3D SPATIAL ORGANIZATION AND NETWORK-GUIDED COMPARISON OF MUTATION PROFILES IN GLIOBLASTOMA

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

CANSU DİNÇER

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

AUGUST 2019

Approval of the thesis:

3D SPATIAL ORGANIZATION AND NETWORK-GUIDED COMPARISON OF MUTATION PROFILES IN GLIOBLASTOMA

Submitted by Cansu Dinçer in partial fulfillment of the requirements for the degree of **Master of Science in Health Informatics Department, Middle East Technical University** by,

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

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

Assoc. Prof. Dr. Nurcan Tunçbağ
Supervisor, **Health Informatics, METU** _____

Examining Committee Members:

Prof. Dr. Tolga Can
Computer Engineering, METU _____

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

Assoc. Prof. Dr. Tunca Doğan
Health Informatics, Hacettepe University _____

Assoc. Dr. Mehmet Somel
Department of Biology, METU _____

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, METU _____

Date: **28.08.2019**

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 : CANSU DİNÇER

Signature : _____

ABSTRACT

3D SPATIAL ORGANIZATION AND NETWORK-GUIDED COMPARISON OF MUTATION PROFILES IN GLIOBLASTOMA

Dinçer, Cansu

MSc., Department of Bioinformatics

Supervisor: Assoc. Prof. Dr. Nurcan Tunçbağ

August 2019, 73 pages

Glioblastoma multiforme (GBM) is the most aggressive and heterogeneous type of brain tumor. The heterogeneity of GBM is the main obstacle to develop effective treatment strategies. In this study, we aimed to decrease the heterogeneity among GBM patients from The Cancer Genome Atlas (TCGA), classify the patients and propose therapeutic hypothesis for patient groups by using patient mutation profiles. We therefore implemented a systems level approach to mutations using their biophysical characteristics and organization in patient-specific subnetworks. While 3D mutation patches decrease the heterogeneity among patients, network guided analysis classified patients into five groups. Since each patient group carries a set of signature 3D mutation patches, we collected GBM cell line mutation and drug sensitivity data to link GBM patient group to GBM cell lines and eventually drug responses through 3D patch mutations. Therefore, we can propose drug responses of specific patient group for specific drugs. As an example, by targeting CSF1R, Pazopanib can be effective for Group 3 yet Group 2 can be resistant to inhibition of ATM which is a mediator of PTEN phosphorylation. As a conclusion, from mutations to protein interaction networks and eventually to therapeutic data, this study is a new perspective for precision medicine.

Keywords: Glioblastoma, 3D mutation patch, protein interactions, patient-specific network modeling, network-guided tumor grouping

ÖZ

GLIOBLASTOMA HASTALARI MUTASYON PROFİLLERİNİN 3 BOYUTLU UZAMSAL ORGANİZASYONU VE AĞ GÜDÜMLÜ KARŞILAŞTIRILMASI

Dinçer, Cansu

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Doç. Dr. Nurcan Tunçbağ

Ağustos 2019, 73 sayfa

Glioblastoma multiforme (GBM), en agresif seyirli, heterojen beyin tümörü çeşididir. Bu heterojenlik verimli tedavi yöntemlerinin geliştirilmesinin önündeki temel engeldir. Bu çalışmada, TCGA' de bulunan GBM hastalarının mutasyon profillerini kullanarak, hastalar arasındaki çeşitliliği azaltmayı, hastaları gruplara ayırıp, gruplara has ilaçların ve grupların ilaca tepkilerinin çıkarımını yapabilmeyi amaçlıyoruz. Bu nedenle, hasta mutasyonlarına onların biyofiziksel karakterlerini ve hastaya özgü ağlardaki organizasyonlarını kullanan sistemsel bir yaklaşım uyguladık. Üç boyutlu mutasyon birikim bölgelerini belirleyerek hasta çeşitliliğini azaltırken, ağ temelli yaklaşımla hastaları beş farklı gruba ayırdık. Bu hasta gruplarından her biri belirli üç boyutlu mutasyon birikim bölgeleri taşıdıklarından, bu mutasyon birikim bölge bilgisini GBM hücre hattı mutasyon ve ilaç hassasiyet verileri ile birleştirerek, hasta gruplarımızı hücre hatlarına ve en son ilaçlara bağladık. Böylece hasta gruplarımızın belirli ilaçlara verebilecekleri tepkileri önerdik. Örneğin, CSF1R' ı hedefleyen Pazopanib Grup 3 için etkili olabilecekken, Grup 2 ATM' yi engelleyici ilaçlara karşı dirençli olabilir. Sonuç olarak, mutasyon, protein etkileşim ağları ve ilaç verileri kullanarak yapmış olduğumuz bu çalışma kişiye özel, hassa tıp çalışmalarına yeni bir bakış açısı olabilir.

Anahtar Sözcükler: Glioblastoma, 3 Boyutlu mutasyon birikim bölgeleri, protein etkileşimleri, hastaya özgü ağ modelleme, ağ güdümlü tümör gruplama

To My Family,

ACKNOWLEDGMENTS

First of all, I would like to thank my dear advisor Assoc. Prof. Dr. Nurcan Tunçbağ for all of her endless support and time for me. This work has been accomplished thanks to her guidance. I have always felt blessed and honored to work with Dr. Tunçbağ who has been always wise, patient and professional. I will always be grateful for all the opportunities given to me to open my horizon, improve myself and take a step into the research world.

Furthermore, I would like to thank my thesis committee members: Prof. Dr. Tolga Can, Assoc. Prof. Dr. Yeşim Aydın Son, Assoc. Prof. Dr. Mehmet Somel and Assoc. Prof. Dr. Tunca Doğan for their valuable time, feedbacks and consideration to be in my thesis committee.

I would like to express my gratitude to my dear colleagues Cansu Demirel, Gökçe Senger, Meriç Kınalı, Alperen Taciroğlu, Muazzez Çelebi Çınar, Elif Bozlak and Evrim Fer for making my academic journey as an enjoyable and productive experience, establishing an environment that is based on mutual respect, trust and support, and particularly being also my lifelong friends. Moreover, I am especially thankful to my childhood friends Burcu Baydarlıoğlu, Gizem Kars, Mert Can Özturan, Yağmur Kılıç, Murat Akbaba and Didem Toker for their friendship for more than ten years that have been full of fun, tears and support.

Lastly, I would like to thank my beloved family for their endless love, support and belief. As a special thanks: to my mother, Nurcan Dinçer, for teaching me to dream, resist, succeed, respect and love; to my father, Mustafa Dinçer, for opening up my horizon, letting me asking so many questions and making me curious; to my one and only sister, Yağmur Dinçer, for always standing by my side, supporting and encouraging me when I needed all the time; and to my Vira. I also owe a special thanks to my beloved Yiğit Özyurt with who I grow up, learn and explore, for always being in my life, all of his support and patient. I hope you all know how blessed and happy I feel to have you as a family in my life.

TABLE OF CONTENTS

ABSTRACT iv
ÖZ v
DEDICATION vi
ACKNOWLEDGMENTS vii
TABLE OF CONTENTS ix
LIST OF TABLES xi
LIST OF FIGURES xii
LIST OF ABBREVIATIONS xiv
CHAPTERS 1
1. INTRODUCTION 1
2. LITERATURE REVIEW 5
2.1. Effects of mutations on tumorigenesis 5
2.2. Analysis of mutations by integrating structural information 7
2.3. Analysis of mutations by integrating network-based approaches 8
3. MATERIALS AND METHOD 13
3.1. Overview of the Method 13
3.2. Data 14
3.2.1 The Cancer Genome Atlas (TCGA) 14
3.2.2 The Universal Protein Resource (UniProt) 14
3.2.3 Protein Structure Databases 15
3.2.4 Disease Association of Mutations 17
3.2.5 Cancer Driver Effects of Mutation and Mutated Genes 17
3.2.6 PFAM 19
3.2.7 Cell Model Passports 19
3.2.8 Genomics of Drug Sensitivity in Cancer (CancerRxGene) 19
3.2.9 Interaction Reference Index (iRefWeb) 19

3.3. Identification of the 3D spatial clusters 20
3.4. Identification of protein regions and the effect of the mutations 22
3.5. Reconstruction of patient-specific sub-networks and grouping of the patients. 23
3.5.1 Network reconstruction 23
3.5.2 Network-guided grouping of the patients 24
3.6. Linking the patient groups to drug response 25
4. RESULTS 27
4.1. 3D Spatial Organization of GBM Mutations 27
4.2. Characteristics of the GBM mutations from the structural and chemical perspectives 34
4.2.1. Structural positions and chemical characteristics of the GBM mutations 34
4.2.3. Characteristics of interface mutations 39
4.3. Patient specific sub-networks from mutation profiles and patient groups from pathway similarities 42
4.4. Potential therapeutic targets for each patient group 48
5. DISCUSSION AND CONCLUSION 55
REFERENCES 59

LIST OF TABLES

Table 4.1. Number of mutations mapped to protein structural regions. 34
Table 4.2. Disease association of singleton and patch mutations in the interface region of the hubs and the rest. 37
Table 4.3. Numerical information of mutations in each interface type. 39

LIST OF FIGURES

Figure 3.1. Overview of the methodology. 13
Figure 3.2. Mapping mutated residues on 3D protein structures. 15
Figure 3.3. Identification of 3D spatial organization of mutations on proteins. 21
Figure 3.4. Identification of different protein regions on 3D structures. 22
Figure 3.5. Identification of affected protein-protein interactions and biological pathways. 23
Figure 3.6. Summary of GBM cell lines patient group linkage. 26
Figure 4.1. Mutation profile of GBM patients. 28
Figure 4.2. 3D Patch profile of GBM patients. 28
Figure 4.3. Kaplan-Meier survival curves of the patient groups in the mutation profile. 29
Figure 4.4. Kaplan-Meier survival curves of the patient groups in the patch profile. 29
Figure 4.5. Mapping patches of frequently mutated hub proteins to their functional domains. 30
Figure 4.6. Example of different domains of protein of PIK3R1 gene. 31
Figure 4.7. Histogram of the patch sizes for intra- and inter- patches. 32
Figure 4.8. Examples of 3D mutation patches on protein structures. 33
Figure 4.9. Fraction of core, interface and surface mutations according to their 3D spatial organizations. 35
Figure 4.10. Fraction of chemical property changes of mutated driver proteins according to their physical locations. 36
Figure 4.11. Fraction of mutations according to their PolyPhen-2 disease association in different locations. 38
Figure 4.12. EVmutation disease association score distribution of mutations on different locations. 38
Figure 4.13. Representation of two types of proteins having one or multiple interface regions. 40
Figure 4.14. Network representation of RB1 and PIK3CA interface mutations. 40
Figure 4.15. 3D spatial organization of interface mutations. 42
Figure 4.16. Co-clustering frequency matrix. 43
Figure 4.17. Kaplan-Meier survival plots of the patient groups classified with NMF and consensus clustering. 44
Figure 4.18. Enrichment of KEGG pathways across the patient groups. 47
Figure 4.19. Predominant 3D patches in each patient group. 47
Figure 4.20. Merged network of Group 1. 48
Figure 4.21. Patient groups, cell line and drug linkages. 49

Figure 4.22. Hypothetical therapeutic proposal for Group 2 patients by using ATM as target protein. 50
Figure 4.23. Hypothetical therapeutic proposal for Group 2 patients by using CDK1 as target protein. 51
Figure 4.24. Hypothetical therapeutic proposal for Groups 2 and 5 patients by using CHEK2 as target protein. 51
Figure 4.25. Hypothetical therapeutic proposal for Groups 3 and 5 by using CSF1R and PDGFRB as target proteins, respectively.....52
Figure 4.26. Hypothetical therapeutic proposal for Group 5 patients by using SRC as target protein. 53

LIST OF ABBREVIATIONS

3D	Three Dimensional
3did	Three-Dimensional Interacting Domains
ATM	Ataxia Telangiectasia Mutated
ATP	Adenosine Triphosphate
BIND	The Biomolecular Interaction Network Database
BioGRID	Biological General Repository for Interaction Datasets
CaMP	Cancer Mutation Prevalence
CancerRxGene	Genomics of Drug Sensitivity in Cancer
CDC25	Cell Division Cycle 25
CDK	Cyclin Dependent Kinase
CHASM	Cancer-Specific High-throughput Annotation of Somatic Mutations
CORUM	The Comprehensive Resource of Mammalian Protein Complexes
COSMIC	The Catalogue of Somatic Mutations in Cancer
CRISPR-KO	Clustered Regularly Interspaced Short Palindromic Repeats Knock Out
CSF1	Colony Stimulating Factor 1
CSF1R	Colony stimulating factor 1 Receptor
DIP	Database of Interacting Proteins
DNA	Deoxyribonucleic Acid
DoCM	Database of Curated Mutations
ECLAIR	Ensemble Classifier Learning Algorithm to predict Interface Residues
EGFR	Epidermal Growth Factor Receptor
eQTL	Expression Quantitative Trait Loci
ERK	Extracellular Signal Regulated Kinase
FDR	Fold Discovery Rate
FGF	Fibroblast Growth Factor
GBM	Glioblastoma Multiforme
GSEA	Gene Set Enrichment Analysis
GTP	Guanosine Triphosphate

HPRD	The Human Protein Reference Database
IARC	The International Agency for Research on Cancer
IC50	Half Maximal Inhibitory Concentration
ICGC	International Cancer Genome Consortium
iRefWeb	Interaction Reference Index
JAK/STAT	Janus Kinase/Signal Transducers and Activators of Transcription
KEGG	Kyoto Encyclopedia of Genes and Genomes
ln	Natural Logarithm
MAPK	Mitogen-Activated Protein Kinase
MEMo	Mutual Exclusivity Modules
MI-Score	MINT-inspired score
MINT	Molecular Interaction Database
miRseq	Micro RNA Sequencing
MMPI	Mammalian Protein-Protein Interaction Database
mRNA	Messenger Ribonucleic Acid
MuSiC	Mutational Significance in Cancer
MutSig	Mutation Significance
NBS	Network Based Stratification
NCG	The Network of Cancer Genes
NMF	Non-Negative Matrix Factorization
NMR	Nuclear Magnetic Resonance
NTA	Network Topology Analysis
OncoKB	Oncology Knowledgebase
OPHID	The Online Predicted Human Interaction Database
ORA	Over-Representation Analysis
PCSF	Prize-Collecting Steiner Forest
PDB	The Protein Data Bank
PDGFRB	Platelet Derived Growth Factor Receptor Beta
PI3K	Phosphatidylinositol-3-kinase
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha
PIK3R1	Phosphoinositide-3-Kinase Regulatory Subunit 1
Poly-Phen2	Polymorphism Phenotyping v2
PPI	Protein-Protein Interaction
PRISM	Protein Interactions by Structural Matching
RAP1	Ras-Associated Protein-1
RNA	Ribonucleic Acid

RNAseq	RNA Sequencing
SpacePAC	Spatial Protein Amino Acid Clustering
TAMs	Tumor-Associated Macrophages and Microglia
TCGA	The Cancer Genome Atlas
UnipArc	The Universal Protein Resource Archive
UniProt	The Universal Protein Resource
UniProtKB	The Universal Protein Resource Knowledgebase
UniRef	The Universal Protein Resource Reference Clusters
VEGF	Vascular Endothelial Growth Factor

CHAPTER 1

1. INTRODUCTION

Molecular alterations on genome accumulate through time and may disrupt cellular functions leading to cancer. Continuous accumulations of genomic alterations can result in elimination of cell due to deleterious mutations or turning into cancer cell by gaining advantages to proliferate, adapt, and metastasize (Stratton, Campbell, & Futreal, 2009). High-throughput next generation sequencing technologies as well as small scale targeted sequencing techniques provide vast amount of information for thousands of genes, transcripts and proteins at the same time. Therefore, identification of mutations providing advantages for tumorigenesis is one of the key focus since proteins having these mutations can be the favored molecular targets in drug discovery and cancer therapy. However, extreme heterogeneity in mutation profiles between and within tumors as well as among individuals of the same type of cancer obstructs direct identification. Further, mutations are also divided according to their impact on tumorigenesis, since not all mutations can result in transformation from normal to cancer cells. As an another barrier, cell has complex and dynamic nature and functions by interconnected cellular pathways. In addition to these barriers, epigenetic and post-translational factors are also important in cancer progression and therapeutic resistance. Thus, not all drugs would be effective for all patients. In order to improve the understanding of cancer and its causatives, components of the cellular mechanisms and their relations to each other are needed to be comprehensively analyzed.

Molecular mechanism of cell is mainly carried out by proteins and their interactions. Through proteins functioning by interacting to each other, to nucleic acids and small compounds; a cellular network is created in which nearly every component is connected to each other directly or indirectly. While it is important to know who is connecting with who, questions as through where and how they are connecting are important as well. Accumulation of mutations on the proteins affect the interaction profile by disrupting the stability of the protein or only inhibiting or activating specific interfaces. It has already been reported that disease associated mutations are prone to affect protein interacting sites and perturb interactions (Sahni et al., 2015). If protein stability and folding are affected

dramatically, proteins disappear from rewired cellular network since they are become nonfunctional. Yet in some cases, mutations are happened on interacting site of the proteins which can lead to continuous activation of proteins related to growth, proliferation or resistance to apoptosis; or inhibition of specific interfaces which lead to disappearing of connections between mutated proteins and partners from affected interactions (Sahni et al., 2015). Different sets of mutations from different patients can form distinct profiles of interactions and eventually different phenotypes of disease (An, Gursoy, Gurgey, & Keskin, 2013; Engin, Kreisberg, & Carter, 2016; Ozdemir, Gursoy, & Keskin, 2018; Ozdemir, Halakou, Nussinov, Gursoy, & Keskin, 2019). Therefore, structural positions of mutations and interaction patterns of mutated proteins are critical for meaningful inference about the impact of the mutations on phenotypes.

Genome-wide mutation profiles are promising resource to elucidate the underlying mechanism of cancer, to classify the patients according to their genetic backgrounds and to propose potential therapies when systems level strategies are applied. Considering the amount of omics data, computational approaches are essential for analyzing the effects of mutations on proteins, protein interactions and functional pathways in a patient-specific way. Several studies focus on impact of mutation on PPIs (An et al., 2013; Engin et al., 2016; Ozdemir et al., 2018; Ozdemir et al., 2019), protein structures (Gao et al., 2017; Kamburov et al., 2015; Meyer et al., 2016; Niu et al., 2016; Ryslik et al., 2014; Tokheim et al., 2016) and interaction perturbations (Engin et al., 2016; Kar, Gursoy, & Keskin, 2009; Porta-Pardo, Garcia-Alonso, Hrabe, Dopazo, & Godzik, 2015; Sahni et al., 2015). There are several studies concentrated on structural clustering of mutations by using physical contacts of mutated residues (Gao et al., 2017), distance between mutated residues or significant proximity of mutated residue pairs (Meyer et al., 2016; Niu et al., 2016; Ryslik et al., 2014) to identify the marker mutations or affected pathways which leads to cancer formation and progression. On the other hand, there are also perturbation-based approaches which rely on rewiring of cellular networks of tumorigenic cells. Incorporating different omics data together has been applied to identify the target mutations, proteins or pathways (Acuner Ozbabacan, Gursoy, Nussinov, & Keskin, 2014; Chuang, Lee, Liu, Lee, & Ideker, 2007; Ciriello, Cerami, Sander, & Schultz, 2012; Drake et al., 2016; Dutkowski & Ideker, 2011; Engin, Guney, Keskin, Oliva, & Gursoy, 2013; Hofree, Shen, Carter, Gross, & Ideker, 2013; Kim, Wuchty, & Przytycka, 2011; Vandin, Upfal, & Raphael, 2011; Wu, Dong, & Wei, 2018; Xi, Li, & Wang, 2017). The network-based stratification (NBS) approaches to the heterogeneity problem with an integrative perspective and incorporates mutation profiles and molecular gene networks to classify the patients (Hofree et al., 2013). In order to utilize activity of proteins, Drake et al used mutations, transcriptional and phosphoproteomic data together (Drake et al., 2016). Distinctly, Engin et al integrate structural data with mutation and PPI information to highlight the mechanisms of metastasis (Engin et al., 2013).

One of the deadliest type of brain tumor, Glioblastoma Multiforme (GBM) is well known for its aggressiveness, resistance and heterogeneity, which makes the disease as incurable. The average survival is between twelve to fifteen months despite great development in medicine (Bleeker, Molenaar, & Leenstra, 2012; Zong, Verhaak, & Canoll, 2012). By the

help of next generation sequencing, mutation screens became available, thus they indicated that GBM has enormous genomic heterogeneity which is the main obstacle to develop effective therapies. In this thesis, we aimed to decrease the heterogeneity among GBM patients whom data comes from The Cancer Genome Atlas (TCGA) (Tomczak, Czerwinska, & Wiznerowicz, 2015), to classify them, and to propose potential therapeutics and their responses. We started from finding the spatial arrangement of the mutations which are the clusters having mutated residues both in physical contacts or close proximity, then we continued with reconstruction of patient specific PPI subnetworks for each patient. Since cellular pathways are composed of several proteins interacting each other, we needed intermediate molecules which are connecting mutated proteins. We applied Forest module of Omics Integrator (Tuncbag et al., 2016) for network modelling and reconstructed patient specific subnetworks primarily affected by the set of mutations across patients. Since functional pathways were completed by the algorithm, we reduced networks into significantly enriched sets of pathways which were used to classify the patients into clinically similar groups. Each patient group was associated with the survival profile of the patients in the group significantly (P-value: 0.0408). Ultimately, we used mutation and drug response data of GBM cell lines in order to link patient groups with GBM cell lines, then propose hypothetical therapeutics on the basis of patch distribution of patient groups.

In Chapter 2, we covered what is the impact of mutations on tumor formation and progression, how mutation profiles can be used to identify druggable targets or classify the patients by reviewing strategies in the literature focusing on both structural information and network-based approaches.

In Chapter 3, we described both the data and methodology in detail. We started explaining data, then continued with how spatial organization of the mutations were found, how the structural and physicochemical characteristics of GBM mutations were evaluated, what is the algorithm based on to reconstruct patient sub-networks and how we implemented it, and lastly, how we connected GBM cell lines to patient groups and how we inferred the potential therapeutics.

In Chapter 4, we explained how 3D patch organization of a mutation decreased the inter-patient heterogeneity and how network guided analysis can stratify patients into five groups. Moreover, we described physicochemical and structural consequences of mutations on structures and behavior of cancer related mutations and proteins. Eventually, we explained how potential therapeutics were proposed for patient groups on the basis of 3D patch profiles.

In Chapter 5, we discussed our results and how they would contribute the perspective of mutation analysis. We also indicated the importance of network-based approach for integrating different data to infer biologically meaning outcome.

CHAPTER 2

2. LITERATURE REVIEW

2.1. Effects of mutations on tumorigenesis

Cells are strictly controlled, complex and dynamical systems governed by the genome. Any alteration on DNA has the potential to affect the dynamical system, which leads to different diseases such as Mendelian and complex diseases (Amberger, Bocchini, & Hamosh, 2011; Hindorff et al., 2009) including different types of cancer. These alterations could be hereditary variations, acquisition of exogenous DNA or RNA sequences from viruses that are associated with various cancer types such as human papilloma viruses (HPVs), hepatitis-B virus (HBV), hepatitis-C virus (HCV), Epstein Barr virus (EBV), human T lymphotropic virus 1 (HTLV-1) and human herpesvirus 8 (Talbot & Crawford, 2004; Walboomers et al., 1999), and somatic changes on the cancer genome such as substitutions, indels (insertions or deletions), DNA rearrangements and copy number changes. In addition to genomic alterations, an epigenetic mechanism could provide advantages for tumorigenesis by activating or deactivating cancer related genes through DNA methylation or histone modifications. Through these continuous accumulations of genomic alterations, some cells are eliminated due to deleterious mutations; yet some of them are selected due to their acquired capability to proliferate, adapt, and metastasize (Stratton et al., 2009).

Across all of these mechanisms, somatic mutations are the major causative factor in most human cancers (Weir, Zhao, & Meyerson, 2004). However, not all somatic mutations result in cancer, thus mutations were conventionally divided into two categories as driver and passenger mutations. Driver mutations provide growth or drug resistance advantages to tumor cells while passenger mutations are not seemed beneficial for tumorigenesis or drug resistance. However, lately, another class of mutation has been also defined as latent mutation (Nussinov, Jang, Tsai, & Cheng, 2019). These mutations are passenger mutations until they transform into driver mutations in a specific context or under different factors such as environmental factors or conformational changes; or these mutations have not been discovered as driver yet. These disease causing driver somatic mutations can give advantages to the tumor cells by changing the expression of corresponding proteins, disrupting folding or stability of the protein or perturbation in the interaction profile of the

protein, which can be loss of all or specific interactions or rarely gains of interactions (Sahni et al., 2013). While activation of oncogenes by mutations is advantageous for tumorigenesis, repression of tumor suppressor genes leaves cells uncontrolled and unprotected for resistance to cell death, excessive proliferation, enhancement of invasiveness, and among others. In order to understand these complex molecular changes in cancer and design effective therapies, driver and passenger mutations should be distinguished.

High-throughput next generation sequencing technologies, unlike small scale targeted sequencing techniques, provide information for thousands of genes, transcripts and proteins at the same time. The large-scale cancer genome sequencing projects including The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) (Zhang et al., 2011) provide a large amount of data that can help the development of different approaches for omics research. Rather than analyzing effects of individual mutation for genomic characterization, nowadays comprehensive mutation profiles of the patients can be obtained, and this brings a new precision medicine approach as treatments according to the genetic background of patients rather than only disease types (Drew, 2016; T, M, Jose, Chandran, & Zachariah, 2009). Since molecular heterogeneity in mutation profiles between and within tumors as well as among individuals of the same type of cancer is enormous, classification of the patients is crucial for a better prognosis and an effective treatment.

One mechanism through which cellular organization and tissue homeostasis are altered in the context of cancer is by disrupting interactions of proteins. Since proteins are not in isolation but function by interacting with each other rather than in isolation, studying individual mutations is not enough to understand their role in cancer formation and progression (Sahni et al., 2013). Additionally, cancer does not result from a specific mutation but a set of mutations that are changing the interaction profile (interactome) of the cell to become more proliferative and more resistant to apoptosis or drug. Therefore, cancer cells use different signaling pathways which lead to the same outcome. As a result, genome-wide mutation profiles from high-throughput sequencing technology is a promising resource to understand the underlying mechanism of cancer, classify the patients according to their genetic make-ups and propose effective therapies when systems level strategies are applied to analyze the overall changes in the dynamic signaling mechanism of the cell.

The major questions in mutation analysis to guide the design of new therapeutics are how we can distinguish driver mutations from passenger mutations and how can we identify perturbed cellular pathways which lead to tumorigenesis. The most conventional approach for identification of driver mutations is a significant recurrence of the mutation based on the hypothesis that frequency of driver mutations will be higher than the random passenger mutations (Dees et al., 2012; Greenman et al., 2007; Wood et al., 2007). Dees et al uses this hypothesis and developed Mutational Significance in Cancer (MuSiC) algorithm to separate driver and passenger mutations by using the mutation rates (Dees et al., 2012). Additionally, in (Wood et al., 2007) they calculate cancer mutation prevalence (CaMP)

score which defined as the probability that the number of mutations in a gene having a frequency that is higher than that expected for each gene. They further continued with the genes having CaMP scores higher than 1 and labeled them as candidate cancer genes. While it is reported that driver genes have increased mutation frequency than passenger genes, yet a small range of mutations represents driver genes. By using only the frequency of mutation accumulation, functional but rare mutations are lost. However, researches based on more detailed and larger scale analysis of mutations are more informative in terms of identification of driver genes and mutations and the functional pathways that are affected. In the following sections, structural and network-based approaches and their applications to cancer studies are comprehensively explained.

2.2. Analysis of mutations by integrating structural information

Whereby large-scale sequencing analysis, an enormous amount of genomic data is available. Moreover, development in crystallography and microscopy, as well as prediction methods, us to obtain a large amount of structural information for proteins. Merging three-dimensional (3D) structural information with large-scale mutation information has the potential to enlighten the impacts of cancer mutations on the biological functions of proteins and their physical interactions. In this section, studies focusing on structural analysis for the identification of mutation effects and their related pathways will be detailed.

Several studies concentrated on the identification of cancer related mutations and pathways by utilizing structural aspects of mutational proteins coming from large-scale databases. There are plenty of researches based on the hypothesis that driver mutations may accumulate specific functional site on proteins and leads to tumorigenic cellular activities (Gao et al., 2017; Kamburov et al., 2015; Meyer et al., 2016; Niu et al., 2016; Ryslik et al., 2014; Tokheim et al., 2016). Conventional methods are mostly focused on frequency-based approaches. As an example, Buljan et al. collected all the mutations from TCGA and ICGC patients for 40 different cancer types to count how many mutations occur for each residue on protein sequences for all patients. By this way, they found that hotspot residues which can provide significant affinity to the interaction than other residues, accumulate a significantly higher number of mutations than their surrounding residues. Then, they found that these hotspot mutations are often located on interfaces which indicates that any changes in interaction profile could be the reason for tumorigenic switching in the cell (Buljan, Blattmann, Aebersold, & Boutros, 2018). Besides from sequence-based recurrence approaches, several studies extended this analysis, they focused on the accumulation of the mutations on protein structures. Gao et al. mapped mutations onto 3D structures of corresponding proteins in order to get mutation clustering. If residues are physically connected and both of them are mutated, then they can form a cluster in which recurrence is analyzed. By this method, they identified potential driver genes from rare mutations (Gao et al., 2017). With a similar perspective, Meyer et al. developed mutation3D method in order to identify genes in cancer formation and progression by using structural details of proteins. They again mapped mutated residues

on protein structures. Rather than physical contacts, they used the distance between alpha-(α)-carbon atoms of mutated residues and complete-linkage clustering strategy which is a hierarchical clustering method. Within a specified maximum dissimilarity value which is the permitted maximum distance between α-carbon atoms, all mutated single elements are merging with their nearest neighboring clusters (Meyer et al., 2016). Additionally, another similar concept is also applied in (Ryslik et al., 2014) which is called Spatial Protein Amino Acid Clustering (SpacePAC). In this study, they used mutations from the Catalogue of Somatic Mutations in Cancer (COSMIC) and structures from only Protein Data Bank (PDB). They created three non-overlapping spheres with different radii corresponding protein, in which there are as many as mutations. After the normalization step, they obtained the most significant clusters which can identify the cancer related mutations. Niu et al. use spatial clustering of mutations by first calculating significant pairwise proximity of each residue in proteins (Niu et al., 2016). Then, they used these significant proximal pairs as seed nodes and iteratively added mutated residues if they are significantly paired with other mutated residues in these seed nodes. With this approach, they identified both mutation and mutation-drug clusters on 3D structures and proposed druggable mutations. Moreover, in order to better explain the molecular consequences of driving mutations on tumor cells, COSMIC (Futreal et al., 2004) started an effort as COSMIC-3D (Malhotra et al., 2019). The main difference in COSMIC-3D, they concentrated interface regions of protein-protein, protein-nucleic acid and protein-ligand interactions on which they analyzed the impact of mutations. They observed that the most recurrent mutations were clustered in binding sites which can make inhibitors less effective, reduces the DNA/RNA binding activity and change the cell signaling by changing interactions. Stehr et al. spatially clustered COSMIC mutation from eight cancer types (breast, prostate, stomach, colon, pancreas, thyroid, kidney, lung) (Stehr et al., 2011). In order to understand the mechanism of gain and loss of functions, they functionally analyze the structural impact of the mutations on tumor suppressor and oncogenes and found that tumor suppressor lost their function generally by mutation destabilization on core regions yet, gain of function of oncogenes is often related with specific mutations on the functional sites of the proteins which are generally ATP or GTP binding sites. These methods can answer the questions asking which mutations are important and which residues are targetable. Since a cellular network composed of interacting molecules, a comprehensive understanding of interaction perturbations is critical for enlightening altered signaling mechanisms. Therefore, this understanding needs both structural aspects of proteins and their interaction, and network-based approaches. In the next section, different perspectives of network-based approaches will be discussed.

2.3. Analysis of mutations by integrating network-based approaches

Cellular mechanism consists of direct and indirect interactions as we called interactome between proteins, nucleic acids, and metabolites. In the perspective of protein-protein interactions (PPIs), interactome is a network whose nodes are proteins and edges are the physical interactions between them and composed of functional cross-linked signaling

pathways. Characterization of these networks is critical for fully understanding the contribution of individual mutation to phenotype in a specific context since a mutation on a gene can have an impact larger than impact only on the protein by affecting cellular pathways. As an example, a mutation on Met receptor tyrosine kinase which is activated by hepatocyte growth factor and mediates the mitogen-activated protein kinase (MAPK) and phosphatidylinositol-3 kinase (PI3K) pathways, results in loss of Cbl E3-ligase tyrosine kinase binding site in the juxtamembrane region and thus disruption of the ubiquitination process of the receptor. Mutation in this region induces continuous activation of MAPK and PI3K pathways in lung cancer which indicates significant tumor growth *in vivo* (Kong-Beltran et al., 2006; Pawson & Warner, 2007). There are different approaches using network-based strategy to reveal genotype-phenotype relation such as integrative and perturbation-based approaches. While perturbation studies focus on the rewiring of the cellular networks in order to highlight the affected key pathways, integrative studies try to elucidate hidden mechanisms by incorporating as much as data together. Moreover, focusing on interaction changes, it is inevitable to use structural details of the mutated proteins. Therefore, this section begins with studies based on integrative methods, then continue with the studies merging network-based approaches with structural information.

Integrative approaches can complete the hidden components of the network by using mutation, clinical, gene interaction, gene expression or phosphorylation data (Acuner Ozbabacan et al., 2014; Chuang et al., 2007; Ciriello et al., 2012; Drake et al., 2016; Dutkowski & Ideker, 2011; Engin et al., 2013; Hofree et al., 2013; Kim et al., 2011; Vandin et al., 2011; Wu et al., 2018; Xi et al., 2017). As an example, the network-based stratification (NBS) tried to overcome the heterogeneity problem by integrating mutation profiles with molecular gene network. After constructing a binary matrix for mutation profiles of each patient as 1 and 0 representing mutated and non-mutated genes, respectively, these matrices were projected on a gene interaction network. By network propagation followed by non-negative matrix factorization and consensus clustering, patients were classified. According to this classification, the group showing the worst survival has 20 genes for fibroblast growth factor (FGF) signaling pathway, which causes resistance for platinum and anti-VEGF (Vascular Endothelial Growth Factor) therapy. Therefore, besides classification, network-based stratification method can identify the affected cellular pathways which can be used for drug targets (Hofree et al., 2013). Further, Wu et al. use another network-based method to integrate mutation and gene expression data in GBM for elucidating dysregulated pathways in the disease and found there are two main dysregulated pathways (epidermal growth factor receptor related pathways and TP53 associated pathways) both represent different subtypes of GBM (Wu et al., 2018). In another integrative network-based approach, mutations, transcriptional and phosphoproteomic data were used together to model patient-specific pathways in prostate cancer. By using control and metastatic prostate cancer, they identified the activity of a protein from differential expression and phosphorylation of its targets. Then these activated regulators integrated with somatic mutation data from various resources in order to understand if these regulators are significantly related with the mutated genes and found that kinases are nearby in the common pathways with mutated genes. They also

found sub networks for each group which is enriched in AKT/mTOR/MAPK signaling, nuclear receptor signaling and the cell cycle (Drake et al., 2016). Kim et al. utilized the integration of phenotypic, genomic and interaction data to identify disease associated genes and dysregulated pathways by applying network-based strategy. They first selected target genes which are differentially expressed genes, then found the associations between mutated genomic loci and expression level changes of target genes with an expression quantitative trait loci (eQTL) analysis. Additionally, Engin et al, integrated PPIs, mutations and structural details of the protein interfaces together in order to enlighten genotype-phenotype association of metastasis process. They applied reverse engineering by using guilt-by-association principle and built the phenotype specific PPIs for both breast and lung cancer metastasis from the primary tumor in breast cancer. In order to understand the mutation effects, mutations were also mapped on these phenotype specific networks. As a result, they found lung metastasis progression has a relationship with the immune system and infectious diseases, yet this association was not found in brain metastasis. This is a good example of reverse engineering approach with integrative strategy (Engin et al., 2013). Ciriello et al. identified driver networks rather than driver mutations or driver genes by using Mutual Exclusivity Modules (MEMo) algorithm which can merge copy number variation and somatic mutation information (Ciriello et al., 2012). These driver networks composed of proteins having mutation recurrently, function in the same biological process and have mutually exclusive genetic alteration. Application of this algorithm to GBM data highlighted several core modules involving TP53, RB, PI3K signaling. Finally, they tried to find putative causal genes by utilizing pathways between causal and target genes through molecular interaction network created by PPIs, phosphorylation, and gene regulatory networks (Kim et al., 2011). Consequently, in order to overcome the obstacles of heterogeneity in tumors and develop personalized therapeutic strategies, reverse engineering from mutations to networks and eventually pathways is one of the promising approach.

Mutations in the same protein may result in different interaction profiles and eventually different disease phenotypes. While mutations changing the stability of the protein can result in severe alterations in its overall interactions, mutations affecting only one interface of a protein having several interfaces can change the interaction profile of the protein by lost and gained interaction partners. Both Meyer et al. (Meyer et al., 2018) and Mosca et al (Mosca, Ceol, & Aloy, 2013) developed structurally enriched protein interface databases. While Interactome3D (Mosca et al., 2013) has 3D coordinate information, Interactome Insider (Meyer et al., 2018) only provides Uniprot indices of interacting residues. Comprehensive structural information of protein-protein interaction networks provides analysis of mutation impact on interactions and cellular pathways. Moreover, it is reported in (Sahni et al., 2015) that disease associated mutations are generally affecting protein interactions leading to changes in the cellular functions. Thereof, several studies focused on the impact of disease associated mutations on interaction profiles (Engin et al., 2016; Kar et al., 2009; Porta-Pardo et al., 2015; Sahni et al., 2015). For the perturbation analysis, it is critical to know the interface regions, thus studies based on perturbation approach enriches network-based approaches with structural information. Sahni et al. give importance to rewiring nature of the network (Sahni et al., 2015). They compare the

mutated and non-mutated interactome and found that nearly 60% of disease-associated missense mutations perturb protein-protein interactions by complete or partial loss of interactions. While mutations can induce complete loss of interactions by affecting the folding and stability of the proteins, they can also affect only one interface site of the protein and the other interfaces can still bind to their partners which they called this phenomenon as “edgotype”. Thus, different mutations on the gene can affect different interfaces on corresponding protein, which creates different interaction profiles and thus different disease phenotypes. Therefore, affected interactions should be taken into account for an appropriate evaluation of the rewiring of cellular networks of tumor cells. As an additional benefit of considering edgotype, this analysis can elucidate specific targets for both prognosis and personalized therapy. In the study of Engin et al. (Engin, Hofree, & Carter, 2015), they tried to create a method which can find the candidate cancer pathways by using mutations, protein structures, and PPIs. They first mapped the mutation on protein structures and took the mutations on interface and core regions. According to the location of the mutation, they erased the edges of the protein. If a mutation is on the core, they assumed that it will affect the protein stability and deleted all the edges of the protein from structurally resolved protein-protein interaction network. Yet they erased only the edges on which interface mutation is found if a mutation is not on core region. With this study, the researchers showed that somatic mutations can have differential consequences even in the same protein.

CHAPTER 3

3. MATERIALS AND METHOD

3.1. Overview of the Method

GBM patients have extremely heterogeneous mutation profiles which is an obstacle to develop effective therapies. In order to overcome the heterogeneity, to group the GBM patients and to propose potential targeted therapies for the patient groups; we applied a systems level approach using (i) three dimensional (3D) spatial organization of the mutations, (ii) organization of mutated proteins in patient specific networks and (iii) drug responses of the GBM cell lines. We proceed in two ways: first one is to calculate mutation patches (3D spatial clusters) for all mutated proteins, and second is to reconstruct patient specific networks and group patients according to their pathway similarities. Then, we found the signature 3D mutation patches for each patient group and used this information to link GBM cell lines and patient groups to infer drug responses of patient groups. The overview of the method is indicated in **Figure 3.1**.

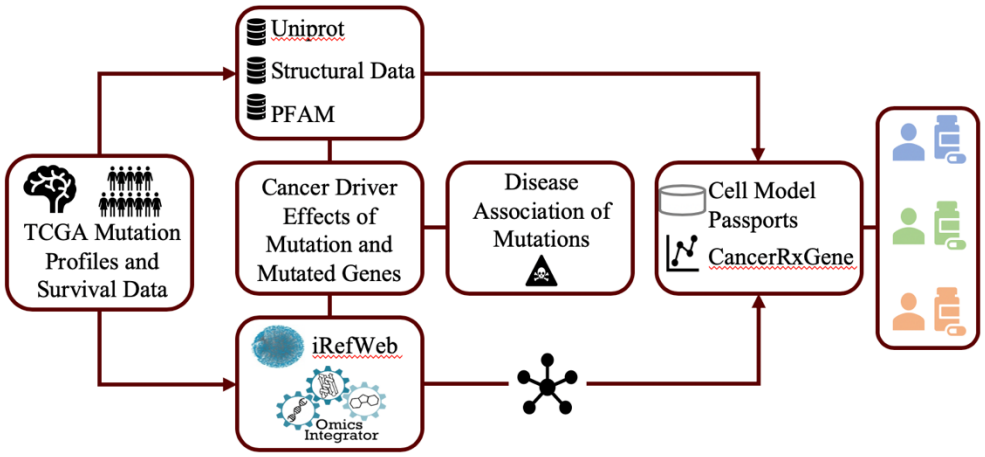


Figure 3.1. Overview of the methodology. TCGA-GBM mutation profiles of the patients were retrieved. The 3D organization of each mutation was found. Each cancer related driver protein having at least one mutation was used to reconstruct patient-specific sub-

networks. Finally, the sub-networks were used to classify the patients, to find signature patches in each patient group and to demonstrate the help of 3D organization in overcoming heterogeneity. Eventually, we investigated whether patient groups have any association with the clinical outcome by using cell line drug sensitivity data.

3.2. Data

In this section, we described all the information used for this study in detail.

3.2.1 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research et al., 2013) is a project which generates the cancer -omics data such as genomic, epigenomic, transcriptomic and proteomic data in 33 cancer types. In this study, we downloaded protected maf files including mutation profiles of 290 GBM patients from TCGA-GBM project (Brennan et al., 2013). We filtered 15399 missense, nonsense and frameshift mutations and their changes on the protein sequences in order to use in our further analysis. Additionally, we collected survival data of each patient in order to perform survival analysis between patient groups.

3.2.2 The Universal Protein Resource (UniProt)

The Universal Protein Resource (UniProt) (UniProt, 2008) is a platform having protein sequence and annotation such as function, subcellular localization, interaction, 3D structure etc. Its databases are the UniProt Reference Clusters (UniRef), the UniProt Archive (UniParc), the UniProt Knowledgebase (UniProtKB) and Uniprot Proteomes. UniRef was created by clustering protein sequences or sequence fragments from both UniprotKB and UniParc in all organisms deposited in Uniprot as sets of sequences containing 100%, 90% and 50% sequence similarities (UniRef100, UniRef90, UniRef50, respectively) to the longest sequence in order to decrease the database size and increase the transfer speed. UniParc stores all the protein sequences form variety of sources to prevent the redundancy by giving an unique identifier for all the identical sequences regardless of being proteins of different species. As an important information, Uniprot also deposists complete sets of proteins in organisms which is called proteome of the organism. This, Uniprot Proteome stores 228,004 proteomes for different species including all the information from UniprotKB. From them, 16,485 proteomes were manually and algorithmically selected as reference to cover well-studied organisms. Lastly, UniProtKB deposits the annotation data about proteins from amino acid sequence to structural data coming from different publications and databases. It consists of two different sections: UniProtKB/Swiss-Prot which is manually curated and other is UniProtKB/TrEMBL which consists of computationally analyzed information. We retrieved *Homo sapiens* proteome in 20.11.2018 to get amino acid sequences, related gene

names and their synonyms, and also their status if corresponding Uniprot ID is reviewed or not (Proteome ID: UP000005640).

3.2.3 Protein Structure Databases

Since amino acid sequences of the proteins fold to be stable and functional with covalent and non-covalent bonds, every protein has its own 3D structure. Moreover, proteins work by interacting each other by their specific sites called "interfaces" and these interacting proteins also have unique structures. The structures of proteins and interacting proteins deposited in the databases and they can be both experimental or computationally predicted. While experimental structures are more accurate, the ratio of proteins having experimentally determined structures to proteins in the human proteome is very low. Also, experimental structures may not cover all the corresponding sequences fully. Therefore, we need computational methods to predict structures for filling the gaps. In this study, we retrieved experimental structures from Protein Data Bank (PDB) and computationally predicted structures from various resources. Since residue indices of the structure files generally do not follow the indices of Uniprot sequences, we aligned protein sequences of all structure files having at least one mutation to corresponding reference proteins' canonical sequences in order to obtain the structural positions of the mutated residues for further analysis (**Figure 3.2**).

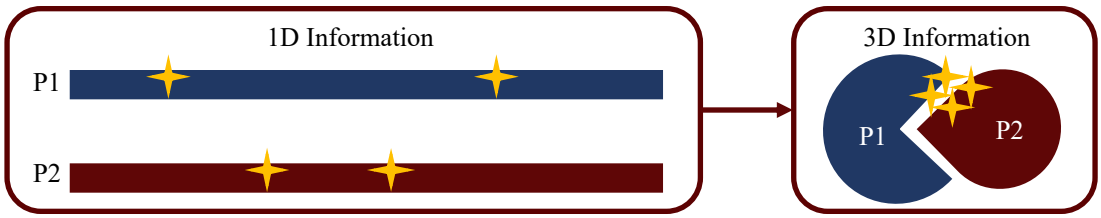


Figure 3.2. Mapping mutated residues on 3D protein structures. Mutations that are far away from each other in sequence may have in close proximity in 3D structure. Therefore, we mapped sequence indices of mutated residues on 3D structure of the corresponding protein.

a) RCSB Protein Data Bank (PDB)

Protein Data Bank (Berman, Kleywegt, Nakamura, & Markley, 2014) is an open achieve stored experimentally determined 3D structures of biological molecules/macromolecules. It consists of more than 150000 structures of proteins, nucleic acids, protein-nucleic acid interactions, and also protein-small molecule complexes. 3D structures are the results of mainly X-ray diffraction, nuclear magnetic resonance (NMR) and cyro-electron microscopy experiments. In this study, we collected 1865 experimentally determined PDB structures having at least one GBM mutation.

b) Modbase

Modbase (Pieper et al., 2011) is a database deposits computationally predicted protein structures. It uses Modeller algorithm (Fiser & Sali, 2003) based on homology (comparative) modelling which takes the sequence of target protein and searches for sequence similarity between all possible template proteins having 3D structures in order to identify the best structural template. This template is then used for formation of 3D structural model of the target protein. In this study, we retrieved 2430 computationally predicted structures from Modbase database.

c) Interactome3D

Interactome3D (Mosca et al., 2013) is a database for single and interacting protein structures constructed by both retrieving structural data from various databases and modeling protein interaction structures with its workflow. While it stores structures from both PDB and Modbase for single proteins, for interacting proteins it deposits all available PDB structures and predicts the missing structures by homology modeling using globular (PDB) and domain-domain (3did- Database of three-dimensional interacting domains (Mosca, Ceol, Stein, Olivella, & Aloy, 2014)) templates. Totally, it consists of more than 12000 structures for interacting proteins. In this study, we collected all PDB, Modbase and Interactome3D data from Interactome3D, only 74 structures of interacting proteins come from Interactome3D modeling algorithm.

d) Protein Interactions by Structural Matching (PRISM)

Protein Interactions by Structural Matching (PRISM) (Tuncbag, Gursoy, Nussinov, & Keskin, 2011) is a method for prediction of protein-protein interactions and also construction of their structures computationally. While most homology modeling method hypothesizes that protein structures are more conserved than protein sequences, PRISM hypothesizes that interacting site of proteins (which are called “interfaces”) are more conserved than globular structures of the proteins. Therefore, it assumes that proteins can have similar interface motifs while they have different globular structures. The algorithm has four steps which are surface extraction of all target proteins, structural alignment of target surfaces and template interacting surfaces to find the best representative interface for target proteins, transformation for creation of structural file of interacting proteins, then filtering, flexible refinement and energy calculation both to eliminate structures having colliding residues and optimize the structure according to its energy. PRISM can be used in Linux environment for specified target proteins or pre-runned structures can be downloaded from PRISM web server. In this study, we retrieved 60 precalculated structures. Since their residue indices do not track with the indices of Uniprot, we did alignment to obtain Uniprot indices of interface residues.

e) Interactome Insider

Interactome Insider (Meyer et al., 2018) is a tool which connects genomic variants to structural interactome. They retrieved 3D structure information of interactomes of seven species and calculated their interacting sites, interfaces, in order to annotate whether or not a given genomic variant is on interface. Experimental and homology modeling structures, however, are limited. Thus, they developed Ensemble Classifier Learning Algorithm to predict Interface Residues (ECLAIR) classifier which is a machine learning algorithm based on ensemble of eight random forest classifiers for interface prediction. The distinction from the above algorithms is that ECLAIR does not produce structure, it only predicts sequence indices of interacting residues. They enriched the interactome with 185,957 more interaction having known interface by the help of ECLAIR. In this study, we collected Uniprot indices of interfaces in the database which covers PDB (340), Interactome3D (74) and ECLAIR (283).

3.2.4 Disease Association of Mutations

a) Evmutation

EVmutation (Hopf et al., 2017) is a statistical method which predicts the effects of mutations using epistatic information with evolutionary conservation. The method gives a damage score for which each position in a Uniprot sequence is substituted by the remaining 19 amino acid. The more negative values of the calculated score means the more damaging the corresponding mutation. In this study, we collected precalculated data.

b) Polymorphism Phenotyping v2 (PolyPhen-2)

Polyphen2 (Adzhubei et al., 2010) is a tool which uses a learning based strategy to predicts the effect of mutations and then, interprets results as benign, possibly damaging or probably damaging. It uses structural and evolutionary considerations by using eight sequence- and three structure-based features. We retrieved again the precalculated mutation effect data for this study.

3.2.5 Cancer Driver Effects of Mutation and Mutated Genes

a) The Network of Cancer Genes (NCG)

The Network of Cancer Genes (Repana et al., 2019) is a database for genes whose mutations have cancer driver effect to cells. The database consists of manually curated

information from publications and cancer sequencing screens. In this study, we collected only the known cancer genes and their tumor suppressor/oncogene annotations.

b) The Catalogue of Somatic Mutations in Cancer (COSMIC)

The Catalogue of Somatic Mutations in Cancer (COSMIC) (Tate et al., 2019) is a repository of somatic mutation effect in cancer. However, COSMIC includes data from targeted screens and genome screens for mutation data, structural genomic rearrangements, fusion, copy number variation, methylation etc. In this study, we only focused on COSMIC Cancer Gene Census (Sondka et al., 2018) project and its data which is a list of all cancer genes in COSMIC database.

c) Cancer Genome Interpreter

Cancer Genome Interpreter (Tamborero et al., 2018) is a platform which takes cancer genome and annotates its alterations for identifying if they have any driver effect for tumorigenesis or if they have any impact on response of treatments. Cancer Genome Interpreter consists of two parts: one is analysis which users can submit their list of alterations and interested cancer type and, obtain the results; other is the pre-runned datasets including Cancer Biomarkers, Cancer Genes (both prediction and literature-based), Cancer Bioactivities and Validated Oncogenic Mutations. In this study, we only collected validated oncogenic mutations which was created by combining the data in Database of Curated Mutations (DoCM) (Ainscough et al., 2016), ClinVar (Landrum et al., 2016), Oncology Knowledge Base (OncoKB) (Chakravarty et al., 2017), The International Agency for Research on Cancer (IARC) (Petitjean et al., 2007) databases, and published experimental assays. We, then add the genes having validated oncogenic mutations to our cancer driver gene lists to enrich it.

d) FireBrowse of Broad Institute

Firebrowse of the Broad Institute is a platform providing Firehose analysis pipeline results on TCGA data. For each TCGA cancer type, FireBrowse gives results for clinical, copy number, methylation, micro RNA sequencing (miRseq), messenger ribonucleic acid (mRNA), RNA sequencing (RNAseq), mutation and pathway analyses. In this study, we used the results of mutation analysis on GBM cancer type from three different mutation analysis methods as Mutation Significance (MutSig) 2CV (Lawrence et al., 2013), Mutation Assessor (Reva, Antipin, & Sander, 2011) and Cancer-Specific High-throughput Annotation of Somatic Mutations (Wong et al., 2011) (CHASM) 1.0.5. MutSig 2CV calculates the gene significance by taking number of non-silent and silent mutations in the gene and, covariant space of neighboring genes. In this project, we took the genes whose P-value is smaller than 0.05 as significant genes. The more significant P-value means more probability to be a cancer driver gene. In this study, we only considered genes having

P-value smaller than 0.05 as cancer driver genes. Mutation Assessor calculates the functional impact scores for missense mutations by using evolutionary conservative patterns and provides these functional impacts as high, medium, neutral and low. In our analysis, we only considered genes having high and medium functional impact as significant genes. CHASM uses machine learning strategy to distinguish between driver and passenger missense mutations and gives the probability for each mutation according to selective survival advantage that is provided to the cancer cells by the mutation. In this analysis, we only considered mutations having P-value smaller than 0.05.

3.2.6 PFAM

Pfam (El-Gebali et al., 2019) is a depository for functional domains. Each protein consists of domain information for each index of its Uniprot sequence and also annotation of the functional domains. In this study, we retrieved domain information of each Uniprot sequence index of corresponding proteins.

3.2.7 Cell Model Passports

Cell model passports (van der Meer et al., 2019) is a database for 1634 cell lines and organoids from various type of cancers. It provides clinical and genomic data such as mutation profiles, gene expression (microarray/RNAseq), copy number variation, Clustered Regularly Interspaced Short Palindromic Repeats-Knock Out (CRISPR-KO) based essentiality, fusion, drug response and methylation data of models. While user can download the processed raw data from Cell Model Passports, they can also be directed to the raw data. In this study, we collected processed mutation profiles of GBM models.

3.2.8 Genomics of Drug Sensitivity in Cancer (CancerRxGene)

Genomics of Drug Sensitivity in Cancer (Yang et al., 2013) developed by screening models (cell lines) with different compounds to find drug response data and genomic markers of sensitivity of cancer models. For drug response, they use half maximal inhibitory concentration (IC50) which indicates in which concentration drug can reduced the activity of the cell in half. Additionally, they applied Z-transformation on the natural logarithm (ln) of IC50 values in each cell line screen. Therefore, each drug has a specific Z-value for each cell line. In this study, we downloaded drug screening results of GBM models and use Z-scores of drugs for corresponding cell lines.

3.2.9 Interaction Reference Index (iRefWeb)

iRefWeb (Turner et al., 2010) is a reference protein interaction data (interactome). It collects protein interaction information from various database: The Biomolecular Interaction Network Database (BIND) (Bader et al., 2001), Biological General Repository

for Interaction Datasets (BioGRID) (Stark et al., 2006), The Comprehensive Resource of Mammalian Protein Complexes (CORUM) (Ruepp et al., 2008), Database of Interacting Proteins (DIP) (Salwinski et al., 2004), IntAct (Hermjakob et al., 2004), The Human Protein Reference Database (HPRD) (Peri et al., 2003), Molecular Interaction Database (MINT) (Chatr-aryamontri et al., 2007), MPact (Guldener et al., 2006), Mammalian Protein-Protein Interaction Database (MPPI) (Pagel et al., 2005) and The Online Predicted Human Interaction Database (OPHID) (Brown & Jurisica, 2005). The weight of interaction represents by MINT-inspired score (MI score) which based on evidence that the interaction has such as publications, experimental method used for identification of the interaction. In this study, we used iRefWeb as weighted interactome in network reconstruction and filtered interactions having a MI score less than 0.4 and also the proteins such as UBC, APP, ELAVL1, SUMO2, CUL3 and the proteins huge in size (TTN, MUC16, SYNE1, NEB, MUC19, CCDC168, FSIP2, OBSCN, GPR98) to limit the noise coming from random mutations on these proteins as in (Hristov & Singh, 2017). Since they are huge and have very high degrees, they prone to have large number of mutations by chance and it may affect network construction. Moreover, we further filtered the interactions if they have structural information or not.

3.3. Identification of the 3D spatial clusters

After collecting all cancer related/driver protein structures and protein complexes having at least one GBM mutation, we constructed a residue-residue interaction network for each structure. In order to construct residue-residue network, we first defined each amino acid of the structure as a node and added peptide bond between each of them as edge. Due to the nature of protein folding, residues who are distant to each other in sequence could be in close proximity in 3D space. Considering this situation, we calculated the distance between all atoms of each residue to all atoms of another residue to find the residues who are in close proximity with a distance formula **(Equation 1)**. If the calculated distance is less than 5A, we considered these residues as interacting residues and added edge between them. Thus, we created residue-residue networks R(v,e) of all driver proteins having structure information and at least one GBM mutation as shown in **Figure 3.3**. In order to identify the 3D spatial clusters/patches, we mapped all mutated residues of a protein from all patients on each corresponding residue-residue network and searched for shortest paths between each mutated residue pairs with a length less than 3. After gathering all shortest paths in each residue network, we revealed mutation clusters which we named as patches where mutated residues are connected each other either directly or by the help of one residue between them, and singleton mutations which stay outside of any mutation clusters.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \quad (1)$$

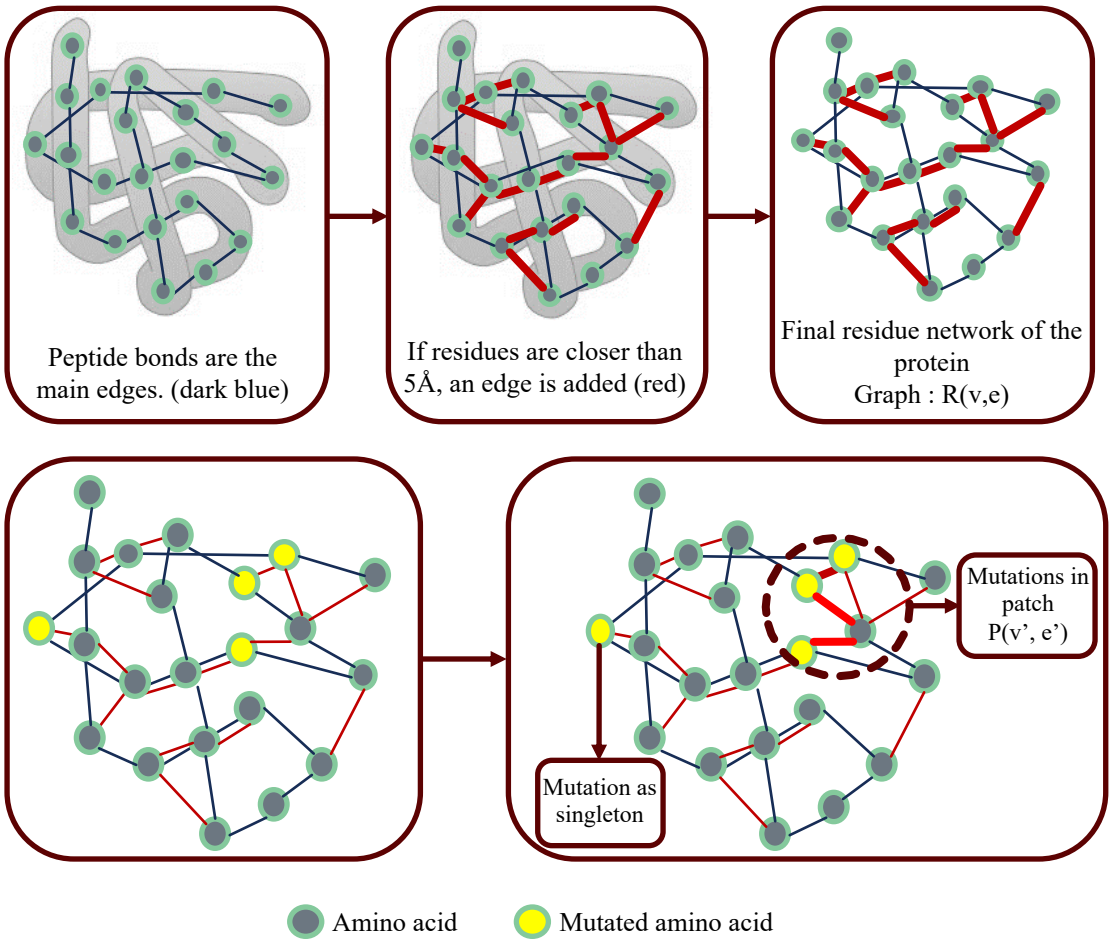


Figure 3.3. Identification of 3D spatial organization of mutations on proteins. We started with considering all amino acid residues as nodes in residue-residue network of the protein. Then, we calculated the distance between each atom in each residue. If any of the calculated distance between amino acids is lower than 5Å, we added an edge between corresponding amino acids in residue-residue network in addition to peptide bond between amino acids. Afterwards, we mapped all mutated amino acids on corresponding protein across all the patients on to these constructed residue-residue network. Lastly, we searched shortest pathways between each mutated residue on the network in order to identify 3D spatial organization of the mutations: patch or singleton.

In order to analyze preliminary if mutation patches decrease the heterogeneity, we grouped patients according to presence of each patch in these patients. First, we ranked the patches from most frequent one to lowest. Then, we took the most frequent patch and patients having mutation in this patch as our first group. Other groups were formed in the same way following the ranking order yet there are no common patients across the groups. When all the group has at least ten patients, grouping stopped.

3.4. Identification of protein regions and the effect of the mutations

Proteins have unique 3D structures having different regions as core, surface and interface. While core regions buried inside, surfaces are the regions covering structures. Yet interfaces are on the surfaces, they are the regions where proteins interact with other proteins or nucleic acids. To identify where GBM mutations located, we firstly collected interface information from Interactome Insider which includes PDB, Interactome3D and ECLAIR data, and PRISM and found interface mutations. We then used FreeSASA software (Mitternacht, 2016), which can calculate solvent accessible surface area at residue level, in order to differentiate surface and core regions of the proteins. We labelled residues as surface residues if calculated relative solvent accessible surface area of the residue in its monomer state is greater than or equal to 5% and labelled as core residues if not. Since interfaces are on the surface regions, we also excluded interface residues from surface residue set. **Figure 3.4** indicates the steps for identification of the protein regions. In this study, we only considered the structure files whose length greater than 50 residues for identification of protein regions. After finding location of the mutations, we classified the patches as intra- and inter-patches. While former intra-patches do not include interface mutations, inter-patches have at least one interface mutations. As an alternative scenario for the inter-patches is that mutations in the patches could elongate to partner protein by interface residues of both proteins.

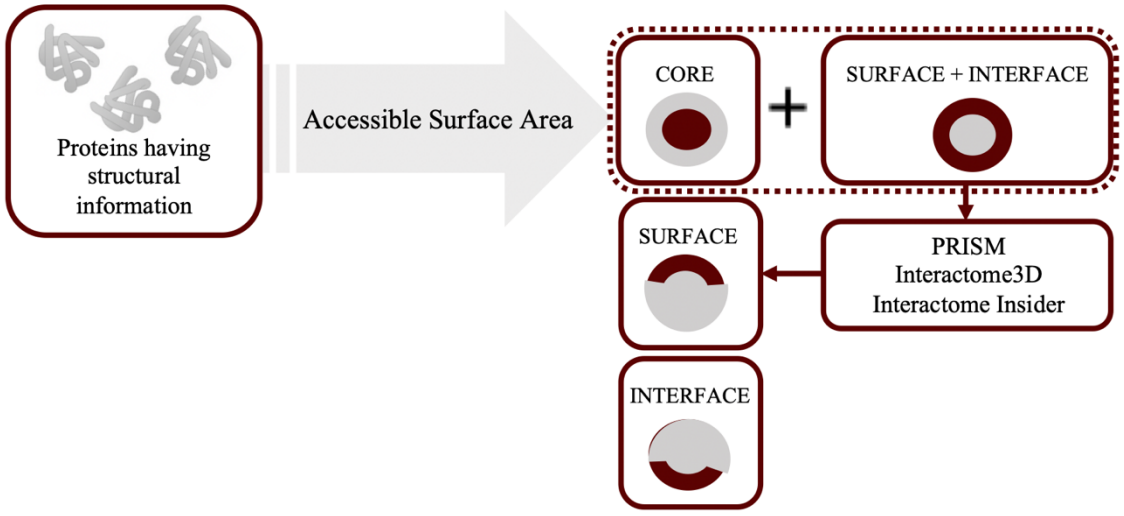


Figure 3.4. Identification of different protein regions on 3D structures. Firstly, we collected interface residues of all mutated proteins. Then, we used FreeSASA algorithm in order to differentiate buried residues than non-buried residues. While buried residues are core residues, non-buried of them could be surface or interface residues. When we excluded interface residues coming from various databases, we differentiated surface residues from the interface residues.

While non-synonymous mutations change the protein sequence and, thus the structure and function, impact of these mutations varies. Position of the mutation, cancerogenicity of the gene having the mutation or environmental factor can affect the disease association of the mutation. In order to calculate the effect of mutations as disease causing or neutral, we used EVmutation and PolyPhen-2 and compared the damage of the mutations based on their localization, spatial organization and their role in the cancer progression (tumor suppressors or oncogenes).

3.5. Reconstruction of patient-specific sub-networks and grouping of the patients

3.5.1 Network reconstruction

In this study, we used Forest module of Omics Integrator (Tuncbag et al., 2016) in order to construct patient specific sub-networks. It solves the prize-collecting steiner forest problem for a given reference graph G(V, E, w) where V is the node set {v|v ∈V}, E is the edge set {e|e ∈ E} and w is the edge weights. Prize collecting steiner forest problem tries to take as much as terminal nodes with additional proteins to connect them and as less as unreliable edge to produce optimum network from predefined input and parameters (**Figure 3.5**). The algorithm firstly uses the prize and cost functions in **(Equation 2)** and **(Equation 3)** respectively. **(Equation 2)** gives each node a prize on the basis of predefined weight p(v), beta (β), mu (μ) parameters, and the degree of the node in assigned reference interactome degree (v). By prize function, the algorithm can decrease the dominance of the hub proteins. **(Equation 3)** takes the interactome weight of each interaction as edge confidence p(e) and gives its edge cost c(e).

$$p'(v) = \beta.p(v) - \mu.degree(v) \quad (2)$$

$$c(e) = 1 - p(e) \quad (3)$$

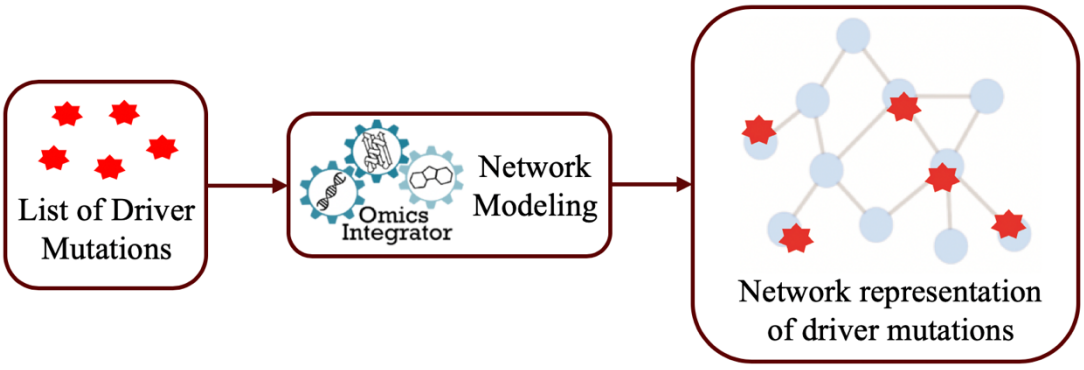


Figure 3.5. Identification of affected protein-protein interactions and biological pathways.

After determining prize and cost functions, forest tries to minimize its objective function which is **(Equation 4)**. The main idea is to pay penalty for each terminal if it is not included in the final network and pay the cost for each edge to connect nodes depends on its cost. Thus, algorithm tries to balance selecting and eliminating a set of nodes to decrease the penalty and the cost of their edges. Moreover, since Omics Integrator uses the Prize-Collecting Steiner Forest (PCSF) problem for a given set of terminal nodes with predefined prizes/weights, it is also important to adjust the function with suitable parameters which are μ (mu), ω (omega), β (beta) and D (depth). While μ is for scaling factor for hub proteins, ω is for tuning the number of trees in the final network. While β is scaling factor for adjusting dominance of terminal nodes in the final networks, D determines the number of edges from the root to the leaf nodes.

$$f'(F) = \sum_{v \notin VF} p'(v) + \sum_{e \in EF} c(e) + \omega.\kappa \qquad (4)$$

As an input of the algorithm, we provided filtered iRefWeb probability weighted protein-protein interactions as the reference interactome for all patients, yet for each patient we prepared a list of driver genes having at least one non-synonymous mutation in the patient as terminal nodes. We also weighted each terminal node in each patient with the number of passenger and driver mutation it has since it is reported in (Burke, Perisic, Masson, Vadas, & Williams, 2012; Kan et al., 2010; Porta-Pardo et al., 2017), not all mutation on a driver gene have the same effect. For each passenger mutation, the weight increased with 0.5, however for each driver mutation we added 1 to the weight. In this study, different combinations of parameters were tried in order to include highest fraction of terminal (input) nodes in the final networks for each patient. Then, we used the parameter set as ω (omega) = 10.0, depth (D) = 6 and β (beta) = 10 and μ (mu) = 0.005 and μ = 0.01. Thus, we used two different μ values to recover the canonical pathways and more specific ones. Then we merged the node and edge set of the reconstructed networks to come up with a single network for each patient.

3.5.2 Network-guided grouping of the patients

Each patient specific sub-network consists of nodes and edges constructed by Omics Integrator. Rather than using mutated genes on each patient, we used node list of the reconstructed networks for pathway enrichment analysis with WebgestaltR (Wang, Vasaikar, Shi, Greer, & Zhang, 2017) package. WebgestaltR executes different enrichment analysis algorithm such as Over-Representation Analysis (ORA), Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) and Network Topology Analysis (NTA). It also uses different enrichment databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000), Reactome (Joshi-Tope et al., 2005) and Wikipathways (Slenter et al., 2018). In our case, we used node list of our patient networks as input gene lists in "rnk" format and KEGG knowledgebase for enrichment database, and program gave overrepresented KEGG pathways and their enrichment scores for each patient. From them, we filtered the pathways which have False Discovery Rate (FDR) less than 0.1 as enriched pathways in the sub-network. We then eliminated disease pathways

including infections, cancer, addiction related pathways from resulting list. After having enriched pathways and their enrichment scores, we prepared a matrix where rows are union set of enriched pathways from all patients, columns are patient labels, and the entries are the enrichment scores of corresponding pathways in the corresponding patient. If the pathway is not enriched than 0 was put to that entry. We then used matrix for implementing non-negative matrix factorization (NMF) (Paatero & Tapper, 1994) which is an unsupervised approach to decompose matrix X into individual elements W and H **(Equation 5)**.

$$X \approx W.H \tag{5}$$

With a p x n dimensional matrix, X, has components: W (p x r) which represents basis elements in its columns and H (r x n) indicating the coordinates of the data points for the basis elements in W. H indicates the way to reconstruct the approximation as an linear combination of basis elements. We implemented this algorithm without a network regularizer and then consensus clustering from pyNBS package which is a Python implementation of Network Based Stratification (NBS) (Huang, Jia, Carlin, & Ideker, 2018). In this package, NMF is applied for thousand times on subsamples of the real data. Sampling is random through 80% of the patients and genes are selected without replacement. The latent feature is the number of dimensions we reduced the initial matrix. In our study, we used it as 5. Then, from a thousand clustering results, a list of H data frames are created. The consensus clustering (Monti, Tamayo, Mesirov, & Golub, 2003) uses these data frames and forms co-clustering matrix. This co-clustering represents the similarity of the patients which has been used to classify the patients. In order to determine if patient groups and survival information of their patients has a significant relationship, we implement survival analysis based on cox-proportional hazards model (Andersen & Gill, 1982) providing a P-value from comparison of whole model. Lastly, the patient groups were searched for whether any identified spatial patch has tendency to represent a group by using Hypergeometric Test. With this statistical approach, we will get how significantly mutations in specific patch occur in corresponding patient group.

3.6. Linking the patient groups to drug response

In order to predict therapeutic responses of patient groups, we linked patient groups to GBM cell lines and then to drug responses. Firstly, we connected patient groups to cell lines using mutation information of cell lines from Cell Model Passports. If at least one mutation belonging to a predominant patch in a group is also present in the GBM cell line then the patient group is associated with that cell line. We then connected patient groups to drugs and responses. Yet, in order to connect them, we followed two steps: for the first step, we connected cell lines to drugs using drug sensitivity data from CancerRxGene; for the second one, we also collected target proteins for each drug, and connected patient groups to drug targets. If any drug target is also significantly enriched in sub-network of any patient group, and if the drug has already been associated with any cell line that has been associated with the patient group, then we linked the drug with the patient group. In

this study, we inferred drug responses of patient groups from the drug responses of the connected cell lines. We used Z-score of natural logarithm of (ln) IC50 values of each cell line to specified drugs. Due to the nature of Z distribution, we accepted values outside of (-1.96, 1.96) interval as significant values with a 95% significance. Drug responses of patient groups predicted from their corresponding cell line drug partners, whose Z-scores of ln(IC50) values below than -1.96 were labelled as sensitive and above than 1.96 labelled as resistant. This step is summarized in **Figure 3.6**.

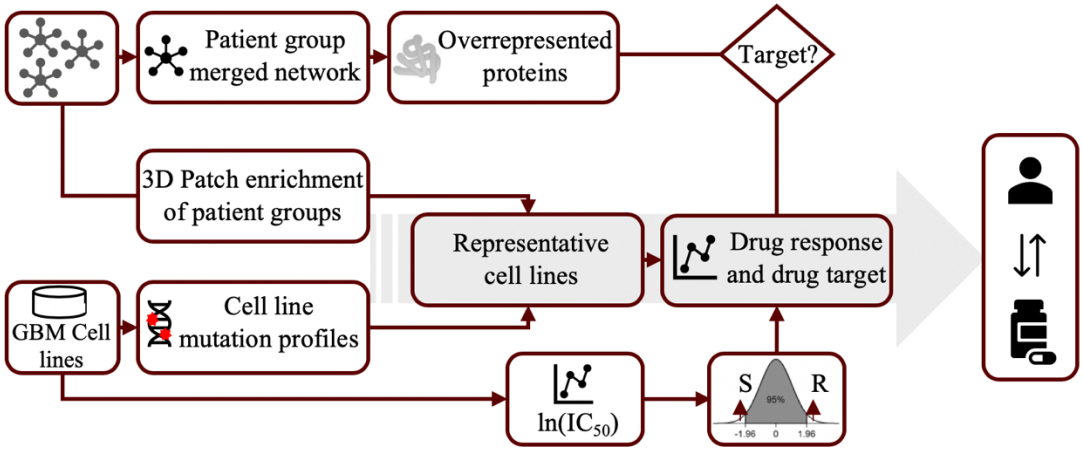


Figure 3.6. Summary of GBM cell lines patient group linkage. Patient specific subnetworks were used to identify 3D patch enrichment in each patient group. Additionally, merged networks of patient groups were used to obtain the overrepresented proteins in each group. Mutation data of GBM cell lines and signature 3D patches of patient groups were used to link the patient groups to cell lines. After having representative cell lines for each group, drug targets were used to connect patient groups to drugs. If a patient group is linked with a cell line and target of a drug which has a response information from the representative cell line, is overrepresented in merged network of patient group, then, the hypothetical therapeutics can be proposed.

CHAPTER 4

4. RESULTS

4.1. 3D Spatial Organization of GBM Mutations

Out of 15,399 unique non-synonymous (missense, nonsense and frameshift) mutations across 290 TCGA-GBM patients, only 14,308 mutations are mapped on manually curated canonical UniProt entries of *Homo sapiens* proteome from which 697 of them are frameshift mutations and 13,611 of them are nonsense and missense mutations. We used UniProt sequences as the references for each protein in order to map mutations on to 3D structures. Since indices in protein 3D structure files do not track with UniProt sequences, we aligned each 3D structure sequence to corresponding reference UniProt sequence and found 4702 mutations were aligned to at least one protein structure either from PDB or from ModBase. Through residue-residue networks of 3D structures having at least one mutation, the spatial organization of 4702 mutations were calculated as described in 3.3. and, as a result we obtained 220 3D spatial patches comprised of 580 mutations and 4122 singleton mutations which stay outside of patches. Thus, most mutations were found as singletons, while approximately 10% of the mutations are in close proximity to each other and form mutation patches. Additionally, we grouped precalculated patches as 160 intra-patches which do not include any interface mutation and 60 inter-patches which have at least one interface mutation.

We initially checked whether 3D spatial organization of mutations (patches) can affect the heterogeneity of mutation profiles by calculating and comparing the frequencies of mutations alone and in patches, separately. Despite the fact that the mean value of mutations in each patient is 50.43, only 213 mutations are present in at least two patients. If we increase the number to three patients, the common mutation number decreased to 44. Thus, approximately 5% of the patients have the most common mutations which are 289th position of EGFR from Alanine to Valine and 132nd position of IDH1 form Arginine to Histidine. On the other hand, we also tried to find the commonalities resulting from mutation patches in the patients. If a patient has a mutation in any precalculated patch, we evaluated that the patient has the patch mutation. While the mean value of patients sharing a mutation is 1.13, the value increases to 3.5 patients with at least one mutation in the same patch. When we sorted all patches based on their patient frequencies,

we found that TP53 and PTEN patches are the most prevalent among 20 percent of all patients resulting in better detection of commonality. Moreover, we also sorted all mutations in 3D patches, and we found that 289th and 598th positions of EGFR from Alanine to Valine and Glycine to Valine respectively are the most common the patients. In **Figure 4.1** and **Figure 4.2** that are the mutation and patch profiles of the patients, respectively; each column indicates a patient and each row shows a mutation in mutation profile and a patch in patch profile that are present in at least 2 percent of patients in both figures.

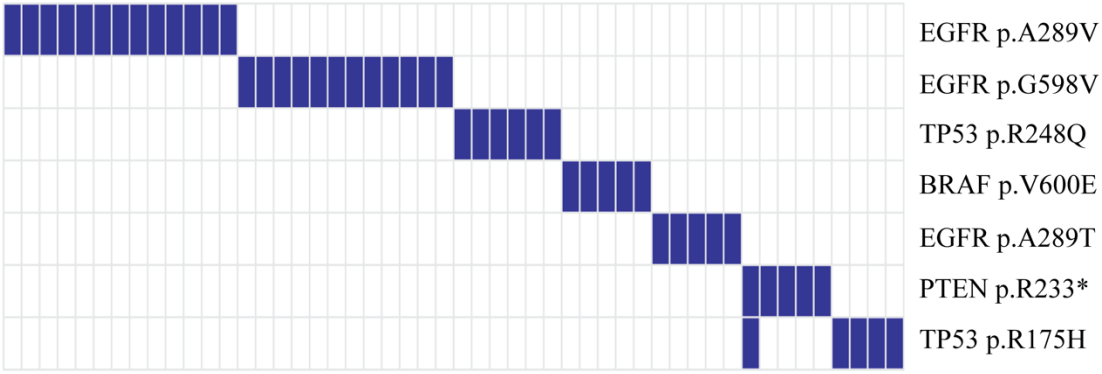


Figure 4.1. Mutation profile of GBM patients. While each column represents a GBM patient, each row indicates mutation in any 3D patch that is present at least 2% of patients.

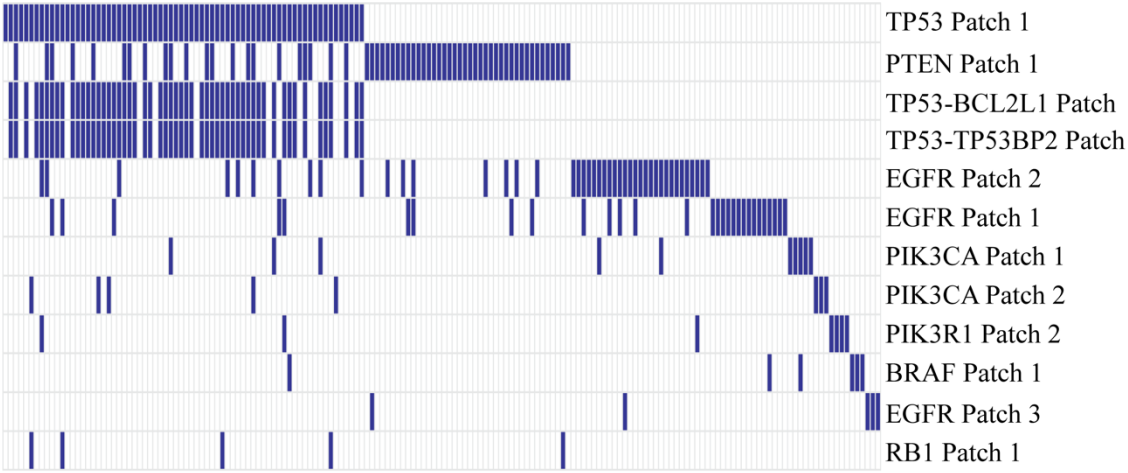


Figure 4.2. 3D Patch profile of GBM patients. While each column represents a GBM patient, each row indicates 3D patch that is present in at least 2% of patients.

On the basis of their most common mutations and patches, we divided patients into mutually exclusive groups and tried to assess the significance of association between patient groups and their survival data. While grouping based on mutation profile cannot show any significant result (P-value: 0.5115, **Figure 4.3**), grouping based on patches

indicates the advantages of spatial organization with a P-value as 0.0001 as shown in **Figure 4.4**.

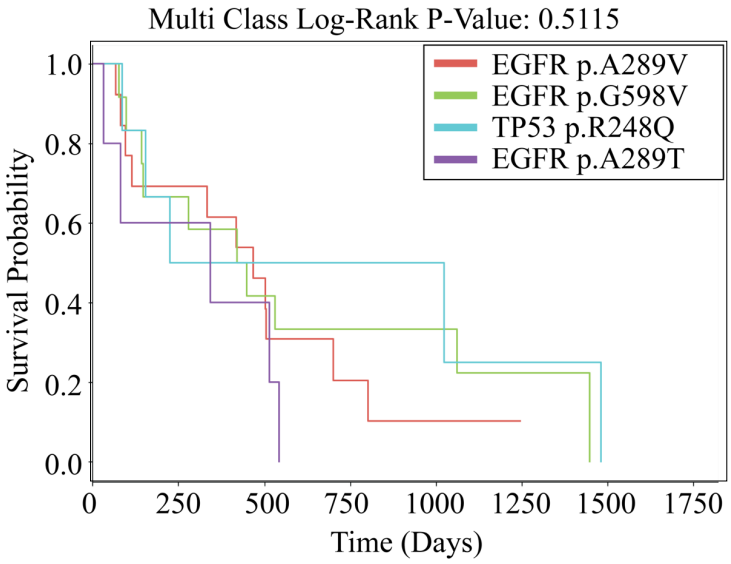


Figure 4.3. Kaplan-Meier survival curves of the patient groups in the mutation profile in **Figure 4.1**.

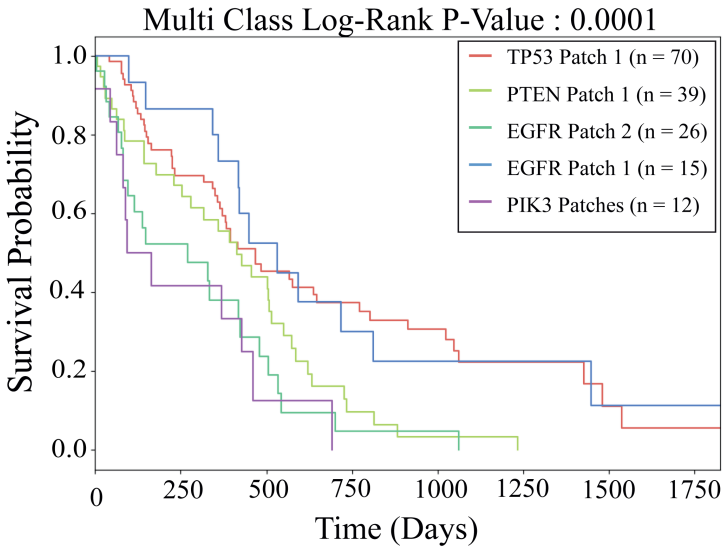


Figure 4.4. Kaplan-Meier survival curves of the patient groups in the patch profile in **Figure 4.2**.

Afterwards, we mapped mutation patches of frequently mutated hub proteins to their functional domains in order to interpret the significance of association between clinical data and patient groups. As shown in **Figure 4.5**, different patches are mainly located in different functional domains of corresponding proteins.

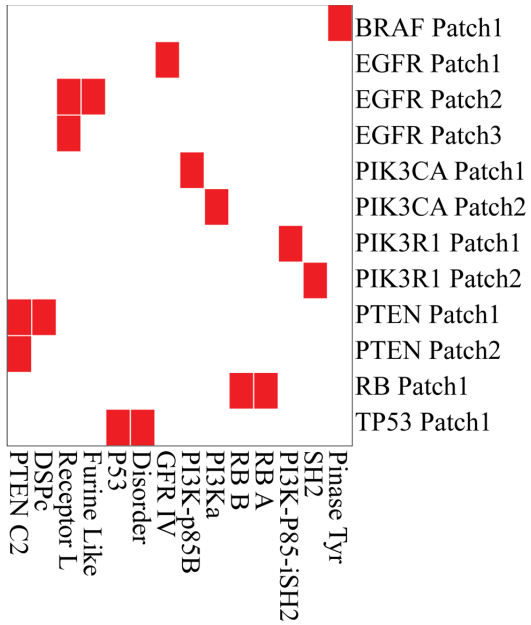


Figure 4.5. Mapping patches of frequently mutated hub proteins to their functional domains. Red colors represent the overrepresentation of patches in the corresponding domains.

For example, P85α which is the protein encoded by phosphoinositide-3-kinase regulatory subunit 1 (PIK3R1) gene has two patches on different domains having different biological functions. While PIK3R1 Patch 1 is on inter-SH2 (iSH2) domain having the inhibitory function on phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (PIK3CA) by binding p110 catalytic domain, Patch 2 is on SH2 domain where the protein binds to phosphorylated residues of activated tyrosine kinases which causes conformational change to relieve the inhibition of p110 (Liu, Knapp, & Ahmed, 2014; Thorpe et al., 2017) as shown in **Figure 4.6**. Clustered mutations on iSH2 domain of p85α was reported in GBM patients that can weaken the inhibition and resulting in a gain of function in PI3K activity (Sun, Hillmann, Hofmann, Hart, & Vogt, 2010). Another important tumor suppressor, PTEN has also two patches and these two patches are on different functional units. PTEN Patch 1 is on Dual specificity phosphatase (DSPc) which is Serine-Threonine and Tyrosine protein phosphatases and the catalytic domain. On the other hand, PTEN Patch 2 is on C2 domain where PTEN binds the phospholipid membrane. Phosphatase domain can get into contact with plasma membrane and then be activated by the help of C2 domain. Studies reported that mutations on C2 domain affecting PTEN-membrane affinity suppress the inhibition function of PTEN in GBM (Lee et al., 1999) and mutations in DSPc domain cause tumor progression by disrupting

the phosphatase function which results in PIP3 accumulation (Chalhoub & Baker, 2009; Georgescu, 2010) in cells and thus the activation of AKT pathway. As a last example, PIK3CA which is a very well-known oncogene, has four different domains. However, two of them are mainly located in functional domains of PIK3CA. Patch 1 is located in P85 binding domain and Patch 2 is on accessory domain (PIK domain) of PIK3CA. P85 binding domain is important in inhibition of PIK3CA by regulatory unit. However, role of PIK domain has not been unresolved yet it is suggested that it may have a role in substrate presentation (Flanagan et al., 1993).

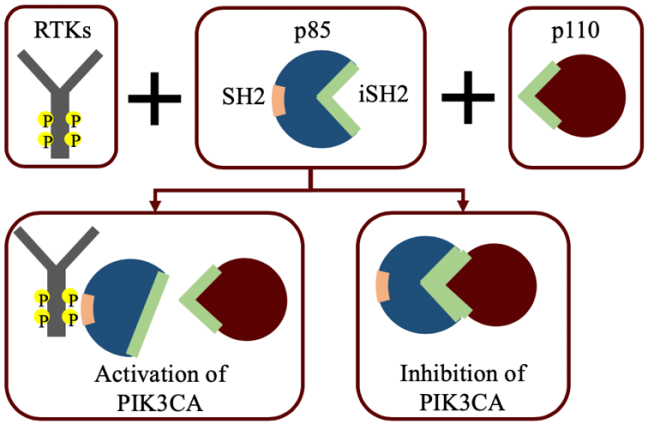


Figure 4.6. Example of different domains of protein of PIK3R1 gene. P85 which is the protein of PIK3R1 has two functional domains as SH2 and iSH2. iSH2 domain can bind to p110 catalytical domain of PI3KCA and inhibits its function, while SH2 domain binds to phosphorylated residues of activated tyrosine kinases resulting in conformational change in the protein and relieves the inhibition of p110.

Heretofore, we used all missense and nonsense mutations for identification of patches without differentiate driver and passenger mutations. Since driver genes are the propulsive force of tumor formation and progression, we prioritized only the mutations of driver genes and eliminated the noise of passenger mutations. We herewith collected cancer gene and mutation information from various sources as detailed in 3.2.5. We retrieved 6270 driver mutations and 3789 driver genes in total. Yet, only 6278 of TCGA GBM mapped mutations are in list of collected driver genes and 2072 of them are on proteins having structural information. When we only focused on these driver genes having GBM mutations, we obtained 112 intra- and 32 inter-patches.

Moreover, we investigated whether there is any association between properties of genes and patches on them. As in **Figure 4.7**, there are many intra-patches formed by less amount of mutations, yet the central proteins generally have larger inter-patches such as TP53 Patch 1 with 41, PTEN Patch 1 with 43 residues. Additionally, in **Figure 4.8**, PTEN, EGFR and PIK3CA patches are shown with their patch residues. While some proteins have patches on their own structures, some of them have patches shared by partner

proteins as in the case of PIK3CA and PIK3R1. We further compared the differences between tumor suppressors and oncogenes in terms of 3D spatial organization of their mutations. Analyzing spatial organization of tumor suppressors and oncogenes are important since they act oppositely such as oncogenes provide advantages to tumor cells by gain-of-function, yet tumor suppressors lose their ability to inhibit tumor cell formation or progression by loss-of-function. We found that driver mutations of tumor suppressors are significantly located in patches whereas diver mutations of oncogenes choose to stay as singletons (P-value = 8.33 x 10^{-6}/ Fisher's Exact Test). As an example of well-known tumor suppressors, PTEN has two patches consisting of 43 and 2 residues, TP53 has only one patch with 41 residues. On the other hand, oncogenes PIK3CA has three patches with 10, 3, 2 and residues, respectively, EGFR has also three patches composed of 11, 14 and 5 resides. While tumor suppressors have bigger patches, oncogenes tend to have relatively smaller patches and many singletons. This can be explained by the differences of underlying logic of loss-of-function and gain-of-function terms. The gain-of-function mutations are limited to very specific sites, whereas proteins can be inactivated variety of ways. These results agree with the previous studies. As an example, in (Fujimoto et al., 2016), it is discussed that mutations in oncogenes tend to be clustered in a few regions but mutations in tumor suppressor can be distributed over functional domains. Moreover, we compared the types of non-synonymous mutations between oncogenes and tumor suppressors and found that frameshift and nonsense mutations are significantly more frequent in tumor suppressors (P-value = 1.84x10^{-15}/ Fisher's Exact Test). Both frameshift and nonsense mutations may result in abnormal proteins. In case of tumor suppressors, these types of mutations can disrupt the suppressor proteins completely, making cells more vulnerable to cancer.

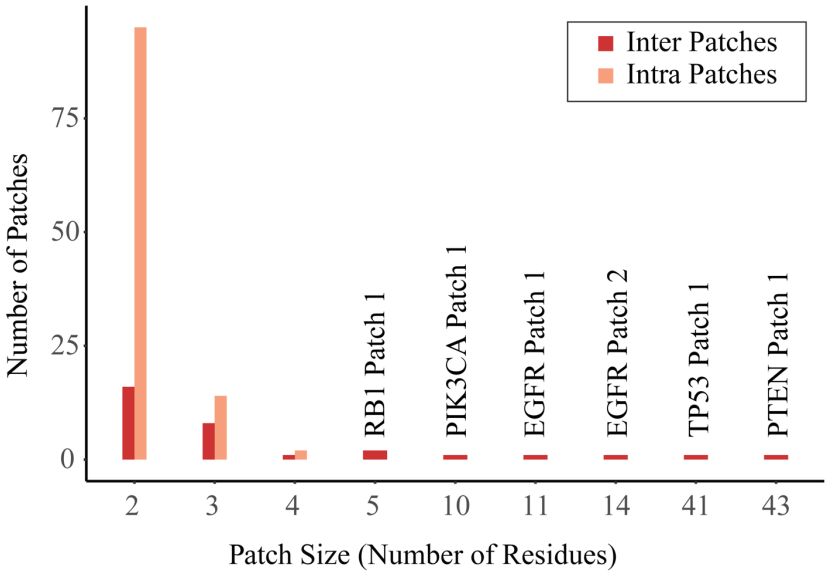
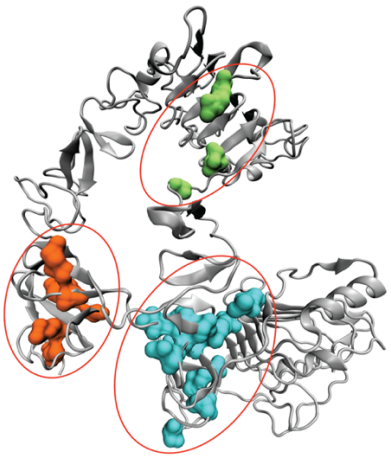
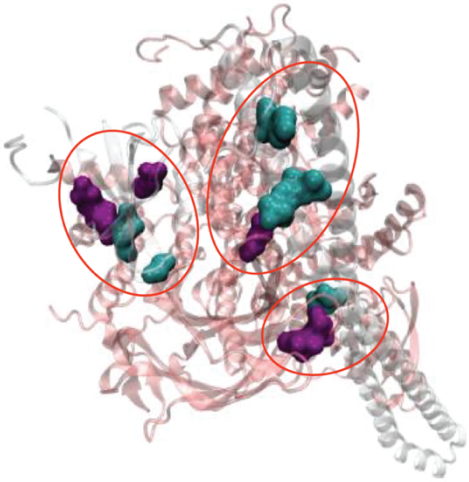


Figure 4.7. Histogram of the patch sizes for intra- and inter- patches. While many intra-patches formed by less amount of mutations, the central proteins generally have inter-patches consisting of larger number of mutations.

PTEN
PDB ID: 5bug:B



EGFR
PDB ID: 3qwq:A



PIK3CA-PIK3R1
PDB ID: 4jps: A-B

Figure 4.8. Examples of 3D mutation patches on protein structures. PTEN Patch 1 has 43 residues and it is an example for mutations forming a residue network from the surface to the core. EGFR has three patches with different sizes: 11, 14 and 5 residues, respectively and EGFR is an example for protein having multiple patches. Lastly, PIK3R1-PIK3CA complex has three inter-patches and they are an example for proteins having common patches.

4.2. Characteristics of the GBM mutations from the structural and chemical perspectives

4.2.1. Structural positions and chemical characteristics of the GBM mutations

Structural locations and physicochemical properties of the mutated residues can enlighten the molecular mechanisms which are affected by them in tumor cells. We thus divided mutations according to their locations into three groups (surface, interface and core) as detailed in 3.4. While surface and interface are on the regions where there is solvent accessibility, core regions are buried inside of the structure and there is no solvent accessibility. The difference of interface over surface is having residues which are physically contacting to a partner protein. In our study, we found that most of the mutations (65.6%) are located in surface regions of corresponding structures although only 18.3% and 16.1% of the mutations are on core and interface respectively. The detailed numbers of mutations mapped to protein structural regions are indicated in **Table 4.1**. Since interface mutations can change the interaction profile of proteins and thus change the wiring of cell networks, we also investigated whether mutations of interface region significantly tend to form spatial patches or stay as singleton. Then, we found that interface mutations are statistically more populated in spatial patches than non-interface mutations with a P-value smaller than 0.0001 (Fisher's Exact Test). A graphical representation for fraction of patch and singleton mutations according to their locations is shown in **Figure 4.9**.

Table 4.1. Number of mutations mapped to protein structural regions.

Number of mutations from TCGA: 15399			
Structural Region	All	PDB	Models
Core	861	372	ModBase: 489
Surface	3084	1153	ModBase: 1931
Interface	757	340	Interactome3D: 74 PRISM: 60 éclair: 283
Total	4702	1865	2837

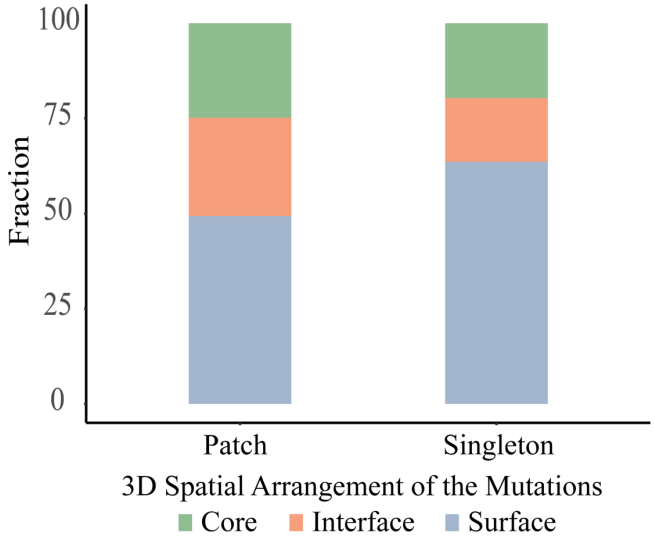


Figure 4.9. Fraction of core, interface and surface mutations according to their 3D spatial organizations.

Chemical properties of the residues in an amino acid sequence determine the final folded structure of protein. These properties were defined as hydrophobic, charged and polar. Any changes in the DNA sequence may change the amino acid in protein and thus the chemical property of the residue which may result in disruption of overall or partial protein function. For example, while proteins with larger core region are generally robust to mutation (Faure & Koonin, 2015), it is reported in (Guo, Choe, & Loeb, 2004) that core region is sensitive to non-hydrophobic changes due to their disruptive effect on the structure. Since the changes are important in protein function, we also analyzed switches or preservation of wild type chemical classes in each mutated residue on driver proteins according to their 3D locations. As indicated in **Figure 4.10**, surface and interface mutations demonstrate similar pattern, yet core mutations are slightly different from those. Overall, surface and interface residues are more open to change their chemical properties with a P-value as 1.81×10^{-11} (Chi-square test). While core residues mostly preserve their hydrophobic character, surface and interfaces are more prone to change their character from charged to charged and to polar. There are also changes hydrophobic to polar and to charged, yet they are less frequent. While **Figure 4.10** indicates a high frequency in changes from hydrophobic to hydrophobic in surface and interface residues, this is less than expected frequency with a P-value as 2.29×10^{-42} (Chi-square test). Additionally, we found that mostly changed interface residues are the charged ones and in (Nishi et al., 2013), they found that GBM mutations are mostly altering the electrostatic component of binding energy with a destabilizing effect. As a result, chemical class changes are

expected to be functionally critical and alter the protein binding or solubility characteristics.

Figure 4.10. Fraction of chemical property changes of mutated driver proteins according to their physical locations.

While positions of mutated residues can be same between patients, they can be mutating to different amino acids. We also investigated these kinds of alterations between patients and we found 70 unique positions. As an example, Proline appeared to be mutated to Arginine at 596th position of EGFR in one patient while it is mutated to Leucine or Serine in other patients. We also checked whether their chemical properties are also changing differently, or they are mainly located in specific protein location. We found that they are mostly located in surface and interface residues and 63% of all them were originally charged amino acids which are altering generally to polar or preserving charged class.

4.2.2. 3D mutation patches and disease association

Heretofore, we analyzed locations and chemical properties of mutated residues, Since all alterations do not have the same impact on protein structures, we further assessed the effect of mutations on their disease-causing potential. We applied two different methods to the mutations by separating them according to their locations. Evmutation is the first method which uses an unsupervised statistical method inferring the impacts of mutations via taking into account co-evolution and epistasis. The latter one is PolyPhen-2 which uses a learning-based strategy to infer the impact of mutation by incorporating sequence and

structure based properties. Since these methods have not been trained for cancer mutations, we first applied them on to all mutations whether they can differentiate cancer driver gene mutations than passenger mutations and we found that both methods classified the mutations on driver genes as more damaging (P-value according to EVMutation = 6 $x10^{-44}$/Student's t-test, P-value according to Polyphen2 = 2 $x10^{-64}$/Chi-square test). In addition to driver genes, we also found that mutations on tumor suppressors are slightly more damaging than oncogenes with a P-value as 0.015 (Student's t-test) according to EVmutation.

We, then focused on impacts of mutations based on their structural locations and found that core and interface mutations are more damaging than surface mutations. As shown in **Figure 4.11** (PolyPhen-2) and **Figure 4.12** (EVMutation), both methods identified that surface mutations are the least damaging mutations yet only PolyPhen-2 found mutations in core region as the most damaging ones across all of them (P-value = 0.004/Chi-square). When we integrated the spatial organization information with 3D location, we also found that interface mutations in patches are more damaging compared to singleton mutations with a P-value as 0.002 (Chi-square test). Afterwards, we analyzed 487 interface mutations on driver genes and having PolyPhen-2 results. Since some proteins are more central than other ones and mutations on their interface regions can affect much more interactions, we further divided driver proteins having interface mutations into two classes which are frequently mutated proteins and the rest. As detailed in **Table 4.2**, 81% and 93% of singletons and patch mutations are damaging, respectively. While both percentages are relatively high, the impacts of mutations on frequently mutated proteins and the rest are different. Patch mutations are more damaging in frequently mutated proteins with a P-value as 0.00028 (Chi-square test) yet singleton mutations are more damaging in the rest of the proteins with a P-value as 0.00029 (Chi-square test). As a result, cancer mutations on interface region of central proteins are significantly located in mutation patches while interface mutations in other proteins stay as singletons.

Table 4.2. Disease association of singleton and patch mutations in the interface region of the hubs and the rest.

	Frequently mutated proteins		Rest		Total	
	Patch	Singleton	Patch	Singleton	Patch	Singleton
Benign	1	5	7	67	8	72
Possibly Damaging	11	0	2	63	13	63
Probably Damaging	69	1	21	240	90	241
Total	81	6	30	370	111	376

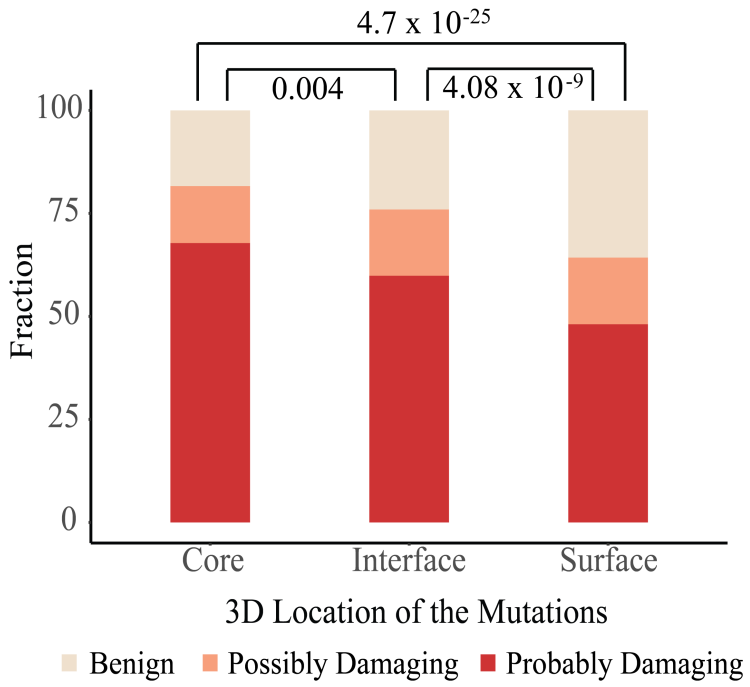


Figure 4.11. Fraction of mutations according to their PolyPhen-2 disease association in different locations (Chi-square test).

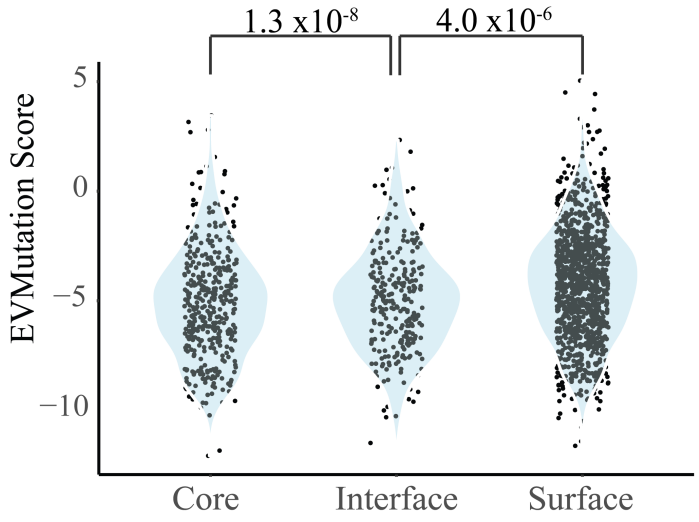


Figure 4.12. EVmutation disease association score distribution of mutations on different locations. The more negative score implies that the mutation is more damaging (Student's t-test).

4.2.3. Characteristics of interface mutations

While some proteins interact with only one partner protein, others can interact with several proteins through using same or different interfaces. We classified interfaces into 3 groups as namely 'one interface, one partner', 'one interface, shared by multiple partners' and 'multiple interfaces, used by different subsets of partners. Corresponding numbers of each group are represented in **Table 4.3**. In our study, most of the proteins including interface mutations have multiple partners and affect 6144 interactions in total. When we investigated tendency of their spatial organizations, we found that mutations on proteins having multiple interfaces are mostly located in patches yet interface mutations on proteins with a single interface stay as a singleton (P-value = 5.32×10^{-43}/Chi-aquare test).

Two of the interface groups are shown as examples in **Figure 4.13**. The first one represents 'multiple interfaces used by different subsets of partners' with PTPN11 protein interacting with both GRB2 and ERBB2 proteins through different interfaces. There are two different mutations on 510th and 69th position of PTPN11 on different interface surfaces. On the other hand, in the second one, we represented 'one interface used by one partner' group with TP53 protein which interacts both TP53BP2 and BCL2L1 through a shared interface. There is a mutation on 178th position of TP53 which can affect both of the interactions.

In order to better demonstrate the affected interfaces, we used network representation as in **Figure 4.14**. In this figure, there are two types of edges: one is between mutations which corresponds that they are on the same or overlapping interfaces, the second one is between protein and mutation which represents the interface of the protein. Therefore, mutations on 436th and 395th position of RB1 are on the same interface yet, mutation on 556th position is on different interface. While interface 1 of PIK3CA has so many mutations, interface 3 of PIK3CA has only one mutation. However, none of the mutations in PIK3CA and RB1 is used exclusively for binding to a single partner.

Table 4.3. Numerical information of mutations in each interface type.

	Number of mutations	Number of genes	Number of cancer genes
One interface, one partner	241	216	117
One interface, shared by multiple partners	361	325	206
Multiple interfaces, used by different subsets of partners	155	35	23

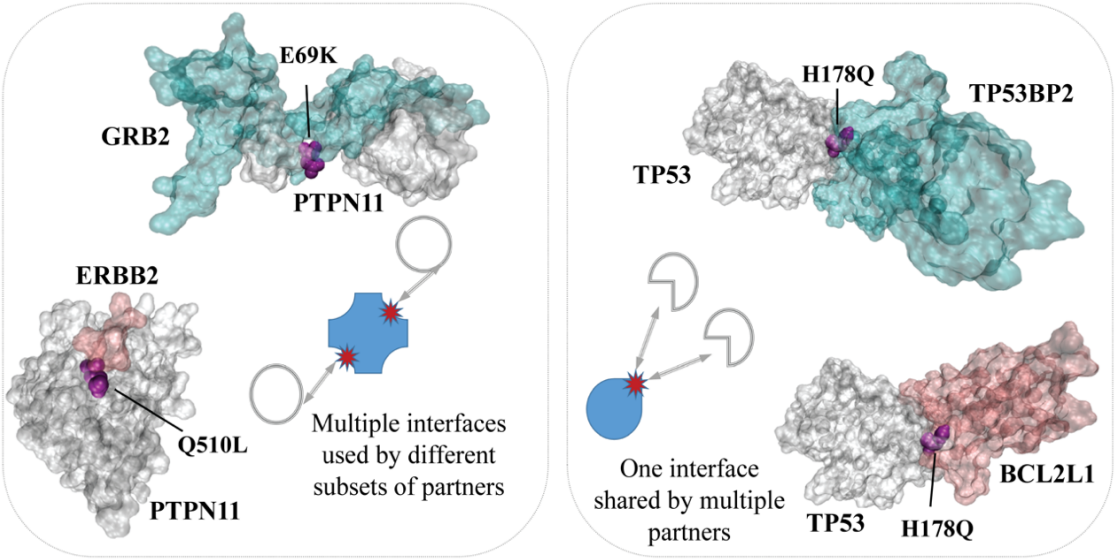


Figure 4.13. Representation of two types of proteins having one or multiple interface regions. The left part represents proteins having multiple interfaces (PTPN11) with different subset of partners and the right part represents proteins having one interface (TP53) shared by different partners.

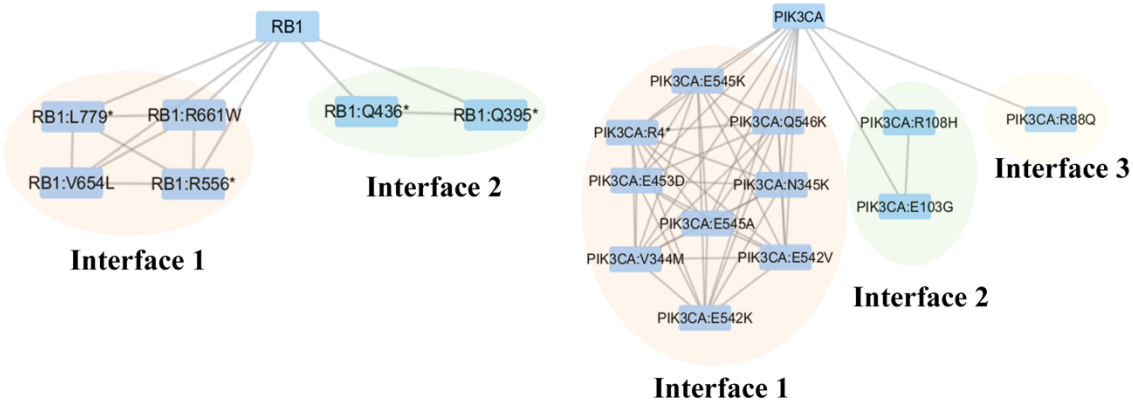
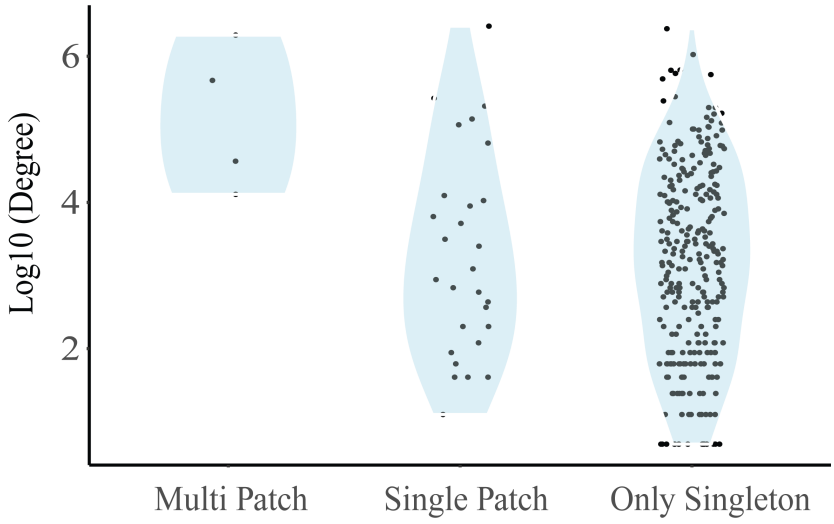


Figure 4.14. Network representation of RB1 and PIK3CA interface mutations. In the networks, mutations and proteins are nodes and edges symbolize both shared partner between mutations, interface between mutations and proteins.

Several previous studies have been reported that interface mutations are more likely to disrupt protein interactions in different diseases. Chen et al. studied interface mutations in autism disorder and compared siblings with and without the disease (Chen et al., 2018). They found that interaction disrupting mutations in sibling with autism have generally impact on central proteins and affect significantly high number of interactions compared to unaffected sibling. Raimondi et al. found that there are also distinct differences in these disrupting interactions between different cancer and histological subtypes (Raimondi et al., 2016). Additionally, Sahni et al. also reported that mutations in healthy individuals rarely affect interactions yet approximately 65% of disease related variants perturb protein-protein interactions (Sahni et al., 2015). Besides this information, they also integrated mutation information with protein networks and found that these variants are generally affecting only some of the interactions rather than disrupt all the interactions. Therefore, these kinds of integrative approaches have potential to enhance genotype-to-phenotype relations. Thus, we also checked the GBM mutations having structure information and we found that GBM mutations are also significantly more frequent in the interface region than the rest with a P-value far smaller than 0.0001. We then analyzed if there is also any tendency for hub protein interactions. In our dataset, there is 87 highly connected proteins with 1263 interactions which is 14.5 interaction in average yet remaining 3013 proteins only have 4412 interactions with an average as 1.47 per protein. Our result represents that these highly connected hub proteins (TP53, EGFR, PTEN, PIK3CA) are likely to have multiple patches in their interface regions and interface mutations are mostly located in the patches of these hubs as indicated in **Figure 4.15**.

A



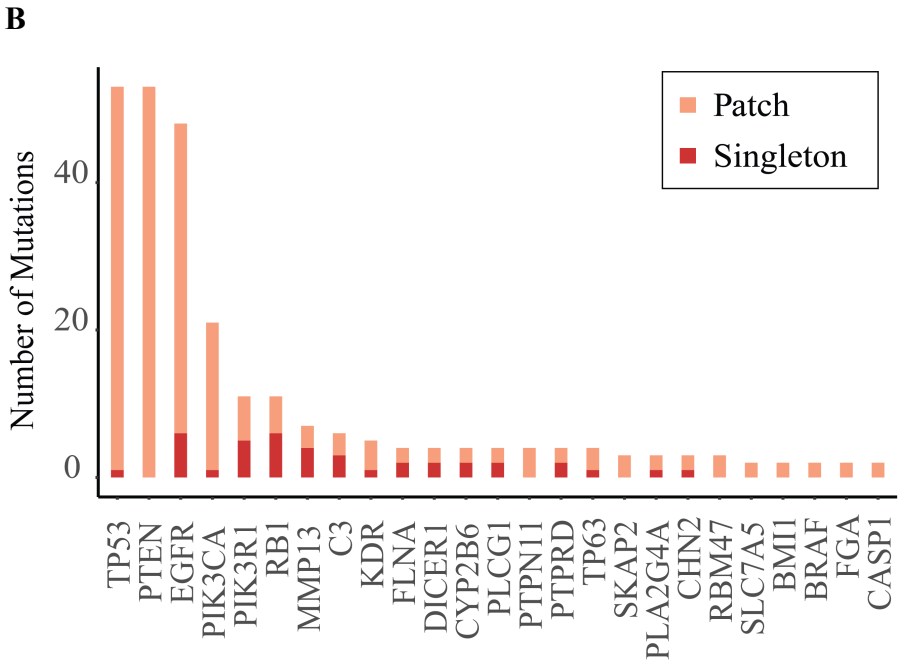


Figure 4.15. 3D spatial organization of interface mutations. **A.** Proteins having at least one interface mutation are classified as "proteins having multiple patches", "proteins having single patch" and "proteins without any patch". High degree proteins (hubs) in the interactome prone to include multiple patches on their interface regions. **B.** Hub proteins prone to have interface mutations locating on patches. For this graph, only the proteins having at least one patch on their interface regions are shown.

4.3. Patient specific sub-networks from mutation profiles and patient groups from pathway similarities

Mutations on distinct proteins may alter same pathways yet mutations on same protein but on different interfaces may change different pathways. This behavior can alter the wiring of interactions between proteins thus may change the signaling propagation. Since functional pathway information cannot be uncovered by comparing only mutated proteins, we used network modeling approach by implementing Forest module of Omics Integrator to reconstruct patient specific sub-networks. By prioritizing mutated proteins, the algorithm connects them to intermediate ones and thus reveals affected biological pathways. As a result, sub-networks for 205 patients were reconstructed and reduced into 137 unique KEGG pathways in total. We then grouped the patients according to corresponding set of overrepresented pathways and their enrichment scores. In this way, we have used affected biological functions which reveal the disease phenotype. When we applied consensus clustering (**Figure 4.16**) followed by non-negative matrix factorization (NMF), patients were classified into five groups on the basis of their pathway similarities.

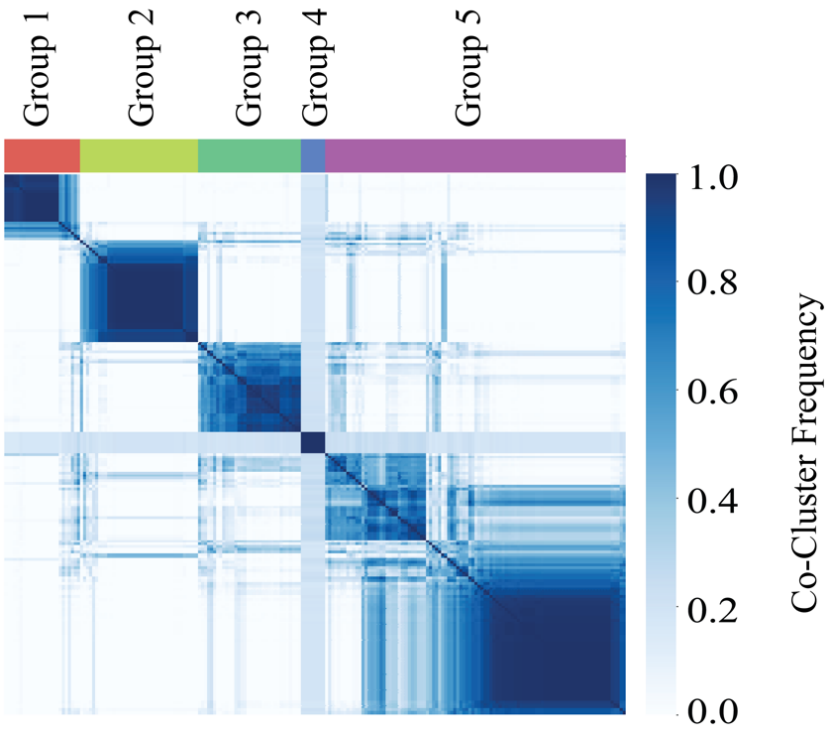


Figure 4.16. Co-clustering frequency matrix.

In addition to enrichment analysis on patient specific subnetworks, we performed enrichment analysis for each patient to their list of mutated proteins in order to indicate the advantages of network-guided analysis. As a result of this analysis, we obtained six significantly enriched pathways for only 11 patients. These pathways include EGFR signaling and Glioma pathways which are not informative enough for further analysis to stratify the patients.

In order to assert an association between patient groups and clinical outcome, we performed survival analysis for each group. As indicated in **Figure 4.17**, there is a significant difference between the survival plot of each groups with a multi class log-rank P-value as 0.0408. Across patient groups, Group 5 has the highest survival with 450.09 days yet, survival of Group 4 is only 259.75 days as the lowest. Additionally, we compared the three known GBM transcriptomic subtypes which are classical, preneural and mesenchymal (Q. Wang et al., 2017), only Group 2 shows significant enrichment in classical subtype (P-value = 0.016/ Hypergeometric test).

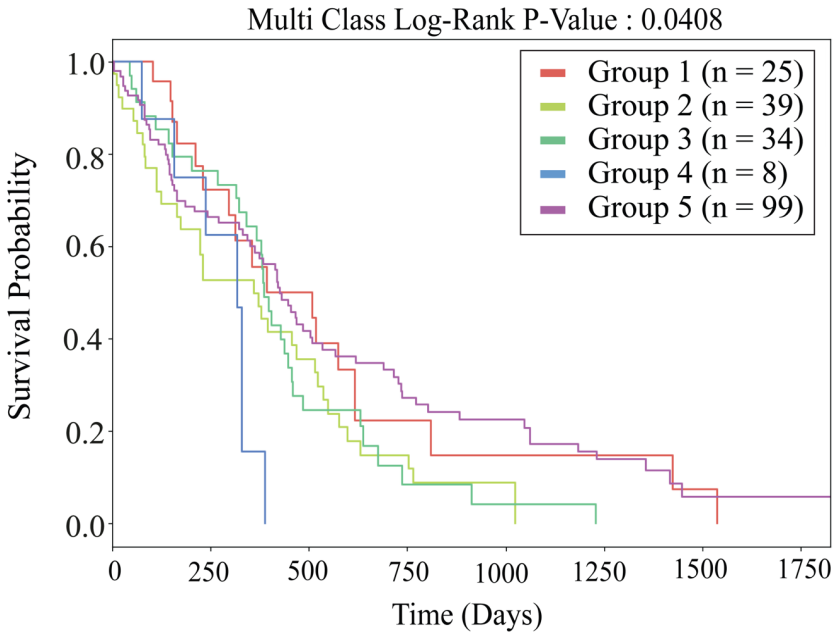


Figure 4.17. Kaplan-Meier survival plots of the patient groups classified with NMF and consensus clustering.

While some pathways are significantly active across all patient groups, some of them are enriched in specific patient groups or specific set of patient groups since groups were classified according to their overrepresented pathways. Predominant pathways for each patient group were indicated in **Figure 4.18** except Group 4 which did not enriched in any pathway. While Rap1, EGFR, and TNF signaling pathways are common in all groups, TGF-beta signaling and Hippo signaling are predominant in Groups 5 and Group 2, respectively. Jak-Stat pathway is present in all except Group 5. While mTOR and Hif-1 signaling pathways are enriched in Groups 2 and 3. In addition to enriched pathways, we calculated overrepresented 3D patches in each patient group as indicated in **Figure 4.19**. PTEN Patch 1 and TP53 Patch 1 are found in all patient groups except Group 3 and Group 4, respectively. While there are less amount of patches are found in two patient group concurrently, a high fraction of patches are mainly found in specific patient groups such as EGFR Patch 3, BRAF Patch 1, PTEN Patch 2, PIK3R1 Patch 1, RB1 Patch 1 and Patch 1,3 and 4 of PIK3CA. Moreover, we observed that different patches of same protein can be enriched in different patient groups as in the case of EGFR and PTEN patches. Additionally, BRAF mutations V600E and G596D form a 3D patch in only Group 5.

When we focused on interface mutations, we found that totally 318 interactions in patient networks were affected. For each patient group, number of affected interactions are 23 for Group 1, 82 for Group 2, 36 for Group 3, 8 for Group 4 and lastly 223 for Group 5. When we deeply analyzed, we found interaction of EGFR to MAPK8IP1, CAV1, RIN1 and SHC1 are the most common ones in 34 patients.

Thereafter, we merged the patients in each patient group in order to obtain union networks. We illustrated a sample merged network of Group 1 in **Figure 4.20**. Since groups were formed according to pathway similarity, mutated proteins in each patient are not the same. We indicated this difference by using pie chart for each node. This pie charts indicate the ratio of being mutated as red and not mutated as blue. Some nodes are fully blue since they were added to the network as intermediate nodes to connect the mutated proteins. For example, NFKBIA is intermediate proteins that do not have a mutation in any patient in Group 1, yet it connects mutated proteins including IKBKB, TP53, NFKB1. Another example is CTNNB1 which is an intermediate protein found in Group 3 to connect proteins such as PIK3R1, AKT1, LRP2. In this way, mutated and intermediate proteins complete the signaling propagation thus pathways can be detected in enrichment analysis. Additionally, not every protein is included in each patient group, hence we indicated the frequency of the protein through size of the node. Similarly, edge thickness represents the frequency of that edge in group.

When we investigated common nodes in each union network, TP53 which is a very well-known hub protein comes to the forefront. However, other central proteins such as IKBKG and MDM2 are specific to Group 1 and Group 2, respectively. In general, there are 971 total proteins in the union networks. From them only 17 proteins are common in all groups (Group 4 was excluded due to its small size) whereas 685 proteins are present only in one group.

4.4. Potential therapeutic targets for each patient group

According to result of section 4.3 represented in **Figure 4.17**, there is a significant relation among patient groups and survivals which indicates that patient groups may have similar clinical background and thus similar response to therapies. In order to analyze drug response of patient groups, we collected 37 GBM cell lines deposited in Cell Model Passports which both have mutation and drug sensitivity data. These cell lines have 13,243 unique mutations and only 16 of the mutations are located in pre-calculated 3D spatial patches on proteins such as RB1, BRAF, EGFR, PTEN and TP53 which are also present significantly in one or more patient groups. However, the reduction in the quantity of remaining mutations decreases the number of cell lines that we can use to link patient groups to cell line to 17. Moreover, we collected 73 drugs, their targets and responses of these 17 cell lines to them. In order to infer potential therapeutic responses of patient groups, first we found the best representative set of cell lines for each patient group and, all drugs that are targeting proteins which are significantly present on corresponding patient group's network and have drug sensitivity data for at least one representative cell line.

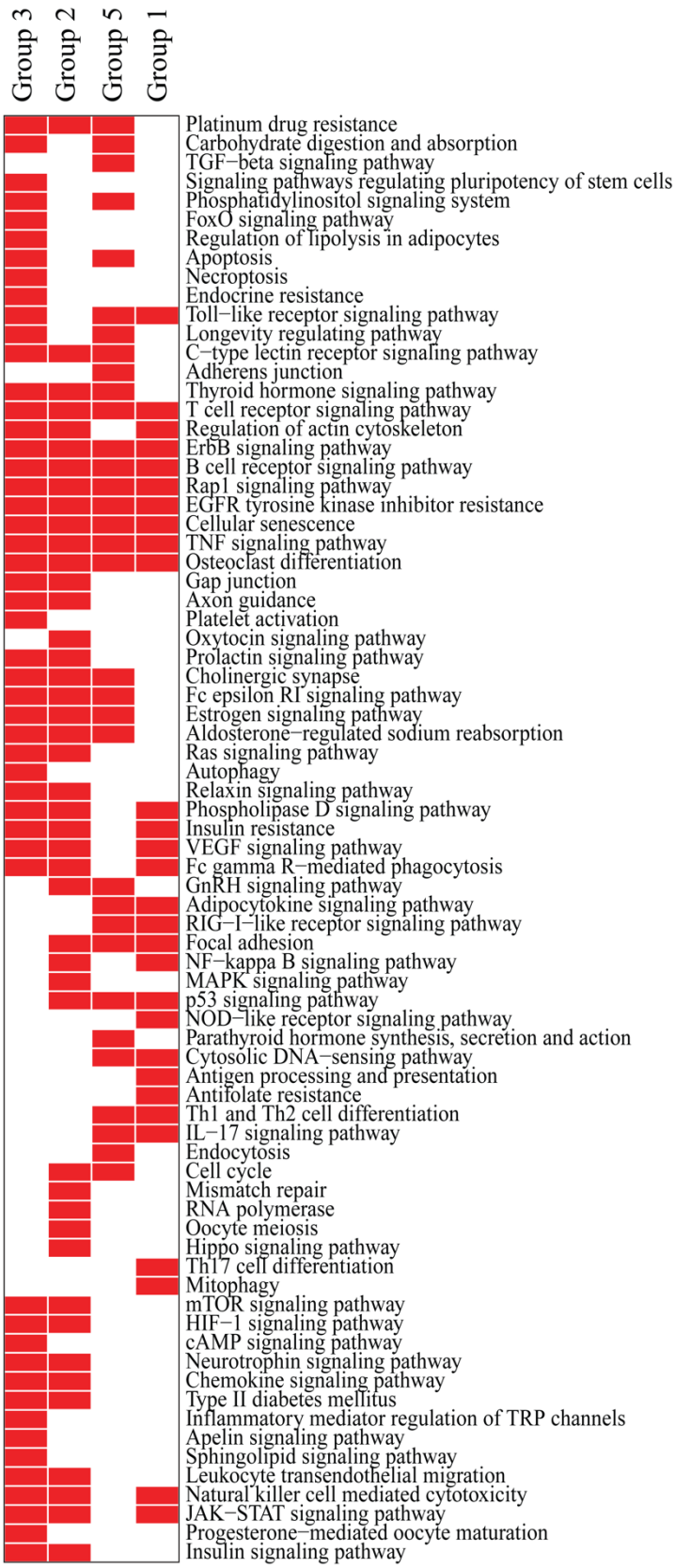

Group 3
Group 2
Group 5
Group 1
Platinum drug resistance
Carbohydrate digestion and absorption
TGF−beta signaling pathway
Signaling pathways regulating pluripotency of stem cells
Phosphatidylinositol signaling system
FoxO signaling pathway
Regulation of lipolysis in adipocytes
Apoptosis
Necroptosis
Endocrine resistance
Toll−like receptor signaling pathway
Longevity regulating pathway
C−type lectin receptor signaling pathway
Adherens junction
Thyroid hormone signaling pathway
T cell receptor signaling pathway
Regulation of actin cytoskeleton
ErbB signaling pathway
B cell receptor signaling pathway
Rap1 signaling pathway
EGFR tyrosine kinase inhibitor resistance
Cellular senescence
TNF signaling pathway
Osteoclast differentiation
Gap junction
Axon guidance
Platelet activation
Oxytocin signaling pathway
Prolactin signaling pathway
Cholinergic synapse
Fc epsilon RI signaling pathway
Estrogen signaling pathway
Aldosterone−regulated sodium reabsorption
Ras signaling pathway
Autophagy
Relaxin signaling pathway
Phospholipase D signaling pathway
Insulin resistance
VEGF signaling pathway
Fc gamma R−mediated phagocytosis
GnRH signaling pathway
Adipocytokine signaling pathway
RIG−I−like receptor signaling pathway
Focal adhesion
NF−kappa B signaling pathway
MAPK signaling pathway
p53 signaling pathway
NOD−like receptor signaling pathway
Parathyroid hormone synthesis, secretion and action
Cytosolic DNA−sensing pathway
Antigen processing and presentation
Antifolate resistance
Th1 and Th2 cell differentiation
IL−17 signaling pathway
Endocytosis
Cell cycle
Mismatch repair
RNA polymerase
Oocyte meiosis
Hippo signaling pathway
Th17 cell differentiation
Mitophagy
mTOR signaling pathway
HIF−1 signaling pathway
cAMP signaling pathway
Neurotrophin signaling pathway
Chemokine signaling pathway
Type II diabetes mellitus
Inflammatory mediator regulation of TRP channels
Apelin signaling pathway
Sphingolipid signaling pathway
Leukocyte transendothelial migration
Natural killer cell mediated cytotoxicity
JAK−STAT signaling pathway
Progesterone−mediated oocyte maturation
Insulin signaling pathway

Figure 4.18. Enrichment of KEGG pathways across the patient groups. Reds indicate the enrichment of corresponding KEGG pathway in corresponding patient group (except Group 4 which does not have any KEGG pathway dominantly enriched in its patients).

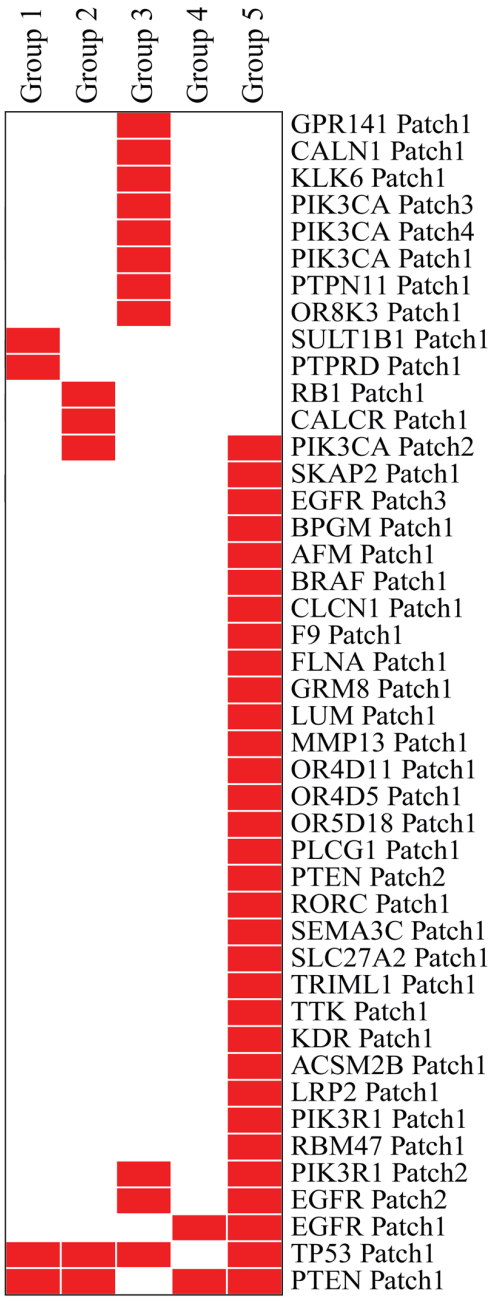


Figure 4.19. Predominant 3D patches in each patient group. Red indicates overrepresentation of the patches in corresponding patient group.

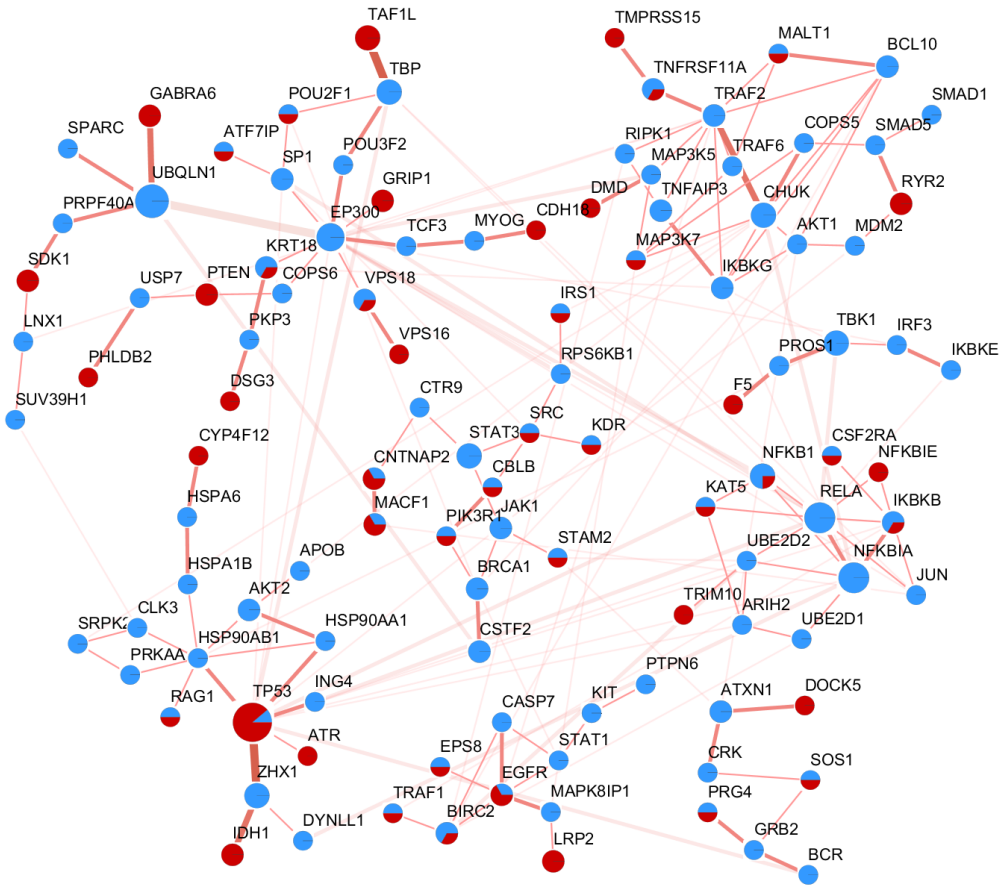


Figure 4.20. Merged network of Group 1. While red color represents how many patients in the patient group has corresponding protein as mutated, and blue color represent how many patients in the patient group has corresponding protein as non-mutated which means being an intermediate protein connecting mutated proteins. Node size and edge thickness are the frequency of the corresponding node and edge in the patient networks in the group

In **Figure 4.19**, all the overrepresented 3D patches are indicated for each patient group such that RB1, TP53, PTEN, patches are enriched in Group 2 yet TP53, PTEN patches without RB1 are significantly present in Group 1. At the same time, PTEN, TP53 and BRAF have a strong tendency to be present in Group 5. Moreover, two patches of PTEN are significantly represented in different groups. Through these signature 3D patches intersected with cell line mutations, we linked patient groups to cell lines. According to our results as represented in **Figure 4.21**, Group 1 is connected to cell lines which are linked to TP53 Patch 1 and PTEN Patch 1, while Group 3 and Group 4 are connected to cell lines that have at least one mutation in TP53 Patch 1 and PTEN Patch 1, respectively. Group 2 linked to cell lines using TP53 Patch1, PTEN Patch 1 and RB Patch 1 patches yet Group 5 does not use RB1 Patch 1. After obtaining drugs which were exposed to linked cell lines and targets of them, we first checked if these targets are present in networks of

patient groups. As represented in **Figure 4.21**, only Group 2, Group 3 and Group 5 include drug targets in their networks. With this study, from mutation patches and group networks to drug responses, we can connect patient groups to drug with specifying their response information. Below each connection will be detailed.

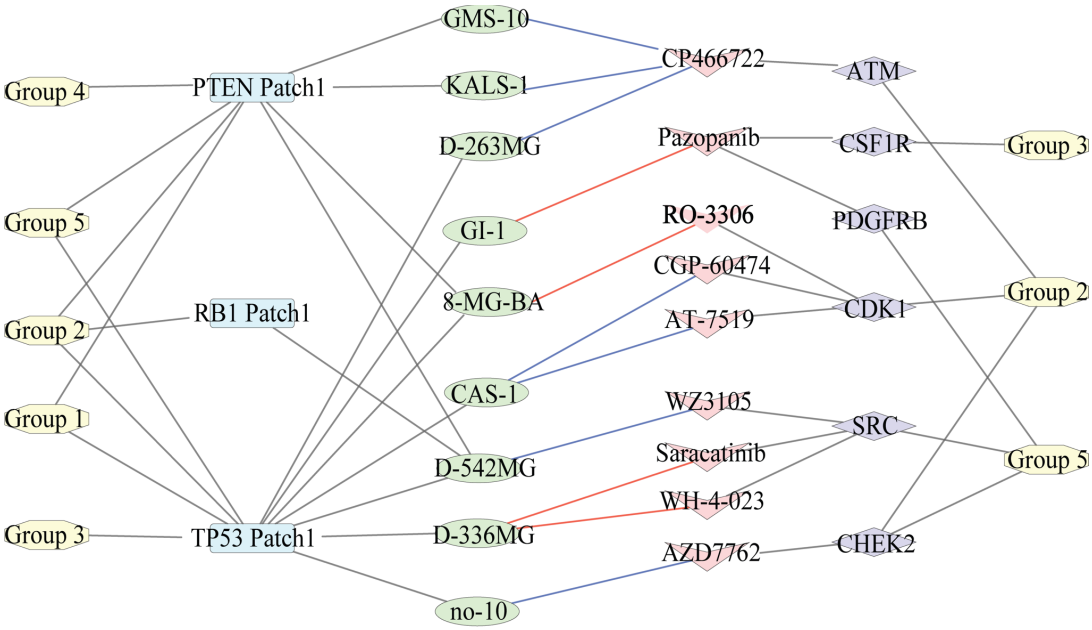


Figure 4.21. Patient groups, cell line and drug linkages. Yellow box represents patient groups, pink triangles are drug representations. Blue box is shared patched, green eclipse connected cell lines and purples are targets of the drugs. The color of the edges between cell lines and drugs represent the drug response: blue as resistant and red as sensitive.

We connected Group 2 to CP466722 by ATM, to RO3306, CGP-60474 and AT7519 by CDK1 and, to AZD7762 by CHEK2. Moreover, we linked Pazopanib with Group 3 and Group 5 via CSF1R and PDGFRB, respectively. Additionally, Group 5 can also be linked to AZD7762 with CHEK2 and, to WZ3105, Saracatinib and WH-4-023 by SRC protein.

CP466722 is an inhibitor of ATM (ataxia telangiectasia-mutated) protein kinase which has a role in repairment of double-strand break induced by ionizing irradiation. Since the blood brain barrier limits the chemotherapeutic options, radiation therapy is considered as a primary option. However, most of the patients develop resistance to radiation and disease recurrence happens. It is reported that ATM may be linked with radiotherapy resistance (Estiar & Mehdipour, 2018), and several studies have indicated that ATM inhibition can make cell sensitive to radiation (Li et al., 2017; Rainey, Charlton, Stanton, & Kastan, 2008). Li et al. (Li et al., 2017) used shRNA to inhibit ATM expression with radiation therapy on glioma stem cells and found weakening in cell proliferation and lowering in survival. Yet in our study, we found that Group 2 may be resistance to ATM

inhibition as indicated in **Figure 4.22**. In overrepresented pathways, Group 2 also shows a resistance profile to platinum-based drugs which create breaks on DNA and lead to apoptosis in tumor cells. As a result of these profiles, Group 2 GBM patients may develop another DNA repair mechanism to escape from damages.

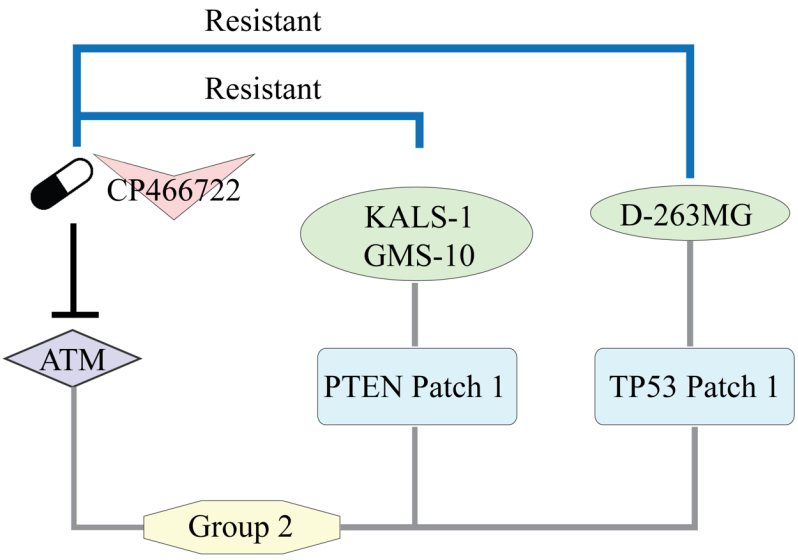


Figure 4.22. Hypothetical therapeutic proposal for Group 2 patients by using ATM as target protein.

Additionally, RO3306, CGP-60474 and AT7519 are both cyclin dependent kinase (CDK) inhibitors. While RO3306 is specifically inhibits CDK1 which is responsible for the G2/M phase transition, others are not specific to any CDK. Moreover, Group 2 only sensitive to RO3306 (**Figure 4.23**), yet resistant to CGP-60474 and AT7519. In cell, CDK1 binds to cyclin B and accumulates during G2 phase in an inactive phosphorylated position. G2/M transition is mediated by cell division cycle 25 (CDC25) phosphatase which activates CDK1-cyclin b complex by dephosphorylation and thus starts mitosis in the cell (Vassilev, 2006). Selectively inhibiting CDK1 is important since non selective inhibiting has been reported as causing significant cytotoxic effects resulting from loss of CDK7 and CDK9 in vivo (Bose, Simmons, & Grant, 2013). Based upon the information form recent study (Jorda et al., 2018) which compared several inhibitors and found RO3306 as the most suitable compound for inhibition of CDK1, we may assert that RO3306 can be a potential therapeutic for Group 2 patients due to its selectivity to CDK1.

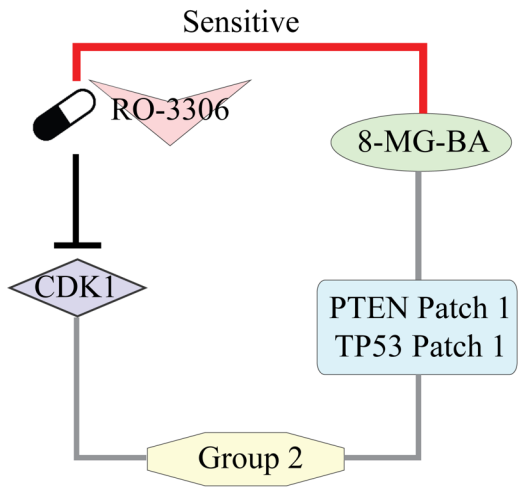


Figure 4.23. Hypothetical therapeutic proposal for Group 2 patients by using CDK1 as target protein.

Further, AZD7762 is a selective checkpoint kinase inhibitor (CHK1 and CHK2) (Zabludoff et al., 2008). CHK2 is the product of CHEK2 gene and acts as tumor suppressor by inhibiting CDC25 which activates CDKs to enter mitosis phase, in case of double stranded breaks on DNA. It has role in DNA repair, cell cycle arrest and apoptosis (Sallinen, Ikonen, Haapasalo, & Schleutker, 2005). However, both Group 2 and Group 5 represented a profile as resistant to AZD7762 (**Figure 4.24**).

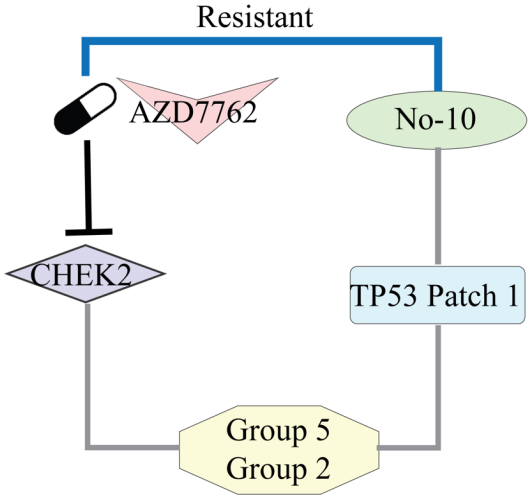


Figure 4.24. Hypothetical therapeutic proposal for Groups 2 and 5 patients by using CHEK2 as target protein.

Pazopanib which is a multi-targeted receptor tyrosine kinase inhibitor, is linked to Group3 and Group 5 through its targets Colony Stimulating Factor 1 Receptor (CSF1R) and Platelet Derived Growth Factor Receptor Beta (PDGFRB), respectively. Colony Stimulating Factor 1 (CSF1) binds to CSF1R and activates several signaling pathways,

including Ras/Raf/Mitogen-Activated Protein Kinase (Ras/Raf/MAPK), Phosphatidylinositol-3-Kinase (PI3K) and Janus Kinase/Signal Transducers and Activators of Transcription (JAK/STAT) pathways which are important mainly in proliferation and dysregulation of these pathways results in tumor formation/progression. Moreover, CSF1R and CSF1 have a role in migration, differentiation, and survival of Tumor-Associated Macrophages and Microglia (TAMs) which have tumor permissive and immunosuppressive characteristics and they are highly available in glioma microenvironment. CSF1, the ligand of CSF1R, is responsible for the differentiation of TAMs to pro-tumorigenic. Therefore, inhibition of CSF1R results in the differentiation of the macrophages and makes them more anti-tumorigenic (Cannarile et al., 2017; Ries et al., 2014). As another target of Pazopanib, Platelet Derived Growth Factor Receptor Beta (PDGFRB) protein is a receptor tyrosine kinase and functions as a cell surface receptor. It activates cell proliferation and survival. Additionally, it is proven that PDGFRB is overexpressed in GBM cells and very important for self-renewal (Papadopoulos & Lennartsson, 2018). Therefore, as shown in **Figure 4.25**, we suggest that Group 3 and Group 5 might be sensitive to a treatment based on Pazopanib.

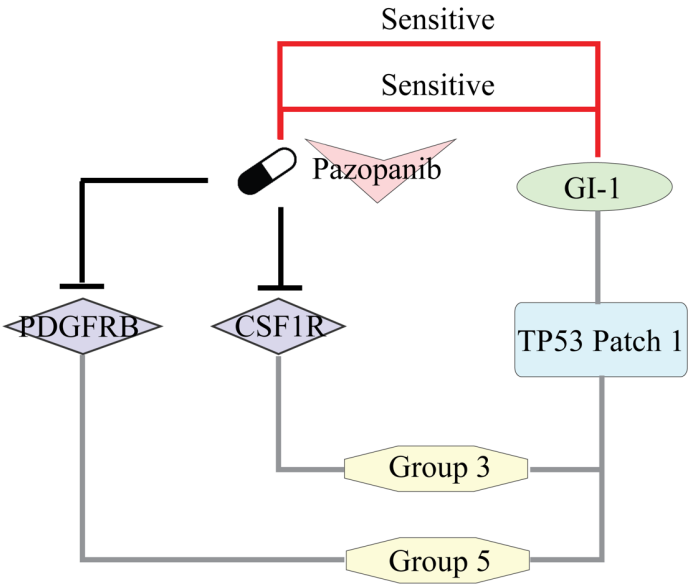


Figure 4.25. Hypothetical therapeutic proposal for Groups 3 and 5 by using CSF1R and PDGFRB as target proteins, respectively.

As a last example, SRC protein is a target of WZ3105, Saracatinib and WH-4-023 in Group 5, and a non-receptor protein tyrosine kinase playing an important role in growth, adhesion, and differentiation (Roskoski, 2015). It is also a component of several cell signaling pathways including Epidermal Growth Factor Receptor (EGFR), ERBB, and Ras-Associated Protein-1 (Rap1) signaling pathways. WZ3105 which is a kinase-inhibitor, targets SRC. The cell line D-452MG connecting with Group 5 is resistant to this compound, thus we suggest that Group 5 might be possibly resistant to WZ3105 (**Figure 4.26**).

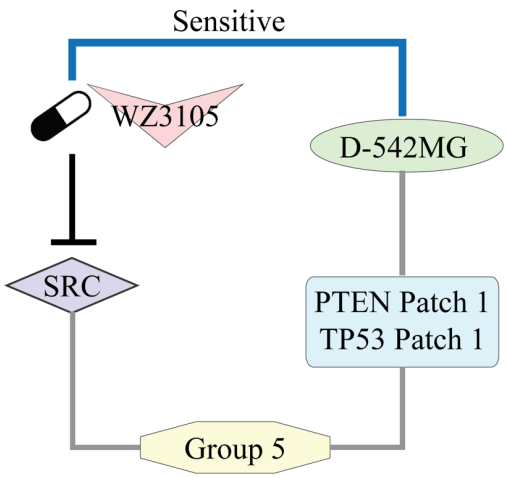


Figure 4.26. Hypothetical therapeutic proposal for Group 5 patients by using SRC as target protein.

CHAPTER 5

5. DISCUSSION AND CONCLUSION

GBM has remained an incurable form of brain tumor by the reason of molecular heterogeneity which is the main obstacle to development of efficient therapies for each particular patient context. In this study, we concentrated on organization of GBM mutations both in protein structures and protein-protein interaction networks to uncover differences and commonalities across patients. Thus, we used a systems level approach from mutation profiles to patient-specific subnetworks and clinical outcome. Firstly, we analyzed individual mutations of 290 GBM patients coming from TCGA on basis of their spatial organization on 3D structures of proteins and interactions, physicochemical characteristics, oncogenic properties, and disease associations. Then, we continued with unearthing the affected functional pathways from network arrangements of mutated proteins, and classification of the patients based on their overrepresented pathways. Lastly, we linked each patient group to related drugs and responses of these drugs in order to propose hypothetical therapeutics.

From structural point of view, out of 15399 mutations, 4702 mutations have structural information and only 10% of them are on spatial 3D groups. Although there is a small portion of all mutations, we realized that different 3D patches of a protein are located in distinct domains that could have distinct functional consequences and also different phenotypic impacts for patients. By using this information, we grouped the patients into mutually exclusive groups based on their most common patches. We reduced the heterogeneity across patients through 3D spatial organization according to statistically significant association between patient groups and their survival curves. While patient group included at least one mutation in TP53 patch or EGFR patch shows a better survival than patient group having at least one mutation in PI3K patches. Therefore, the strong association between patient groups and their survivals indicates that patients with similar 3D spatial organization in their proteins may have similar disease phenotypes which may represent similar affected functions and pathways in the tumor cells and this result suggests that 3D spatial organization of mutations can help to overcome the obstacle resulting from molecular heterogeneity.

Moreover, we found an association between protein oncogenicity and spatial organization or their mutations. We found that driver mutations of tumor suppressors and oncogenes act statistically different. While tumor suppressor driver mutations tend to be in patches, driver mutations in oncogenes tend to be singletons. Further, patches of tumor suppressor are statistically larger than the patches on oncogenes. This result agrees with the idea that tumor suppressors can be functionally damaged in several ways thus their mutations could be distributed however, making a protein active needs specific alteration. When we analyzed the types of mutations in different oncogenic proteins, we found that nonsense and frameshift mutations are more frequent in tumor suppressors which could make them unfunctional.

By using structural information of the mutations, we divided them as core, surface and interface mutations and found that GBM mutations are more frequently located on interface regions. We further found that these interface mutations are mostly populated in spatial patches. As a physicochemical point of view, changes in the chemical properties is functionally critical and alter the protein binding or solubility characteristics. Thus, we also analyzed switches or preservation of wild type chemical classes in each mutated residue on proteins according to their 3D locations. While core mutations tend to preserve their hydrophobic characters, interface and surface mutations are significantly more tend to change. Charged interface mutations are more prone to change and our study agrees with the result of previous study about interface mutations and their impacts on altering the electrostatic component of binding energy with a destabilizing effect (Nishi et al., 2013). As disease association point of view, the most damaging mutations are the core ones followed by interface and the least damaging one is the surface mutations. When we integrated spatial organization, interface mutations in patches show more damaging characteristics compared to singleton mutations.

Further, we investigated the proteins having one or multiple interfaces. Interface mutations in proteins with a single interface tend to be singletons while interface mutations in proteins having multiple interfaces are mostly located in patches. Moreover, interfaces can be used specifically by one partner or shared by multiple partners. Proteins having multiple partners are called hubs and these highly connected hub proteins (TP53, EGFR, PTEN, PIK3CA) are likely to have multiple patches in their interface regions. These patches are prone to be larger in tumor suppressor and smaller in oncogenes. When we analyzed the disease association of mutations on hub proteins, we found that patch mutations on hub proteins are more disease-causing although singleton mutations of rest of the proteins are more disease-causing.

Applying network-based approach to driver mutated proteins, we reconstructed patient-specific subnetworks and reduced each network into enriched pathways to uncover the potentially affected pathways in patients. With this strategy, we stratified the patients into 5 groups and each patient group has a set of signature 3D spatial patches and significant association between survivals of the patients in corresponding group. Among these patient groups, Group 5 has the highest survival while Group 4 has the lowest one. Additionally, some pathways are commonly overrepresented in a set of patient groups, some of them

are specific for one patient groups. For example, Rap1, EGFR, and TNF signaling pathways are common in all groups, NOD-like receptor signaling pathway and Hippo signaling pathway are specific to Group 1 and Group 2, respectively. In addition to this analysis, we also applied the same pathway enrichment analysis to mutated gene lists and did not get a meaningful result which highlighted the importance of a network-based approach to enlighten the hidden molecular mechanisms of the cell.

Since there are functional similarities between patients in the same group and differences between inter groups, we integrated available drug treatment data to our patient groups by using mutation profiles of GBM cell lines. Each group of patients is linked to each GBM cell line through its predominant patches. A multi-targeted receptor tyrosine kinase inhibitor, Pazopanib is linked to Group 3 patients through its target CSF1R and both Group 3 and GI-1 GBM cell line have TP53 Patch 1 as the connection marker. By this way, we hypothetically proposed Pazopanib as an effective therapeutic for Group 3 patients.

As a conclusion, we integrated non-synonymous mutations of patients to structural information through a network-based approach which reduced the molecular heterogeneity across patients. Our approach from mutations to protein interactions and eventually to signaling networks and pathways let us to connect pharmacological information of cell lines to hypothetical clinical outcomes for patients. We believe that this study indicates a new perspective for application of network-based analysis for the precision medicine.

REFERENCES

Acuner Ozbabacan, S. E., Gursoy, A., Nussinov, R., & Keskin, O. (2014). The structural pathway of interleukin 1 (IL-1) initiated signaling reveals mechanisms of oncogenic mutations and SNPs in inflammation and cancer. *PLoS Comput Biol, 10*(2), e1003470. doi:10.1371/journal.pcbi.1003470

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Ainscough, B. J., Griffith, M., Coffman, A. C., Wagner, A. H., Kunisaki, J., Choudhary, M. N., . . . Mardis, E. R. (2016). DoCM: a database of curated mutations in cancer. *Nat Methods, 13*(10), 806-807. doi:10.1038/nmeth.4000

Amberger, J., Bocchini, C., & Hamosh, A. (2011). A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). *Hum Mutat, 32*(5), 564-567. doi:10.1002/humu.21466

An, O., Gursoy, A., Gurgey, A., & Keskin, O. (2013). Structural and functional analysis of perforin mutations in association with clinical data of familial hemophagocytic lymphohistiocytosis type 2 (FHL2) patients. *Protein Sci, 22*(6), 823-839. doi:10.1002/pro.2265

Andersen, P. K., & Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *Ann. Statist., 10*(4), 1100-1120. doi:10.1214/aos/1176345976

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., & Hogue, C. W. (2001). BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res, 29*(1), 242-245. doi:10.1093/nar/29.1.242

Berman, H. M., Kleywegt, G. J., Nakamura, H., & Markley, J. L. (2014). The Protein Data Bank archive as an open data resource. *J Comput Aided Mol Des, 28*(10), 1009-1014. doi:10.1007/s10822-014-9770-y

Bleeker, F. E., Molenaar, R. J., & Leenstra, S. (2012). Recent advances in the molecular understanding of glioblastoma. *J Neurooncol, 108*(1), 11-27. doi:10.1007/s11060-011-0793-0

Bose, P., Simmons, G. L., & Grant, S. (2013). Cyclin-dependent kinase inhibitor therapy for hematologic malignancies. *Expert Opin Investig Drugs, 22*(6), 723-738. doi:10.1517/13543784.2013.789859

Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., . . . Network, T. R. (2013). The somatic genomic landscape of glioblastoma. *Cell, 155*(2), 462-477. doi:10.1016/j.cell.2013.09.034

Brown, K. R., & Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics, 21*(9), 2076-2082. doi:10.1093/bioinformatics/bti273

Buljan, M., Blattmann, P., Aebersold, R., & Boutros, M. (2018). Systematic characterization of pan-cancer mutation clusters. *Mol Syst Biol, 14*(3), e7974. doi:10.15252/msb.20177974

Burke, J. E., Perisic, O., Masson, G. R., Vadas, O., & Williams, R. L. (2012). Oncogenic mutations mimic and enhance dynamic events in the natural activation of phosphoinositide 3-kinase p110alpha (PIK3CA). *Proc Natl Acad Sci U S A, 109*(38), 15259-15264. doi:10.1073/pnas.1205508109

Cancer Genome Atlas Research, N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., . . . Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet, 45*(10), 1113-1120. doi:10.1038/ng.2764

Cannarile, M. A., Weisser, M., Jacob, W., Jegg, A. M., Ries, C. H., & Ruttinger, D. (2017). Colony-stimulating factor 1 receptor (CSF1R) inhibitors in cancer therapy. *J Immunother Cancer, 5*(1), 53. doi:10.1186/s40425-017-0257-y

Chakravarty, D., Gao, J., Phillips, S. M., Kundra, R., Zhang, H., Wang, J., . . . Schultz, N. (2017). OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol, 2017*. doi:10.1200/PO.17.00011

Chalhoub, N., & Baker, S. J. (2009). PTEN and the PI3-kinase pathway in cancer. *Annu Rev Pathol, 4*, 127-150. doi:10.1146/annurev.pathol.4.110807.092311

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., & Cesareni, G. (2007). MINT: the Molecular INTeraction database. *Nucleic Acids Res, 35*(Database issue), D572-574. doi:10.1093/nar/gkl950

Chen, S., Fragoza, R., Klei, L., Liu, Y., Wang, J., Roeder, K., . . . Yu, H. (2018). An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. *Nat Genet, 50*(7), 1032-1040. doi:10.1038/s41588-018-0130-z

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol, 3*, 140. doi:10.1038/msb4100180

Ciriello, G., Cerami, E., Sander, C., & Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res, 22*(2), 398-406. doi:10.1101/gr.125567.111

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., . . . Ding, L. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res, 22*(8), 1589-1598. doi:10.1101/gr.134635.111

Drake, J. M., Paull, E. O., Graham, N. A., Lee, J. K., Smith, B. A., Titz, B., . . . Stuart, J. M. (2016). Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer. *Cell, 166*(4), 1041-1054. doi:10.1016/j.cell.2016.07.007

Drew, L. (2016). Pharmacogenetics: The right drug for you. *Nature, 537*(7619), S60-62. doi:10.1038/537S60a

Dutkowski, J., & Ideker, T. (2011). Protein networks as logic functions in development and cancer. *PLoS Comput Biol, 7*(9), e1002180. doi:10.1371/journal.pcbi.1002180

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., . . . Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res, 47*(D1), D427-D432. doi:10.1093/nar/gky995

Engin, H. B., Guney, E., Keskin, O., Oliva, B., & Gursoy, A. (2013). Integrating structure to protein-protein interaction networks that drive metastasis to brain and lung in breast cancer. *PLoS One, 8*(11), e81035. doi:10.1371/journal.pone.0081035

Engin, H. B., Hofree, M., & Carter, H. (2015). Identifying mutation specific cancer pathways using a structurally resolved protein interaction network. *Pac Symp Biocomput*, 84-95.

Engin, H. B., Kreisberg, J. F., & Carter, H. (2016). Structure-Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. *PLoS One, 11*(4), e0152929. doi:10.1371/journal.pone.0152929

Estiar, M. A., & Mehdipour, P. (2018). ATM in breast and brain tumors: a comprehensive review. *Cancer Biol Med, 15*(3), 210-227. doi:10.20892/j.issn.2095-3941.2018.0022

Faure, G., & Koonin, E. V. (2015). Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. *Phys Biol, 12*(3), 035001. doi:10.1088/1478-3975/12/3/035001

Fiser, A., & Sali, A. (2003). Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol, 374*, 461-491. doi:10.1016/S0076-6879(03)74020-8

Flanagan, C. A., Schnieders, E. A., Emerick, A. W., Kunisawa, R., Admon, A., & Thorner, J. (1993). Phosphatidylinositol 4-kinase: gene structure and requirement for yeast cell viability. *Science, 262*(5138), 1444-1448. doi:10.1126/science.8248783

Fujimoto, A., Okada, Y., Boroevich, K. A., Tsunoda, T., Taniguchi, H., & Nakagawa, H. (2016). Systematic analysis of mutation distribution in three dimensional protein structures identifies cancer driver genes. *Sci Rep, 6*, 26483. doi:10.1038/srep26483

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., . . . Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer, 4*(3), 177-183. doi:10.1038/nrc1299

Gao, J., Chang, M. T., Johnsen, H. C., Gao, S. P., Sylvester, B. E., Sumer, S. O., . . . Sander, C. (2017). 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med, 9*(1), 4. doi:10.1186/s13073-016-0393-x

Georgescu, M. M. (2010). PTEN Tumor Suppressor Network in PI3K-Akt Pathway Control. *Genes Cancer, 1*(12), 1170-1177. doi:10.1177/1947601911407325

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., . . . Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature, 446*(7132), 153-158. doi:10.1038/nature05610

Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H. W., & Stumpflen, V. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res, 34*(Database issue), D436-441. doi:10.1093/nar/gkj003

Guo, H. H., Choe, J., & Loeb, L. A. (2004). Protein tolerance to random amino acid change. *Proc Natl Acad Sci U S A, 101*(25), 9205-9210. doi:10.1073/pnas.0403255101

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., . . . Apweiler, R. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Res, 32*(Database issue), D452-455. doi:10.1093/nar/gkh052

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A, 106*(23), 9362-9367. doi:10.1073/pnas.0903103106

Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat Methods, 10*(11), 1108-1115. doi:10.1038/nmeth.2651

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Scharfe, C. P., Springer, M., Sander, C., & Marks, D. S. (2017). Mutation effects predicted from sequence co-variation. *Nat Biotechnol, 35*(2), 128-135. doi:10.1038/nbt.3769

Hristov, B. H., & Singh, M. (2017). Network-Based Coverage of Mutational Profiles Reveals Cancer Genes. *Cell Syst, 5*(3), 221-229 e224. doi:10.1016/j.cels.2017.09.003

Huang, J. K., Jia, T., Carlin, D. E., & Ideker, T. (2018). pyNBS: a Python implementation for network-based stratification of tumor mutations. *Bioinformatics, 34*(16), 2859-2861. doi:10.1093/bioinformatics/bty186

Jorda, R., Hendrychova, D., Voller, J., Reznickova, E., Gucky, T., & Krystof, V. (2018). How Selective Are Pharmacological Inhibitors of Cell-Cycle-Regulating Cyclin-Dependent Kinases? *J Med Chem, 61*(20), 9105-9120. doi:10.1021/acs.jmedchem.8b00049

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., . . . Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res, 33*(Database issue), D428-432. doi:10.1093/nar/gki072

Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., . . . Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci U S A, 112*(40), E5486-5495. doi:10.1073/pnas.1516373112

Kan, Z., Jaiswal, B. S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H. M., . . . Seshagiri, S. (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature, 466*(7308), 869-873. doi:10.1038/nature09208

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30. doi:10.1093/nar/28.1.27

Kar, G., Gursoy, A., & Keskin, O. (2009). Human cancer protein-protein interaction network: a structural perspective. *PLoS Comput Biol, 5*(12), e1000601. doi:10.1371/journal.pcbi.1000601

Kim, Y. A., Wuchty, S., & Przytycka, T. M. (2011). Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput Biol, 7*(3), e1001095. doi:10.1371/journal.pcbi.1001095

Kong-Beltran, M., Seshagiri, S., Zha, J., Zhu, W., Bhawe, K., Mendoza, N., . . . Yauch, R. (2006). Somatic mutations lead to an oncogenic deletion of met in lung cancer. *Cancer Res, 66*(1), 283-289. doi:10.1158/0008-5472.CAN-05-2749

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., . . . Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res, 44*(D1), D862-868. doi:10.1093/nar/gkv1222

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., . . . Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature, 499*(7457), 214-218. doi:10.1038/nature12213

Lee, J. O., Yang, H., Georgescu, M. M., Di Cristofano, A., Maehama, T., Shi, Y., . . . Pavletich, N. P. (1999). Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. *Cell, 99*(3), 323-334. doi:10.1016/s0092-8674(00)81663-3

Li, Y., Li, L., Wu, Z., Wang, L., Wu, Y., Li, D., . . . Wang, D. (2017). Silencing of ATM expression by siRNA technique contributes to glioma stem cell radiosensitivity in vitro and in vivo. *Oncol Rep, 38*(1), 325-335. doi:10.3892/or.2017.5665

Liu, S., Knapp, S., & Ahmed, A. A. (2014). The structural basis of PI3K cancer mutations: from mechanism to therapy. *Cancer Res, 74*(3), 641-646. doi:10.1158/0008-5472.CAN-13-2319

Malhotra, S., Alsulami, A. F., Heiyun, Y., Ochoa, B. M., Jubb, H., Forbes, S., & Blundell, T. L. (2019). Understanding the impacts of missense mutations on structures and functions of human cancer-related genes: A preliminary computational analysis of the COSMIC Cancer Gene Census. *PLoS One, 14*(7), e0219935. doi:10.1371/journal.pone.0219935

Meyer, M. J., Beltran, J. F., Liang, S., Fragoza, R., Rumack, A., Liang, J., . . . Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. *Nat Methods, 15*(2), 107-114. doi:10.1038/nmeth.4540

Meyer, M. J., Lapcevic, R., Romero, A. E., Yoon, M., Das, J., Beltran, J. F., . . . Yu, H. (2016). mutation3D: Cancer Gene Prediction Through Atomic Clustering of Coding Variants in the Structural Proteome. *Hum Mutat, 37*(5), 447-456. doi:10.1002/humu.22963

Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res, 5*, 189. doi:10.12688/f1000research.7931.1

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning, 52*(1), 91-118. doi:10.1023/a:1023949509487

Mosca, R., Ceol, A., & Aloy, P. (2013). Interactome3D: adding structural details to protein networks. *Nat Methods, 10*(1), 47-53. doi:10.1038/nmeth.2289

Mosca, R., Ceol, A., Stein, A., Olivella, R., & Aloy, P. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res, 42*(Database issue), D374-379. doi:10.1093/nar/gkt887

Nishi, H., Tyagi, M., Teng, S., Shoemaker, B. A., Hashimoto, K., Alexov, E., . . . Panchenko, A. R. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. *PLoS One, 8*(6), e66273. doi:10.1371/journal.pone.0066273

Niu, B., Scott, A. D., Sengupta, S., Bailey, M. H., Batra, P., Ning, J., . . . Ding, L. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet, 48*(8), 827-837. doi:10.1038/ng.3586

Nussinov, R., Jang, H., Tsai, C. J., & Cheng, F. (2019). Precision medicine review: rare driver mutations and their biophysical classification. *Biophys Rev, 11*(1), 5-19. doi:10.1007/s12551-018-0496-2

Ozdemir, E. S., Gursoy, A., & Keskin, O. (2018). Analysis of single amino acid variations in singlet hot spots of protein-protein interfaces. *Bioinformatics, 34*(17), i795-i801. doi:10.1093/bioinformatics/bty569

Ozdemir, E. S., Halakou, F., Nussinov, R., Gursoy, A., & Keskin, O. (2019). Methods for Discovering and Targeting Druggable Protein-Protein Interfaces and Their Application to Repurposing. *Methods Mol Biol, 1903*, 1-21. doi:10.1007/978-1-4939-8955-3_1

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics, 5*(2), 111-126. doi:10.1002/env.3170050203

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., . . . Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics, 21*(6), 832-834. doi:10.1093/bioinformatics/bti115

Papadopoulos, N., & Lennartsson, J. (2018). The PDGF/PDGFR pathway as a drug target. *Mol Aspects Med, 62*, 75-88. doi:10.1016/j.mam.2017.11.007

Pawson, T., & Warner, N. (2007). Oncogenic re-wiring of cellular signaling pathways. *Oncogene, 26*(9), 1268-1275. doi:10.1038/sj.onc.1210255

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., . . . Pandey, A. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res, 13*(10), 2363-2371. doi:10.1101/gr.1680803

Petitjean, A., Mathe, E., Kato, S., Ishioka, C., Tavtigian, S. V., Hainaut, P., & Olivier, M. (2007). Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Hum Mutat, 28*(6), 622-629. doi:10.1002/humu.20495

Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., . . . Sali, A. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res, 39*(Database issue), D465-474. doi:10.1093/nar/gkq1091

Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., & Godzik, A. (2015). A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput Biol, 11*(10), e1004518. doi:10.1371/journal.pcbi.1004518

Porta-Pardo, E., Kamburov, A., Tamborero, D., Pons, T., Grases, D., Valencia, A., . . . Godzik, A. (2017). Comparison of algorithms for the detection of cancer drivers at subgene resolution. *Nat Methods, 14*(8), 782-788. doi:10.1038/nmeth.4364

Raimondi, F., Singh, G., Betts, M. J., Apic, G., Vukotic, R., Andreone, P., . . . Russell, R. B. (2016). Insights into cancer severity from biomolecular interaction mechanisms. *Sci Rep, 6*, 34490. doi:10.1038/srep34490

Rainey, M. D., Charlton, M. E., Stanton, R. V., & Kastan, M. B. (2008). Transient inhibition of ATM kinase is sufficient to enhance cellular sensitivity to ionizing radiation. *Cancer Res, 68*(18), 7466-7474. doi:10.1158/0008-5472.CAN-08-0763

Repana, D., Nulsen, J., Dressler, L., Bortolomeazzi, M., Venkata, S. K., Tourna, A., . . . Ciccarelli, F. D. (2019). The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol, 20*(1), 1. doi:10.1186/s13059-018-1612-0

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res, 39*(17), e118. doi:10.1093/nar/gkr407

Ries, C. H., Cannarile, M. A., Hoves, S., Benz, J., Wartha, K., Runza, V., . . . Ruttinger, D. (2014). Targeting tumor-associated macrophages with anti-CSF-1R antibody reveals a strategy for cancer therapy. *Cancer Cell, 25*(6), 846-859. doi:10.1016/j.ccr.2014.05.016

Roskoski, R., Jr. (2015). Src protein-tyrosine kinase structure, mechanism, and small molecule inhibitors. *Pharmacol Res, 94*, 9-25. doi:10.1016/j.phrs.2015.01.003

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., . . . Mewes, H. W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res, 36*(Database issue), D646-650. doi:10.1093/nar/gkm936

Ryslik, G. A., Cheng, Y., Cheung, K. H., Bjornson, R. D., Zelterman, D., Modis, Y., & Zhao, H. (2014). A spatial simulation approach to account for protein structure when identifying non-random somatic mutations. *BMC Bioinformatics, 15*, 231. doi:10.1186/1471-2105-15-231

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., . . . Vidal, M. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell, 161*(3), 647-660. doi:10.1016/j.cell.2015.04.013

Sahni, N., Yi, S., Zhong, Q., Jailkhani, N., Charloteaux, B., Cusick, M. E., & Vidal, M. (2013). Edgotype: a fundamental link between genotype and phenotype. *Curr Opin Genet Dev, 23*(6), 649-657. doi:10.1016/j.gde.2013.11.002

Sallinen, S. L., Ikonen, T., Haapasalo, H., & Schleutker, J. (2005). CHEK2 mutations in primary glioblastomas. *J Neurooncol, 74*(1), 93-95. doi:10.1007/s11060-005-5953-7

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res, 32*(Database issue), D449-451. doi:10.1093/nar/gkh086

Slenter, D. N., Kutmon, M., Hanspers, K., Riutta, A., Windsor, J., Nunes, N., . . . Willighagen, E. L. (2018). WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res, 46*(D1), D661-D667. doi:10.1093/nar/gkx1064

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer, 18*(11), 696-705. doi:10.1038/s41568-018-0060-1

Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res, 34*(Database issue), D535-539. doi:10.1093/nar/gkj109

Stehr, H., Jang, S. H., Duarte, J. M., Wierling, C., Lehrach, H., Lappe, M., & Lange, B. M. (2011). The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. *Mol Cancer, 10*, 54. doi:10.1186/1476-4598-10-54

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature, 458*(7239), 719-724. doi:10.1038/nature07943

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A, 102*(43), 15545-15550. doi:10.1073/pnas.0506580102

Sun, M., Hillmann, P., Hofmann, B. T., Hart, J. R., & Vogt, P. K. (2010). Cancer-derived mutations in the regulatory subunit p85alpha of phosphoinositide 3-kinase function through the catalytic subunit p110alpha. *Proc Natl Acad Sci U S A, 107*(35), 15547-15552. doi:10.1073/pnas.1009652107

T, P. A., M, S. S., Jose, A., Chandran, L., & Zachariah, S. M. (2009). Pharmacogenomics: the right drug to the right person. *J Clin Med Res, 1*(4), 191-194. doi:10.4021/jocmr2009.08.1255

Talbot, S. J., & Crawford, D. H. (2004). Viruses and tumours--an update. *Eur J Cancer, 40*(13), 1998-2005. doi:10.1016/j.ejca.2003.11.039

Tamborero, D., Rubio-Perez, C., Deu-Pons, J., Schroeder, M. P., Vivancos, A., Rovira, A., . . . Lopez-Bigas, N. (2018). Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med, 10*(1), 25. doi:10.1186/s13073-018-0531-8

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., . . . Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res, 47*(D1), D941-D947. doi:10.1093/nar/gky1015

Thorpe, L. M., Spangle, J. M., Ohlson, C. E., Cheng, H., Roberts, T. M., Cantley, L. C., & Zhao, J. J. (2017). PI3K-p110alpha mediates the oncogenic activity induced by loss of the novel tumor suppressor PI3K-p85alpha. *Proc Natl Acad Sci U S A, 114*(27), 7095-7100. doi:10.1073/pnas.1704706114

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D. M., Kim, R., Ryan, M., . . . Karchin, R. (2016). Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. *Cancer Res, 76*(13), 3719-3731. doi:10.1158/0008-5472.CAN-15-3190

Tomczak, K., Czerwinska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn), 19*(1A), A68-77. doi:10.5114/wo.2014.47136

Tuncbag, N., Gosline, S. J., Kedaigle, A., Soltis, A. R., Gitter, A., & Fraenkel, E. (2016). Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLoS Comput Biol, 12*(4), e1004879. doi:10.1371/journal.pcbi.1004879

Tuncbag, N., Gursoy, A., Nussinov, R., & Keskin, O. (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat Protoc, 6*(9), 1341-1354. doi:10.1038/nprot.2011.367

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., . . . Wodak, S. J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford), 2010*, baq023. doi:10.1093/database/baq023

UniProt, C. (2008). The universal protein resource (UniProt). *Nucleic Acids Res, 36*(Database issue), D190-195. doi:10.1093/nar/gkm895

van der Meer, D., Barthorpe, S., Yang, W., Lightfoot, H., Hall, C., Gilbert, J., . . . Garnett, M. J. (2019). Cell Model Passports-a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res, 47*(D1), D923-D929. doi:10.1093/nar/gky872

Vandin, F., Upfal, E., & Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol, 18*(3), 507-522. doi:10.1089/cmb.2010.0265

Vassilev, L. T. (2006). Cell cycle synchronization at the G2/M phase border by reversible inhibition of CDK1. *Cell Cycle, 5*(22), 2555-2556. doi:10.4161/cc.5.22.3463

Walboomers, J. M., Jacobs, M. V., Manos, M. M., Bosch, F. X., Kummer, J. A., Shah, K. V., . . . Munoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol, 189*(1), 12-19. doi:10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F

Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. (2017). WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res, 45*(W1), W130-W137. doi:10.1093/nar/gkx356

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., . . . Verhaak, R. G. W. (2017). Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell, 32*(1), 42-56 e46. doi:10.1016/j.ccell.2017.06.003

Weir, B., Zhao, X., & Meyerson, M. (2004). Somatic alterations in the human cancer genome. *Cancer Cell, 6*(5), 433-438. doi:10.1016/j.ccr.2004.11.004

Wong, W. C., Kim, D., Carter, H., Diekhans, M., Ryan, M. C., & Karchin, R. (2011). CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics, 27*(15), 2147-2148. doi:10.1093/bioinformatics/btr357

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., . . . Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science, 318*(5853), 1108-1113. doi:10.1126/science.1145720

Wu, H., Dong, J., & Wei, J. (2018). Network-based method for detecting dysregulated pathways in glioblastoma cancer. *IET Syst Biol, 12*(1), 39-44. doi:10.1049/iet-syb.2017.0033

Xi, J., Li, A., & Wang, M. (2017). A novel network regularized matrix decomposition method to detect mutated cancer genes in tumour samples with inter-patient heterogeneity. *Sci Rep, 7*(1), 2855. doi:10.1038/s41598-017-03141-w

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., . . . Garnett, M. J. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res, 41*(Database issue), D955-961. doi:10.1093/nar/gks1111

Zabludoff, S. D., Deng, C., Grondine, M. R., Sheehy, A. M., Ashwell, S., Caleb, B. L., . . . White, A. M. (2008). AZD7762, a novel checkpoint kinase inhibitor, drives checkpoint abrogation and potentiates DNA-targeted therapies. *Mol Cancer Ther, 7*(9), 2955-2966. doi:10.1158/1535-7163.MCT-08-0492

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., . . . Kasprzyk, A. (2011). International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database (Oxford), 2011*, bar026. doi:10.1093/database/bar026

Zong, H., Verhaak, R. G., & Canoll, P. (2012). The cellular origin for malignant glioma and prospects for clinical advancements. *Expert Rev Mol Diagn, 12*(4), 383-394. doi:10.1586/erm.12.30