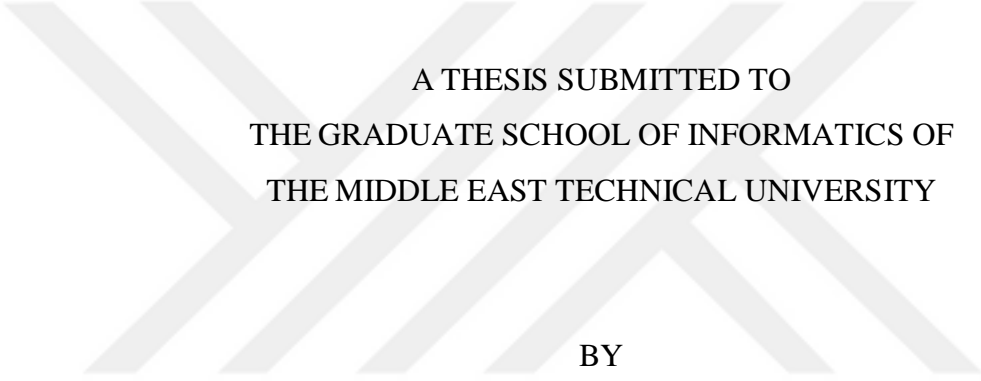


MODELING THE TUMOR SPECIFIC NETWORK REWIRING BY INTEGRATING
ALTERNATIVE SPLICING EVENTS WITH STRUCTURAL INTERACTOME



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

BY

HABİBE CANSU DEMİREL

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

JULY 2019

Approval of the thesis:

**MODELING THE TUMOR SPECIFIC NETWORK REWIRING BY INTEGRATING
ALTERNATIVE SPLICING EVENTS WITH STRUCTURAL INTERACTOME**

Submitted by HABİBE CANSU DEMİREL in partial fulfillment of the requirements for the degree
of **Master of Science in the Bioinformatics Program, 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 Dept., METU

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

Assist. Prof. Dr. Aybar Can Acar
Health Informatics Dept., METU

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

Assist. Prof. Dr. Ceren Sucularlı
Graduate School of Health Sciences, Hacettepe University

Date: 31.07.2019



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 : HABİBE CANSU DEMİREL

Signature : _____

ABSTRACT

MODELING THE TUMOR SPECIFIC NETWORK REWIRING BY INTEGRATING ALTERNATIVE SPLICING EVENTS WITH STRUCTURAL INTERACTOME

Demirel, Habibe Cansu

MSc., Department of Bioinformatics

Supervisor: Assoc. Prof. Dr. Nurcan Tunçbağ

July 2019, 54 pages

Alternative splicing is a post-transcriptional regulation which is important for the diversity of the proteome and eventually the interactome. It enables the production of multiple proteins from a single gene with different structures. In a network point of view, these structural changes can introduce new interactions or cause the loss of the existing ones. The variations in this mechanism has been associated with various diseases including cancer. In this study, we reconstructed patient specific networks with tumor specific protein isoforms by integrating the protein structures and the interaction losses they bring with. For this purpose, we collected 400 breast cancer tumors and 112 normal RNA-seq data from the Cancer Genome Atlas (TCGA) and found the transcripts that show increased expression patterns in tumor cells. We mapped these transcripts to their available protein isoforms found in UniProt. Additionally, we compiled a structural human interactome from multiple sources and aligned the missing residues on isoforms with the known/predicted protein interfaces to find potential interaction losses. At the end, we constructed two interactomes for each sample; one filtered based on the lost interfaces as a result of predominant isoforms (called “terminal set”) and one filtered based on the expression. Then, we used the same terminal set with Omics Integrator to model two sets of networks based on the two patient-specific interactomes. Finally, we compared the resulting two networks and all tumor specific networks simultaneously to reveal pathway, protein-protein interaction and protein patterns that can cluster the tumors according to their similarities. The results of our analysis will contribute to the elucidation of tumor mechanisms and will help for target selection and developing therapeutic strategies.

Keywords: Alternative Splicing, Network Modelling, Multi-omics data

ÖZ

TÜMÖRE ÖZGÜ ETKİLEŞİM AĞLARININ ALTERNATİF UÇ BİRLEŞTİRME OLAYLARI VE YAPISAL İNTERAKTOM KATKISIYLA MODELLENMESİ

Demirel, Habibe Cansu

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Doç. Dr. Nurcan Tunçbağ

Temmuz 2019, 54 sayfa

Alternatif uç birleştirme proteomun ve sonrasında interaktomun çeşitliliğine katkı sağlayan önemli transkripsiyon sonrası mekanizmalardan biridir. Tek bir genden, farklı yapılara sahip birden çok proteinin üretilmesine imkan verir. Etkileşim ağları açısından bakıldığında ise bu yapısal değişiklikler yeni etkileşimlerin kazanılmasına ya da var olan etkileşimlerin kaybedilmesine yol açabilir. Alternatif uç birleştirme olaylarındaki değişimler kanser de dahil olmak üzere farklı hastalıklarla ilişkilendirilmiştir. Bu çalışmada, tümöre özgü protein izoformlarını ve bu izoformların sebep olduğu etkileşim kayıplarını dahil ederek hastaya özel etkileşim ağları oluşturduk. Bu amaçla, Kanser Genom Atlası'ndan (TCGA) elde ettiğimiz 400 meme kanseri ve 112 sağlıklı doku RNA-seq verisini kullanarak tümör örneklerinde artmış ekspresyon gösteren transkriptleri bulduk. Bu transkriptleri UniProt'ta bulunan izoformlarla eşleştirdik. Ayrıca, birkaç kaynaktan alınan verilerle bir yapısal interaktom oluşturduk ve izoformlarda bulunan eksik bölgeleri bilinen ya da tahmin edilmiş protein ara yüzleriyle karşılaştırarak potansiyel etkileşim kayıplarını çıkardık. Böylece, her bir örnek için, biri baskın izoformların (terminal seti) yol açtığı etkileşim kayıplarına göre diğeryse ekspresyona göre filtrelenmiş iki interaktom elde ettik. Sonrasında, Omics Integrator aracını her iki interaktom için aynı terminal setiyle çalıştırarak her bir örnek için iki farklı etkileşim ağı seti elde ettik. Son olarak çıkan iki farklı etkileşim ağını ve oluşturulan tüm etkileşim ağlarını karşılaştırarak hastaları benzerliklerine göre gruplandırabilecek yolak, etkileşim ve protein düzenlerini açığa çıkardık. Bu çalışmanın sonuçlarının tümör gelişimiyle ilgili mekanizmaları aydınlatması ve ayrıca kanserde hastaya özgü tedavi yöntemleri ve hedef seçimi ile ilgili çalışmalara katkıda bulunması beklenmektedir.

Anahtar Sözcükler: Alternatif Uç Birleştirme, Omik Veri, Etkileşim Ağı Modellemesi



To My Family

ACKNOWLEDGMENTS

First, and above all, I would like to express how grateful I am for being able to study under the supervision of Assoc. Prof. Dr. Nurcan Tunçbağ. She was the best mentor and the best example I could ever imagined. I sincerely thank her for everything from her continuous support and patience to invaluable guidance and encouragement.

I would also like to thank my thesis committee members Prof. Dr. Tolga Can, Assoc. Prof. Dr. Yeşim Aydın Son, Assist. Prof. Dr. Aybar Can Acar and Assist. Prof. Dr. Ceren Sucularlı for their valuable time and useful comments with suggestions to improve this study.

I would like to acknowledge the support from Scientific and Technological Research Council of Turkey (TUBITAK) under the project number 117E192.

I would also like to express how lucky I am to meet the loveliest colleagues one could ever have and to spend my master years in a web of friendship that never lacks love, kindness and laughs. For this reason, I would like to thank Cansu Dinçer, Elif Bozlak, Evrim Fer, Gökçe Senger, Meriç Kınalı and Muazzez Çelebi Çınar. I would also thank Fatma Cankara, Gökçe Abay and Alperen Taciroğlu for their amusing and valuable friendships. Moreover, I would like to thank my precious old friends Bilge Çalışkan, Gülnihal Uçarkuş and Merve Kaya Göker for their eternal love and support which never cease to empower me. Finally, I would also like to thank Nagehan Çağlar for her ever-growing love and strong friendship.

Last but not least, I would like to express my deepest gratitude to my family. My parents Mukaddes Demirel and Gürel Demirel never stopped supporting me even in the hardest and darkest times. I would also thank to my brother Mustafa Berk Demirel and sister Ayşe Halise Demirel as they are the sweetest siblings combined with our two mini-sized family members Zuko and Momo. Nothing would be possible without them.

TABLE OF CONTENTS

| | |
|--|------|
| ABSTRACT | iv |
| ÖZ..... | v |
| DEDICATION | vi |
| ACKNOWLEDGMENTS..... | vii |
| TABLE OF CONTENTS | viii |
| LIST OF TABLES | x |
| LIST OF FIGURES..... | xi |
| LIST OF ABBREVIATIONS | xii |
| CHAPTERS | |
| 1.INTRODUCTION..... | 1 |
| 2.LITERATURE REVIEW..... | 5 |
| 2.1. Alternative Splicing and Disease Relationship | 5 |
| 2.2. RNA-seq Analysis for Alternative Splicing Quantification | 6 |
| 2.3. Protein – Protein Interactions and Interfaces..... | 7 |
| 2.4. Network Modelling | 8 |
| 3.MATERIALS AND METHODS | 11 |
| 3.1. Overview of the Pipeline..... | 11 |
| 3.2. Datasets | 13 |
| 3.2.1. Data from TCGA..... | 13 |
| 3.2.2. Data from UniProt..... | 13 |
| 3.2.3. Data for Interfaces..... | 14 |
| 3.3. Transcriptome Assembly..... | 14 |
| 3.4. Parsing GTF Files..... | 15 |
| 3.5. Mapping the Transcripts to Isoforms with Missing Residues | 16 |
| 3.6. Finding the Potential Isoforms that Cause Interaction Losses | 17 |
| 3.7. Finding the Interaction Losses and Modifying the Patient Interactomes | 18 |

| | | |
|--------|---|----|
| 3.8. | Network Modelling with Omics Integrator | 19 |
| 3.9. | Finding Drug Binding Interfaces and Interaction Losses..... | 21 |
| 3.10. | Finding DNA Binding Interfaces and Interaction Losses | 23 |
| 3.11 | Enrichment Analysis and Clustering | 23 |
| 4. | RESULTS | 25 |
| 4.1. | Network Modeling Analysis | 25 |
| 4.1.1. | Analysis of Interactions Belonging to Networks in which Lost Interactions Removed to Inspect Rewiring..... | 25 |
| 4.1.2. | An Example Merged Network from a Patient Obtained via Modeling | 26 |
| 4.2. | Results as Numbers | 29 |
| 4.3. | Analysis of Terminal Nodes..... | 30 |
| 4.3.1. | The Frequency of Terminals | 30 |
| 4.3.2. | Overrepresentation Enrichment Analysis (ORA) Results of Terminal Sets | 32 |
| 4.4. | The Analysis of Lost Interactions Across All Tumor Samples..... | 34 |
| 4.4.1. | Protein – Protein Interactions..... | 34 |
| 4.4.2. | Protein – Drug Interactions | 35 |
| 4.4.3. | Protein – DNA Interactions..... | 38 |
| 5. | DISCUSSION | 39 |
| | REFERENCES..... | 41 |

LIST OF TABLES

Table 4.1: The counts of transcripts and lost interactions from a subset of 20 samples. .29



LIST OF FIGURES

| | |
|--|----|
| Figure 3.1: Overall representation of the methodology. | 12 |
| Figure 3.2: The node statistics of a set of Omics Integrator runs for a selected patient. . | 19 |
| Figure 4.1: The most commonly rewired interactions. | 26 |
| Figure 4.2: Interaction network filtered by expression and lost interactions. | 27 |
| Figure 4.3: Interaction network filtered by expression. | 28 |
| Figure 4.4: The network obtained by merging the two separate condition networks. | 28 |
| Figure 4.5: Counts of the most common terminal nodes across the samples. | 31 |
| Figure 4.6: The results of the Overrepresentation Enrichment Analysis in the form of a clustered heatmap. | 33 |
| Figure 4.7: A clustered patient sub-group. | 34 |
| Figure 4.8: Most common lost protein-protein interactions. | 35 |
| Figure 4.9: The most commonly lost protein -drug interactions across samples. | 36 |
| Figure 4.10: HDAC2 in complex with Vorinostat. | 37 |
| Figure 4.11: The counts of proteins which lost at least one interaction across 400 tumor samples. | 38 |

LIST OF ABBREVIATIONS

| | |
|----------------|--|
| CPTAC | Clinical Proteomic Tumor Analysis Consortium |
| GDC | Genomic Data Commons |
| GSEA | Gene Set Enrichment Analysis |
| ORA | Overrepresentation Enrichment Analysis |
| PDB | Protein Data Bank |
| RNA-Seq | RNA Sequencing |
| TCGA | The Cancer Genome Atlas |
| WHO | World Health Organization |

CHAPTER 1

1.INTRODUCTION

Cancer is a complex disease. Although the in vitro studies in the previous decades enabled us to have a strong knowledge about the nature of the cancer, many were short of revealing the variable interactions between the cancer cells with its normal environment. Now, we know that the tumors contain different cell types that communicate to create the heterogeneous nature of cancer (Hanahan & Weinberg, 2000). Tumor heterogeneity can be observed in a tumor with the presence of multiple genotypes and phenotypes which is called intra-tumor heterogeneity. Inter-tumor heterogeneity, on the other hand, is observed when the different profiles found in different tumors belonging to the same tissue type (Fisher, Puztai, & Swanton, 2013). According to World Health Organization (WHO), cancer is the second leading cause of death globally in 2018 with an estimated 9.6 million mortalities. Among the cancer types, breast cancer is one of the most common and deadliest cancers. To decrease the severity of these statistics, providing adequate therapies for the cancer patients is crucial. However, the heterogeneity of the tumors makes this task challenging as it necessitates the consideration of the different regions or cell types at the same time. Finding mutual biomarkers or mutual biological pathways that can cluster a subgroup in a given tumor is an approach that can lead promising results (Verhaak et al., 2010). As a comprehensive database that provide cancer data including genomic, epigenomic, proteomic and transcriptomic analysis belonging to 33 tumor types from thousands of patients, TCGA (The Cancer Genome Atlas) is a useful starting point for such approaches (Tomczak, Czerwinska, & Wiznerowicz, 2015) (Kelemen et al., 2013).

Alternative splicing is a process which enables the production of more than one mRNA variants from a single gene. Estimated to occur in more than 90 % of the human genes, it is one of the major contributors to the complexity of human proteome since the mRNA variants often translated into protein isoforms (E. T. Wang et al., 2008) (Pan, Shai, Lee, Frey, & Blencowe, 2008). With the differential inclusion or exclusion of exons, existing protein regions may be lost, or new regions may be gained. As a result, proteins with distinct structures and functions may be produced from a single gene (Nilsen & Graveley, 2010; Romero et al., 2006). In a broader view, the changes that alternative splicing

brings with can impact a wide range of mechanisms including protein interactions, cellular properties, protein localization and enzymatic properties (Kelemen et al., 2013). Over the years, alternative splicing has been associated with various diseases by acting through direct causing, severity modifications and determination of the susceptibility (G. S. Wang & Cooper, 2007). Among these diseases, cancer has been a frequently studied interest which resulted in many examples of its relationship with alternative splicing in terms of proliferation, motility, drug response, metastasis and oncogenesis (Skotheim & Nees, 2007). p53, PTEN, BRCA1 and CDC25 proteins in breast cancer (Okumura, Yoshida, Kitagishi, Nishimura, & Matsuda, 2011; Omenn, Yocum, & Menon, 2010; Tammaro, Raponi, Wilson, & Baralle, 2012) GRB7 protein in ovarian cancer (K. Wang et al., 2010), KLF6 in hepatocellular carcinoma (Hanoun et al., 2010) are among the example proteins that has been associated with cancer through alternative splicing events (Tang et al., 2013). Hence, differences in alternative splicing can be a potential biomarker and play a role in cancer therapy and prognosis (Yi & Tang, 2011).

Protein interaction networks tend to include proteins that are high in abundance, conserved and from particular cellular localizations (von Mering et al., 2002). In addition, most of the time alternative splicing data is not represented in disease networks as only the reference isoform is included in networks (Corominas et al., 2014). However, various studies have indicated the importance of alternative splicing in interactome studies and to have a better understanding of diseases. For example, in a study focusing on autism, researchers constructed disease specific interaction networks by incorporating alternatively spliced variants and revealed that half of the interactions and 30% of the newly discovered interacting proteins were obtained from isoforms (Corominas et al., 2014). In another study, tissue specific alternative splicing events were analyzed, and it is revealed that such events increase the diversity of protein interaction networks in tissues (Buljan et al., 2012). As a result, illuminating the effects of alternative splicing on protein networks is a necessity for disease studies as the changes that isoforms of a gene cause on a network can be as drastic as the difference between proteins from different genes (Yang et al., 2016).

Proteins interact with each other through the interface region. The interface can be defined as the area where two proteins get in contact via non-covalent atomic interactions in their complex state. Currently, only a small portion of the known PPIs has at least one experimentally resolved complex state in PDB (Berman et al., 2000) for which the interface can be calculated using atomic distances. For the rest, computationally efficient and accurate methods are frequently used to model interactions and interfaces including PRISM (Baspinar, Cukuroglu, Nussinov, Keskin, & Gursoy, 2014) (Tuncbag, Gursoy, Nussinov, & Keskin, 2011) and Interactome INSIDER (Meyer et al., 2018) for interface predictions

or Interactome3D (Mosca, Ceol, & Aloy, 2013) for interaction predictions. The interface residues are important since they are responsible for the interactions to happen. The disturbances in these residues can lead to the loss of the interactions as exploited in the drug design studies which target protein-protein interactions.

Since patient-specific analysis require detailed information if interaction networks, integrating different types of omics data with them is very important. Using reverse engineering methods to combine multiple data could allow a wider point of view and enable the discovery of various hidden target proteins. Methods that utilize such methods have been previously used to unravel cancer pathways (S. s. C. Huang & Fraenkel, 2009; Torkamani & Schork, 2009; Yeager-Lotem et al., 2009). As one of these methods, Omics Integrator creates high confidence sub networks from a given interactome and a set of important proteins that are scored according to their significance such as expression values.

Considering these aspects, the effect of alternative splicing on tumor specific interaction networks is yet to be elucidated despite some efforts. Yet, it is clear that using tumor specific protein isoforms to model interactions profiles and interaction networks could be quite important in elucidating disease mechanisms. In this work, we combined expression data with the interaction information coming from structural interactomes to reveal the network rewiring of tumor specific interaction networks after adapting to interaction losses that protein variants cause. For this purpose, we collected 400 breast cancer tumors and 112 normal RNA-seq data from the Cancer Genome Atlas (TCGA) and found the differentially expressed transcripts and found the isoforms at protein level. Additionally, we compiled a structural human interactome from multiple sources and aligned the isoforms with the known/predicted protein structures. At the end, we constructed a tumor-specific interactome for each sample based on the lost interfaces as a result of predominant isoforms. Then, we used the set of proteins coming from the differentially expressed transcripts with Omics Integrator to create two interaction networks for each sample by network modelling. One set of networks were created based on a reference interactome that is filtered by expression values while the other set were modelled using interactomes filtered by the lost interactions in addition to expression. Finally, we compared the networks coming from two conditions for each patient and also compared all tumor specific networks simultaneously to reveal pathway, protein-protein interaction and protein patterns that can cluster the tumors according to their similarities.

In Chapter 2, we present a detailed literature review by starting with alternative splicing events and its relevance with the diversity of proteome and interactome and eventually diseases and disease networks. Then, we review a set of the most commonly used transcriptome analysis tools and continue with protein-protein

interactions, interfaces and prediction methods. At the end of the chapter, we recover the approaches for network modelling and the analysis of the disease networks.

In Chapter 3, we present the steps of our methodology. We start by explaining the data we gathered by including the information about sequencing, isoform and interaction datasets. Then we continue with the methods of transcriptome assembly and the parsing of the resulting files. At the same time, we clarify the method we used to assess protein-protein interaction losses based on missing regions found in protein isoforms caused by alternative splicing events when those regions overlap with interaction interfaces. We finish the chapter by explaining the network construction by incorporating omics data and the analysis of the resulting networks.

In Chapter 4, we explain our results obtained from the study. We present the results of the gene set enrichment analyses to reveal the involvement of the terminal sets in biological pathways that can cluster the patients. Discovered gene sets include pathways related to telomeres, gene silencing and cell cycle. We also include the network analyses to elucidate the rewiring patterns between two sets of networks created for the patients and present the most commonly rewired genes into the networks. The frequencies of the lost interactions between protein - protein, protein – DNA and protein – drug interactions are presented too to discover similarities between samples.

Finally, we conclude the study in Chapter 5 with a short overview of the study and we discuss the results obtained in the previous chapter by comparing them with the findings obtained from literature when available. Potential future plans to carry this study one step further are also discussed including the proteomics analyses of the same samples to support our findings and extending the number of tumors by incorporating other cancer types to reveal similarities and differences between them.

CHAPTER 2

2.LITERATURE REVIEW

2.1. Alternative Splicing and Disease Relationship

Splicing is a process in which introns are removed from a pre-mRNA and the remaining exons are combined. However, variations in this process that cause the production of different mRNAs are very common and they are called alternative splicing. Alternative splicing can occur in many ways through various combinations of the exons and introns. Alternative 3' or 5' splice sites, cassette alternative exons, mutually exclusive exons, alternative promoter usage and intron retention are among the examples (Blencowe, 2006). The number of mRNA variants produced as a result of alternative splicing can be as low as 2 while it can climb as high as thousands as in the extreme example of DSCAM gene from *Drosophila melanogaster* that can produce 38,016 different mRNA isoforms (Nilsen & Graveley, 2010; Schmucker et al., 2000). Estimated to occur in more than 90 % of the human genes, it is one of the major contributors to the complexity of human proteome along with alternative transcription start sites, alternative polyadenylation, RNA editing and post translational modifications (Nilsen & Graveley, 2010; Pan et al., 2008; E. T. Wang et al., 2008). Through the changes it causes in the mRNA sequences, alternative splicing has a potential to change many biological dynamics including protein structures, functions, interactions and localizations. Hence, the relationship between alternative splicing and diseases such as cancer has been a frequently studied interest over the years. For example, neurological diseases are among the mostly associated diseases with alternative splicing. Studies show that brain is the organ that have the highest number of alternatively spliced genes and that these genes are often linked to neurodegenerative disorders (Johnson et al., 2009; Mills & Janitz, 2012; Yeo, Holste, Kreiman, & Burge, 2004). In a recent study, Raj et al. analyzed 450 brain tissue samples from two different age groups to reveal hundreds of aberrant splicing events that were associated with Alzheimer's disease. They also assessed the role of alternative splicing levels and the expression in disease by finding 21 significantly related genes among which some genes were already known to be associated with Alzheimer's disease while some were novel genes in terms of relevance (Raj et al., 2018). Among these diseases, cancer has been a frequently studied interest which resulted in many

examples of its relationship with alternative splicing in terms of proliferation, motility, drug response, metastasis and oncogenesis (Skotheim & Nees, 2007). It is known that some alternatively spliced variants of a gene are only produced in cancer samples and not found in healthy tissues (Kim, Goren, & Ast, 2008). One famous example of the alternative splicing on cancer is SR protein (serine/arginine-rich proteins) SF2/ASF which is a splicing factor whose overexpression has been associated with cancer transformation. SF2/ASF is found as upregulated in some cancer types including lung, colon and breast cancer. The researchers showed that the overexpression of SF2/ASF causes the increased production of isoform-2 mRNA of protein S6 kinase- β 1 (S6K1) which in turn acts as a proto-oncogene and enables transformation in mice. In addition, they used knockdown experiments to illustrate that the transformation effect can be reversed if the isoform-2 or the SF2/ASF were targeted (Cooper, Wan, & Dreyfuss, 2009; Karni et al., 2007). p53, PTEN, BRCA1 and CDC25 proteins in breast cancer (Okumura et al., 2011; Omenn et al., 2010; Tammara et al., 2012), GRB7 protein in ovarian cancer (K. Wang et al., 2010), KLF6 in hepatocellular carcinoma (Hanoun et al., 2010) are among the example proteins that has been associated with cancer through alternative splicing events (Tang et al., 2013). Hence, the changes in alternative splicing patterns can be targeted as potential biomarkers for cancer therapy and prognosis (Yi & Tang, 2011).

One of the most common mechanisms that cause aberrant alternative transcripts is mutations. It is estimated that 15% of the all mutations that cause genetic diseases are associated with mRNA splicing (Srebrow & Kornblihtt, 2006). The point mutations that occur at splice sites can enable the production of different variants depending on its location on the gene. Exon skipping is one of the most frequent type of events that is caused by point mutations in splice sites. The skipping of the exon can cause the early introduction of stop codon due to a frameshift which may lead to a shorter mRNA that generally ends up being degraded due to nonsense mediated decay. If the loss of exon leads to a shorter mRNA without a frameshift, resulting mRNA can be translated to a shorter protein. Genes LKB1 (Hastings et al., 2005), KIT (Chen et al., 2005), CDH17 (X. Q. Wang et al., 2005), KLF6 (Narla et al., 2005) and BRCA1 (Pettigrew et al., 2005) are among the example genes which are associated with cancer due to presence of mutations in their splice sites (Anna & Monika, 2018).

2.2. RNA-seq Analysis for Alternative Splicing Quantification

With the common use of RNA-seq and the vast amount of data it provides, the analysis of RNA-seq data using computational methods has gained popularity along with the tools that inspect alternative splicing. Currently, there are two

main approaches that are utilized to study alternative splicing using RNA-seq data. The first set of methods are based on the quantification of mRNA isoforms from sequence reads with the help of sequence alignment. The second set of methods, on the other hand, depend on the assessment of alternative splicing data from the quantification of events (Park, Pan, Zhang, Lin, & Xing, 2018) such as MISO (Katz, Wang, Airoidi, & Burge, 2010), SpliceTrap (Wu et al., 2011) and rMATS (Shen et al., 2014). Both group of methods have its own flows and advantages as they can be utilized depending on the aim of the study. For example, methods based on transcriptome assembly falls short when the input transcriptome is insufficient as it highly depends on the present information (Conesa et al., 2016). Yet, tools that can use a reference or enable the discovery of novel products at the same time are also available including Cufflinks (Trapnell et al., 2012) and StringTie (Pertea et al., 2015). StringTie is a fast and efficient transcriptome assembler which claims to give better results than its most commonly used counterparts including Cufflinks on both real and artificial data. It assembles the transcripts and produces the expression results for each transcript in terms of different quantification statistics using normalization methods including FPKM and TPM.

2.3. Protein – Protein Interactions and Interfaces

Proteins are not functional on their own, rather they interact with each other. Almost 80% of the proteome interact to contribute in a process or a pathway or to form a cellular machinery (Berggard, Linse, & James, 2007). Some protein interactions are malfunctioning and pathogenic that may be associated with diseases such as cancer, neurodegenerative diseases. Proteins interact with each other through the interface region. The interface can be defined as the area where two proteins get in contact via non-covalent atomic interactions in their complex state. In the sequence-structure-function cascade, structure is more conserved than the sequence. Not all residues in the binding site contribute equally to the binding free energy. Rather, some residues contribute significantly more to the binding free energy, called “hot spots” and provide targets for the drug design (Bogan & Thorn, 1998; Clackson & Wells, 1995).

The interface region can be determined using different approaches such as calculation of the atomic distances, identification of the buried surface area after the complex formation and the Voronoi diagrams (Janin, Bahadur, & Chakrabarti, 2008). Currently, only a small portion of the known PPIs has at least one experimentally resolved complex state in PDB. For the rest, computationally efficient and accurate methods are frequently used. These predictive methods can be classified as blind docking, knowledge-based

modeling, evolutionary coupling (Hopf et al., 2014). In the knowledge-based modeling, usually a template is used to find the structural model of a PPI which limits the search space to predict the structural orientation of the proteins in their complex state. PRISM (Baspinar et al., 2014; Tuncbag et al., 2011) is a knowledge-based approach which depends on the assumption that if the complementary parts of an interface structurally match to the regions on the surfaces of two given target proteins, then these proteins are said to be interacting with each other. In evolutionary coupling approach, co-evolving residues in sequence are assumed to be spatially in close proximity; so that the three-dimensional structures of the protein interactions can be identified with an acceptable accuracy. In blind docking, all possible orientations of the protein pairs are sampled if there is no prior knowledge.

Interactome INSIDER (Meyer et al., 2018) (integrated structural interactome and genomic data browser) is a tool which collects the interaction data from multiple databases and finds the best interface for the specific interaction. If available, co-crystallized proteins from PDB or homology models from Interactome3D (Mosca et al., 2013) are used to calculate the protein-protein interfaces. If not, INSIDER uses a machine learning based approach called ECLAIR to predict the interfaces. For human interactions, it holds a vast amount of data by offering 121,575 high confidence interfaces.

2.4. Network Modelling

Although the recent developments enabled the production of a vast amount of data regarding omics elements from interactomes to transcriptomes, integrating different types of these elements to be able to obtain meaningful results still remains as an important task. To overcome this problem, many approaches that combine these elements have been developed. Approaches that are based on matrix factorization, correlation, Bayesian methods and network related methods are among the commonly used bases for the integration of multi-omics data (S. Huang, Chaudhary, & Garmire, 2017). Among these methods, network models enable the analysis of the high-throughput data from a systems perspective. By integrating molecule level information with the interaction networks, network modelling provides a focused point of view in which connection between important hits, hidden contributors or specific pathways could be revealed (Kedaigle & Fraenkel, 2018). Omics Integrator (Tuncbag et al., 2016) is a software package which is composed of two tools. While Garnet tool identifies transcription factors associated with expression changes by incorporating epigenetic changes nearby expressed genes, Forest tool solves the prize-collecting Steiner forest problem to create subnetworks by focusing on the omics hits or terminal nodes provided by the user. By using the terminal list with

an interactome, Forest tries to create reliable subnetworks that connect the elements in the terminal list with the help of the prizes given to the terminals and the weights in the interactome. The main idea behind the algorithm is including as many as possible terminals and to connect these terminals with additional proteins while avoiding the unreliable interactions as possible.





CHAPTER 3

3.MATERIALS AND METHODS

In this chapter, we explain the methods and materials that we used to find predominant isoforms that cause interaction losses, to integrate them with omics data to reconstruct interaction networks.

3.1. Overview of the Pipeline

In this study, we reconstructed patient specific signaling networks with tumor specific protein isoforms by integrating their 3D structures and altered interactions to reveal how the tumor networks are rewired. This section is dedicated to summarizing the overall pipeline. First of all, we retrieved processed RNA-seq data of 400 breast cancer tumors and 112 normal samples from The Cancer Genome Atlas (TCGA) program of NCI Genomic Data Commons (GDC) (Grossman et al., 2016). After the detection of expressed transcripts using StringTie (Pertea et al., 2015), we calculated the log fold changes between tumor and an average of pooled normal samples for each patient. At the same time, we mapped the transcripts to protein isoforms with missing regions to detect the lost regions differing from the canonical protein and collected structural interactome from multiple sources (PDB, Interactome3D (Mosca et al., 2013), Interactome Insider (Meyer et al., 2018), PRISM (Baspinar et al., 2014; Tuncbag et al., 2011)). If a missing region in an isoform involves a known protein-protein interaction interface, the interaction is accepted as lost for the interactome of the samples where that isoform is present without its canonical counterpart. The same analysis was also performed on protein – drug and protein – DNA interactions. After finding the lost protein-protein interactions, we ended up with two interactomes for each tumor sample. The first one was filtered based on the expression levels. For the second one, if a missing region in an isoform involves a known protein-protein interaction interface, the interaction is accepted as lost for the interactome of the samples where that isoform is present without its canonical counterpart. In this way, we prepared tumor specific interactomes for 400 patients. Then, we used the set of proteins which lost at least one interaction as the set of important targets (called “terminal set”) of each tumor samples. We used Omics Integrator to find the optimal

network that represents the input data best. Omics Integrator reconstructs signaling networks by including as many of the altered proteins as possible and by keeping the network small since it avoids using unreliable protein-protein interactions. In our study, it integrates the terminal sets with the previously generated patient specific interactomes.

In this way, we mapped patient specific structural data onto classical pathways for a rational network analysis and obtained high confidence subnetworks which connect the isoforms by also integrating the expression data. Finally, we compared the two networks created for each sample to reveal pathway, protein-protein interaction and protein patterns that can cluster the tumors according to their similarities.

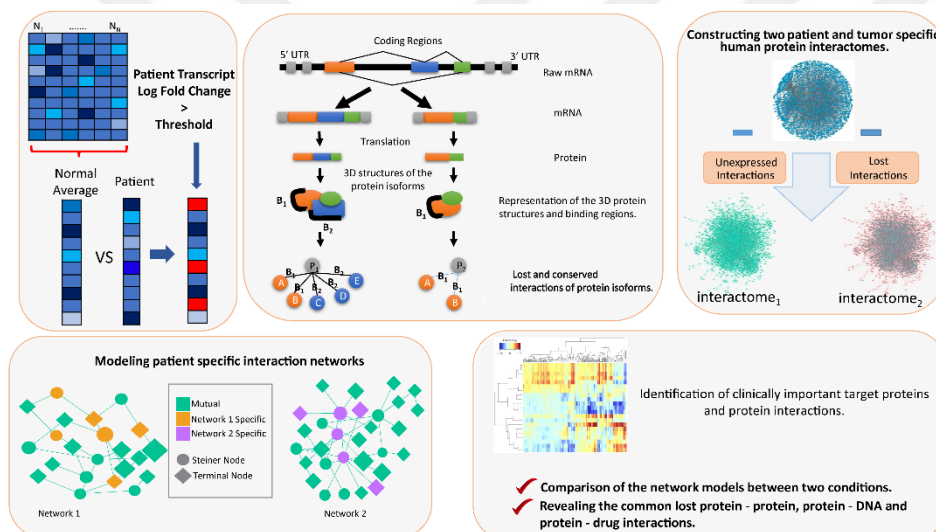


Figure 3.1: Overall representation of the methodology.

3.2. Datasets

3.2.1. Data from TCGA

The Cancer Genome Atlas (TCGA) is a public funded collaboration project which resulted in a database covering the data of more than 30 cancer types. A variety of data types are available in the atlas including raw high-throughput sequencing, gene and exon expression, DNA methylation, mutations and structural variations. TCGA data is available on The NCI's Genomic Data Commons (GDC) repository along with some other major programs (Grossman et al., 2016). For our study, we downloaded 400 RNA-Seq BAM (binary alignment map) files belonging to 400 breast cancer tumor samples along with 112 RNA-seq BAM files belonging to normal tissues of some of the same patients. All samples were coming from female patients and the same study of TCGA-BRCA. The BAM files were created by GDC using the STAR (Dobin et al., 2013) two-pass alignment procedure with GRCh38 reference genome after a quality control step. The final quality assessments were completed using Picard Tools. Since the aligned reads were under the controlled access data type, we applied for an authorization to access the data through NIH (National Institutes of Health) database of Genotypes and Phenotypes (dbGaP) and obtained a "General Research Use" access.

3.2.2. Data from UniProt

The Universal Protein Resource (UniProt) is a protein sequence and annotation database which holds almost 120 million protein entries belonging to different organisms. One of the three main databases that UniProt consists of is UniProt Knowledgebase (UniProtKB) and it has two arms. While the entries under UniProtKB/TrEMBL group are unreviewed and automatically annotated, the entries belonging to UniProtKB/Swiss-Prot group are curated and reviewed; hence, more credible. Second database is UniProt reference Clusters (UniREF) which includes clustered sets of sequences. The last of the three databases is The UniProt Archive (UniParc) that holds the protein sequences. Altogether with its core databases, UniProt provides a large amount of information about proteins including sequences, functions, interactions, locations and variants with many more others (UniProt, 2019). Apart from separate entries, UniProt also provides proteomes to enable the accession of all proteins belonging to one species. The results table obtained after a selection can be customized to include different characteristics of the selected proteins at the same time. For our analysis, we selected the human proteome and limited the proteins with reviewed status (Swiss-Prot) to obtain a reliable dataset which included 20404 entries. Then, we

added multiple columns to the resulting table. The selected columns were “Ensembl”, “Gene names (primary)” and “Gene names (synonym)”, “Alternative products (isoforms)” and “Alternative sequence”. This way, we obtained a dataset which includes all gene names and all Ensembl transcript IDs that map to a reviewed protein. In addition, if there are isoforms available for a protein, their IDs, isoform names, the sequence changes with respect to the canonical isoform and the transcript names that match each isoform are also included in the dataset. This information is crucial for our study since we need to map the Ensembl transcript IDs found in the transcriptome files obtained from StringTie to protein isoforms in UniProt.

3.2.3. *Data for Interfaces*

We collected the interface data of protein-protein interactions from Interactome INSIDER (integrated structural interactome and genomic data browser)(Meyer et al., 2018) (Interactome INSIDER is a tool which collects the interaction data from multiple databases and finds the best interface for the specific interaction. If available, co-crystallized proteins from PDB or homology models from Interactome3D (Mosca et al., 2013) are used to calculate the protein-protein interfaces. If not, INSIDER uses a machine learning based approach called ECLAIR to predict the interfaces. We only included the 121,575 high confidence interfaces in our study. In addition to INSIDER, we also integrated the predicted interface residues obtained from PRISM (Protein Interactions by Structural Matching) (Baspinar et al., 2014; Tuncbag et al., 2011). PRISM predicts the protein-protein interactions with interfaces by using a knowledge-based algorithm in which it tries to match the regions of two input proteins from PDB with the opposite sides of a known interface. In the cases where the same interaction is present in both INSIDER and in PRISM, we merged the interface residues. From these two sources, we collected 123,182 protein-protein interfaces as a list of UniProt residues for each interactor in each interaction. Only the interfaces which are longer than 5 residues are considered in the analysis

3.3. **Transcriptome Assembly**

Having 400 tumor samples to be compared with 112 normal samples came with some disadvantages since the lack of enough normal samples combined with the lack of replicates made a standard differential expression analysis almost impossible. As the files were already aligned, the next step in the downstream analysis was to assemble the transcriptomes for each patient. Starting with aligned files prevented us from using some powerful transcriptome assemblers

including Kallisto (Bray, Pimentel, Melsted, & Pachter, 2016) that require the input files to be raw sequence file. Although the gene expression values and counts were already calculated using HTSeq in GDC as a part of the mRNA analysis pipeline, we could not use them since we needed transcript counts in order to continue with protein isoforms.(Anders, Pyl, & Huber, 2015) To calculate the transcript expression values, we used StringTie (Pertea et al., 2015). StringTie is a fast and efficient transcriptome assembler which claims to give better results than its most commonly used counterparts including Cufflinks on both real and artificial data. StringTie can provide different outputs according to the needs of the usage. The output options include a list of assembled transcripts where abundances using different normalization methods such as TPM and FPKM are available, gene abundances, input files for downstream analysis and merged GTF files. Although the only mandatory input of StringTie is the alignment file in SAM or BAM format, a reference gene annotation file in GTF/GFF3 format is highly recommended. In our case, it was also required since we needed the official transcript IDs which can be mapped to UniProt protein isoforms. As a reference, we used the comprehensive gene annotation GTF file which only covers the reference chromosomes from Gencode (Harrow et al., 2012) release 22 (GRCh38.p2). This reference file was selected since it was the reference used in RNA-Seq alignments by HTSeq in GDC pipeline and release 22 was the source of index files for the alignments completed with STAR.

We assembled the transcriptomes for each of 512 BAM files using the -e, -G and -p 36 options. -e option enables reference to skip the novel transcripts which do not match with the reference annotation, -G option denotes the usage of a reference annotation and -p option sets the number of threads to be used. The remaining parameters were not specified so they were used as the default values. As a result of these runs, we obtained a Gene Transfer Format (GTF) file for each BAM file.

3.4. Parsing GTF Files

After obtaining the GTF files from StringTie for each patient, the next step was finding the expressed transcripts. Each GTF file contains 9 columns of information which are “seqname”, “source”, “feature”, “start”, “end”, “score”, “strand”, “frame” and “attributes” for each line which denotes an element. Among these, we were firstly interested in “feature” column which denotes the type of the element such as exon, transcript and mRNA. If the “feature” column of a line is “transcript”, we continued with the “attributes” column which holds a variety of information about the element in that line from identification numbers to expression values. In case of transcripts, the available information included “gene_id”, “transcript_id”, “REF_gene_name”, “cov”, “FPKM” and

“TPM”. Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) (Trapnell et al., 2010) and Transcripts Per Million (TPM) (Li, Ruotti, Stewart, Thomson, & Dewey, 2010) are among commonly used RNA abundance measures. FPKM is a slightly modified form of reads per kilobase per million reads (RPKM) (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008) and it is better at estimating the abundance of paired-end reads (Pertea et al., 2015). Although both TPM and FPKM use the same normalization methods for RNA length and for the sequencing depth, the order or normalizations make TPM more comparable between samples as it shows the proportion of the reads mapped to a gene/transcript with respect to total number of reads. On the other hand, FPKM shows proportions that differ between samples which cause it to be inconsistent. Hence, in the following steps, we continued with TPM values.

To find whether a transcript is expressed, after finding the “transcript” keyword in the “feature” column, we checked whether the TPM value of that specific transcript is higher than 0.1. If a transcript’s TPM value is higher than 0.1, we accept the transcript as expressed. This limit is set to eliminate the inclusion of transcripts that are detected due to measurement or biological noises. We calculated the logarithm to the base 2 values of TPM values belonging to the transcripts which surpassed the threshold.

For 112 normal samples, we calculated the average of the logarithm to the base 2 values of TPM values for each transcript. We preferred logarithm values over TPM values to be able to reduce the effects of highly expressed transcripts on the averages. These averages are then used to compare the values coming from tumor samples versus normal samples. For each GTF file belonging to 400 tumor samples, we extracted the list of expressed transcripts with corresponding gene names. Then, we calculated the log-fold-changes by subtracting the average logarithm value of the transcript of the normal samples from the logarithm value of the transcript in the sample.

3.5. Mapping the Transcripts to Isoforms with Missing Residues

To be able to map the transcripts to protein isoforms, we started with the list of transcripts with corresponding gene names and UniProt isoform IDs available if the transcript matches with a known alternative sequence. Isoform IDs are in the format of PDB ID – Number as in the example of P46736-1 and P46736-2. Additionally, each isoform is also given a name such as Isoform 1, Isoform Alpha-2 and Isoform Long. The isoform names do not necessarily match with the isoform IDs as the name of the isoform O75110-1 is “Isoform Long” while the name of the isoform P46100-1 can be “Isoform 4”. For this reason, the numbers or the isoform names do not give the correct information about whether

an isoform is canonical or not. Among the alternative sequences available for a protein, separating the canonical sequence from the others is important since all the sequence changes are stated with respect to the canonical sequence. The canonical isoform is selected by UniProt depending on a couple of criteria which consider the prevalence, similarity to orthologous sequences, the description of the constitution of the isoform and the length. We distinguished the canonical isoform from the remaining sequences by using the order of the isoforms in the data file the first isoform found in the alternative products section is the canonical isoform.

For each protein, the event type that is responsible for the isoforms were also available. The event keywords consisted of “Alternative splicing”, “Alternative promoter usage”, “Alternative initiation” and “Ribosomal frameshifting”. If a protein did not cover the “Alternative splicing” keyword, we excluded the protein isoforms from the further analysis. The missing sequences of an isoform with respect to the canonical isoform is shown in the format “VAR_SEQ Start End Missing (in isoform X)” where Start and End indicates the beginning and the ending of the missing sequence and X denotes the isoform name. There may be multiple missing fragments for an isoform and a missing fragment can be valid for multiple isoforms. Apart from the missing sequences, “VAR_SEQ” subsection can also have the information about the sequence variations in the form of substitutions or insertions, but they are not included in this study. By parsing every line in the data file, we created a mapping of Ensembl transcript IDs to UniProt isoform IDs and gene names. The isoforms both include the canonical sequences and the alternative sequences which have missing residues.

3.6. Finding the Potential Isoforms that Cause Interaction Losses

After creating a mapping of transcripts to isoforms as a reference, we examined the transcripts in our tumor samples to find the list of the isoforms in each patient. To do so, we iterated over each expressed transcript extracted from the GTF file of a tumor sample. The isoforms that belong to the transcripts which comply with the conditions below are selected as the final list of isoforms and the interaction losses that came with the missing residues were calculated for them. The conditions are as follows:

- I. The Ensembl transcript ID of the patient transcript should be found in the transcript IDs obtained from UniProt and map to an isoform which has missing sequences with respect to the canonical isoform.
- II. The log-fold-change value of the transcript should be higher than 1 which corresponds to a 2-fold difference between the original values.

- III. If expressed, the transcript belonging to the canonical isoform should not have a log-fold-change value higher than 1.

With this filtration, for each patient, we obtained the list of potential isoforms which can cause the loss of a protein-protein interactions due to the missing interface residues.

3.7. Finding the Interaction Losses and Modifying the Patient Interactomes

As the reference interactome, we used the iREFIndex (Razick, Magklaras, & Donaldson, 2008)(version 13.0) network which was previously prepared for Omics Integrator (Tuncbag et al., 2016) to only include the gene names as interactors. In this interactome, every interaction is weighted using an MIScore which is a scoring method based on experimental data to assess the reliability of an edge. The scores are used by Omics Integrator program when patient specific interaction networks are generated. Self-interactions are removed from the iREFIndex interactome. Additionally, proteins including UBC, APP, ELAVL1, SUMO2 and CUL3 are excluded due to their nonspecific interactions with high number of proteins. At the same time, proteins that are extremely big in size including TTN, MUC16, SYNE1, NEB, MUC19, CCDC168, FSIP2, OBSCN and GPR98 are also removed from the reference interactome (Hristov & Singh, 2017).

The first step in customizing the interactomes is filtering the interactions based on gene expression. For each tumor sample, we removed the genes and their interactions from the reference interactome where the gene does not have any transcript that has a TPM value higher than 0.1 and is matching to the transcript IDs obtained from UniProt dataset. This filtration is applied so that we only include the genes from protein coding and expressed transcripts in the interactome. The interactomes generated in this step were used as the first set of input interactomes for the Omics Integrator runs.

The second step in customizing the interactomes comes with mapping the missing regions of isoforms to the protein-protein interaction interfaces to remove the lost interactions. For each isoform found for each tumor sample in the previous step, we compared the interface residues belonging to the canonical sequence of the isoform with the residues that isoform is missing. If at least 5 residues from an interface was missing in the isoform, we accepted that interaction as lost in case when the interactor of the isoform is also expressed in the patient. Such interactions are removed from the interactome obtained in the

first step to create a further customized interactome for each tumor sample to be used as the second set of input interactomes for the Omics Integrator runs.

3.8. Network Modelling with Omics Integrator

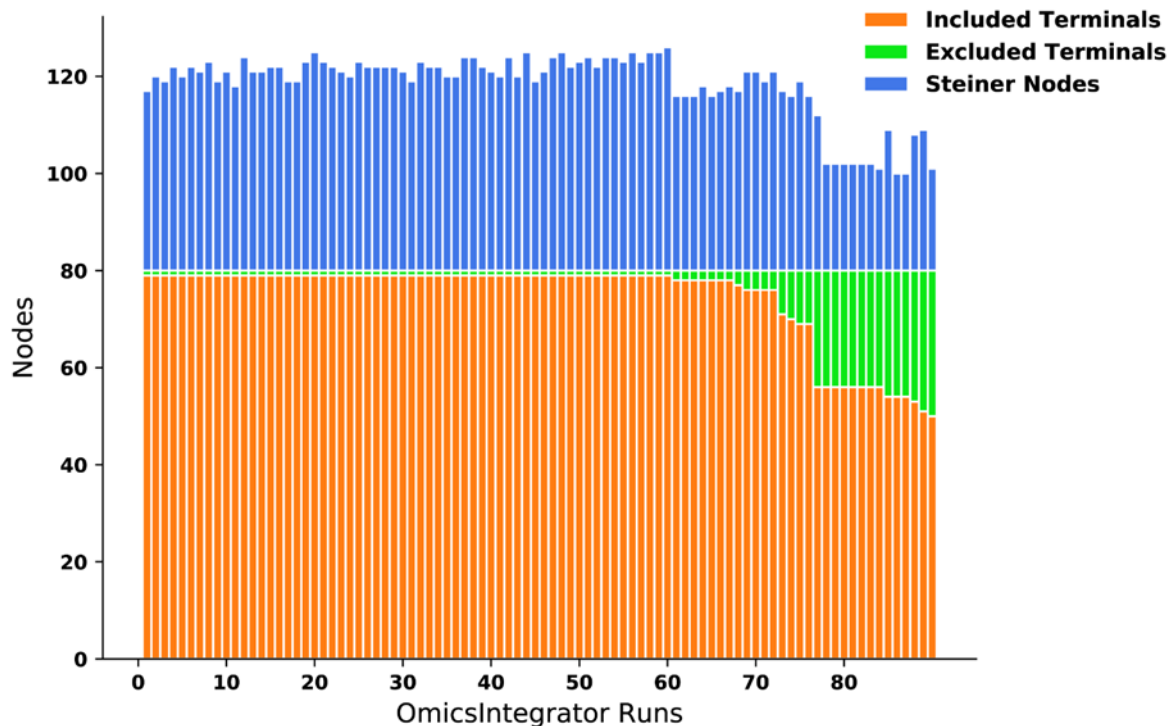


Figure 3.2: The node statistics of a set of Omics Integrator runs for a selected patient. Orange bars represent the number of included terminals in each resulting network while the green ones represent the terminals that are excluded. The blue bars illustrate the number of Steiner nodes that are added to the resulting network by Omics Integrator.

We used the Forest module of Omics Integrator tool to reveal hidden proteins in subnetworks that may have a role in the disease formation and to observe the network rewiring after the removal of additional interactions by integrating the isoforms found in previous steps. After completing many runs for each patient, we ordered the resulting statistics of the runs according to included number of terminals, sum of negative prizes and the number of hubs to get the results where the terminal inclusion was high while the sum of negative prizes and the number of hubs were low. Since the resulting top networks were similar in terms of the elements mentioned in the filters up to a point in both conditions as shown in Figure 3.2 for a selected sample, we merged the networks to be able to extract

an optimal consensus from the results of the runs. We merged the best 50 networks and we included the proteins that are found in at least the half them with the interactions in between. The merging step was completed for both conditions for each tumor sample.

Forest tool solves the prize-collecting Steiner forest problem to create subnetworks by focusing on the omics hits or terminal nodes provided by the user. By using the terminal list with an interactome, Forest tries to create reliable subnetworks that connect the elements in the terminal list with the help of the prizes given to the terminals and the weights in the interactome. The main idea behind the algorithm is including as many as possible terminals and to connect these terminals with additional proteins while avoiding the unreliable interactions as possible. Forest uses the prize function stated below (**Equation 1**) to assign a prize to each node based on the significance value obtained from the user and the number of edges that node has in the interactome. While the significance increases the chance of a node to be included, number of edges decreases this probability as the bias coming from the hub nodes is unwanted in most cases.

$$p'(v) = \beta \cdot p(v) - \mu \cdot \text{degree}(v) \quad (1)$$

In this function, $p(v)$ stands for terminal node prize where v is node and $\text{degree}(v)$ is the number of connections that a node v has in the interactome. As scaling factor β increases, the chance of a terminal node to be included in the final network also increases as the impact of the node prize is boosted. On the other hand, second scaling factor μ decreases the chance of node's involvement as a negative weight is given to the node proportional to its number of connections.

In addition to the number of interactions and prize values, Forest also considers the weights available for each interaction in the interactome by converting them to costs using the function below (**Equation 2**):

$$c(e) = 1 - p(e) \quad (2)$$

where cost $c(e)$ of an edge is inversely correlated with its weight and hence confidence. These two functions are used in the main objective function of the algorithm (**Equation 3**) whom Forest tries to minimize.

$$f'(F) = \sum_{v \in VF} p'(v) + \sum_{e \in EF} c(e) + \omega \quad (3)$$

In the objective function, v denotes each node belonging to node set V and e denotes each edge from the edge set E where the prize and cost functions $p'(v)$ and $c(e)$ are stated above, respectively. k is the number of trees in the final network and ω is a parameter to tune the edge cost between a dummy node and a node in the node set. Dummy nodes are the artificial nodes introduced in the beginning of the run which are utilized to solve the optimization problem and removed in the final network. D is an additional required parameter which controls the depth of the final network by limiting the maximum path length. With all these functions and parameters, the objective function implies that the algorithm has to pay a penalty for each terminal that is not included in the final network while it has to pay the cost of each edge that it decides to add.

In our case, the terminals are the genes belonging to proteins that lose at least one protein-protein interaction due to the presence of an isoform with a missing sequence. As the prizes of the terminals, we used the log-fold-change values calculated for each transcript. If a gene has multiple transcripts that surpassed the log-fold-change limit explained in section 3.6, we took the average of the log-fold-change values. As the results of the algorithm depends on the chosen parameters, the recommended usage of Forest is completing multiple runs with different parameters to obtain the optimal result. For this reason, among the tuning parameters, D is set as 10 as suggested, ω is selected in the interval $[1, 5]$, β is in the interval $[1, 10]$ and μ is set as 0.005 and 0.01. With these parameters, we completed 100 x 2 runs for each of 400 patients using two different interactomes for each 100x runs. One of the interactomes were only filtered according to gene expression results while the other one filtered according to the lost interactions caused by isoforms with missing residues in addition to expression values.

3.9. Finding Drug Binding Interfaces and Interaction Losses

In addition to the losses of protein-protein interactions due to the missing protein regions resulting from alternative splicing, we incorporated drug-protein interaction data to be able to observe potential interaction losses. To do so, we used to information provided by DrugPort(de Beer, Berka, Thornton, & Laskowski, 2014) to select the molecules to be included. DrugPort is a branch of PDBSum (Laskowski, Jablonska, Pravda, Varekova, & Thornton, 2018) that holds the information about drug molecules and nutraceuticals from DrugBank (Wishart et al., 2018; Wishart et al., 2006) with their target proteins and their presence in PDB (Berman et al., 2000). To obtain the drug and nutraceutical IDs available in the DrugPort, we parsed the webpage at https://www.ebi.ac.uk/thornton-srv/databases/drugport/data/appdrugs_pdb.dat. After obtaining the IDs, we used them to reach the molecule specific webpages

created for each molecule in the DrugPort. These pages list many properties of the molecules including generic names, brand names, targets, the presence of the drug, target and the drug-target structures in found in PDB. Then, we analyzed the information webpages of each drug and nutraceutical to obtain the names of the PDB structures that involve a drug or nutraceutical bound with its target protein. Among 586 drugs and 64 nutraceuticals which are found in PDB database, 193 molecules had such structures. 356 human Uniprot entries were in contact with a drug or nutraceutical in 1319 PDB structures.

After finding the names of the molecules, targets and the presence of the drug – protein complex in PDB, we continued with PDBSum to get the list of protein residues where the interaction occurs. PDBSum is a web server which holds the data about different characteristics and the analysis results of the PDB structures in a text-based or visual format including secondary structures, protein domains, protein – ligand interactions and protein – DNA interactions. Since we were interested in protein – drug interactions, we obtained the list of protein - ligand interactions for each PDB structure with the specific drug / nutraceutical that we found in the previous step is bound. If more than one of the same ligands are bound to a protein, all binding residues are considered as a single group. We did not include the interfaces that are shorter than 3 amino acids long.

After obtaining the amino acid residues which are bound to a drug or nutraceutical for each PDB structure, we mapped these PDB residues to their equivalents on the Uniprot protein sequences in order to match them with the missing residues on alternative sequences. To do so, we used a PDB/UniProt Mapping server called “pdbsws” (Martin, 2005) which maps the PDB residues to corresponding residues in UniProt entries by aligning them. For each PDB structure, we modified the following URL to access the mapping directly from our scripts: “<http://www.bioinf.org.uk/cgi-bin/pdbsws/query.pl?plain=1&qtype=pdb&id=1yqv&all=yes>”. The accessed webpage holds the list of residues from PDB for each chain with the corresponding residues in UniProt and their indices in the sequence. After mapping the sequences, we again compared the drug binding residues with the lost regions in the protein isoforms where the log-fold-change value of the transcript belonging to the isoform is higher than 1. In addition, the transcript belonging to the canonical isoform should not have a log-fold-change value higher than 1. As a result, we obtained the lost protein – drug or protein – nutraceutical interactions caused by alternative splicing. Moreover, we also found potentially lost interactions caused when the interactor protein is not expressed in the tumor sample.

3.10. Finding DNA Binding Interfaces and Interaction Losses

Apart from protein – protein and protein – drug interactions, DNA – protein interactions constitute another important part of the interaction possibilities of the proteins. To be able to analyze this type of interactions, we obtained 993 structures from PDB by using “Macromolecule Type” search criteria to select the files in which only proteins and DNA molecules are present without any RNA molecules or DNA/RNA hybrids. All structures belong to Homo sapiens only. Although the residues where protein – DNA interaction occurs were also present in PDBSum similar to protein – drug interactions, the results there were in visual format, thus we were unable to parse the data. To overcome this problem, we used the same program that is used in PDBSum, NUCPLOT (Luscombe, Laskowski, & Thornton, 1997). NUCPLOT is a program that takes PDB files as inputs and produces both visual and text-based results about the residues which are involved in protein – DNA interaction. It calculates the hydrogen bonds, covalent interactions and the van der Waals contacts. In our study, we are only interested in the non-bonded van der Waals contacts. To calculate them, Nucplot uses a simple distance calculation method and accepts the interaction if the distance between an atom from the protein residue and an atom from DNA molecule is smaller than 3.9 Å. We obtained the text-based interaction results for our 993 PDB structures and matched the PDB file residues with corresponding Uniprot entry sequences using pdbsws in order to be able to map the sequence losses to protein – DNA interactions. PDB entries having chains that match more than one Uniprot ID (for example fusion proteins), interactions between hetero atoms and DNA atoms and small peptides have been excluded from the results. After mapping the sequences, we again compared the DNA binding residues with the lost regions in the protein isoforms where the log-fold-change value of the transcript belonging to the isoform is higher than 1. In addition, the transcript belonging to the canonical isoform should not have a log-fold-change value higher than 1. As a result, we obtained the lost protein - DNA interactions caused by alternative splicing. Moreover, we also found potentially lost interactions caused when the interactor protein is not expressed in the tumor sample

3.11 Enrichment Analysis and Clustering

To be able to observe the distribution of the terminal sets in known gene sets, we performed Overrepresentation Enrichment Analysis (ORA) using WebGestaltR (Liao, Wang, Jaehnig, Shi, & Zhang, 2019) for each tumor sample. WebGestaltR is an R package that is implemented from the gene analysis tool kit named WebGestalt. As an input set, we used the gene names that have at least one protein coding transcript whose log-fold-change value is higher than 1 and

missing residues caused at least one interaction loss. At the same time, the log-fold-change value of the transcript coding for the canonical isoform had to be smaller than 1. We used "ORA" as enrichment method, "hsapiens" as organism, 'KEGG' and "Reactome" pathways together with "Geneontology Biological Process" and "Geneontology Molecular Function" as enrichment databases, "mean" as collapse method, 3 as minimum number and "fdr" as significance method.

After completing the ORA runs for 400 tumor samples, we extracted the resulting gene sets in which a subgroup of the terminals is enriched with a value smaller than 0.05. We removed the gene sets that are found in less than 25 patients to be able to obtain a better representation of the gene sets and patients. After the removal of these gene sets, some patient samples ended up with no enriched gene sets, hence, they are removed too. Then we clustered the enriched gene sets using "pheatmap" package of R. (Kolde, 2012)

CHAPTER 4

4. RESULTS

Alternative splicing is one of the processes which are responsible for the diversity of the proteome by enabling the production of multiple mRNAs to potentially code for multiple proteins. The changes and abnormalities in this process have been associated with diseases including cancer. In this work, we integrated the missing residues in protein isoforms with structural interactome information to observe network rewiring patterns of breast cancer tumor samples. We identified common protein – protein, protein – drug and protein - DNA interaction losses across most of the patients which may indicate tumor related patterns. We also created two sets of networks by using Omics Integrator to extract the high confidence subnetworks with interactomes that are personalized for each sample. We then compared the networks from these two sets to identify the proteins that are added into networks to compensate the interaction losses caused by protein isoforms.

4.1. Network Modeling Analysis

4.1.1. Analysis of Interactions Belonging to Networks in which Lost Interactions Removed to Inspect Rewiring

To be able to observe the effects of network rewiring after the removal of interactions lost due to isoforms, we extracted the interactions that are only found in the second set of merged networks. The results are shown in Figure 4.1 where the counts represent the number of patient samples that each interaction is observed. The most frequent interaction that is missing from the networks that are only filtered according to expression without the integration of isoform information is the *LIG1 – RGS2* interaction. *LIG1* is a gene that codes for human DNA ligase 1 protein. Ligases play crucial roles in DNA replication, recombination, and repair. Although there are multiple ligases, DNA ligase 1 stands out with a higher participation rate in ligase activity in proliferating cells than its counterparts (Lindahl & Barnes, 1992). Multiple studies showed that

DNA ligase 1 expression rates are considerably high in malignant tumor samples with respect to normal benign cells where other ligases do not follow this trend (Jessberger et al., 1997; Montecucco et al., 1992; Signoret & David, 1986; Sun et al., 2001). In addition, they illustrated that the inhibition of DNA ligase 1 could even prevent the tumor cell growth (Sun et al., 2001).

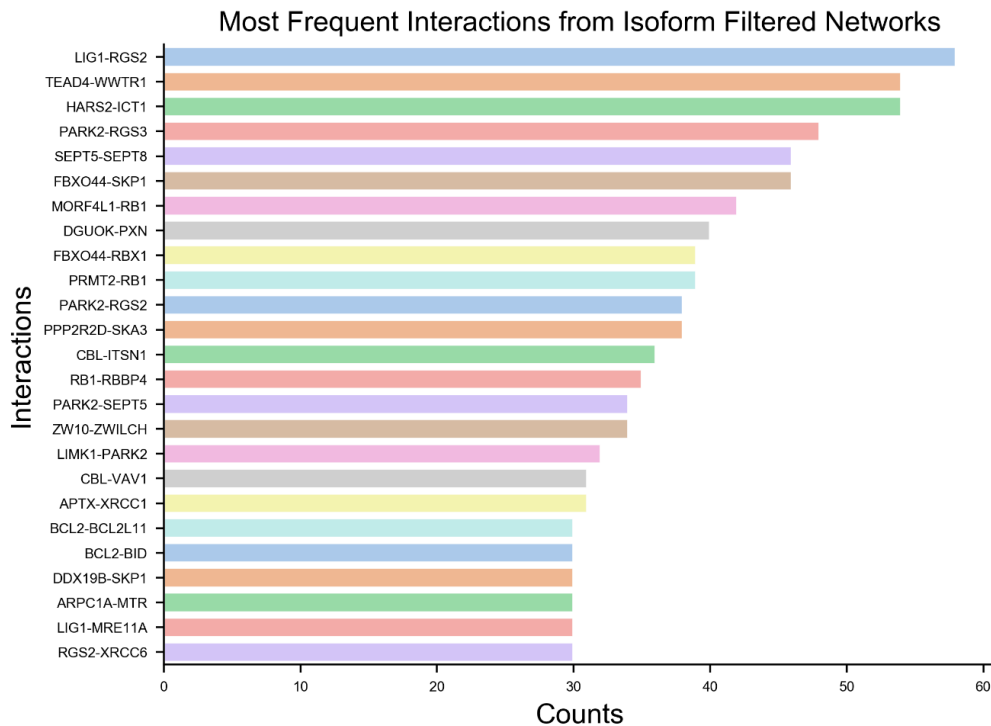


Figure 4.1: The most commonly rewired interactions. These are the counts of the interactions which are added to the network models after we removed the interaction losses from the interactomes.

4.1.2. An Example Merged Network from a Patient Obtained via Modeling

After we completed two sets of Omics Integrator runs for each patient, we merged the resulting gene sets respectively to only include the genes that are robust to changes in parameters. In the merged networks, we only included the interactions where the nodes are observed in at least the half of the runs. Then, we merged the two union interactomes coming from the previous step to discover the nodes that are unique for each network or nodes that are found in both networks.

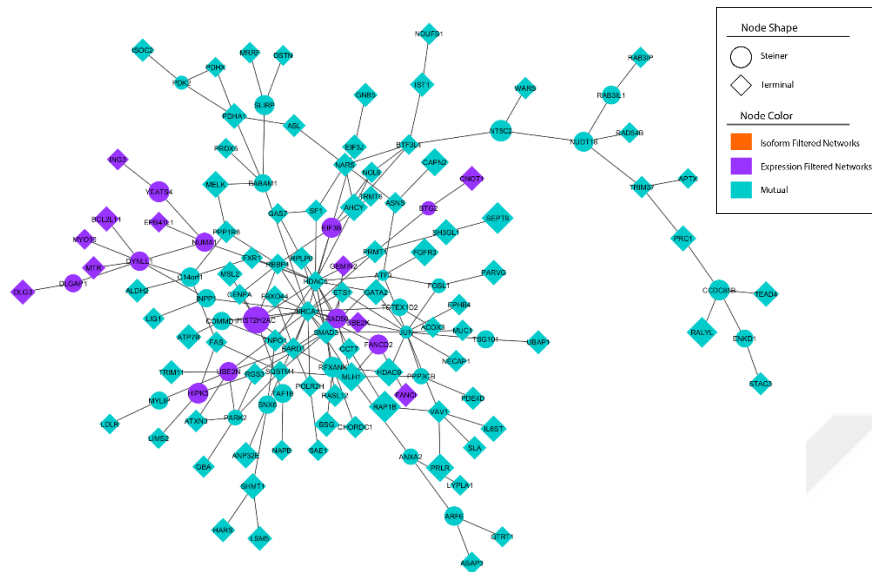


Figure 4.3: Interaction network filtered by expression. The purple color represents the nodes that are specific for this network.

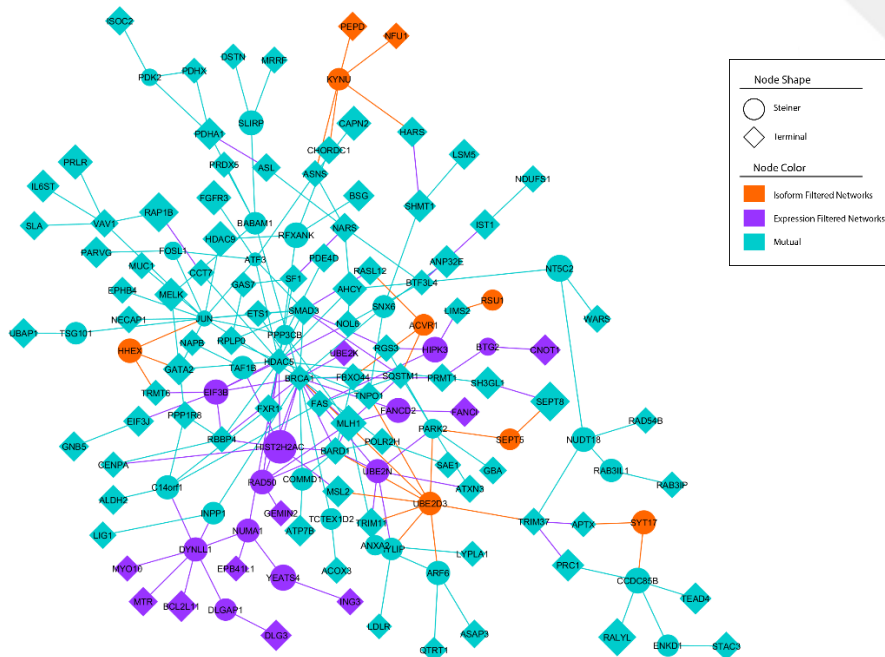


Figure 4.4: The network obtained by merging the two separate condition networks. Orange and purple colors represent the source network specific nodes. Blue nodes are mutual in both source networks.

4.2. Results as Numbers

The results of the isoform analysis and the interaction losses indicate that different samples show similar trends with respect to the counts. Table 4.1 illustrates a list of results where 20 patient samples are randomly chosen.

In this table, “Canon Genes” stands for the number of genes having transcripts belonging to canonical isoform and have a log-fold-change value higher than 1. On the other hand, “Isoform Genes” depicts the number of genes which only have the isoforms that have missing residues that show a higher log-fold-change value than 1. “Lost PPI” stands for the number of lost protein-protein interaction and “Terminals” are the gene names belonging to such proteins that lost interactions. And finally, “Lost DNA” depicts the number of proteins that lost at least one protein – DNA interaction.

Table 4.1: The counts of transcripts and lost interactions from a subset of 20 samples.

| Canon Genes | Isoform Genes | Lost PPI | Terminals | Lost Drug | Lost DNA |
|--------------------|----------------------|-----------------|------------------|------------------|-----------------|
| 2580 | 655 | 254 | 59 | 3 | 4 |
| 2119 | 542 | 254 | 54 | 4 | 5 |
| 2356 | 766 | 482 | 82 | 3 | 7 |
| 2344 | 580 | 303 | 64 | 2 | 5 |
| 3277 | 748 | 617 | 105 | 7 | 10 |
| 2776 | 675 | 472 | 84 | 9 | 2 |
| 2363 | 508 | 354 | 55 | 4 | 4 |
| 2493 | 651 | 466 | 88 | 3 | 11 |
| 1977 | 532 | 253 | 60 | 2 | 4 |
| 3123 | 735 | 707 | 90 | 5 | 6 |
| 2969 | 726 | 454 | 85 | 7 | 11 |
| 2311 | 659 | 526 | 78 | 3 | 8 |

Table 4.1 (cont.)

| | | | | | |
|------|-----|-----|-----|----|----|
| 3438 | 695 | 545 | 81 | 3 | 8 |
| 2210 | 585 | 439 | 74 | 4 | 6 |
| 3022 | 742 | 699 | 104 | 10 | 8 |
| 1812 | 562 | 326 | 66 | 4 | 4 |
| 1459 | 422 | 185 | 49 | 5 | 3 |
| 2926 | 731 | 675 | 86 | 5 | 4 |
| 2703 | 699 | 588 | 99 | 4 | 10 |
| 2939 | 768 | 690 | 96 | 6 | 9 |

4.3. Analysis of Terminal Nodes

4.3.1. The Frequency of Terminals

For each patient, we determined a group of genes called terminal sets. To be included in this set, a gene must have at least one transcript that codes for an isoform having missing residues that cause an interaction loss. At the same time, the log-fold-change value of that transcript should be higher than 1 while the log-fold-change value of the transcript coding for the canonical isoform must be smaller than 1 in case it is expressed. We used these terminal sets as inputs for Omics Integrator runs. To be able to find the genes that are most commonly selected as terminal nodes, we calculated the frequency of each terminal. In the Figure 4.5, we illustrated the counts of the terminals where they are present in at least 150 tumor samples. According to the results, *DLG3* is the most common gene included in the terminal sets with a count of 317. *DLG3* (Discs Large MAGUK Scaffold Protein 3) is a membrane-associated guanylate kinase-family gene. Multiple studies have pointed out that it is down-regulated in several cancer types including glioblastoma, lung and colon cancer (Fukuhara et al., 2003; Hanada et al., 2000; Liu et al., 2014; Makino et al., 1997). Its overexpression, on the other hand, has been associated with programmed cell death and mitotic cell cycle arrest which prevents proliferation and migration (Liu et al., 2014). The two available non-canonical isoforms of *DLG3* protein in UniProt have wide missing regions both more than 300 amino acids. The

finding that DLG3 is the most frequent terminal in our samples suggest that these isoforms showed an increased expression in tumor samples while their canonical isoform could not pass the log-fold-change threshold 1. These results may indicate a down regulation of the protein with an upregulation of potentially less functional or dysfunctional isoforms as suggested by the literature findings.

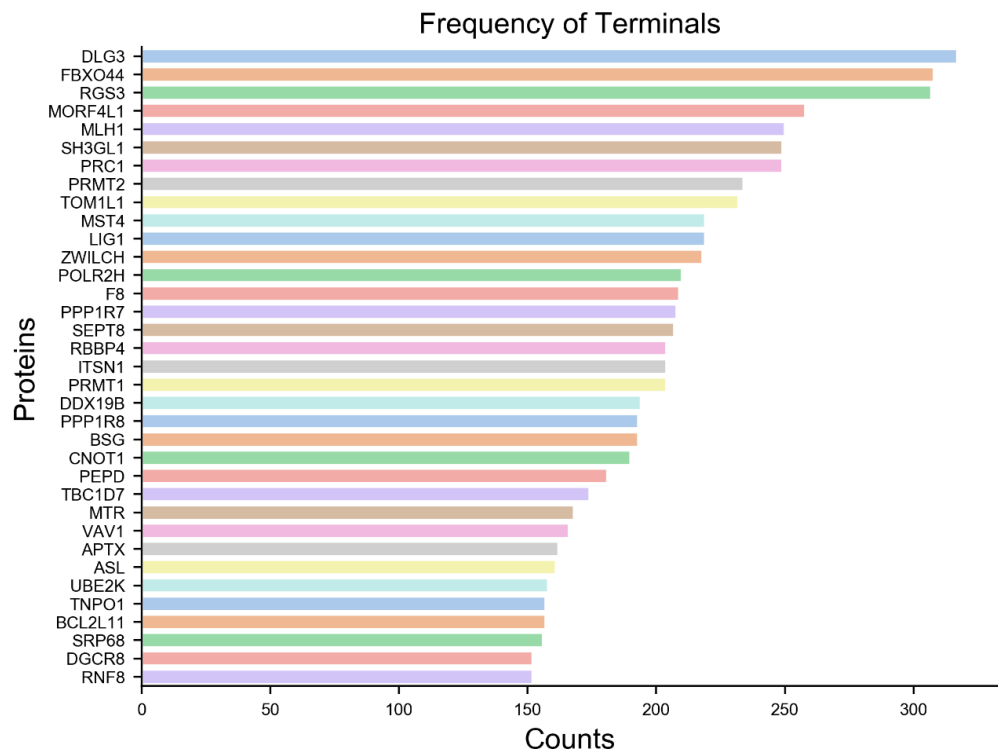


Figure 4.5: Counts of the most common terminal nodes across the samples.

4.3.2. *Overrepresentation Enrichment Analysis (ORA) Results of Terminal Sets*

Since the gene names in terminal sets indicate a higher expression rate of isoforms and potential interaction and function losses, we are also interested in their inclusion in biological pathways. To be able to observe the distribution of the terminal sets in known gene sets, we performed Overrepresentation Enrichment Analysis for each tumor sample. As an input set, we used the gene names that have at least one protein coding transcript whose log-fold-change value is higher than 1 and missing residues caused at least one interaction loss. After completing the ORA runs for 400 tumor samples, we extracted the resulting gene sets in which a subgroup of the terminals is enriched with a value smaller than 0.05. We removed the gene sets that are found in less than 25 patients to be able to obtain a better representation of the gene sets and patients. After the removal of these gene sets, some patient samples end up with no enriched gene sets, hence, they are removed too. In the Fig 4.6, there are 311 patient samples and 123 gene sets which are clustered both in row-wise and column-wise. The results indicate that the enriched gene sets of some patients show very similar trends as some clusters can be pointed in the created heatmap.

For example, there is a distinct cluster of patients that have very similar terminal patterns towards to the bottom of the heatmap as pointed out in Figure 4.7. When we inspect the close-up figure, we observe multiple gene sets related to telomeres and their regulation with protein and RNA localizations. Among the results, “Positive Regulation of Telomere Maintenance via Telomerase” is an interesting result as telomerases are often associated with cancer. Studies also suggest that telomerases are not produced in most of the normal cells and while they are not accepted as the cancer drivers, their presence enables the constant growth in most cancer types (Shay & Wright, 2011). In addition to the cluster related to telomeres, gene silencing and cell cycle related gene sets are also enriched in the heatmap and they are found in multiple patients in similar patterns to cluster them.

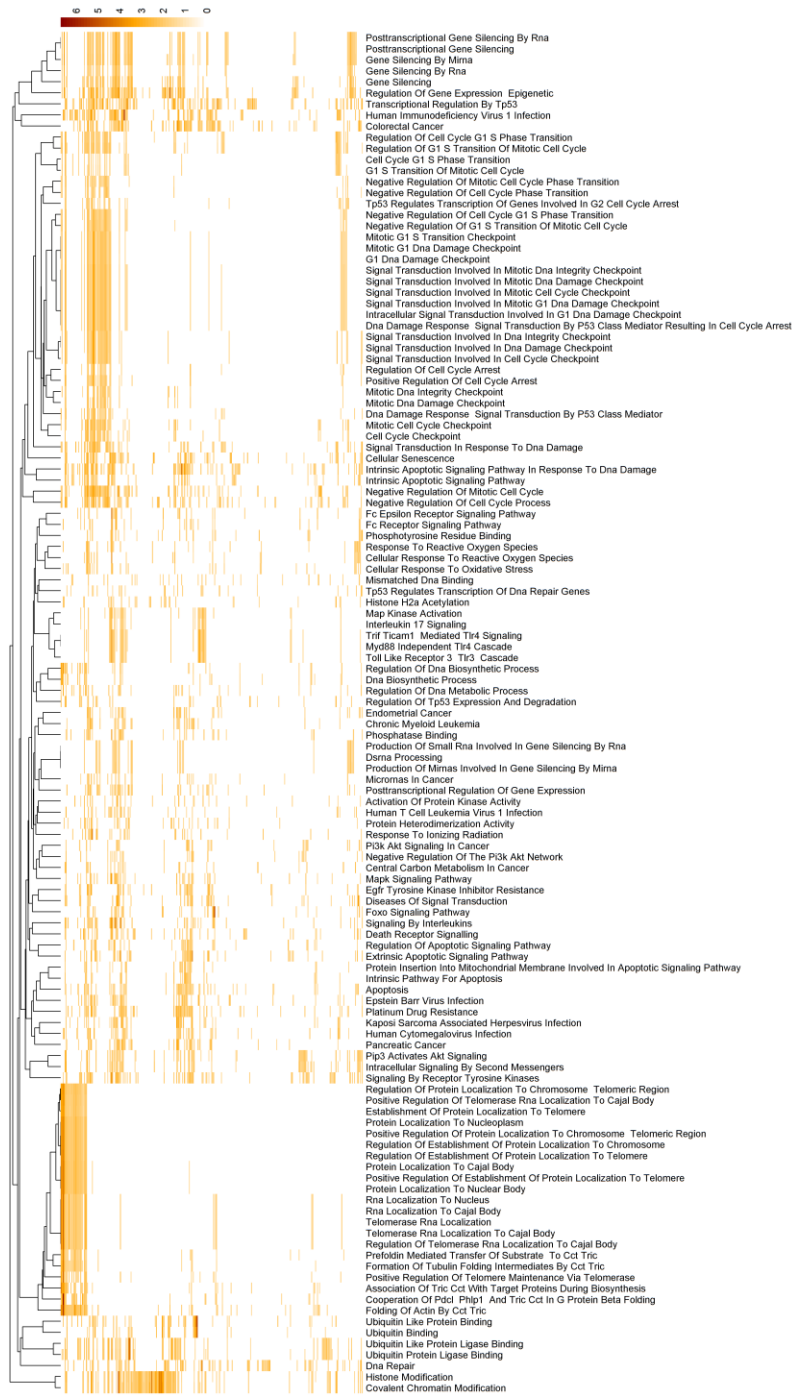


Figure 4.6: The results of the Overrepresentation Enrichment Analysis in the form of a clustered heatmap. The intensity of the color depicts the significance of the enrichment as it is proportional to the negative logarithm of base 10 of FDR results.

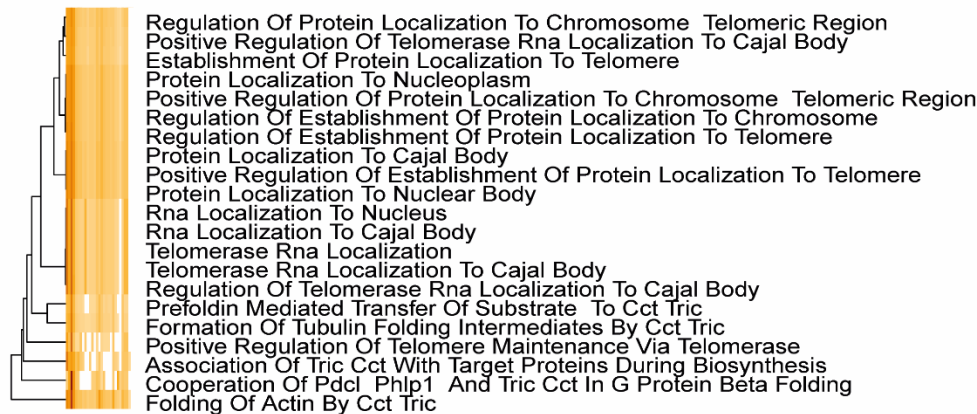


Figure 4.7: A clustered patient sub-group. The group is enriched in telomere and telomerase related genes from the results of overrepresentation enrichment analysis.

4.4. The Analysis of Lost Interactions Across All Tumor Samples

4.4.1. Protein – Protein Interactions

We labelled protein- protein interactions as lost when the interface of an interaction was disrupted due to missing residues in either side of the interaction. Moreover, the log-fold-change value of the transcript coding for the isoform with the missing residue should be higher than 1 indicating an increased expression with respect to a pool of normal samples. If an isoform meets these expectations and its missing residues matches with a protein – protein interface, the expression of the canonical isoform decides whether the interaction will be accepted as “lost”. When the log-fold-change value belonging to canonical isoform is lower than 1, we accepted the interaction as “lost”, otherwise, the interaction was retained in the interactomes.

As a result of the assessment of lost interactions for all samples, we found that 12125 protein – protein interactions were classified as lost. Among these interactions, we were interested in the ones that are common in a high number of patient samples. We calculated the frequency of every lost protein-protein interaction and illustrated the results in Figure 4.8 for the interactions when the loss is present in at least 150 patient samples.

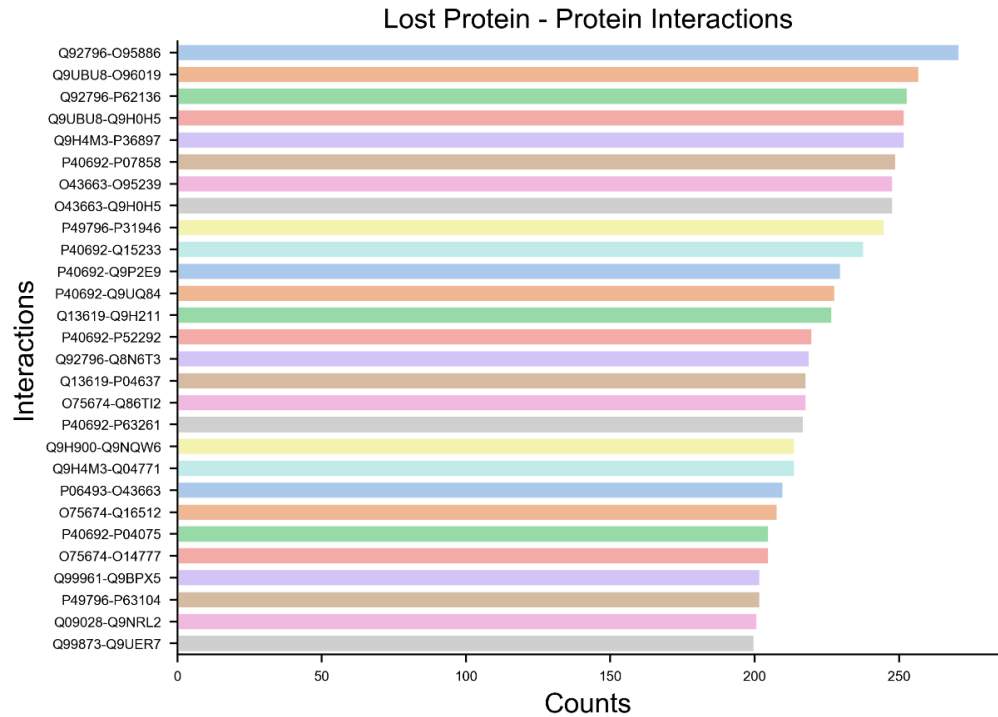


Figure 4.8: Most common lost protein-protein interactions.

4.4.2. Protein – Drug Interactions

After extracting the residues where protein-drug or protein-nutraceutical interaction occurs, we obtained the residues of 351 Uniprot proteins which are found bound to a drug in PDB. After finding the lost drug – protein interactions caused by the missing regions, to be able to find whether there is a similar pattern in the lost protein – drug interactions across the patients, we calculated the frequencies of each lost interaction. Since the number of drug-protein interactions that we obtained from DrugPort was relatively low, we expected to see smaller number of lost interactions in case of drugs with respect to protein – protein interactions. As a result, we obtained a total number of 53 unique protein – drug interaction losses. We only kept the interactions that are lost in at least ten tumor samples and visualized the count results in Figure 4.9 which is available below. Results indicate there are very common interaction losses found in more than the half of the patients which are P23526 – IPA, Q92769 – SHH, and Q94760 – CIR.

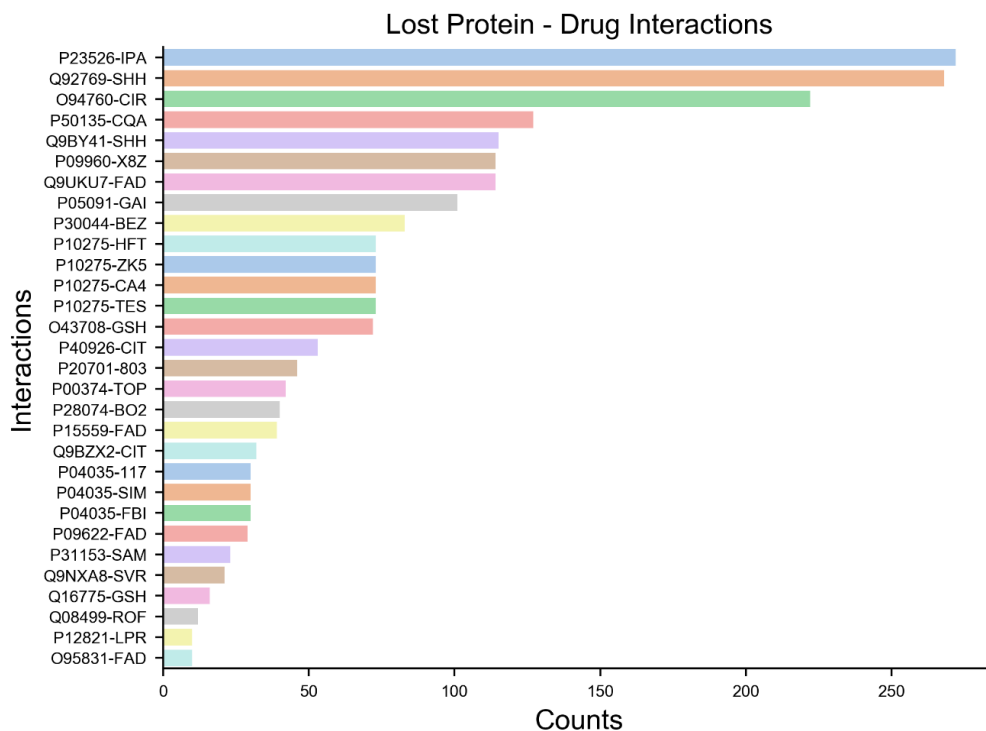


Figure 4.9: The most commonly lost protein -drug interactions across samples.

Q92769 – SHH interaction is an interesting result as SHH or Vorinostat is an approved histone deacetylase inhibitor designed for the treatment of cutaneous T cell lymphoma (CTCL) (Bubna, 2015). Histone deacetylase 2 (HDAC2) is a protein belonging to histone deacetylase (HDAC) family which are responsible for the removal of acetyl groups from the lysine residues of histones proteins. Their role in deacetylation is important as they can play a role in transcriptional regulation, cell cycle progression and developmental events (Conte et al., 2015). The overexpression of *HDAC2* has been previously observed in multiple cancer types. Inhibitors that target HDAC family such as Vorinostat have been shown to provide antitumor effects by enabling a range of mechanisms including growth arrest and apoptosis (Marchion & Munster, 2007).

In our results, the interactions of two HDAC proteins HDAC2 (Q92769) and HDAC8 (Q9BY41) with Vorinostat are expected to be disrupted in more than 250 and 100 tumor samples respectively. This result also shows that the log-fold-change values of the isoforms belonging to these proteins were higher than 1 for each protein which implies an overexpression in general. In UniProt dataset, there is only 2 isoforms available for HDAC2. While one of them is the

canonical isoform, the other variant lacks the first 30 residues compared to the canonical.

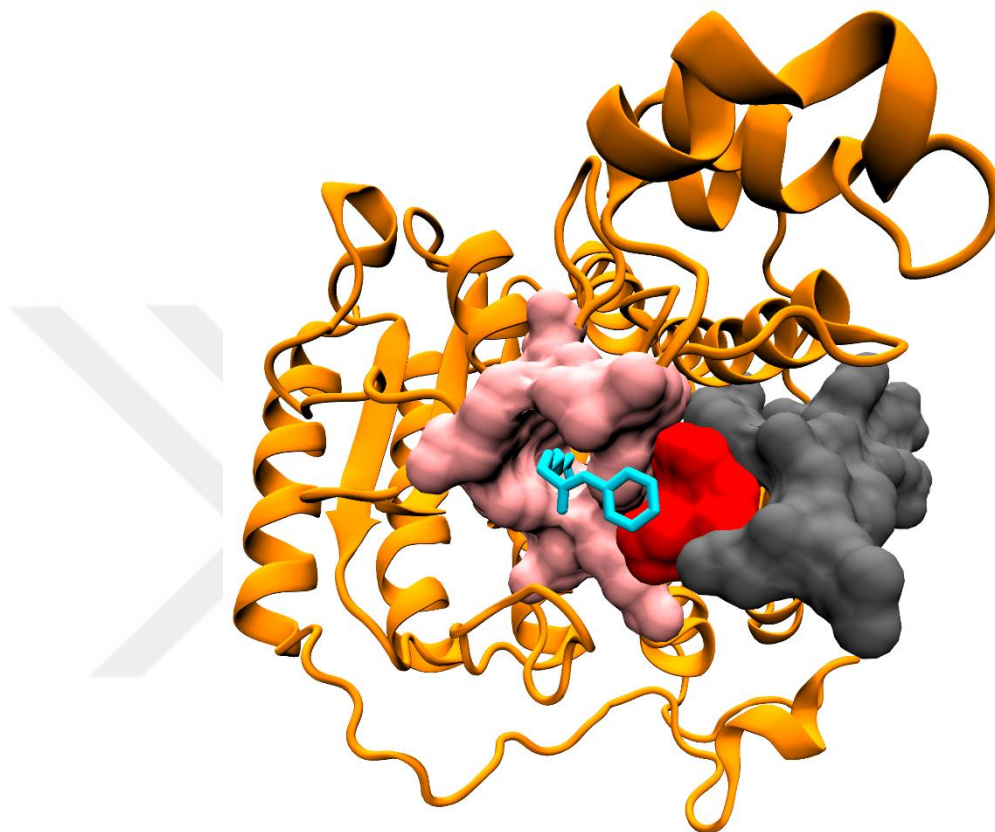


Figure 4.10: HDAC2 in complex with Vorinostat. The part shown in pink surface representation constitutes the interface of HDAC2 protein with Vorinostat. In the non-canonical isoform of the protein, the region colored gray and red will not be present where red region is a 2 amino acid long part of the 13 amino acid long interface. Illustrated from the A chain of 4LXZ PDB structure using VMD (Humphrey, Dalke, & Schulten, 1996).

According to the data available in DrugPort, there is only one PDB structure available which harbors HDAC2 protein bound to Vorinostat with an ID 4LXZ (Lauffer et al., 2013). The binding site of HDAC2 protein with Vorinostat from PDBSum constitutes 13 amino acid residues among which 2 of them are found in the missing region in non-canonical isoform. We present the 3D structure of HDAC2 in complex with Vorinostat in the Figure 4.10.

4.4.3. Protein – DNA Interactions

Like protein – drug interactions, the number of available protein DNA interactions was very low compared to protein – protein interactions. Using the DNA bound proteins available in PDB, we managed to extract the DNA binding regions from 901 PDB structures some of which belong to the same protein. Hence, the number of interactions that are labelled as lost was also quite low respectively. Since the DNA region to which a protein is bound does not have specific identifier like in the drugs or protein interactions, we illustrated the protein – DNA interaction losses using the protein names only. Below is the illustration of the frequency of proteins that lost at least one DNA interaction in 400 tumor samples if the frequency is at least 10.

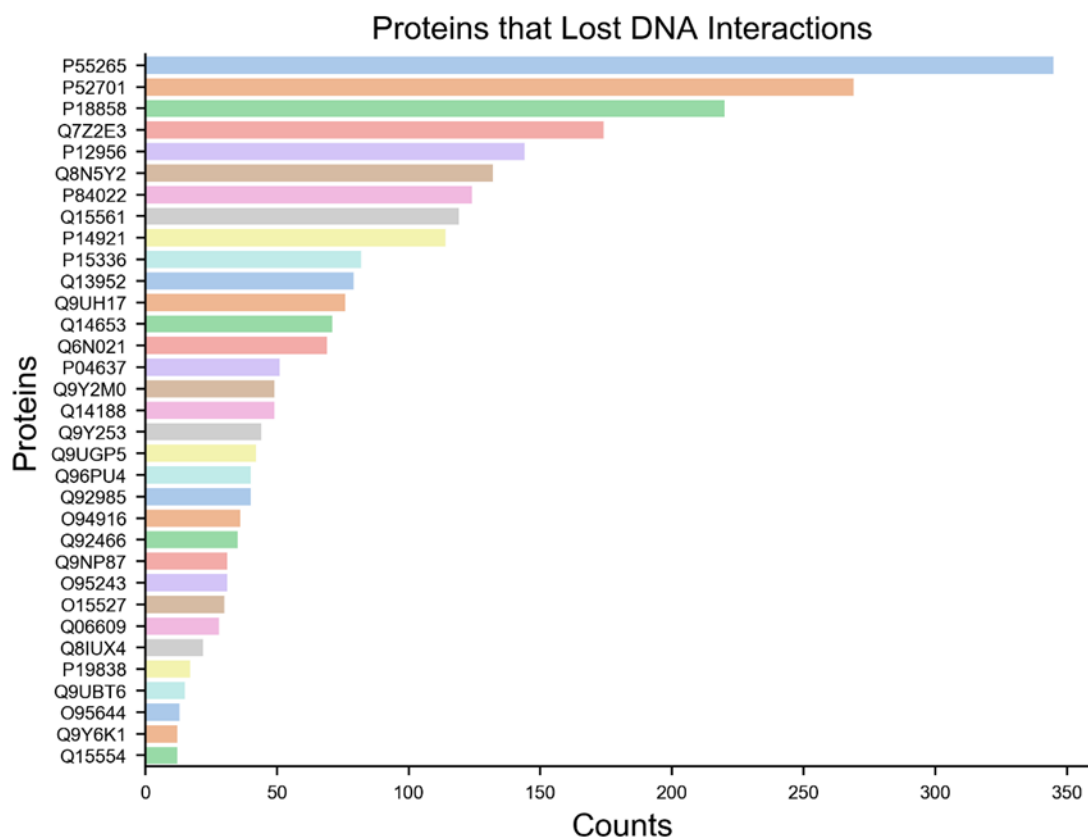


Figure 4.11: The counts of proteins which lost at least one interaction across 400 tumor samples.

CHAPTER 5

5.DISCUSSION

In this study, by integrating the expression and the interface data, we model two different interaction networks to be able to observe the changes between them which hints the rewiring capabilities of networks after they are disturbed. Moreover, we also investigated the distribution of the lost interactions that are caused by the missing residues on isoforms mapping to transcripts that have a two-fold expression rate with respect to an average of normal samples.

After finding the terminal proteins which cause interaction losses due to their predominant isoforms having missing residues, gene enrichment analysis showed that a subgroup of our samples is highly enriched in telomere related genes and can be clustered according to those genes. Cancer is readily associated with telomerase activity since the preservation of telomeres are accepted as a crucial part of the cancer immortality while most of the normal cells cannot even synthesize telomerases (Forsyth, Wright, & Shay, 2002; Shay & Wright, 2011; Wright, Piatyszek, Rainey, Byrd, & Shay, 1996). Their enrichment in our samples indicate an elevated expression trend belonging to non-canonical isoforms. Gene silencing and the cell cycle regulation are also two other apparently enriched gene sets. Finding cell cycle regulation among the result is not surprising as in the case of telomerases since defects in the cell cycle regulation may prevent the controlled growth of a cell and enable overproduction (Foster, 2008). Previous studies have shown that mutations on the tumor suppressor genes could cause such defects in the cell cycle (DeVita, Hellman, & Rosenberg, 1997). As the variants of alternative splicing has the potential to induce the same effect, if a tumor suppressor loses its interactions with its targets, it may result in a reduced suppression impact in tumor cells.

In addition to the terminal proteins, we were also interested in the interaction losses they bring with. To examine further, we calculated the frequencies of the most commonly interrupted interactions between proteins with proteins, drugs and DNA molecules. The results illustrated many interaction losses that are shared between a high number of patients. For example, SHH and its interaction with two histone deacetylases are among the most common lost interactions. Since SHH, or Vorinostat is used for the treatment of cutaneous T cell lymphoma

(CTCL), the loss of an interaction between Vorinostat and its main target HDAC2 could be quite troublesome in terms of therapy (Marchion & Munster, 2007). These findings that show a missing interaction between a cancer drug and its target is especially important for us as it depicts the importance of personalized drug treatments. Since the example interaction is missing in 250 patients, a potential usage of the mentioned drug may fail in treatment for more than half of the patients.

In our study, we created two different sets of interaction networks for each patient to see the way the network continues after some interactions that it depends are no longer available. The comparison between these two sets of interactions showed us, although not as common as the lost interactions explained above, some interactions that are included in the networks after the alternative splicing effect is reflected are more prevalent. We found that the interaction between *LIG1* – *RGS2* was the most common interaction to be included in the networks where *LIG1* is a ligase that is found to be highly expressed in malignant tumors differing from its other ligase counterparts (Jessberger et al., 1997; Montecucco et al., 1992; Signoret & David, 1986; Sun et al., 2001). Moreover, its inhibition could even prevent tumor cell growth according to studies (Sun et al., 2001). Finding a tumor related gene in the most commonly added interaction is an encouraging result although the importance of the specific interaction of *LIG1* – *RGS2* should be further investigated.

In summary, by integrating the interaction losses alternative splicing could cause with the expression information, we found multiple interactions and proteins which can be further examined to elucidate their effects specific to breast cancer. In addition, we found the network rewiring patterns of many samples to observe how the cell can compensate the interaction losses. We believe that the results of our study could enable new starting points for cancer research and personalized treatment strategies. Yet, although findings from different analyses and different interactions provided us interesting results each of which could be further examined, they are still on the transcript level. To be able to obtain more reliable results to support our findings, we will integrate our study with the results of clinical proteomics studies of the same patients from CPTAC (The National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium) (Edwards et al., 2015; Ellis et al., 2013). Moreover, we will also include multiple types of tumor cells in the future researches to be able to detect the differences and similarities across different tumor samples.

REFERENCES

- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2), 166-169. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/25260700>. doi:10.1093/bioinformatics/btu638
- Anna, A., & Monika, G. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet*, 59(3), 253-268. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/29680930>. doi:10.1007/s13353-018-0444-7
- Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O., & GURSOY, A. (2014). PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. *Nucleic Acids Res*, 42(Web Server issue), W285-289. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24829450>. doi:10.1093/nar/gku397
- Berggard, T., Linse, S., & James, P. (2007). Methods for the detection and analysis of protein-protein interactions. *Proteomics*, 7(16), 2833-2842. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17640003>. doi:10.1002/pmic.200700131
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28(1), 235-242. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/10592235>. doi:10.1093/nar/28.1.235
- Blencowe, B. J. (2006). Alternative splicing: new insights from global analyses. *Cell*, 126(1), 37-47. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16839875>. doi:10.1016/j.cell.2006.06.023
- Bogan, A. A., & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol*, 280(1), 1-9. Retrieved from

<https://www.ncbi.nlm.nih.gov/pubmed/9653027>.
doi:10.1006/jmbi.1998.1843

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, *34*(5), 525-527. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/27043002>. doi:10.1038/nbt.3519

Bubna, A. K. (2015). Vorinostat-An Overview. *Indian J Dermatol*, *60*(4), 419. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/26288427>. doi:10.4103/0019-5154.160511

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., & Babu, M. M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell*, *46*(6), 871-883. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22749400>. doi:10.1016/j.molcel.2012.05.039

Chen, L. L., Sabripour, M., Wu, E. F., Prieto, V. G., Fuller, G. N., & Frazier, M. L. (2005). A mutation-created novel intra-exonic pre-mRNA splice site causes constitutive activation of KIT in human gastrointestinal stromal tumors. *Oncogene*, *24*(26), 4271-4280. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/15824741>. doi:10.1038/sj.onc.1208587

Clackson, T., & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, *267*(5196), 383-386. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/7529940>. doi:10.1126/science.7529940

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., . . . Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol*, *17*, 13. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/26813401>. doi:10.1186/s13059-016-0881-8

Conte, M., Dell'Aversana, C., Benedetti, R., Petraglia, F., Carissimo, A., Petrizzi, V. B., . . . Altucci, L. (2015). HDAC2 deregulation in tumorigenesis is causally connected to repression of immune modulation and defense escape. *Oncotarget*, *6*(2), 886-901. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/25473896>. doi:10.18632/oncotarget.2816

- Cooper, T. A., Wan, L., & Dreyfuss, G. (2009). RNA and disease. *Cell*, 136(4), 777-793. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/19239895>. doi:10.1016/j.cell.2009.02.011
- Corominas, R., Yang, X., Lin, G. N., Kang, S., Shen, Y., Ghamsari, L., . . . Iakoucheva, L. M. (2014). Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nat Commun*, 5, 3650. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24722188>. doi:10.1038/ncomms4650
- de Beer, T. A., Berka, K., Thornton, J. M., & Laskowski, R. A. (2014). PDBsum additions. *Nucleic Acids Res*, 42(Database issue), D292-296. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24153109>. doi:10.1093/nar/gkt940
- DeVita, V. T., Hellman, S., & Rosenberg, S. A. (1997). Cancer : principles & practice of oncology. Retrieved from <http://books.google.com/books?id=H6prAAAAMAAJ>.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/23104886>. doi:10.1093/bioinformatics/bts635
- Edwards, N. J., Oberti, M., Thangudu, R. R., Cai, S., McGarvey, P. B., Jacob, S., . . . Ketchum, K. A. (2015). The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J Proteome Res*, 14(6), 2707-2713. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/25873244>. doi:10.1021/pr501254j
- Ellis, M. J., Gillette, M., Carr, S. A., Paulovich, A. G., Smith, R. D., Rodland, K. K., . . . Clinical Proteomic Tumor Analysis, C. (2013). Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov*, 3(10), 1108-1112. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24124232>. doi:10.1158/2159-8290.CD-13-0219
- Fisher, R., Pusztai, L., & Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer*, 108(3), 479-485.

Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/23299535>.
doi:10.1038/bjc.2012.581

Forsyth, N. R., Wright, W. E., & Shay, J. W. (2002). Telomerase and differentiation in multicellular organisms: turn it off, turn it on, and turn it off again. *Differentiation*, 69(4-5), 188-197. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/11841477>. doi:10.1046/j.1432-0436.2002.690412.x

Foster, I. (2008). Cancer: A cell cycle defect. *Radiography*, 14(2), 144-149.

Fukuhara, H., Masuda, M., Yageta, M., Fukami, T., Kuramochi, M., Maruyama, T., . . . Murakami, Y. (2003). Association of a lung tumor suppressor TSLC1 with MPP3, a human homologue of Drosophila tumor suppressor Dlg. *Oncogene*, 22(40), 6160-6165. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/13679854>.
doi:10.1038/sj.onc.1206744

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med*, 375(12), 1109-1112. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/27653561>.
doi:10.1056/NEJMp1607591

Hanada, N., Makino, K., Koga, H., Morisaki, T., Kuwahara, H., Masuko, N., . . . Saya, H. (2000). NE-dlg, a mammalian homolog of Drosophila dlg tumor suppressor, induces growth suppression and impairment of cell adhesion: possible involvement of down-regulation of beta-catenin by NE-dlg expression. *Int J Cancer*, 86(4), 480-488. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/10797259>.
doi:10.1002/(sici)1097-0215(20000515)86:4<480::aid-ijc6>3.0.co;2-6

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1), 57-70. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/10647931>. doi:10.1016/s0092-8674(00)81683-9

Hanoun, N., Bureau, C., Diab, T., Gayet, O., Dusetti, N., Selves, J., . . . Torrisani, J. (2010). The SV2 variant of KLF6 is down-regulated in hepatocellular carcinoma and displays anti-proliferative and pro-apoptotic functions. *J Hepatol*, 53(5), 880-888. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20801538>.
doi:10.1016/j.jhep.2010.04.038

- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., . . . Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*, 22(9), 1760-1774. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22955987>. doi:10.1101/gr.135350.111
- Hastings, M. L., Resta, N., Traum, D., Stella, A., Guanti, G., & Krainer, A. R. (2005). An LKB1 AT-AC intron mutation causes Peutz-Jeghers syndrome via splicing at noncanonical cryptic splice sites. *Nat Struct Mol Biol*, 12(1), 54-59. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/15608654>. doi:10.1038/nsmb873
- Hopf, T. A., Scharfe, C. P., Rodrigues, J. P., Green, A. G., Kohlbacher, O., Sander, C., . . . Marks, D. S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*, 3. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/25255213>. doi:10.7554/eLife.03430
- Hristov, B. H., & Singh, M. (2017). Network-Based Coverage of Mutational Profiles Reveals Cancer Genes. *Cell Syst*, 5(3), 221-229 e224. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/28957656>. doi:10.1016/j.cels.2017.09.003
- Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front Genet*, 8, 84. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/28670325>. doi:10.3389/fgene.2017.00084
- Huang, S. s. C., & Fraenkel, E. (2009). Integrating Proteomic, Transcriptional, and Interactome Data Reveals Hidden Components of Signaling and Regulatory Networks. *Science Signaling*, 2(81), ra40-ra40. doi:10.1126/scisignal.2000350
- Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *J Mol Graph*, 14(1), 33-38, 27-38. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/8744570>.
- Janin, J., Bahadur, R. P., & Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. *Q Rev Biophys*, 41(2), 133-180. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/18812015>. doi:10.1017/S0033583508004708

- Jessberger, R., Schar, P., Robins, P., Ferrari, E., Riwar, B., & Hubscher, U. (1997). Regulation of DNA metabolic enzymes upon induction of preB cell development and V(D)J recombination: up-regulation of DNA polymerase delta. *Nucleic Acids Res*, *25*(2), 289-296. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/9016556>. doi:10.1093/nar/25.2.289
- Johnson, M. B., Kawasaki, Y. I., Mason, C. E., Krsnik, Z., Coppola, G., Bogdanovic, D., . . . Sestan, N. (2009). Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron*, *62*(4), 494-509. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/19477152>. doi:10.1016/j.neuron.2009.03.027
- Karni, R., de Stanchina, E., Lowe, S. W., Sinha, R., Mu, D., & Krainer, A. R. (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat Struct Mol Biol*, *14*(3), 185-193. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17310252>. doi:10.1038/nsmb1209
- Katz, Y., Wang, E. T., Airoidi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*, *7*(12), 1009-1015. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21057496>. doi:10.1038/nmeth.1528
- Kedaigle, A. J., & Fraenkel, E. (2018). Discovering Altered Regulation and Signaling Through Network-based Integration of Transcriptomic, Epigenomic, and Proteomic Tumor Data. *Methods Mol Biol*, *1711*, 13-26. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/29344883>. doi:10.1007/978-1-4939-7493-1_2
- Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S. (2013). Function of alternative splicing. *Gene*, *514*(1), 1-30. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22909801>. doi:10.1016/j.gene.2012.07.083
- Kim, E., Goren, A., & Ast, G. (2008). Alternative splicing and disease. *RNA Biol*, *5*(1), 17-19. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/18388487>. doi:10.4161/rna.5.1.5944
- Kolde, R. (2012). Pheatmap: pretty heatmaps. *R package version*, *61*(926), 915.

- Laskowski, R. A., Jablonska, J., Pravda, L., Varekova, R. S., & Thornton, J. M. (2018). PDBsum: Structural summaries of PDB entries. *Protein Sci*, 27(1), 129-134. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/28875543>. doi:10.1002/pro.3289
- Lauffer, B. E., Mintzer, R., Fong, R., Mukund, S., Tam, C., Zilberleyb, I., . . . Steiner, P. (2013). Histone deacetylase (HDAC) inhibitor kinetic rate constants correlate with cellular histone acetylation but not transcription and cell viability. *J Biol Chem*, 288(37), 26926-26943. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/23897821>. doi:10.1074/jbc.M113.490706
- Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4), 493-500. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20022975>. doi:10.1093/bioinformatics/btp692
- Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*, 47(W1), W199-W205. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/31114916>. doi:10.1093/nar/gkz401
- Lindahl, T., & Barnes, D. E. (1992). Mammalian DNA ligases. *Annu Rev Biochem*, 61, 251-281. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/1497311>. doi:10.1146/annurev.bi.61.070192.001343
- Liu, Z., Niu, Y., Xie, M., Bu, Y., Yao, Z., & Gao, C. (2014). Gene expression profiling analysis reveals that DLG3 is down-regulated in glioblastoma. *J Neurooncol*, 116(3), 465-476. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24381070>. doi:10.1007/s11060-013-1325-x
- Luscombe, N. M., Laskowski, R. A., & Thornton, J. M. (1997). NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucleic Acids Res*, 25(24), 4940-4945. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/9396800>. doi:10.1093/nar/25.24.4940

- Makino, K., Kuwahara, H., Masuko, N., Nishiyama, Y., Morisaki, T., Sasaki, J., . . . Saya, H. (1997). Cloning and characterization of NE-dlg: a novel human homolog of the *Drosophila* discs large (dlg) tumor suppressor protein interacts with the APC protein. *Oncogene*, *14*(20), 2425-2433. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/9188857>. doi:10.1038/sj.onc.1201087
- Marchion, D., & Munster, P. (2007). Development of histone deacetylase inhibitors for cancer treatment. *Expert Rev Anticancer Ther*, *7*(4), 583-598. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17428177>. doi:10.1586/14737140.7.4.583
- Martin, A. C. (2005). Mapping PDB chains to UniProtKB entries. *Bioinformatics*, *21*(23), 4297-4301. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16188924>. doi:10.1093/bioinformatics/bti694
- 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. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/29355848>. doi:10.1038/nmeth.4540
- Mills, J. D., & Janitz, M. (2012). Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases. *Neurobiol Aging*, *33*(5), 1012 e1011-1024. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22118946>. doi:10.1016/j.neurobiolaging.2011.10.030
- Montecucco, A., Biamonti, G., Savini, E., Focher, F., Spadari, S., & Ciarrocchi, G. (1992). DNA ligase I gene expression during differentiation and cell proliferation. *Nucleic Acids Res*, *20*(23), 6209-6214. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/1475182>. doi:10.1093/nar/20.23.6209
- Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, *5*(7), 621-628. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/18516045>. doi:10.1038/nmeth.1226
- Mosca, R., Ceol, A., & Aloy, P. (2013). Interactome3D: adding structural details to protein networks. *Nat Methods*, *10*(1), 47-53. Retrieved from

<https://www.ncbi.nlm.nih.gov/pubmed/23399932>.
doi:10.1038/nmeth.2289

- Narla, G., Difeo, A., Reeves, H. L., Schaid, D. J., Hirshfeld, J., Hod, E., . . . Martignetti, J. A. (2005). A germline DNA polymorphism enhances alternative splicing of the KLF6 tumor suppressor gene and is associated with increased prostate cancer risk. *Cancer Res*, *65*(4), 1213-1222. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/15735005>. doi:10.1158/0008-5472.CAN-04-4249
- Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, *463*(7280), 457-463. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20110989>. doi:10.1038/nature08909
- Okumura, N., Yoshida, H., Kitagishi, Y., Nishimura, Y., & Matsuda, S. (2011). Alternative splicings on p53, BRCA1 and PTEN genes involved in breast cancer. *Biochem Biophys Res Commun*, *413*(3), 395-399. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21893034>. doi:10.1016/j.bbrc.2011.08.098
- Omenn, G. S., Yocum, A. K., & Menon, R. (2010). Alternative splice variants, a new class of protein cancer biomarker candidates: findings in pancreatic cancer and breast cancer with systems biology implications. *Dis Markers*, *28*(4), 241-251. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20534909>. doi:10.3233/DMA-2010-0702
- Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, *40*(12), 1413-1415. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/18978789>. doi:10.1038/ng.259
- Park, E., Pan, Z., Zhang, Z., Lin, L., & Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet*, *102*(1), 11-26. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/29304370>. doi:10.1016/j.ajhg.2017.11.002
- Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*, *33*(3), 290-295.

Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/25690850>.
doi:10.1038/nbt.3122

- Pettigrew, C., Wayte, N., Lovelock, P. K., Tavtigian, S. V., Chenevix-Trench, G., Spurdle, A. B., & Brown, M. A. (2005). Evolutionary conservation analysis increases the colocalization of predicted exonic splicing enhancers in the BRCA1 gene with missense sequence changes and in-frame deletions, but not polymorphisms. *Breast Cancer Res*, 7(6), R929-939. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16280041>. doi:10.1186/bcr1324
- Raj, T., Li, Y. I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., . . . De Jager, P. L. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat Genet*, 50(11), 1584-1592. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/30297968>. doi:10.1038/s41588-018-0238-1
- Razick, S., Magklaras, G., & Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9, 405. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/18823568>. doi:10.1186/1471-2105-9-405
- Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J., . . . Dunker, A. K. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc Natl Acad Sci U S A*, 103(22), 8390-8395. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16717195>. doi:10.1073/pnas.0507916103
- Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., . . . Zipursky, S. L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6), 671-684. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/10892653>. doi:10.1016/s0092-8674(00)80878-8
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11), 2498-2504. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/14597658>. doi:10.1101/gr.1239303

- Shay, J. W., & Wright, W. E. (2011). Role of telomeres and telomerase in cancer. *Semin Cancer Biol*, 21(6), 349-353. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22015685>. doi:10.1016/j.semcancer.2011.10.001
- Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., . . . Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*, 111(51), E5593-5601. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/25480548>. doi:10.1073/pnas.1419161111
- Signoret, J., & David, J. C. (1986). Control of the expression of genes for DNA ligase in eukaryotes. *Int Rev Cytol*, 103, 249-279. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/3528020>.
- Skotheim, R. I., & Nees, M. (2007). Alternative splicing in cancer: noise, functional, or systematic? *Int J Biochem Cell Biol*, 39(7-8), 1432-1449. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17416541>. doi:10.1016/j.biocel.2007.02.016
- Srebrow, A., & Kornblihtt, A. R. (2006). The connection between splicing and cancer. *J Cell Sci*, 119(Pt 13), 2635-2641. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16787944>. doi:10.1242/jcs.03053
- Sun, D., Urrabaz, R., Nguyen, M., Marty, J., Stringer, S., Cruz, E., . . . Weitman, S. (2001). Elevated expression of DNA ligase I in human cancers. *Clin Cancer Res*, 7(12), 4143-4148. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/11751514>.
- Tamaro, C., Raponi, M., Wilson, D. I., & Baralle, D. (2012). BRCA1 exon 11 alternative splicing, multiple functions and the association with cancer. *Biochem Soc Trans*, 40(4), 768-772. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22817731>. doi:10.1042/BST20120140
- Tang, J. Y., Lee, J. C., Hou, M. F., Wang, C. L., Chen, C. C., Huang, H. W., & Chang, H. W. (2013). Alternative splicing for diseases, cancers, drugs, and databases. *ScientificWorldJournal*, 2013, 703568. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/23766705>. doi:10.1155/2013/703568

- Tomczak, K., Czerwinska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*, *19*(1A), A68-77. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/25691825>. doi:10.5114/wo.2014.47136
- Torkamani, A., & Schork, N. J. (2009). Identification of rare cancer driver mutations by network reconstruction. *Genome Res*, *19*(9), 1570-1578. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/19574499>. doi:10.1101/gr.092833.109
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, *7*(3), 562-578. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22383036>. doi:10.1038/nprot.2012.016
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, *28*(5), 511-515. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20436464>. doi:10.1038/nbt.1621
- 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. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/27096930>. 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. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21886100>. doi:10.1038/nprot.2011.367
- UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, *47*(D1), D506-D515. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/30395287>. doi:10.1093/nar/gky1049

- Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., . . . Cancer Genome Atlas Research, N. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, *17*(1), 98-110. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20129251>. doi:10.1016/j.ccr.2009.12.020
- von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, *417*(6887), 399-403. doi:10.1038/nature750
- Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., . . . Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470-476. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/18978772>. doi:10.1038/nature07509
- Wang, G. S., & Cooper, T. A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet*, *8*(10), 749-761. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17726481>. doi:10.1038/nrg2164
- Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., . . . Liu, J. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*, *38*(18), e178. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20802226>. doi:10.1093/nar/gkq622
- Wang, X. Q., Luk, J. M., Leung, P. P., Wong, B. W., Stanbridge, E. J., & Fan, S. T. (2005). Alternative mRNA splicing of liver intestine-cadherin in hepatocellular carcinoma. *Clin Cancer Res*, *11*(2 Pt 1), 483-489. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/15701831>.
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., . . . Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*, *46*(D1), D1074-D1082. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/29126136>. doi:10.1093/nar/gkx1037
- Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., . . . Woolsey, J. (2006). DrugBank: a comprehensive resource for in

- silico drug discovery and exploration. *Nucleic Acids Res*, 34(Database issue), D668-672. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16381955>. doi:10.1093/nar/gkj067
- Wright, W. E., Piatyszek, M. A., Rainey, W. E., Byrd, W., & Shay, J. W. (1996). Telomerase activity in human germline and embryonic tissues and cells. *Dev Genet*, 18(2), 173-179. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/8934879>. doi:10.1002/(SICI)1520-6408(1996)18:2<173::AID-DVG10>3.0.CO;2-3
- Wu, J., Akerman, M., Sun, S., McCombie, W. R., Krainer, A. R., & Zhang, M. Q. (2011). SpliceTrap: a method to quantify alternative splicing under single cellular conditions. *Bioinformatics*, 27(21), 3010-3016. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21896509>. doi:10.1093/bioinformatics/btr508
- Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., . . . Vidal, M. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, 164(4), 805-817. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/26871637>. doi:10.1016/j.cell.2016.01.029
- Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., . . . Fraenkel, E. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet*, 41(3), 316-323. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/19234470>. doi:10.1038/ng.337
- Yeo, G., Holste, D., Kreiman, G., & Burge, C. B. (2004). Variation in alternative splicing across human tissues. *Genome Biol*, 5(10), R74. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/15461793>. doi:10.1186/gb-2004-5-10-r74
- Yi, Q., & Tang, L. (2011). Alternative spliced variants as biomarkers of colorectal cancer. *Curr Drug Metab*, 12(10), 966-974. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21787266>.