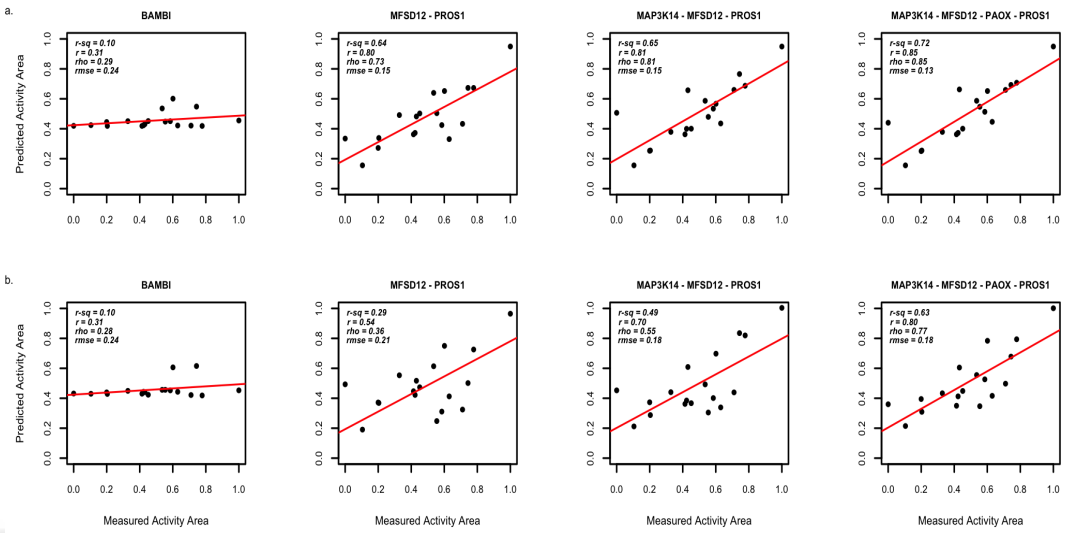

Figure D. 21: The scatterplots which depict the trend of R^2, Pearson r, Spearman rho, and RMSE with varying number of candidate biomarker combinations for TAE684 when the CCLE Activity Area values are used for prediction. (a) Activity Area prediction is performed by the CCLE data. (b) The trained model with the CCLE data is tested by the CGP data.

APPENDIX E

TREND OF CORRELATION SCORES WITH VARYING NUMBER OF COMBINATIONS

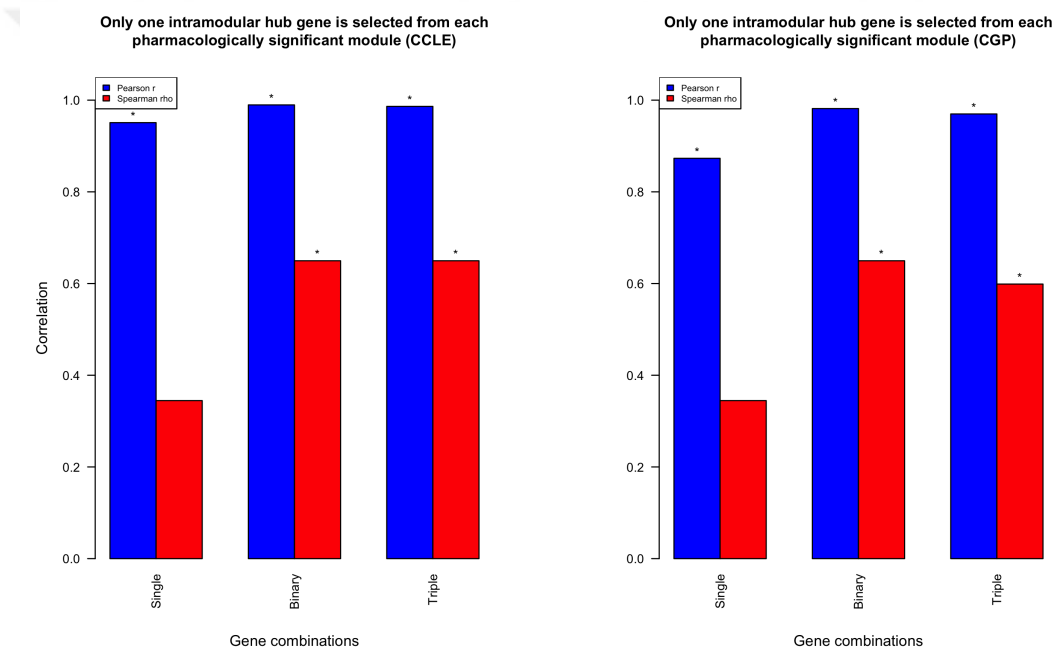

Figure E. 1: The barplots which demonstrate the correlation scores for AZD0530 when the censored IC50 values are included in the CCLE drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

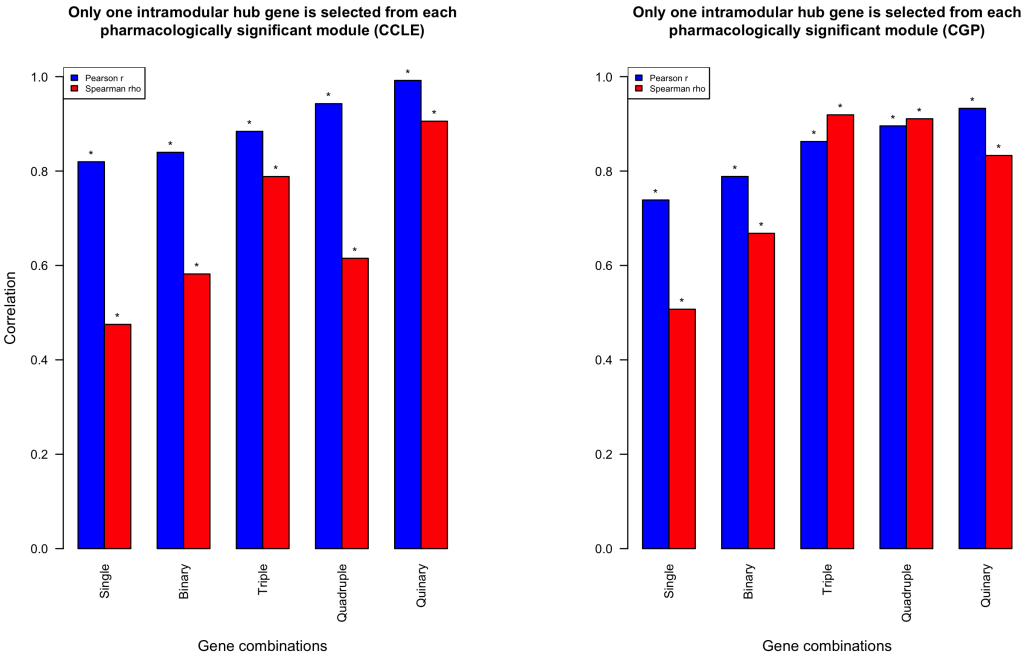

Figure E. 2: The barplots which demonstrate the correlation scores for AZD6244 when the censored IC50 values are included in the CCLE drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

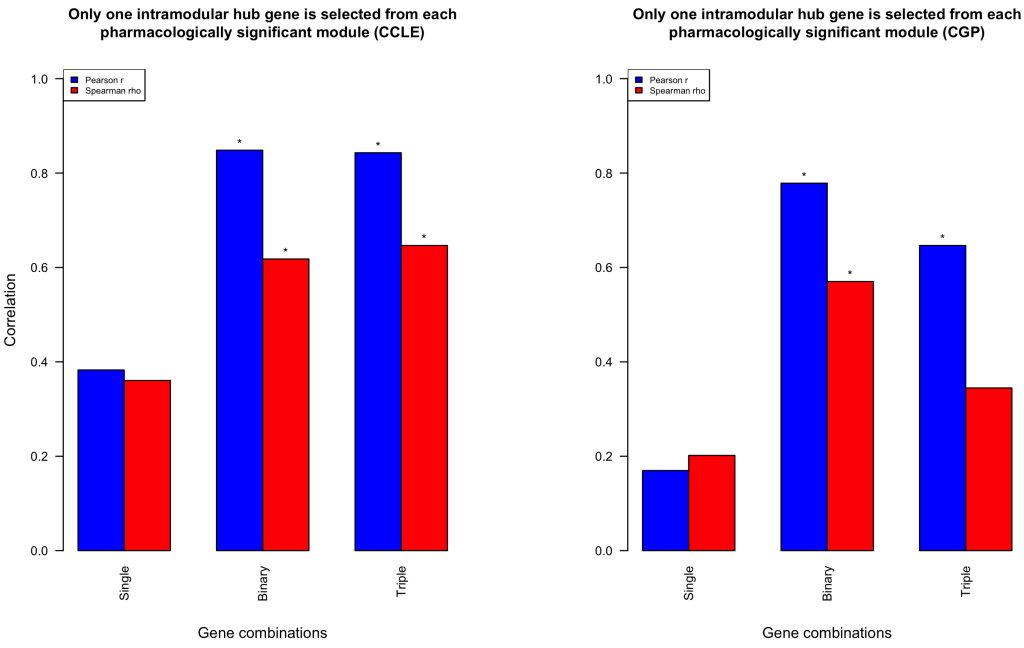

Figure E. 3: The barplots which demonstrate the correlation scores for Erlotinib when the censored IC50 values are included in the CCLE drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

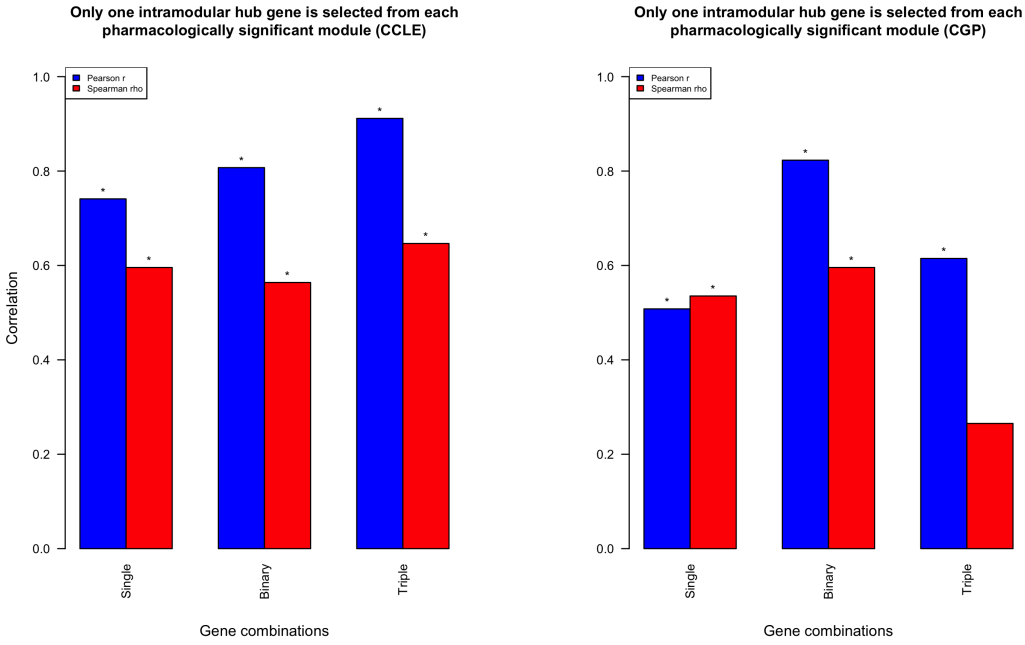

Figure E. 4: The barplots which demonstrate the correlation scores for Lapatinib when the censored IC50 values are included in the CCLE drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

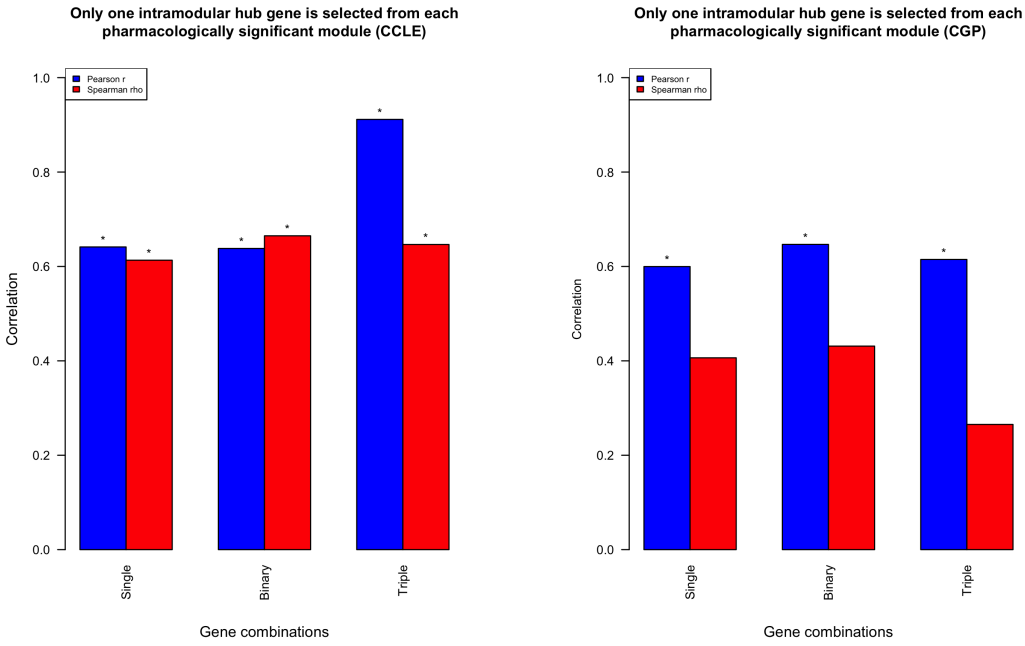

Figure E. 5: The barplots which demonstrate the correlation scores for PD0325901 when the censored IC50 values are included in the CCLE drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

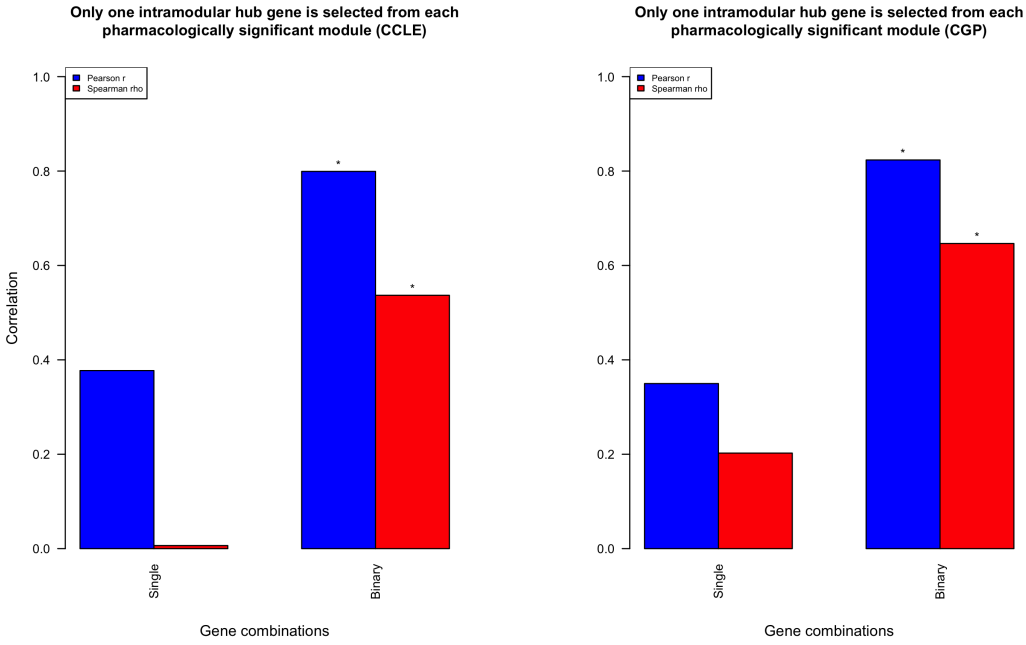

Figure E. 6: The barplots which demonstrate the correlation scores for PF2341066 when the censored IC50 values are included in the CCLE drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

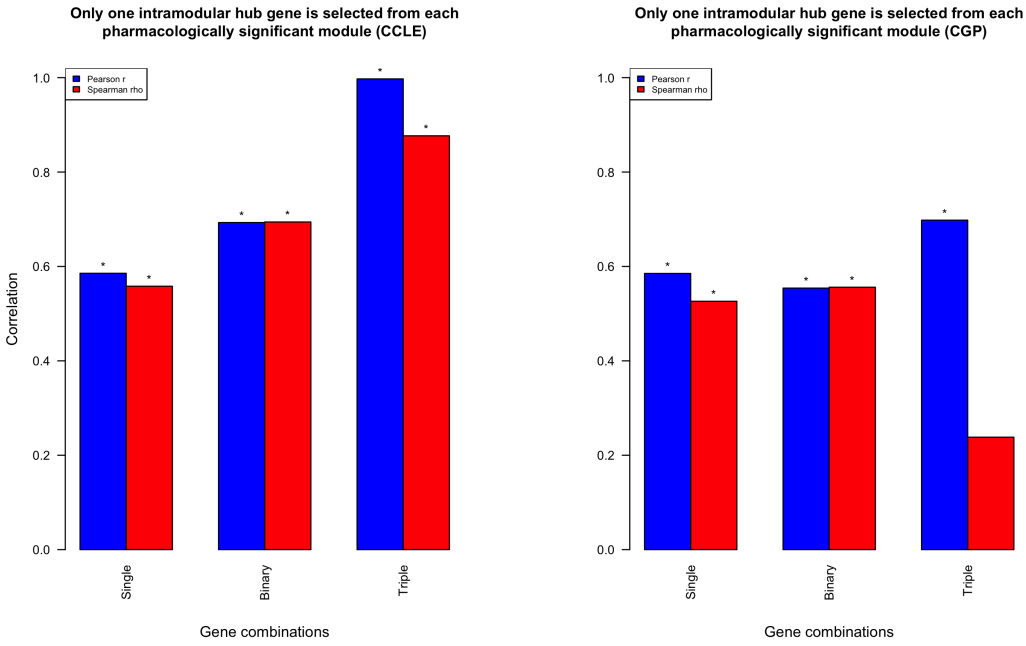

Figure E. 7: The barplots which demonstrate the correlation scores for PLX4720 when the censored IC50 values are included in the CCLE drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

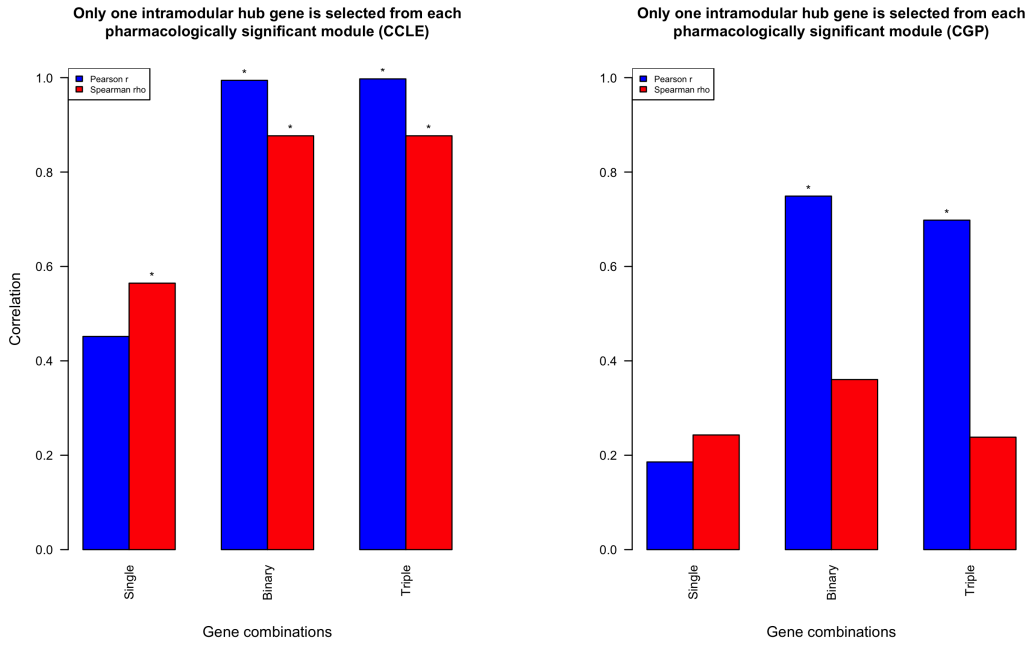

Figure E. 8: The barplots which demonstrate the correlation scores for Sorafenib when the censored IC50 values are included in the CCLE drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

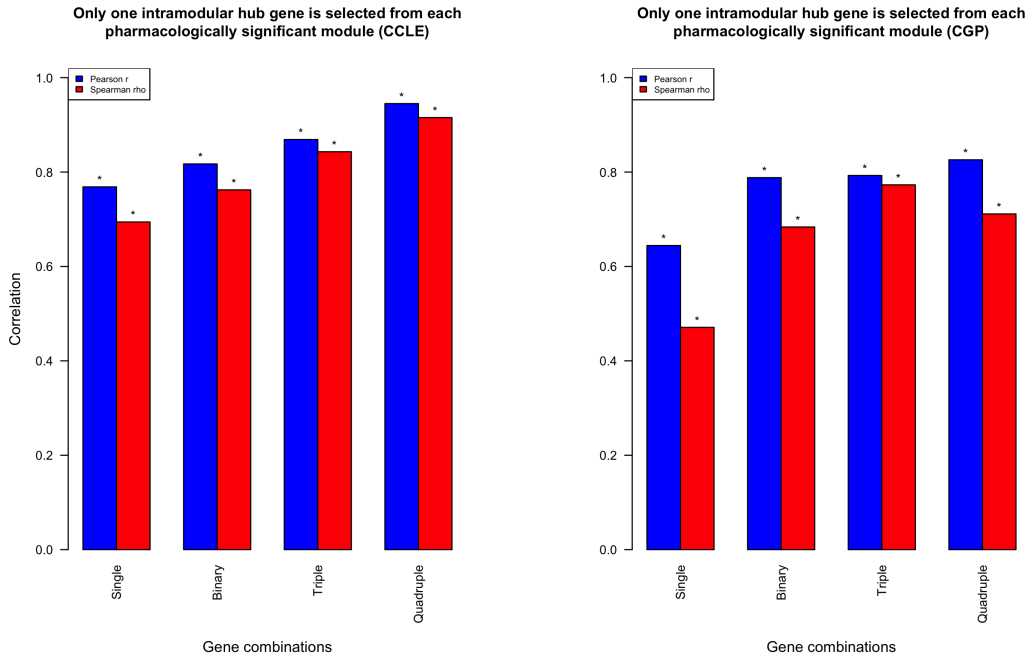

Figure E. 9: The barplots which demonstrate the correlation scores for TAE684 when the censored IC50 values are included in the CCLE drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

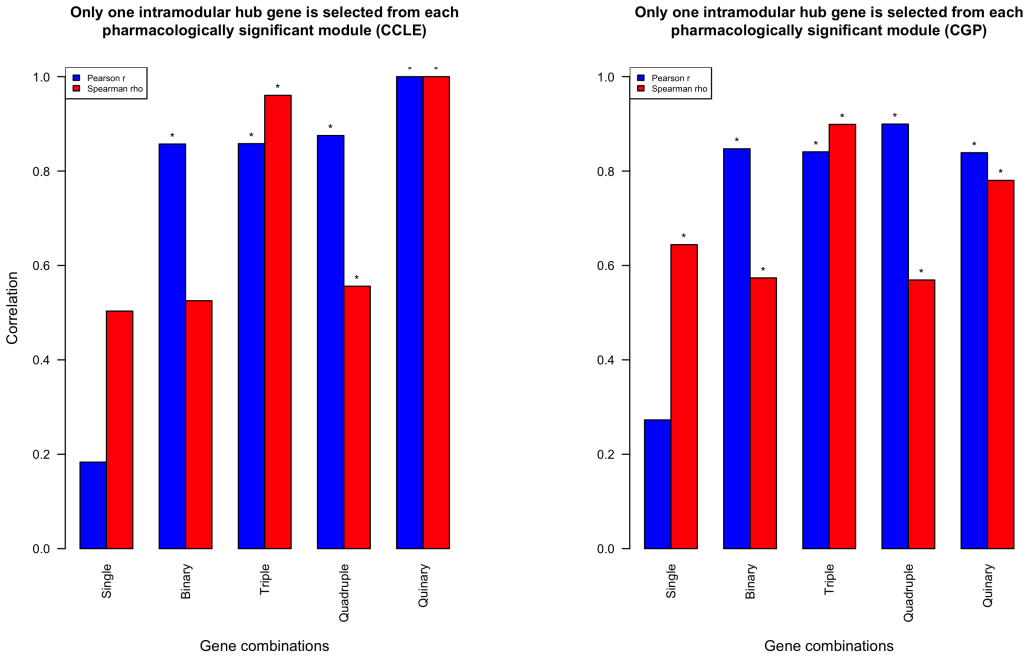

Figure E. 10: The barplots which demonstrate the correlation scores for AZD6244 when both the censored and extrapolated IC50 values are excluded from drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

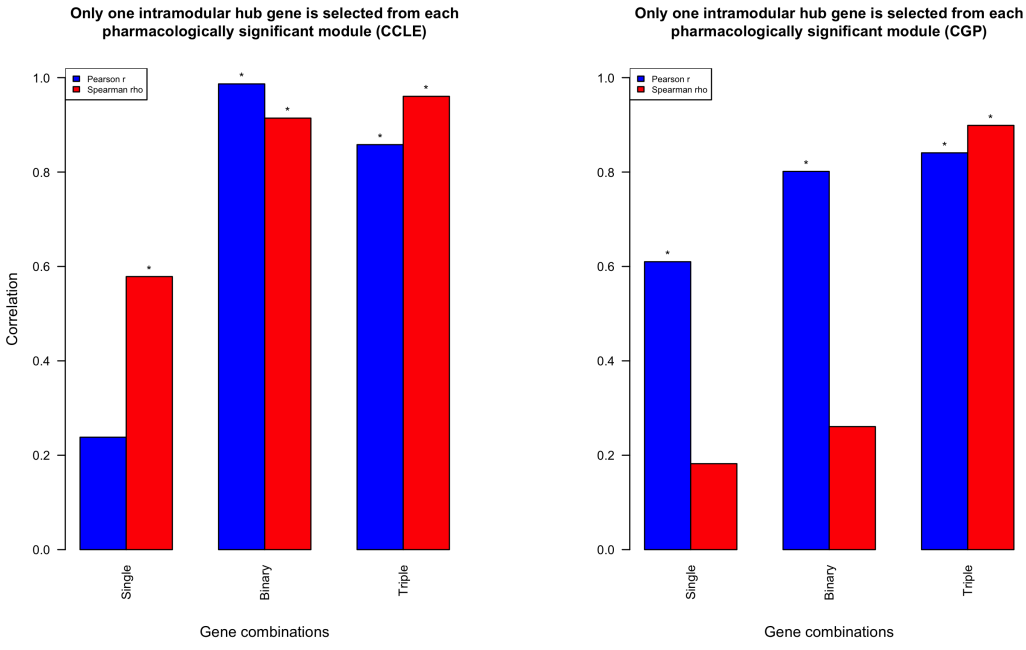

Figure E. 11: The barplots which demonstrate the correlation scores for PD0325901 when both the censored and extrapolated IC50 values are excluded from drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

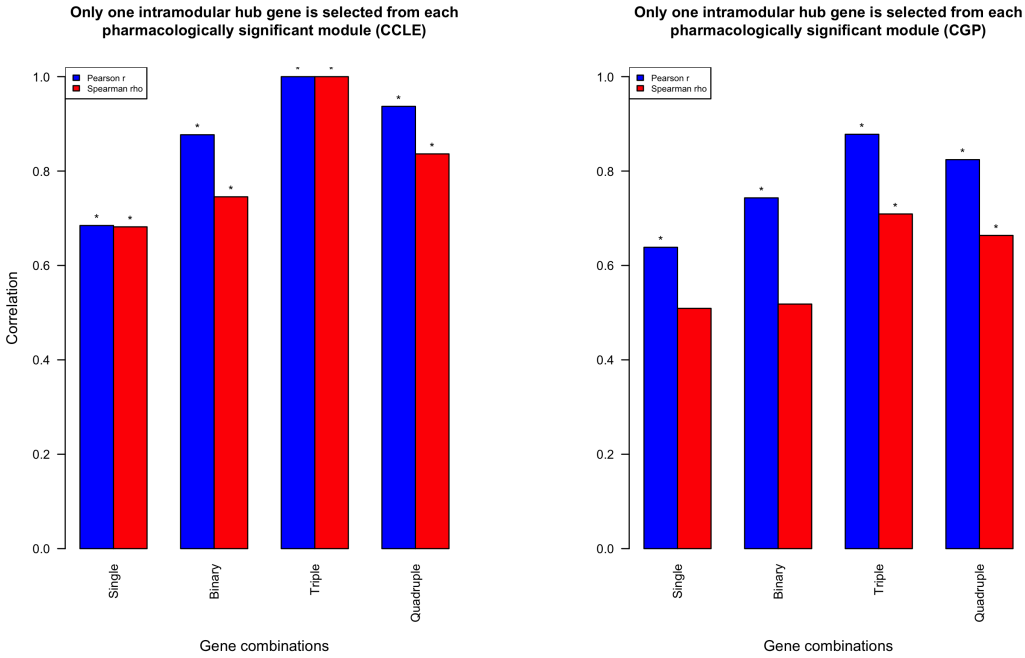

Figure E. 12: The barplots which demonstrate the correlation scores for TAE684 when both the censored and extrapolated IC50 values are excluded from drug sensitivity data. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

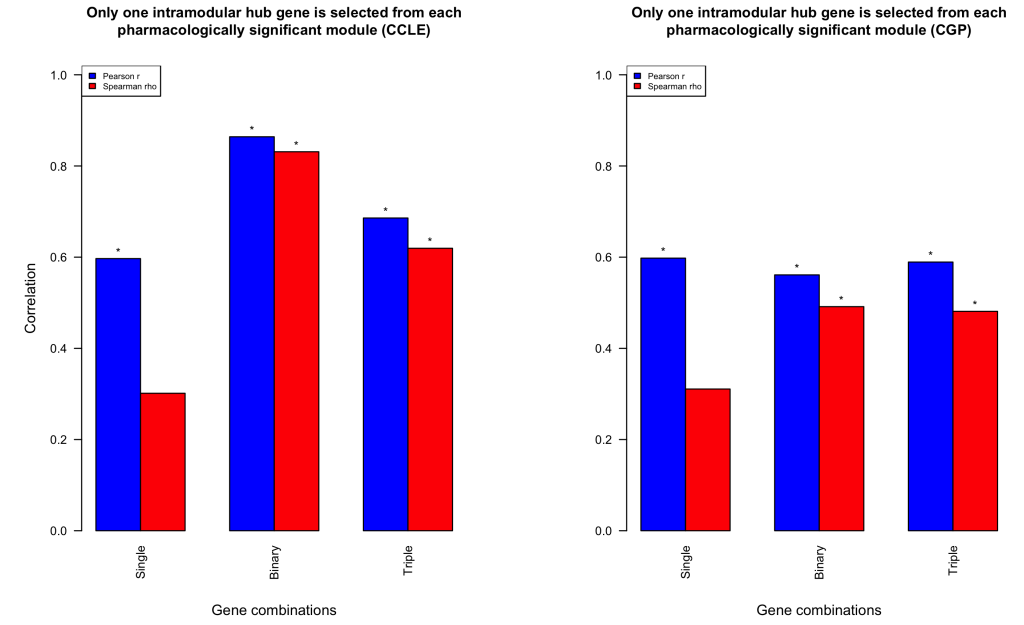

Figure E. 13: The barplots which demonstrate the correlation scores for AZD0530 when the CCLE Activity Area values are used for prediction. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

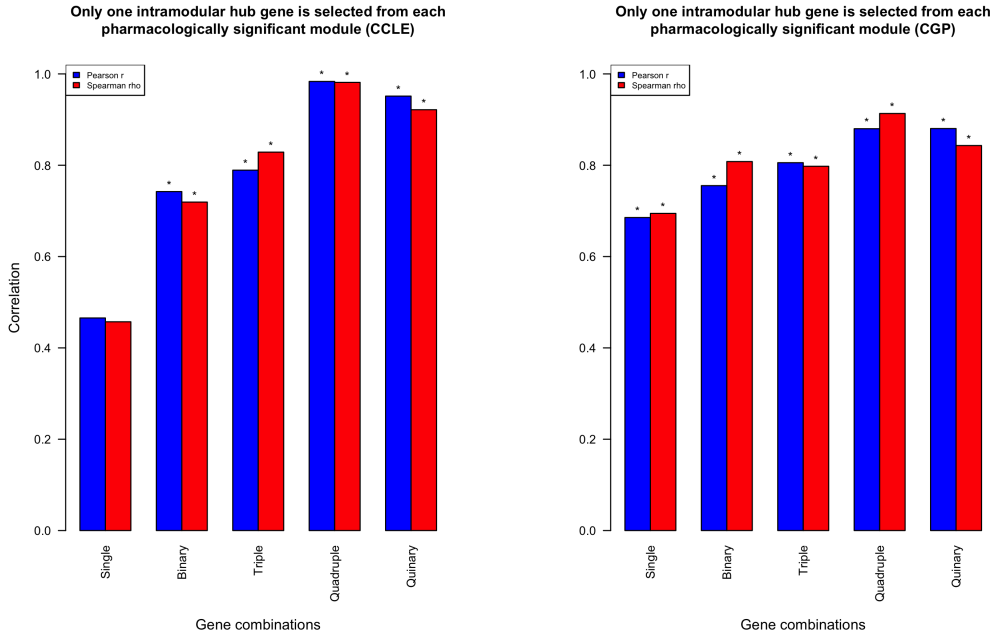

Figure E. 14: The barplots which demonstrate the correlation scores for AZD6244 when the CCLE Activity Area values are used for prediction. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

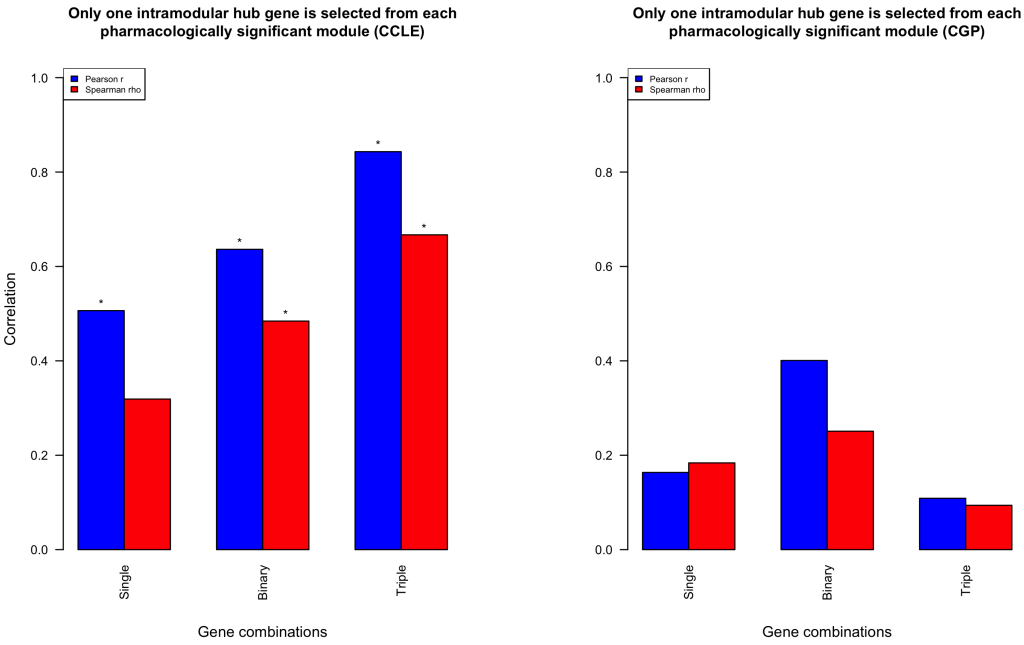

Figure E. 15: The barplots which demonstrate the correlation scores for Erlotinib when the CCLE Activity Area values are used for prediction. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

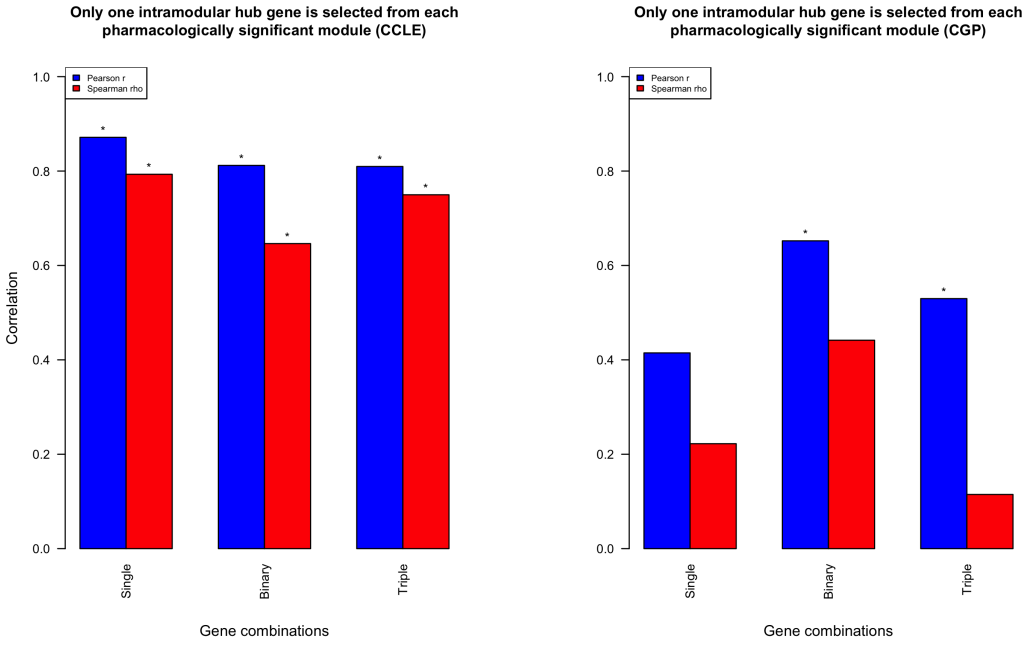

Figure E. 16: The barplots which demonstrate the correlation scores for Lapatinib when the CCLE Activity Area values are used for prediction. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

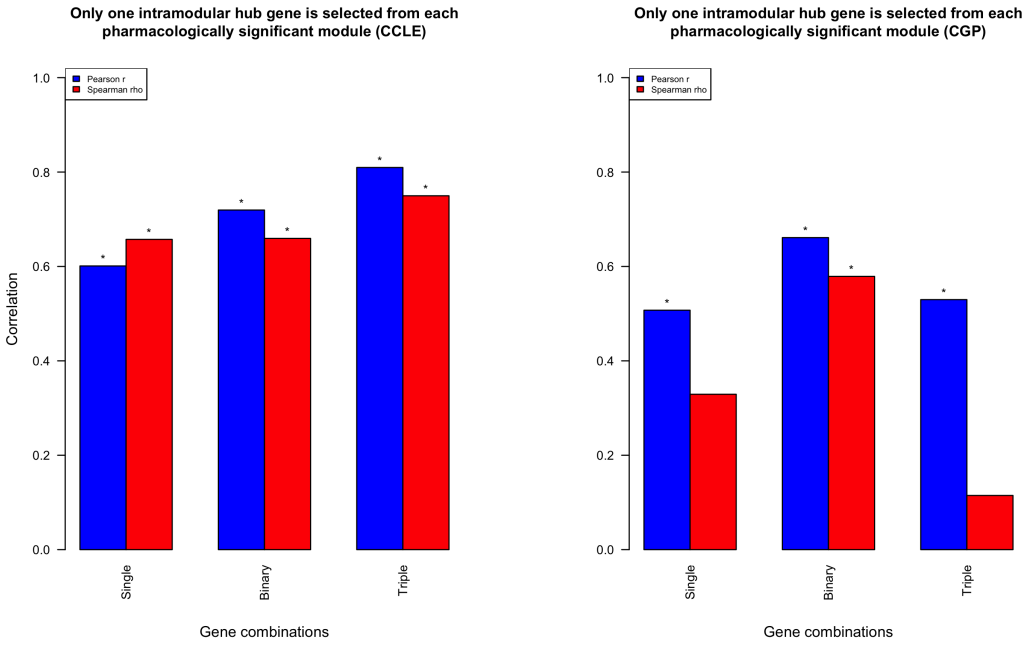

Figure E. 17: The barplots which demonstrate the correlation scores for PD0325901 when the CCLE Activity Area values are used for prediction. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

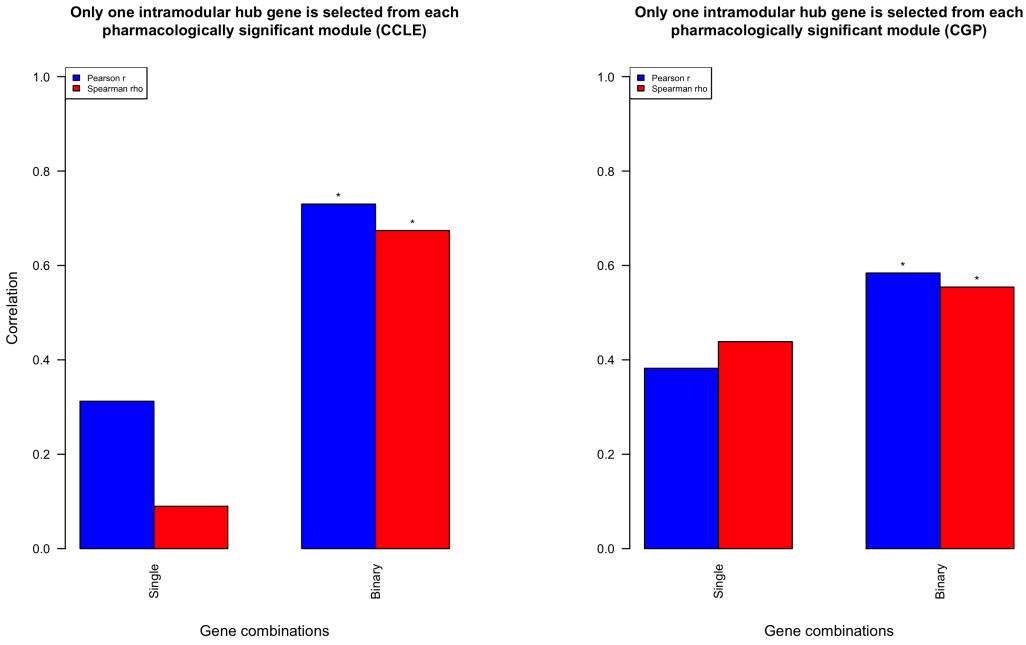

Figure E. 18: The barplots which demonstrate the correlation scores for PF2341066 when the CCLE Activity Area values are used for prediction. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

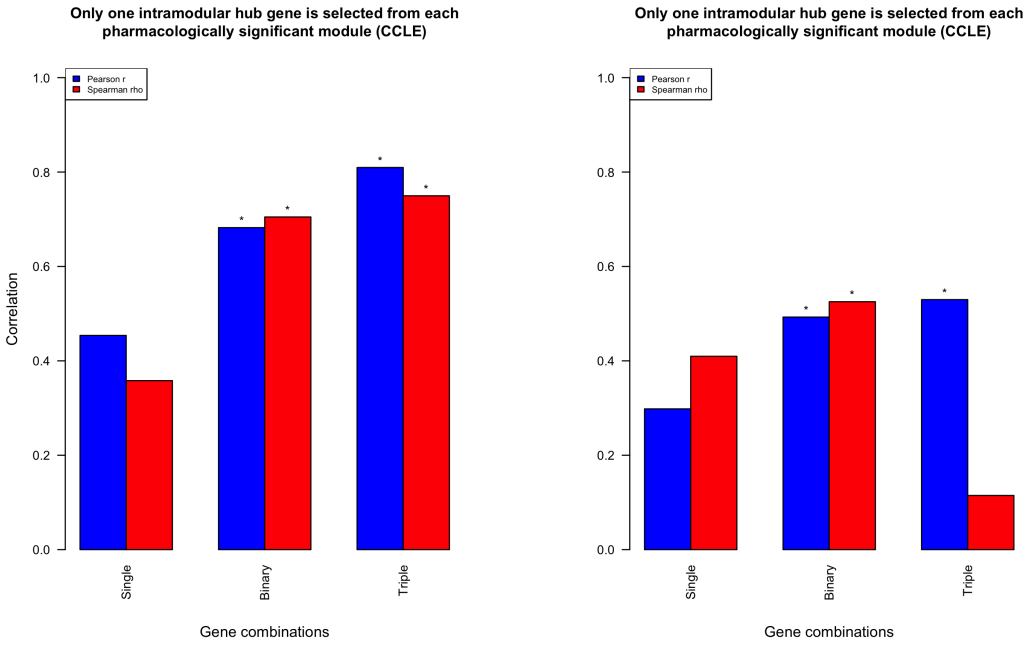

Figure E. 19: The barplots which demonstrate the correlation scores for PLX4720 when the CCLE Activity Area values are used for prediction. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

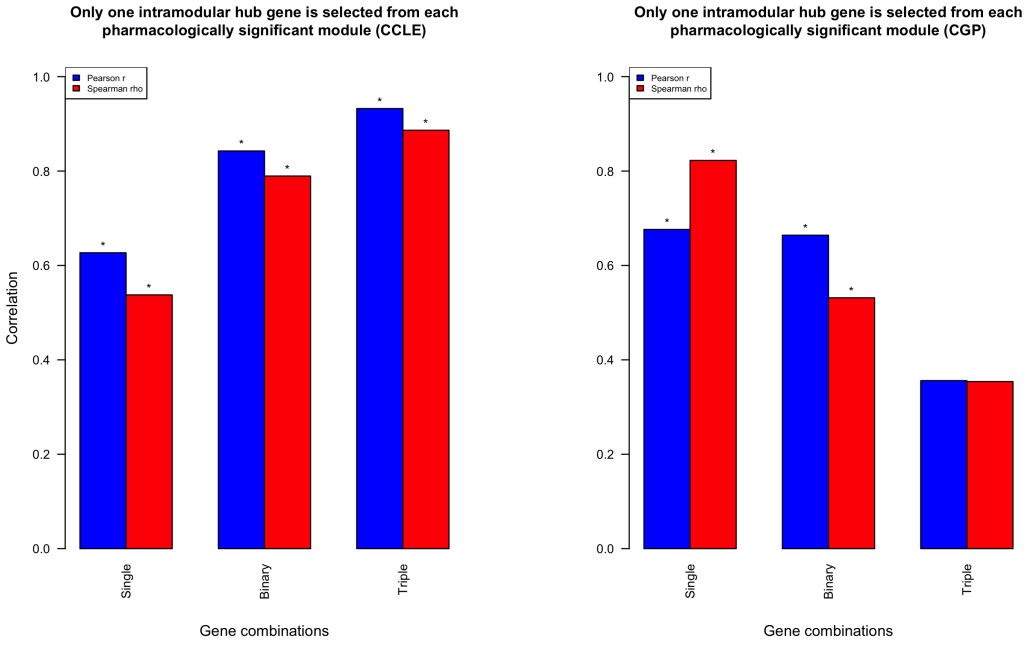

Figure E. 20: The barplots which demonstrate the correlation scores for Sorafenib when the CCLE Activity Area values are used for prediction. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.

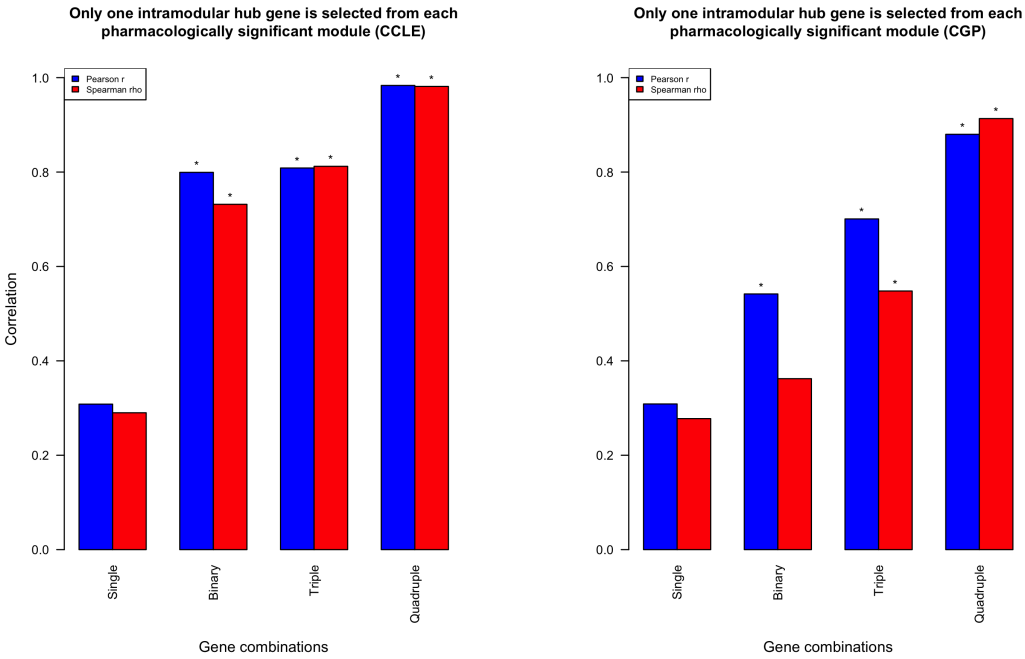

Figure E. 21: The barplots which demonstrate the correlation scores for TAE684 when the CCLE Activity Area values are used for prediction. The barplot on the left side is generated with the CCLE data, while the barplot on the right is generated with the CGP data. Star sign (*) shows that correlation is significant.