COMBINED NETWORK ANALYSIS AND MOLECULAR DYNAMICS SIMULATIONS STUDY FOR CHARACTERIZATION OF PREVALENT SOMATIC MUTATIONS IN BREAST CANCER: SF3B1 CASE STUDY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF
ENGINEERING AND NATURAL SCIENCES
OF ISTANBUL MEDIPOL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
BIOMEDICAL ENGINEERING AND BIOINFORMATICS

By
Asmaa Samy Mohamed Mahmoud
February, 2020

COMBINED NETWORK ANALYSIS AND MOLECULAR DYNAMICS SIMULATIONS STUDY FOR CHARACTERIZATION OF PREVALENT SOMATIC MUTATIONS IN BREAST CANCER: SF3B1 CASE STUDY

By Asmaa Samy Mohamed Mahmoud

February, 2020

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

Assoc. Prof. Dr. M. Kemal Özdemir (Advisor)

Assist. Prof. Dr. Özge Şensoy

Assist. Prof. Dr. Barış Ethem Süzek

Approved by the Graduate School of Engineering and Natural Sciences:

Assoc. Prof. Dr. Yasemin Yüksel Durmaz
Director of the Graduate School of Engineering and Natural Sciences

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: ASMAA SAMY MOHAMED MAHMOUD

Signature : Asmaa Samy

ABSTRACT

COMBINED NETWORK ANALYSIS AND MOLECULAR DYNAMICS SIMULATIONS STUDY FOR CHARACTERIZATION OF PREVALENT SOMATIC MUTATIONS IN BREAST CANCER: SF3B1 CASE STUDY

Asmaa Samy Mohamed Mahmoud
M.S. in Biomedical Engineering and Bioinformatics
Advisor: Assoc. Prof. Dr. Mehmet Kemal Özdemir
February, 2020

Breast cancer has the highest incidence and mortality rates among women. The etiology of the disease has remained elusive because of complex interactions among various factors. The somatic mutations are one of such factors that lead to breast cancer development. Many somatic mutations have been identified in breast cancer. Unfortunately, in many cases, our knowledge about these mutations is limited to their allele frequencies and their relations to cancer deserve further investigation. In this thesis, in silico approach was defined to investigate the impact of somatic mutations in breast cancer by utilizing publicly available databases, bioinformatics, and computational biophysics tools. Firstly, a gene network of 67 genes participate in splicing mechanism was constructed. Analysis of this network reveals that splicing factor 3B subunit 1 (SF3B1) is the central node having higher network metrics and the highest mutation rate among other genes. Then, data and network analyses showed i) impact of aberrant splicing on other biological processes such as regulation of cell proliferation, apoptosis, and transcription and ii) relations among hematologic malignancies and breast cancer that may explain the transformation from one cancer to another. Lastly, the impact of K700E on dynamics and structure of SF3B1 was investigated by performing classical molecular dynamics simulation. Comparative analysis of wild type vs. mutant trajectories showed that the mutation i) decreases the stability of the components of the splicing machinery such as SF3B1, p14, and pre-mRNA, which consequently weakens the interaction formed between pre-mRNA and both K700E and $p14^{RRM}$and ii) distorts the communication among SF3B1 residues. These changes may lead to alternative branch point selection, aberrant splicing of pre-mRNA, and production of abnormal transcripts. This thesis provided

i) insights into complex interactions among genes, pathways, and diseases that may improve the development of new prevention, prognostic and therapeutic approaches for cancer and ii) molecular details to understand the functional consequences of K700E on the spliceosomal machinery, proposing SF3B1 as a potential biomarker and therapeutic target for cancer. In light of these findings, the defined in silico process, which is based on bioinformatics, network analysis, and computational biophysics tools, can be improved and employed in the identification and characterization of highly mutated genes that lead to cancer development and/or prognosis. Consequently, developing new prevention and therapeutic approaches can be improved to combat cancer.

Keywords: Breast cancer; Somatic mutations; pre-mRNA splicing; SF3B1; Gene Networks; Protein Association Networks; Network Analysis; Molecular dynamics simulations.

ÖZET

MEME KANSERİNDE BİR HAYLİ MUTASYONA UĞRAMIŞ BİR GENİ KARAKTERİZE ETMEK İÇİN BİRLEŞİK AĞ ANALİZİ VE MOLEKÜLER DİNAMİK SİMÜLASYONLARI ÇALIŞMASI: SF3B1 ÖRNEK OLAY İNCELEMESİ

Asmaa Samy Mohamed Mahmoud
Biyomedikal Mühendisliği ve Biyoenformatik, Yüksek Lisans
Tez Danışmanı: Doç. Dr. Mehmet Kemal Özdemir
Şubat, 2020

Meme kanseri kadınlar arasında en fazla görünen ve ölüm oranı en yüksek kansetr tipidir. Hastalığa neden olan faktörler çok çeşitli ve karmaşık olduğundan, hastalığa sebep olan gerçek faktörler tam anlaşılamamıştır. Somatik mutasyonlar, meme kanserine yol açan bu faktörlerden biridir. Meme kanserinde birçok somatik mutasyon tespit edilmiştir. Ne yazık ki, birçok durumda, bu mutasyonlar hakkındaki bilgimiz, bu faktörlerin görülme sıklıkları ile sınırlıdır ve bu nedenle kanserle ilişkileri için daha fazla araştırılmaları gerekmektedir. Bu tezde, açık veri tabanları ile biyoinformatik ve hesaplamalı biyofizik araçları kullanılarak somatik mutasyonların meme kanseri üzerindeki etkilerini araştırmak için bir in-siliko yani bilgisayar benzetimleri yaklaşımı tanımlanmıştır. İlk olarak, genleri birbirine ekleyerek 67 genden oluşmuş bir gen ağı oluşturuldu. Bu gen ağının analizleri neticesinde, 3B-alt bir-1 eklentili genin (SF3B1), diğer genlere nazaran daha yüksek ağ metriklerine ve en yüksek mutasyon oranına sahip bir merkezi düğüm noktası olduğu gözlemlenmiştir. Akabinde, veri ve ağ anailizleri neticesinde i) anormal uçbirleştirmenin hücre proliferasyonu, apapotozu ve transkripsiyonu gibi diğer biyolojik süreçleri üzerindeki etkisi, ii) bir kanserden diğerine dönüşümü açıklayabilecek hematolojik maligniteler ile meme kanseri arasında ilişkilerinin olduğu gösterilmiştir. Son olarak da, klasik moleküler dinamik simülasyonu gerçekleştirilerek K700E'nin, SF3B1'in dinamikleri ve yapısı üzerindeki etkisi incelenmiştir. Doğal fenotip ve mutantların izlediği gezinim yörüngülerinin karşılaştırılmalı analizinde mutasyonların i) SF3B1, p14 ve pre-mRNA gibi yapıların birbirine ekli bileşenlerinin stabilitesini azalttığını ve sonuç olarak aralarındaki etkileşimi zayıflattığını, ve ii) SF3B1 kalıntılarının arasındaki etkileşimi bozduğunu göstermiştir.

Bu değişimler alternatif dallanma noktası seçimine, anormal pre-mRNA uçbirleştirmeye ve anormal transkriptlerin üretilmesine yol açabilir. Bu tez, i) kanser için yeni önlemlerin geliştirilmesi, prognostik ve terapötik yaklaşımların gelişimini sağlayacak genler, metotlar ve hastalıklar arasındaki karmaşık etkileşimlere farklı bir sezgi ve ii) K700E'nin bileşenlerinin birbirine eklenmesi kurgusu üzerindeki fonksiyonel sonuçlarını anlamak için SF3B1'i potansiyel bir biyobelirteç ve kanser için terapötik hedef olarak öneren moleküler detayları ortaya çıkarmıştır. Bu bulgular ışığı altında, biyoenformatik, gen ağı analizleri, ve hesaplamalı biyofizik tabanlı araçları kullanan in-siliko yani bilgisayar benzetimleri daha da iyi geliştirilebilir ve bunun neticesinde kansere sebep olan mutasyona uğramış genlerin daha iyi karakterize edilmesine olanak sağlanabilir. Bu şekilde kanserle olan savaşta daha yeni önlemler ve yeni tedavi yaklaşımları geliştirilebilir.

Anahtar sözcükler: Meme Kanseri; Somatik Mutasyonlar; pre-mRNA Uçbirleştirme, SF3B1; Gen Ağı; Protein İlişki Ağı; Ağ Analizi; Moleküler Dinamik Simülasyonları.

Acknowledgement

First and foremost, I would like to express my gratitude and deep appreciation to my supervisor Assoc. Prof. Dr. Mehmet Kemal Özdemir for his continuous guidance, help, and support. Without his support, I would not get my master's degree.

I am grateful to Assist. Prof. Dr. Barış Ethem Süzek and Assist. Prof. Dr. Özge Şensoy to gain this level of knowledge and to learn new skills. Many thanks for their guidance and help since the project started until now.

I would like to express my gratitude to Prof. Reda Alhajj for teaching and guiding me to perform network analysis study in this thesis.

I would like to thank Istanbul Medipol University for offering the High Performance Computing system to perform classical molecular dynamics simulations.

I would like to thank past and present Sensoy lab members (Mehmet Hanifi, Samman Mansoor, Hanife Pekel, and Metehan Ilter) for their help to perform and analyze MD simulations.

I highly appreciate the opportunity that let me have great friends from Medipol. Many thanks for them.

I deeply thank my Egyptian friends for always supporting and motivating me despite far distances.

Last but not least, I dedicate this thesis to my family. I could not have done this without their unconditional love, care, and support.

Contents

1 INTRODUCTION & MOTIVATION **1**

1.1 **Breast Cancer Overview** 1

1.2 **Breast Cancer Etiology** 2

1.3 **Breast Cancer Related Biological Processes** 3

1.4 **Pre-mRNA Splicing Mechanism** 3

1.5 **Motivation** 5

1.6 **Conclusion** 6

2 LITERATURE REVIEW **7**

2.1 **Experimental Studies That Proposed SF3B1 as a Therapeutic Target for Cancer** 7

2.2 **Computational Approaches to Investigate the Impact of Somatic Mutations on Splicing** 8

2.3 **Thesis Contribution** 10

3 IDENTIFICATION OF A HIGHLY MUTATED GENE THAT

CONTRIBUTES TO THE DEVELOPMENT OF BREAST CANCER 13

3.1 Selection of a Biological Process Potentially Involved in Breast Cancer . . . 14

3.2 Identification of Genes Participating in pre-mRNA Splicing 14

3.3 Building a Gene Network . . . 14

3.4 Identifying a Gene with High Network Centrality and Number of Mutations . . . 16

3.5 Conclusion . . . 19

4 DATA & NETWORK BASED ANALYSIS OF SF3B1 20

4.1 Linking SF3B1 to Cancer: A Historical Perspective 21

4.2 Investigating SF3B1-Associated Mutations and the Related Diseases . . . 23

4.3 Identifying Association of SF3B1 Families With Cancer . 31

4.4 Identifying Interactions Between SF3B1 and Different Pathways . . . 36

4.5 Relations Among Subtypes of Hematologic and Breast Cancers . . . 39

4.6 Discussion . . . 41

4.7 Conclusion . . . 43

5 ASSESSMENT OF THE IMPACT OF SF3B1-K700E BY

MEANS OF MOLECULAR DYNAMICS SIMULATIONS 45

5.1 System Setup 46

5.2 Simulation Protocol 47

5.3 Simulation Analysis Tools 48

5.3.1 Root mean square deviation (RMSD) 48

5.3.2 Calculation of number of contacts 49

5.3.3 Principal component analysis (PCA) 49

5.3.4 Dynamic cross correlation map (DCCM) 50

5.4 MD Trajectories Analysis 50

5.4.1 K700E mutation decreases the Stability of SF3B1, pre-mRNA and p14 51

5.4.2 K700E mutation weakens the interactions formed with pre-mRNA 52

5.4.3 K700E mutation affects the global dynamics of SF3B1 53

5.4.4 K700E mutation distorts the communication among SF3B1 residues 55

5.5 Discussion 57

5.6 Conclusion 58

6 CONCLUSION & Future Work 60

List of Figures

1.1.1 This scheme summarizes three classifications (histological, molecular, and functional) of breast cancer, including the most common subtypes of each class. 2

2.3.1 The block diagram above summarizes workflow of this thesis. . . . 11

3.3.1 Depiction of the components of the constructed network. A giant component of the network consists of 54 nodes and 382 edges. Besides, four small isolated components consist of 5, 3, 3 and 2 nodes which are connected by 10, 2, 2 and 1 edges, respectively. The whole network consists of 79 splicing factor proteins, 67 of them are connected together with 397 edges depending on coexpression and/or experimental data, while 12 nodes are disconnected. This network was constructed by STRING web server [1]. 15

3.4.1 Depiction of the extracted network's component. It consists of 54 nodes and 382 edges. In this component, nodes represent splicing factor proteins that are connected together with 382 edges depending on coexpression and/or experimental data. Nodes' color is mapped according to the number of mutations associated whereas nodes' size is mapped according to its degree. The edges are colored depending on the score of experimental interaction. Here, the sizes of the edges represent the score of the coexpression interaction. 19

4.1.1 Distributions plots of (a) time line of articles that included SF3B1 and one or more associated diseases from 2007 to 2018. (b) all countries performed these studies in this time period. Breast cancer (BC), MDS, AML, and CLL are colored in blue, green, red, and orange, respectively 22

4.2.1 Clustered bar charts show distribution of SF3B1 somatic mutations in four cancer types (MDS, AML, CLL, and breast cancer), based on numbers of mutations, samples, and PubMed articles which are colored in blue, red, and green, respectively 26

4.2.2 Clustered bar charts show number of hematologic and breast cancer patients had any of these mutations K700E, K141K, and K666N which are colored in blue, red, and green, respectively. 28

4.2.3 (a) The constructed network consists of 131 diseases are linked to SF3B1 by 472 edges, this list of diseases are reported from 1985 to 2018. (b) A filtered network including 44 diseases that are connected to SF3B1 by 349 edges. 30

4.2.4 The constructed network shows one cyan node represents SF3B1 gene. It is connected to 9 purple nodes represent variants and each of them is connected to one or more pink nodes representing disease. Overall, the network has 25 noes and 37 edges. 31

4.3.1 (a) The gene network contains 103 genes of ARMH family and 188 edges. They compose 16 components where the giant one has 56 genes and the second has 7 genes including SF3B1 are linked to 10 genes including SF3B1. (b) One component of 65 genes of some subfamilies of major spliceosome root family and 1321 edges. (c) One component of 8 genes of B-WICH family and 28 edges. All networks were built using Reactome FI plugin in Cytoscape. . . . 34

4.3.2 A clustered bar chart shows the number of patients who carried mutant genes from each subgroup of major spliceosome family. Breast cancer (BC), MDS, AML, and CLL that are colored in blue, green, red, and yellow, respectively. 36

4.4.1 The constructed network by Reactome FI plugin in Cytoscape shows in total 28 nodes and 15 edges. The network consists of 3 connected components. The giant one contains 7 nodes and two small components consist of 5 and 2 nodes. There are 14 disconnected nodes. Edge's size represents FI score, while the shape represents FI direction. Node's size reflects its degree and node's color is based on FI clustering. 38

4.5.1 Venn diagram shows number of common patients among each possible combination among four cancer types. All data are reported except number of patients among AML, CLL, and breast cancer (BC) which equals to zero. 39

4.5.2 The constructed network by Reactome FI plugin in Cytoscape shows in total 42 nodes and 92 edges. The network consists of 10 components. The giant one contains 37 nodes and 9 components each of them is one node. Edge's size represents FI score, while the shape represents FI direction. Node's size reflects its degree and node's color is based on FI clustering. 40

5.1.1 The depiction of the 3D structure of five domains of the SF3b complex included in this study. They were taken from the cryo-EM structure of human spliceosome complex (PDB:5z56) [2]. Each domain is shown in a different color. SF3B1, SF3B3, SF3B5, SF3B6, SF3B7, and pre-mRNA are shown in blue, red, yellow, grey, orange and green, respectively. The image was created using VMD [3]. . 47

5.4.1 Probability plots of RMSD distributions of (a) SF3B1^{wt} in blue vs. SF3B1^{K700E} in red. (b) pre-mRNA in the wild type structure in blue vs. mutant structure in red. (c) p14 in the wild type structure in blue vs. mutant structure in red. 52

5.4.2 Probability plots of distributions of number of contacts formed between (a) pre-mRNA and side chain of amino acid K700 in wild type shown in blue vs. E700 mutant shown in red, with cutoff distance ≤ 0.3 nm. (b) pre-mRNA and side chains of p14 residues $(20 - 100)$ in wild type shown in blue vs. mutant shown in red, with cutoff distance ≤ 3 nm. 53

5.4.3 2D projection of SF3B1 trajectories of wild type (blue) and two replicates of the mutant system (red and green) along the first two eigenvectors of SF3B1^{K700E} system. 54

5.4.4 RMSF per $C\alpha$ atoms of the first two eigenvectors of SF3B1 obtained for wild type protein (blue) and two replicates of the mutant system (red and green) 55

5.4.5 Dynamic cross-correlation map for SF3B1 is shown in (a) for the wild type system. As can be seen, there is a high positive correlation within the region $455-832$, which is colored in red, and a high negative correlation between two highlighted regions sequences as $455 - 832$ in red and $932 - 1300$ in blue. (b) For the mutant system, the positive correlation decreased drastically within the region $455 - 832$, which is colored in red, and low negative correlation between the two regions $455 - 832$ in red and $932 - 1300$ in blue. 56

5.4.6 3D structure of the two highlighted regions surrounding pre-mRNA in wild type system (a) and mutant system (b). α helical region $(455-832)$ in orange, α helical region $(932-1300)$ in green, both of pre-mRNA and K700E residue are shown in red representing the initial frame of simulation and blue color representing the last frame. The structural image was created using VMD [3]. 57

List of Tables

3.1 Network centrality metrics of six hub genes of the network 18

4.1 Top 10 genes associated with breast cancer, MDS, AML, and CLL based on number of PubMed articles 23

4.2 Top 18 mutated tissues with SF3B1 mutations based on number of mutations, samples, and PubMed articles, collectively. Data are listed in a descending order based on number of mutations. . . . 25

4.3 Top 20 mutations of SF3B1 that are frequently detected in tumor samples. 27

4.4 The association between each of the four diseases and SF3B1 or its family separately based on COSMIC database 33

4.5 The association between each of the four diseases and SF3B1 or its family separately based on DisGeNET and NetworkAnalyzer plugins 35

4.6 Reactome FI analysis for the gene network displayed in Figure 4.4.1 38

4.7 Reactome FI analysis for the gene network displayed in Figure 4.5.2 41

Abbreviations

Akt	Protein kinase B
AML	Acute Myeloid Leukemia
AMPK	AMP-activated protein kinase
ARMH	Armadillo-like helical domain containing
B-WICH	B-WICH chromatin-remodelling complex subunits
BPA	Branch point adenosine
CLL	Chronic Lymphocytic Leukemia
CMML	Chronic Myelomonocytic Leukemia
COSMIC	Catalogue Of Somatic Mutations In Cancer
Cryo-EM	Cryogenic Electron Microscopy
DB4S	DB Browser for SQLite
DCCM	Dynamic Cross Correlation Map
DHX15	DEAH-Box helicase 15
ER	Estrogen receptor
GDA	Gene-diseases association
GRACE	GRaphing and Advanced Computation and Exploration of data

GROMACS	GROningen MAchine for Chemical Simulations
HEAT	Huntingtin, elongation factor 3, a subunit of protein phosphatase 2A, PI3 kinase TOR
HER-2	Hormone ppidermal growth factor receptor 2
HGNC	HUGO Gene Nomenclature Committee
IDC	Invasive ductal carcinoma
ILC	Invasive lobular carcinoma
MAPK	Mitogen-activated protein kinase
MD	Molecular dynamics
MDS	Myelodysplastic Syndrome
NGS	Next-generation sequencing
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PI3K	Phosphatidylinositol 3-kinase
PME	Particle Mesh Ewald
PR	Progesterone receptor
PRPF19	Pre-mRNA-processing factor 19
RMSD	Root Mean Square Deviations
RMSF	Root Mean Square Fluctuation
RNP	Ribonucleoprotein
RRM	RNA Recognition Motif
SF3A2	Splicing factor 3A subunit

SF3B1	Splicing factor 3B subunit 1
snRNPs	Small Nuclear Ribonucleoproteins
SRSF1	Serine/arginine-rich splicing factor 1
SS	Splice site
TNBC	Triple-negative breast cancer
U2AF2	U2 auxiliary factor 2
U2snRNA	U2 spliceosomal RNA
VDA	Variant-diseases association

Chapter 1

INTRODUCTION & MOTIVATION

In this chapter, an overview about breast cancer will be presented. Then pre-mRNA splicing mechanism will be emphasized as one of the related biological processes to breast cancer development and/or prognosis. Lastly, motivation of this research will be explained.

1.1 Breast Cancer Overview

Breast cancer is a type of cancer in which cells of breast tissue grow abnormally. In some cases, breast cells go only under uncontrolled growth, leading to a benign tumor which is not lethal. Unlike other cases when abnormal cells spread out of breast tissue, invading other tissues and forming a malignant tumor (cancerous cells). Breast cancer has many types based on different ways of classifications such as; histological, molecular, and functional [4]. Most common subtypes of each category are shown in Figure 1.1.1.

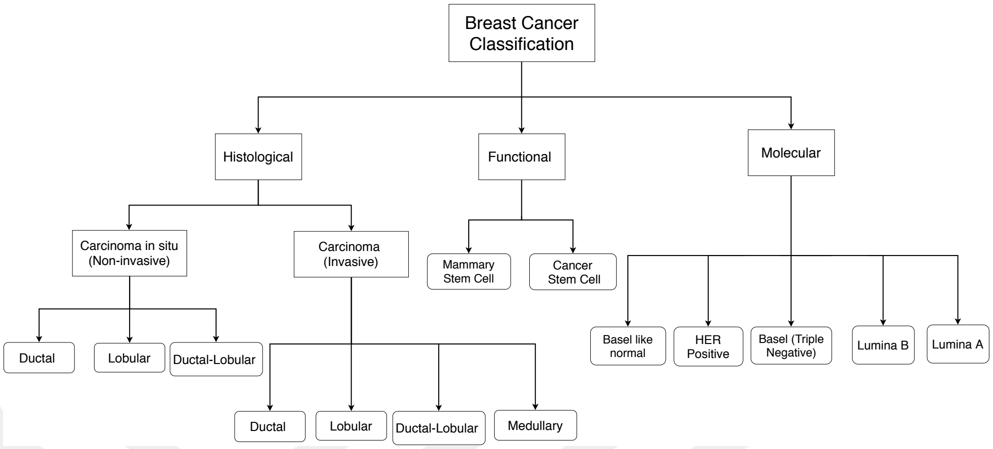


Figure 1.1.1: This scheme summarizes three classifications (histological, molecular, and functional) of breast cancer, including the most common subtypes of each class.

Breast cancer is considered as one of the most frequent and lethal cancer type among women [5] as it accounts for 25% of all cancer types [6] and contributes to 15% of the mortality [7]. The Catalogue Of Somatic Mutations In Cancer (COSMIC) database reported 12,542 samples were diagnosed by breast cancer. Carcinoma is the most common type as it was diagnosed in 11,294 samples. 2,512 and 1009 samples were diagnosed by invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC), respectively.Somatic mutations associated with carcinoma are mostly classified as pathogenic mutations. Unfortunately, in many cases, our knowledge on these mutations is limited to their allele frequencies and their association with formation of malignant tumors deserve further investigation.

1.2 Breast Cancer Etiology

The etiology of breast cancer is complicated as it is caused by many factors such genetic, epigenetic, environmental, lifestyle, and hormonal factors. [6, 8]. Somatic mutations are potentially involved in cancer development and prognosis. After the development of next-generation sequencing (NGS) technology, mutation analysis becomes easy and feasible for clinical studies of breast cancer. The most common harboring genes of somatic mutations are PIK3CA, TP53, GATA3, ESR1, and CDH1 [9–11]. Mutation frequency of each gene varies based on the tissue where breast cancer arises and its hormonal status. For instance, TP53 is the most frequently mutated gene associated with IDC, while CDH1 is the most associated gene with ILC. PIK3CA gene ranks at the second place for both IDC and ILC. Based on hormone status of breast cancer, estrogen receptor (ER) positive is the most common form of breast carcinoma. Additionally, germline and somatic mutations interact together to stimulate carcinogenesis in the case of triple-negative breast cancer (TNBC) in which the cancerous cells are tested negative for ER, progesterone receptors (PR), and hormone epidermal growth factor receptor 2 (HER-2) [12, 13].

1.3 Breast Cancer Related Biological Processes

A lot of research had been conducted to identify and understand many biological processes involved in cancer. Cell cycle proliferation and progression, cell communication, ER and HER2 signaling, and impairment of immune response are considered as main keys for cancer development and progression [14–16]. A pool of proteins interacts with each other and function in a complex system to accomplish a specific mechanism. Disturbing a function of any of them affects other proteins as well, leading to abnormal pathways as in cancerous cells condition. Complexity and diversity of the proteome content in a cell come from alternative splicing mechanism which produces many transcripts from same gene, each encodes for a different functional protein [17].

1.4 Pre-mRNA Splicing Mechanism

Alternative splicing occurs during pre-mRNA splicing mechanism when noncoding regions (introns) are spliced from pre-messenger RNA (pre-mRNA) sequence and coding regions (exons) are joined together to produce mature messenger RNA (mRNA). Splicing is the first step of processing of Capped Intron-Containing Pre-mRNA mechanism. It is followed by two more steps: a) 3'-end processing of pre-mRNA, and b) transportation of mature transcripts to the cytoplasm [18]. Splicing is carried out by a complex ribonucleoprotein (RNP) called spliceosome [19, 20] which consists of five small nuclear ribonucleoproteins (snRNPs) (U1, U2, U4, U5, and U6) and many other related proteins as well [21]. Two transesterification chemical reactions are necessary to catalyze splicing process [22]. During the first step, 2'-OH of branch point adenosine (BPA) attacks phosphodiester bond at 5' splice site (SS), generating free 5' exon. Then, 3'-OH of 5' exon attacks phosphodiester bond at 3' splice site (SS) to ligate both exons and release the intron in between [21]. U2 spliceosomal RNA (U2snRNA) interacts with branch sequence to facilitate the first attack of BPA. SF3b complex is essential at this step to reinforce and stabilize the interaction between U2snRNA and BS by interacting with pre-mRNA near or at BS [23–26]. SF3b complex is the largest component of U2 small nuclear ribonucleoprotein (U2snRNP). It is composed of seven subunits: (SF3B1/SF3b155, SF3B2/SF3b145, SF3B3/SF3b130, SF3B4/SF3b49, SF3B5/SF3b10, SF3B6/p14, and PHF5A/SF3b14b) [22]. Splicing factor 3B subunit 1 (SF3B1) is the largest component of the SF3b complex [27, 28]. It is composed of N-terminal helix-loop-helix and C-terminal of 20 HEAT-repeat domain (Huntingtin, Elongation Factor 3, Protein phosphatase 2A, Target of rapamycin 1) [2, 22, 29]. SF3B1 interacts with other subunits i.e., SF3B3, SF3B7, and SF3B5, all of which maintain the conformation of the protein and facilitate its binding to pre-mRNA. In addition, SF3B1 also interacts with SF3B6 (also known as p14) [22, 30, 31] forming a complex p14-SF3B1 with U2 auxiliary factor U2AF65 to recognize and select the branch splicing site [24, 26, 31]. In particular, the superhelical heat domain of SF3B1 increases accessible surface area of the protein [32], which allows for effective binding of interacting partners and/or RNAs to the protein.

Most of SF3B1 associated point mutations cluster at the C-terminal of heat domain. They are frequently detected in patients who have myelodysplastic syndrome (MDS), chronic lymphocytic leukemia (CLL), acute myeloid leukemia

(AML) [33], uveal melanoma [34], breast cancer [35] and other cancer types. The association between these mutations and aberrant splicing at the molecular level is not fully understood. Aberrant splicing leads to production of abnormal transcripts. Consequently, these transcripts follow two possible pathways: 1) translation of mRNA into proteins with truncated or missing domain(s) which consequently leads to change, loss or gain of function of the protein [36] and 2) degradation by nonsense-mediated mRNA decay (NMD) mechanism which leads to reduction in the proteome content of the cell [37].

1.5 Motivation

Breast cancer has the highest frequency among women compared by other cancers. Hence, billions of dollars had been spent on breast cancer research around the world to find more efficient approaches for prevention, diagnosis, and treatment. However, it is still a challenging disease to be diagnosed and treated due to its complicated etiology. Thousands of somatic mutations had been detected using sequencing techniques and reported in online accessible databases. However, our knowledge about these mutations is limited to their allele frequencies and their relations to cancer deserve further investigation. Somatic mutations, which cause aberrant splicing, are considered as one of the driving events for cancer development and prognosis. It was found that point mutations of SF3B1 cause misrecognition and misselection of branch point splicing via an elusive mechanism, producing abnormal spliced transcripts. Therefore, splicing factor 3B subunit 1 (SF3B1) was selected to be analyzed. Thus, an overview is provided about SF3B1 and its interactions with other genes in complex systems. At the molecular level, the impact of SF3B1^{K700E} was assessed to get mechanistic insights into aberrant splicing. Combined in silico based study in this thesis, though needs validation by experimental studies, can be improved and applied to characterize other highly associated somatic mutations with cancer. Consequently, identification of potential therapeutic targets may be improved.

1.6 Conclusion

In this chapter, breast cancer was briefly introduced, followed by mRNA splicing mechanism as one of the driven events of cancer development and prognosis. Next chapter will present a summary of recent experimental and computational studies about mutations of spliceosome complex's components and their associations with aberrant splicing, leading to cancer. It is infeasible to study the whole spliceosomal machinery due to its complexity as mentioned in Section 1.4. Therefore, in Chapter 3, bioinformatics tools will be utilized to identify one splicing factor as a target for this thesis. After that, the target protein was analyzed by using a combined in silico approach based on i) data and network analysis and ii) computational biophysics as explained in detail in Chapters 4 & 5, respectively. In Chapter 4, different biological networks were constructed to study SF3B1 as a component of complex systems. Besides, connections of SF3B1 with different genes, pathways, and diseases were investigated. In Chapter 5, by utilizing means of molecular dynamics simulations, SF3B1 was studied structurally and dynamically in wild type vs. mutant complexes. Chapter 5 findings provided mechanistic insights into the impact of the K700E mutation on splicing mechanism. Finally, the thesis will be concluded with the final chapter, stating the summary and the future work.

Chapter 2

LITERATURE REVIEW

Many researchers studied some components of spliceosome complex and their association with cancer development and prognosis. They focused on the mechanism of somatic mutations of splicing factors and how it is linked to aberrant splicing. Here, a summary of some recent experimental and computational based studies are presented.

2.1 Experimental Studies That Proposed SF3B1 as a Therapeutic Target for Cancer

In 2015, papasaikas et al. used experimental approach to analyze the impact of knocking down spliceosome components on cell proliferation and apoptosis. Then they used this information to reconstructed a network of functional interactions among splicing factors to highlight therapeutic targets [38] In 2019, a research group performed RNA sequencing analysis for MDS patients. They found that mutant SF3B1 uses 3' alternative splice site (SS). In addition, the research determined that expression level of splicing factor 4 (SUGP1) was reduced in cells containing SF3B1^{K700E} mutation. The group studied other mutations of SF3B1, not only K700E and got similar results of aberrant splicing [39]. In [40], Jiménez-Vacas *et al.* collected samples from Spanish prostate cancer patients to measure and analyze mRNA and protein expression levels of SF3B1. They found SF3B1 was highly expressed which is associated with characteristics of cancerous cells such as metastasis. They confirmed their results by treating prostate cells by pladienolide-B which inhibits the activity of SF3B1, leading to reduction of cell proliferation, migration, tumor formation, and apoptosis. In another paper, the same authors obtained similar results from endocrine-related cancer patients. They found SF3B1 was overexpressed in cell lines of pituitary tumors, pancreatic neuroendocrine tumors, and prostate cancer patients and in silico cohort of liver cancer. In both studies, reducing SF3B1 activity leads to i) inhibition of PI3K/AKT and JNK signaling pathways, ii) modulation of a tumor marker and a splice variant; AR-v7 and In1-ghrelin, respectively, and iii) regulation of expression levels of complexes' components such as spliceosome, SURF, and EJC [41]. Another research group studied the connection between hotspot mutations of SF3B1 and breast cancer development. They analyzed RNA sequencing data of breast epithelial cells with CRISPR knock-in of K700E mutation. They found the same results of using alternative 3' SS, high expression levels of genes involved in NF-$_k$B pathway that controls transcription of DNA and cell survival. Also, they found that SF3B1^{K700E} mutation is associated with mutations of other genes involved in PI3K pathway [42]. All these results propose that SF3B1 can be a new prognostic biomarker and a therapeutic target in many cancer types.

2.2 Computational Approaches to Investigate the Impact of Somatic Mutations on Splicing

Thanks to determination of structures of the yeast spliceosome complex and the human complex in three conformational states using cryogenic electron microscopy (cryo-EM), computational biophysicists were directed to study the spliceosome complex at the molecular level by means of MD simulations. In 2019, Borišek *et al.* built eight models to the spliceosome B^{act} from yeast *Saccharomyces cerevisiae* (PDB: 5GM6) using MD simulations, along with gene sequencing of

SF3B1. They included the following components: Prp8 (same for yeast and Homo sapiens), SF3b components such as Rds3, Ysf3, and Hsh155 (corresponding to PHF5A, SF3B5, and SF3B1 in human, respectively), five RNA sequences (U2, U5, and U6, intron and exon), four Mg2+, and three Zn2+ ions. For mutant systems, they induced K335E, N295D, and L378V mutations in Hsh155. Additionally, they induced non-consensus BPS in the pre-mRNA, A at position -1 with respect to BPA is mutated to C and U at position -2 is mutated to C. Their findings demonstrated Hsh155 can recognize and bind to both consensus and non-consensus BPS that could be possible in alternative splicing mechanism. All possible nucleotides that are considered as possible branch points are recognized by Hsh155 from phosphate group which is the same in all possibilities except BPA (A1) as they found a mismatch in the mutant complex with BPS (A-1C or U-2C). Consequently, these mutations weaken the interaction between intron and U2snRNA. Hsh155 mutations (K335E or N295D) change intron binding in case of non-consensus BPS, while they do not affect in complex with consensus BPS. Additionally, cross correlation among Hsh155 residues showed different pattern, indicating that the dynamics of the protein is changed upon K335E mutation. N295D showed less significant changes but it decreased the stability of the binding between Rds3 (PHF5A in human) and the intron [43]. Another MD study was performed by Borišek *et al.* to investigate the impact of splicing modulators molecules (pladienolides, herboxidienes, and spliceostatins) on SF3b complex. They performed MD simulations for eight models obtained from the human structure. These models include SF3b in two forms (apo and in a complex with splicing modulators), where each of them was represented by wild type and mutant states. They induced three mutations: Y36C/R38C in PHF5A and R1074H in SF3B1. Their analysis showed that i) SMs flow and fit in the narrow region between SF3B1 and PHF5A which is considered as BPA-binding or recognition site. Consequently, they hinder the opening and closing motions of SF3B1, trap it in open state, and prevent BPA recognition; ii) Both cross correlation among SF3B1 residues and its functional plasticity were reduced; and iii) Pladienolide had a more significant impact on SF3B1 structure and dynamics than the other two modulators. These results proposed SF3B1 as a target of these small molecules for cancer therapy [44].

2.3 Thesis Contribution

Previous studies revealed the significant association of SF3B1 with cancer. However, most of them focused on only SF3B1 or few components of spliceosome. Accordingly, in this thesis, SF3B1 was analyzed in different network to get insights into its interactions in complex systems. At the molecular level, the mechanism behind aberrant splicing that occurs due to SF3B1-associated mutation is not fully understood. Therefore, bioinformatics tools were used to provide a wide-scope analysis of a highly mutated gene in cancer (SF3B1). Additionally, computational biophysics tools were used to assess the impact of its most hotspot mutation K700E on the protein's structure and dynamics. The workflow of our work is summarized in a block diagram as shown in Figure 2.3.1. As shown in the diagram, this thesis is mainly divided into three phases as follows:

1. Identification and prioritization of a highly mutated gene that is potentially related to breast cancer development.
2. Data and network analysis of a highly mutated gene by utilizing available databases and bioinformatics tools.
3. Assessment of the impact of the most frequent mutation of a highly mutated gene (SF3B1 ^{K700E}) by means of molecular dynamics simulations.

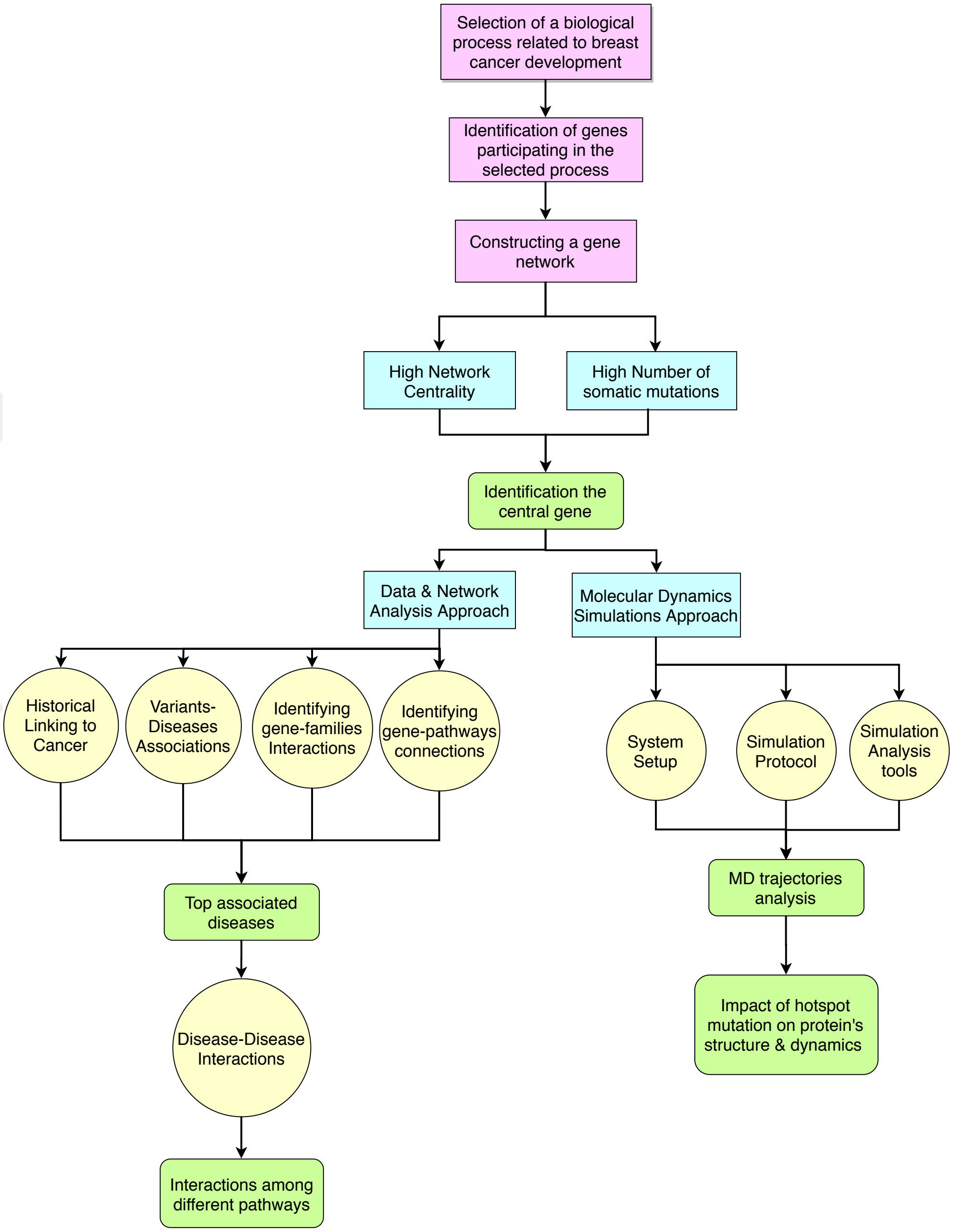

Selection of a biological process related to breast cancer development
Identification of genes participating in the selected process
Constructing a gene network
High Network Centrality
High Number of somatic mutations
Identification the central gene
Data & Network Analysis Approach
Molecular Dynamics Simulations Approach
Historical Linking to Cancer
Variants-Diseases Associations
Identifying gene-families Interactions
Identifying gene-pathways connections
System Setup
Simulation Protocol
Simulation Analysis tools
Top associated diseases
MD trajectories analysis
Disease-Disease Interactions
Impact of hotspot mutation on protein's structure & dynamics
Interactions among different pathways

Figure 2.3.1: The block diagram above summarizes workflow of this thesis.

publicly available databases and open source tools were used to analyze SF3B1 as explained in Chapters 3 & 4:

1. March 2019 release of the Catalogue of Somatic Mutations in Cancer (COSMIC) [45]: to collect somatic mutation data including many attributes such

as; tissue type, nucleotide and amino acid sequence changes, mutation's type, PubMed PMID of study reporting each mutation, patient ID where the tumor sample was taken, etc.

2. The HUGO Gene Nomenclature Committee (HGNC) [46]: to get a list of gene from same family.
3. Reactome database: to collect all genes participating in mRNA splicing mechanism [47].
4. STRING database: to build a gene network based on direct (physical) and indirect (functional) associations [1].
5. DB Browser for SQLite (DB4S): to import tables of COSMIC data and execute SQL queries for data filtration, processing, and combination [48].
6. Cytoscape: an open source bioinformatics software platform to build and visualize different networks such gene-gene interactions and genes-diseases interactions. Also, some available features “plugins" were used [49] such as:
   (a) Reactome FI: a Functional Interaction Network to identify common pathways, biological processes and molecular functions for a set of genes [50].
   (b) DisGeNET: to build gene-diseases associations networks [51]
   (c) NetworkAnalyzer: for subnetwork creation and network analysis by calculating topological parameters and centrality measures [52].

As shown in Figure 2.3.1, work flow has three main phases. Next chapter will present the first phase which starts by selecting a process drives cancer development and ends by identifying the central gene among other genes participating in the same process.

Chapter 3

IDENTIFICATION OF A HIGHLY MUTATED GENE THAT CONTRIBUTES TO THE DEVELOPMENT OF BREAST CANCER

Breast cancer was selected as the first step of the thesis workflow as it is the most frequent cancer type among women. Since there is a significant number of somatic mutations associated with the development of breast cancer, it is not feasible to conduct molecular dynamics simulations to asses the impact of each of these mutations on the function of the target protein. Hence, the following steps were performed to identify a central gene among others participating in the same biological process.

1. Selection of a biological process potentially involved in breast cancer (e.g. harboring highly mutated genes).

2. Identification of genes participating in this biological process.

3. Constructing a gene network
4. Identification of a gene having high network centrality and number of mutations.

3.1 Selection of a Biological Process Potentially Involved in Breast Cancer

Previous studies showed that mutated genes, which are involved in cancer [53], encode abnormal proteins that interfere with various vital biological processes such as cell proliferation, transcription, splicing, immune response [15], cell communication, and apoptosis [16]. Among them, aberrant splicing is considered as one of the most significant driving factors for development of cancer [54], as it leads to production of proteins with altered function. Therefore, “pre-mRNA splicing” is selected as the target biological process in this study.

3.2 Identification of Genes Participating in pre-mRNA Splicing

Reactome database was used [47] to collect all genes participating in pre-mRNA splicing mechanism. First, 3734 results were obtained without using any filter. Then, by specifically indicating the species as "homo sapiens" and the type as "protein", the number of results was decreased down to 86 genes.

3.3 Building a Gene Network

STRING database [1] was used to build a network of genes that code 86 splicing factors obtained in the previous step. It consists of five connected components that have 67 nodes and 12 disconnected nodes along with 397 edges as shown in Figure 3.3.1, where the nodes and edges represent genes and experimental and/or coexpression interactions, respectively.

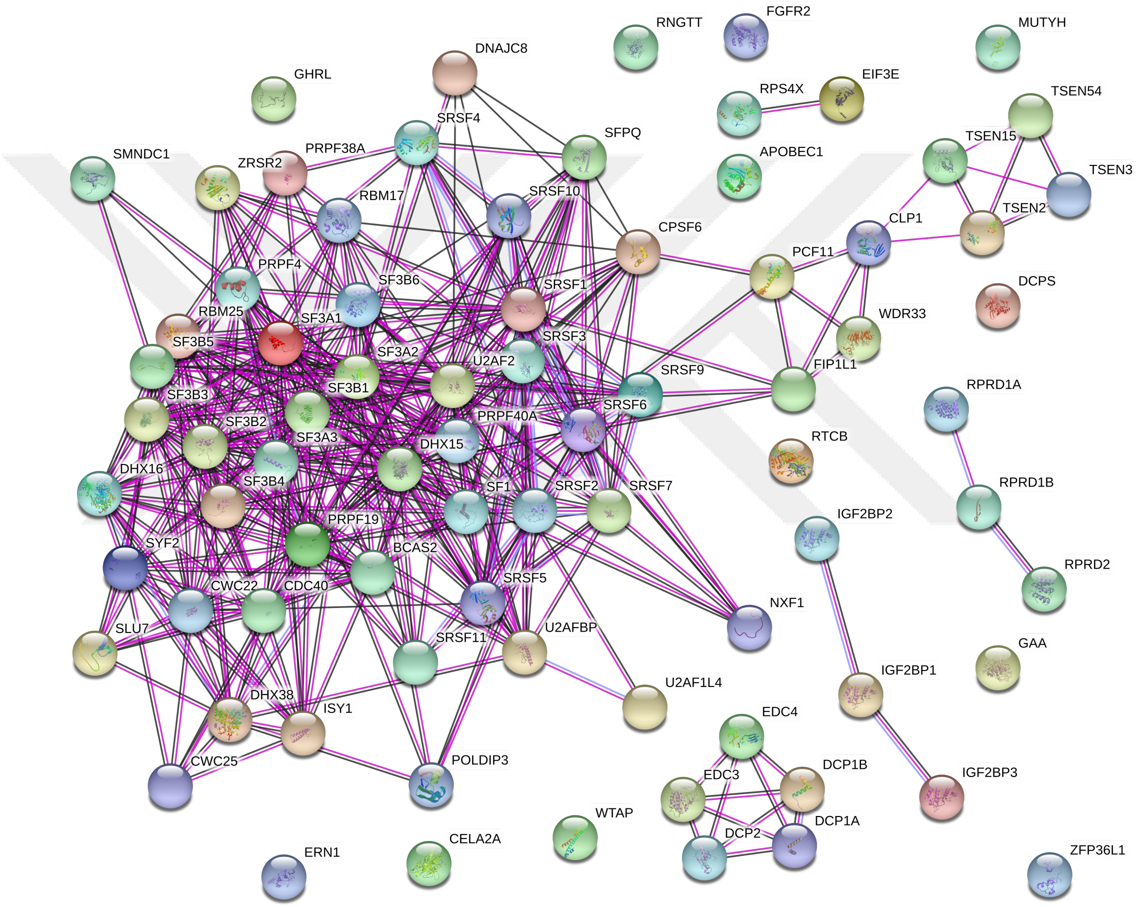


Figure 3.3.1: Depiction of the components of the constructed network. A giant component of the network consists of 54 nodes and 382 edges. Besides, four small isolated components consist of 5, 3, 3 and 2 nodes which are connected by 10, 2, 2 and 1 edges, respectively. The whole network consists of 79 splicing factor proteins, 67 of them are connected together with 397 edges depending on coexpression and/or experimental data, while 12 nodes are disconnected. This network was constructed by STRING web server [1].

3.4 Identifying a Gene with High Network Centrality and Number of Mutations

In order to analyze the resultant network, NetworkAnalyzer plugin was used [52] which is provided by Cytoscape software [49]. NetworkAnalyzer computes topological parameters and centrality measures such as the number of nodes and edges, degree distribution, clustering coefficient, neighborhood connectivity, betweenness, and closeness regarding the network. In order to select the target gene, a giant component that contains a majority of network's nodes was extracted. Some network metrics were considered for network's evaluation such as the degree centrality and measures related to distance such as closeness, radiality, and eccentricity along with number of mutations associated with each gene found in the network. Some genes showed high centrality measures but they are not frequently mutated. Therefore, number of mutations was considered as an essential parameter to identify the hub gene. COSMIC Mutation database (March 2019 release) [45] was used to retrieve corresponding information about each gene that encode splicing factors. To obtain more significant results, the database was filtered based on the following criteria:

1. In COSMIC, there are six histology selections for breast cancer. Among them, the carcinoma was selected as it is the most common histological type [4, 55].

2. According to COSMIC statistics, 82% of breast carcinoma patients have somatic mutations of type "Substitution-Missense".

3. COSMIC uses Functional Analysis through Hidden Markov Models (FATHMM) as a predictive tool for the mutation's impact. FATHMM classifies any mutation as one of the following classes; pathogenic, neutral or not specified. Only pathogenic mutations were included.

4. FATHMM scores are given in the form of p-value that ranges from 0 to 1. The pathogenic mutations have scores greater than 0.5. The threshold was increased to 0.9 and beyond to get the most possible pathogenic mutations.

From the processed database, the corresponding number of mutations associated with each gene in the network's component were retrieved. Then, they were superimposed on the network's component which was analyzed by Cytoscape.

Analysis of the resultant subnetwork revealed that SF3B1 is associated with 25 edges correspond to combined interactions of experimental and coexpression. As distance measures, SF3B1's closeness, radiality, and eccentricity equal to 0.56, 0.87 and 5, respectively as. As shown in Table 3.1, SF3B1 gene ranked in the fifth regarding degree and the sixth order regarding closeness, and radiality. Eccentricity values range from 3 to 6 with an average of 4.9. The other genes that had slightly higher network metrics were splicing factor 3A subunit 2 (SF3A2), DEAH-Box helicase 15 (DHX15), U2 auxiliary factor 2 (U2AF2), Pre-mRNA-processing factor 19 (PRPF19), and Serine/arginine-rich splicing factor 1 (SRSF1). SF3A2 is the only component of SF3a complex that detects the U2/BS duplex [2, 56]. DHX15 is required to catalyze the disassembly of spliceosome complex after the completion of splicing and releasing mature mRNA to cytoplasm [57–59]. U2AF2 recognizes and binds to the polypyrimidine (Py) tract [60]. PRPF19 (also known as PRP19) interacts with other components of the NineTeen complex (NTC) which is essential for NTC stability [61]. NTC complex strengthens the interactions between mRNA and snRNAs [62]. SRSF1 (also known as SF2/ASF) is not a component of spliceosome complex but it regulates pathways of mRNA metabolism mechanisms such as mRNA splicing [63], export [64] and translation [63] and many other biological processes.

Table 3.1: Network centrality metrics of six hub genes of the network

	Degree	**Closeness**	**Radiality**	**Eccentricity**
Min	1	0.210744	0.375817	3
Max	33	0.62963	0.901961	6
Average	11.73077	0.458276	0.783308	5.076923
SF3A2	33	0.62962963	0.9019608	5
PRPF19	26	0.57954545	0.879085	5
SF3B1	25	0.56043956	0.8692811	5
U2AF2	22	0.55434783	0.8660131	5
DHX15	21	0.5483871	0.8627451	5
SRSF1	18	0.56666667	0.872549	4

After superimposing number of associated mutations on the network's component, SF3B1 emerged as the most frequently mutated gene (22 highly pathogenic missense mutations) among the other splicing factors included in the subnetwork. SF3A2, DHX15, U2AF2, PRPF19, and SRSF1 had 1, 13, 2, 6, and 2 highly pathogenic missense mutations, respectively. By considering both network centrality metrics and number of mutations, the Splicing factor 3B subunit 1 (SF3B1) gene emerged as the central node in the network's component as shown in Figure 3.4.1.

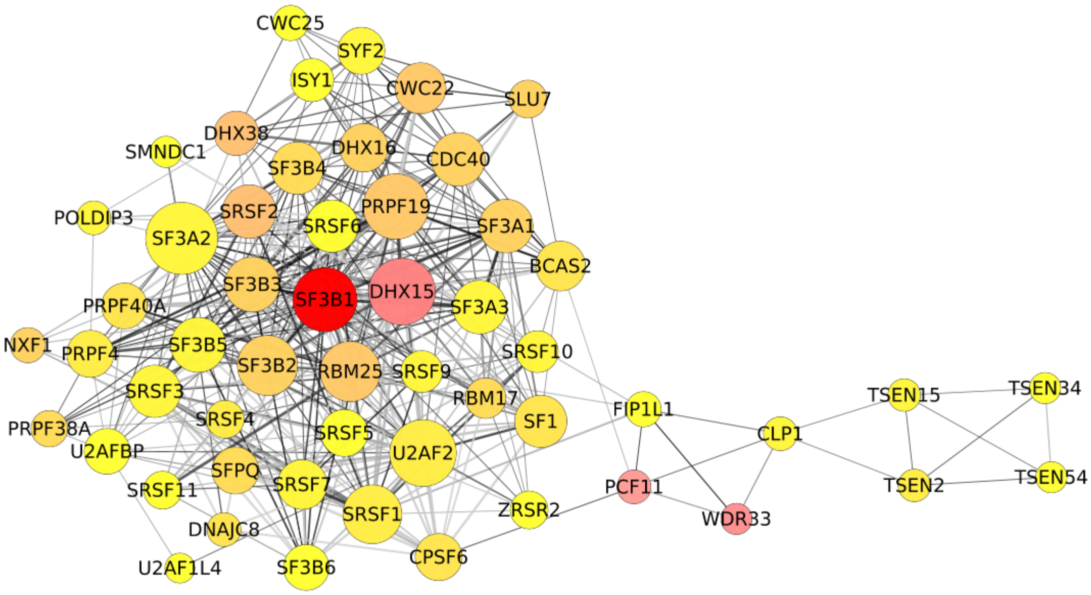


Figure 3.4.1: Depiction of the extracted network's component. It consists of 54 nodes and 382 edges. In this component, nodes represent splicing factor proteins that are connected together with 382 edges depending on coexpression and/or experimental data. Nodes' color is mapped according to the number of mutations associated whereas nodes' size is mapped according to its degree. The edges are colored depending on the score of experimental interaction. Here, the sizes of the edges represent the score of the coexpression interaction.

3.5 Conclusion

By considering both network centrality measures and number of mutations, SF3B1 emerges as a central node among other genes participating in mRNA splicing mechanism. Therefore, bioinformatics and computational biophysics tools will be used in Chapters 4 & 5 to characterize SF3B1 at the phenotypic and molecular levels, respectively.

Chapter 4

DATA & NETWORK BASED ANALYSIS OF SF3B1

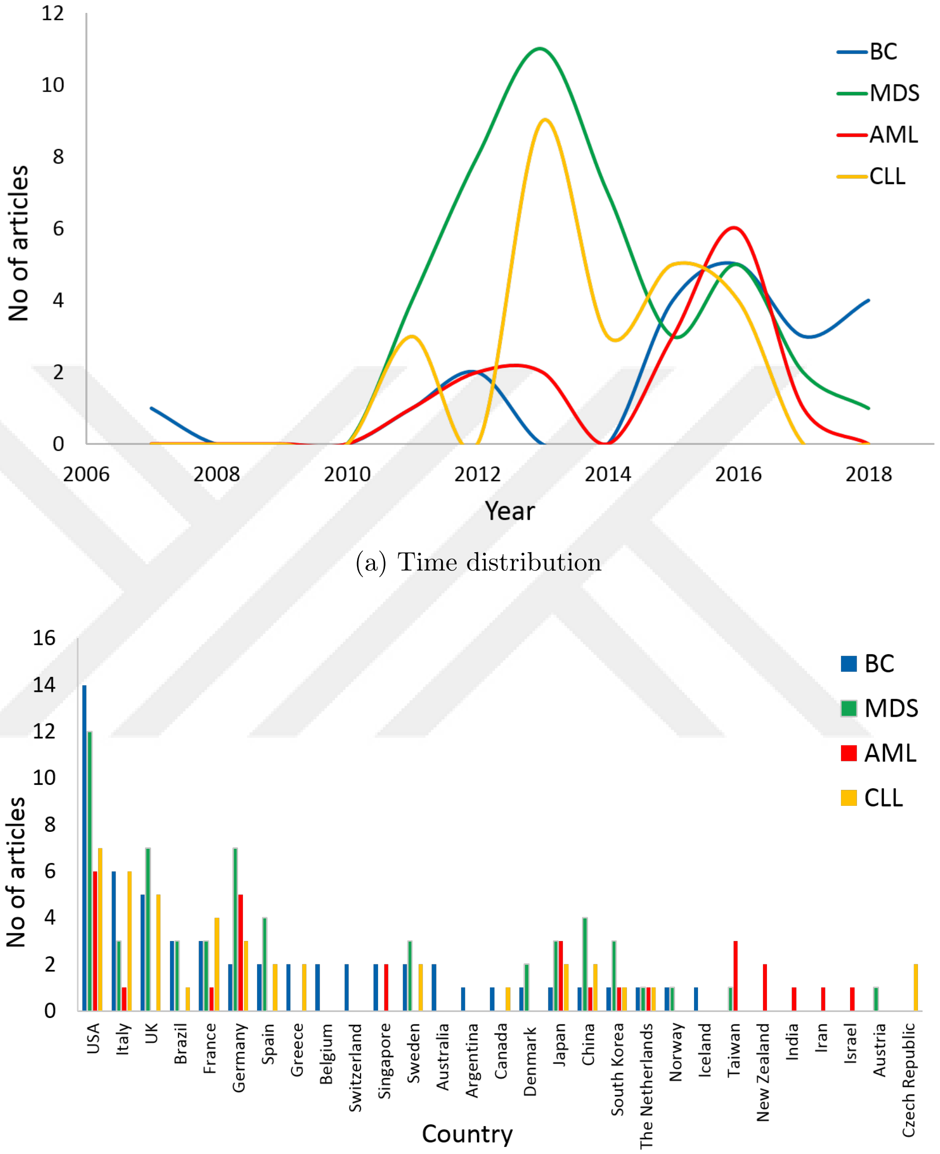
This chapter provides a wide-scope analysis of SF3B1 and its association with two types of cancer: i) breast cancer which is the starting point of this thesis and ii) hematologic cancer as it the most associated cancer type with SF3B1 mutations, including three histological subtypes; Myelodysplastic syndrome (MDS), acute myeloid leukemia (AML), and Chronic lymphocytic leukemia (CLL). The obtained results cover the following aspects:

1. Linking SF3B1 to hematologic and breast cancers from historical perspective.
2. Investigating mutations associated with SF3B1 and the related diseases.
3. Identifying linked families to SF3B1 and comparing SF3B1-cancer association vs. Families-cancer association.
4. Identifying connections between splicing mechanism and other essential pathways in the cell.
5. Identifying relations among MDS, AML, CLL, and breast cancer.

4.1 Linking SF3B1 to Cancer: A Historical Perspective

In this section, COSMIC SF3B1 data were analyzed from a historical perspective (spatial and temporal) for two purposes: i) showing the trend of interest for studying SF3B1 as an associated gene with MDS, AML, CLL, and breast and ii) listing all countries studied this association. Additionally, count of distinct “PubMed_PMID" for each gene associated with each disease was retrieved and ordered descending to compare it with SF3B1. The retrieved data showed that SF3B1 was reported from year 2007 to 2018 as shown in Figure 4.1.1a. In 2007, SF3B1 was mentioned in one article with breast cancer, followed by a steady period with zero articles from 2008 to 2010. In 2011, number of publications increased gradually, especially for MDS and CLL, and reached its peaks in 2013 at 11 and 9 articles, respectively. Regarding breast cancer and AML, their curves had peaks in 2016 at 5 and 6 articles, respectively.

By comparison, SF3B1 received the most attention in literature as an associated MDS gene. It was mentioned in 42 articles from 2011 till 2018, collectively. Then CLL, breast cancer, and AML are placed by 24, 20 and 15 articles, respectively. Additionally, 29 countries contributed to SF3B1 studies as shown in Figure 4.1.1b. Interestingly, only 8 countries (USA, Italy, France, Germany, Japan, China, South Korea, and The Netherlands) study SF3B1 with all four cancers. Overall, authors who affiliated with United States published the highest number of articles. East Asian countries (Japan, China, South Korea, and Taiwan) studied SF3B1 with hematologic cancer in 24 articles and only 3 articles for breast cancer. On the other hand, distribution of European countries is consistent among MDS, BS, and CLL with 31, 30 and 27 articles, while AML-SF3B1 association was studied in 8 articles, representing the lowest interest



(a) Time distribution

(b) Countries' distribution

Figure 4.1.1: Distributions plots of (a) time line of articles that included SF3B1 and one or more associated diseases from 2007 to 2018. (b) all countries performed these studies in this time period. Breast cancer (BC), MDS, AML, and CLL are colored in blue, green, red, and orange, respectively

As shown in Table 4.1, among four cancer types, SF3B1 did not receive equally distributed effort from researchers. Table 4.1 also shows that the overall effort from researchers to study SF3B1 with MDS and CLL seemed significant enough in comparison with top genes. On the other hand, it is placed after 10 genes for both breast cancer and AML, receiving much less attention in literature. TP53 was studied the most for breast cancer and CLL, showing significant association to both cancers. Most of articles studied both MDS and AML, therefore there are 6 common genes (JAK2, NRAS, IDH1, IDH2, DNMT3A, and RUNX1) out of top 10 between both. However, the highest genes for MDS and AML are JAK2 (Janus Kinase 2) and FLT3 (Receptor-type tyrosine-protein kinase), respectively. The first controls growth of red blood cells, white blood cells, and platelets [65], while the latter is involved in regulation of hematopoiesis [66].

Table 4.1: Top 10 genes associated with breast cancer, MDS, AML, and CLL based on number of PubMed articles

Breast Cancer		MDS		AML		CLL	
Gene	**No of PubMed Articles**	**Gene**	**No of PubMed Articles**	**Gene**	**No of PubMed Articles**	**Gene**	**No of PubMed Articles**
TP53	171	JAK2	62	FLT3	469	TP53	32
PIK3CA	118	TET2	46	NPM1	159	NOTCH1	30
PTEN	54	SF3B1	38	NRAS	120	ATM	23
CDH1	48	NRAS	34	IDH2	77	SF3B1	21
AKT1	45	ASXL1	33	CEBPA	74	C11orf65	14
KRAS	45	IDH1	30	KIT	72	BIRC3	12
ESR1	41	DNMT3A	28	IDH1	68	XPO1	12
MED12	41	IDH2	28	DNMT3A	63	BRAF	11
RB1	36	RUNX1	27	JAK2	57	MYD88	9
BRAF	33	TP53	25	RUNX1	54	BTK	6
SF3B1	19			SF3B1	13		

4.2 Investigating SF3B1-Associated Mutations and the Related Diseases

SF3B1 somatic mutations were detected in many cancer types [33]. Accordingly, COSMIC mutation raw data of SF3B1 were processed and analyzed by DB4S to i) compare among all affected tissues, ii) list associated mutations with high rates, and iii) show distribution of hotspots mutations over our target cancers. COSMIC reports 37 attributes in each table, among them, five attributes were counted using “COUNT" clause. These returned numbers represent statistical parameters to evaluate mutations frequencies. These attributes are as the following:

1. AAmutation: the change occurred at the amino acid sequence.

2. SampleName: the cell line where the tumor was detected.

3. PubmedId: PubMed ID for the paper that the sample was reported.

4. PrimaryTissue: the primary tissue/cancer from which the sample originated.

5. HistologySubtype1: level 1 of histological classification of the tumor sample.

COSMIC classified all hematologic cancers in one tissue “Haematopoietic and lymphoid". So, histological classification attribute was used to report specific data for MDS, AML, and CLL separately. According to COSMIC, 586 somatic mutations of SF3B1 were detected in 2552 samples, originated in 30 tissues, and reported in 226 articles. Top 20 tissues were retrieved based on each attribute (count of mutations, samples, and PMID), separately. By combining all outputs, soft tissue and pleura were excluded as they have low number of papers. Similarly, salivary gland and bone excluded as they have high number of papers but low numbers of mutations and samples. The remaining 18 tissues are listed in Table 4.2. Haematopoietic and lymphoid is the top mutated tissue based on three attributes which is consistent with previous section’s results and literature. Breast, skin, lung, and large intestine are placed from 2^{nd} and 5^{th} with different order based on each attribute, separately. Interestingly, eye tissue is ranked second based on number of samples, but it is only mentioned in 9 articles. Also, urinary tract and liver have relatively high number of mutations and patients, however they did not receive enough attention in literature.

Table 4.2: Top 18 mutated tissues with SF3B1 mutations based on number of mutations, samples, and PubMed articles, collectively. Data are listed in a descending order based on number of mutations.

Tissue	**No of mutations**	**No of samples**	**No of PubMed articles**
Haematopoietic and lymphoid	133	1659	99
Large intestine	98	97	14
Lung	79	83	17
Skin	66	97	18
Breast	54	97	21
Urinary tract	32	45	4
Endometrium	31	29	4
Liver	30	37	3
Central nervous system	25	22	8
Kidney	23	22	5
Prostate	21	27	11
Biliary tract	18	23	7
Eye	17	116	9
Pancreas	16	31	6
Oesophagus	14	15	8
Stomach	14	14	5
Thyroid	13	18	5
Upper aerodigestive tract	11	10	7

Haematopoietic and lymphoid tissue has 28 histological classifications, among them, MDS, AML, and CLL are the most common subtypes to be compared with breast cancer. In Figure 4.2.1, frequency of four cancer types with respect to numbers of mutations, samples, and PubMed articles.

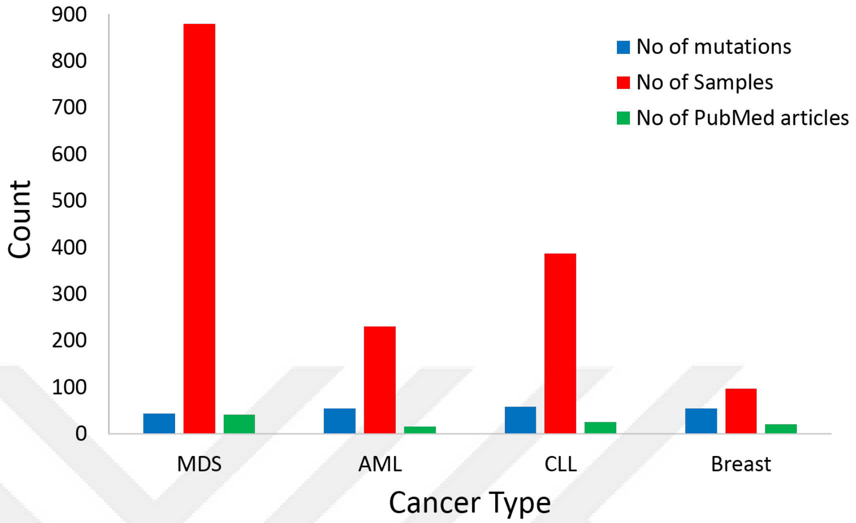


Figure 4.2.1: Clustered bar charts show distribution of SF3B1 somatic mutations in four cancer types (MDS, AML, CLL, and breast cancer), based on numbers of mutations, samples, and PubMed articles which are colored in blue, red, and green, respectively

Total number of entries in SF3B1 table is 2938, while number of samples is 2552. There were 260 samples repeated more than once, making in total 646 entries. These repeated data showed that all repetition except three patients were due to more than one article noted them with multiple mutations were found in same tissue. Haematopoietic and lymphoid was repeat 495 times out of 646, followed by large intestine, breast, lung, and skin that were repeated 42, 22, 14, and 12 times respectively. These values indicate that a group of mutations may be needed to cause hematologic cancer in one patient, while less number of mutations can be enough to cause other cancer types such as breast, lung and skin. The three exceptional samples were common among haematopoietic and lymphoid:

- One patient was diagnosed by chronic myelomonocytic leukemia (CMML) and AML associated with MDS, having p.K700E in both disease.
- One patient was diagnosed by CMML and MDS, having p.K700E and p.K666N associated with each disease, respectively.
- One patient was diagnosed by MDS and AML, having p.K666R and p.H738Y associated with each disease, respectively.

By definition, hotspot mutation means it frequently occurs in tumor samples [67]. Accordingly, 20 hotspot mutations were identified based on number of samples in Table 4.3. Numbers of tissues and articles were also counted, although they were not taken into consideration. Over all tissue, p.K700E is the most hotspot mutated residue (wild type lysine at 700 residue is substituted by mutant glutamic acid) with respect to three attributes. In addition, K666 is considered as a hotspot mutated residue with 6 different types of substitutions. Interestingly, p.K141K ranked second by affecting 114 samples, but it was reported in only 4 articles.

Table 4.3: Top 20 mutations of SF3B1 that are frequently detected in tumor samples.

Mutation	**No of samples**	**No of tissues**	**No of PubMed articles**
p.K700E	782	14	92
p.K141K	114	5	4
p.K666N	110	6	38
p.V1219V	92	3	3
p.G877G	91	2	2
p.R625H	88	9	19
p.R625C	82	11	34
p.G742D	75	6	26
p.E622D	70	5	34
p.H662Q	66	4	29
p.K666T	37	6	28
p.K666R	33	2	17
p.R625L	33	6	24
p.K666E	28	3	16
p.K666Q	20	3	16
p.H662D	17	1	11
p.K666X	17	1	3
p.G740E	16	4	12
p.Y623C	16	3	11
p.E902K	15	2	5

For detailed results, the first three mutations were analyzed with respect to the four cancer types. As shown in Figure 4.2.2, among three hotspots, breast tissue is linked to only K700E mutation. Although K141K, V1219V, and G877G rank in the second, forth, and fifth orders, respectively in Table 4.3, they were detected only in 5, 3, 2 tissues, respectively.

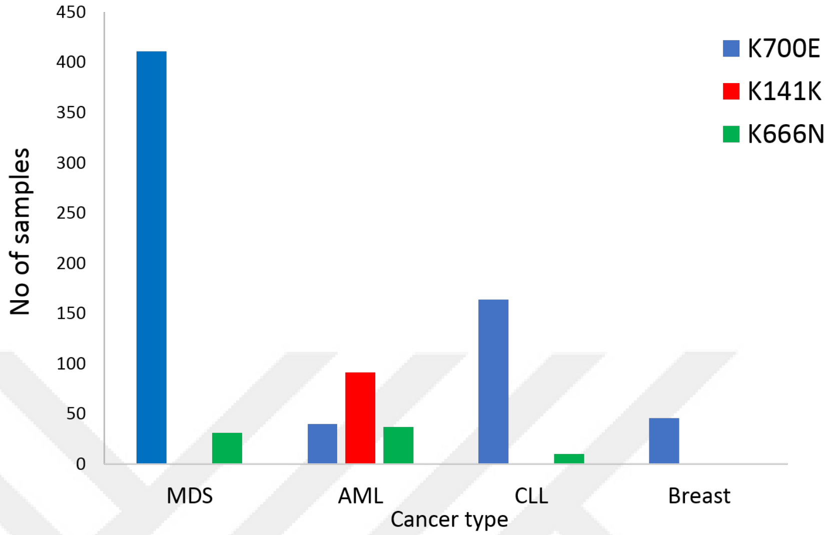


Figure 4.2.2: Clustered bar charts show number of hematologic and breast cancer patients had any of these mutations K700E, K141K, and K666N which are colored in blue, red, and green, respectively.

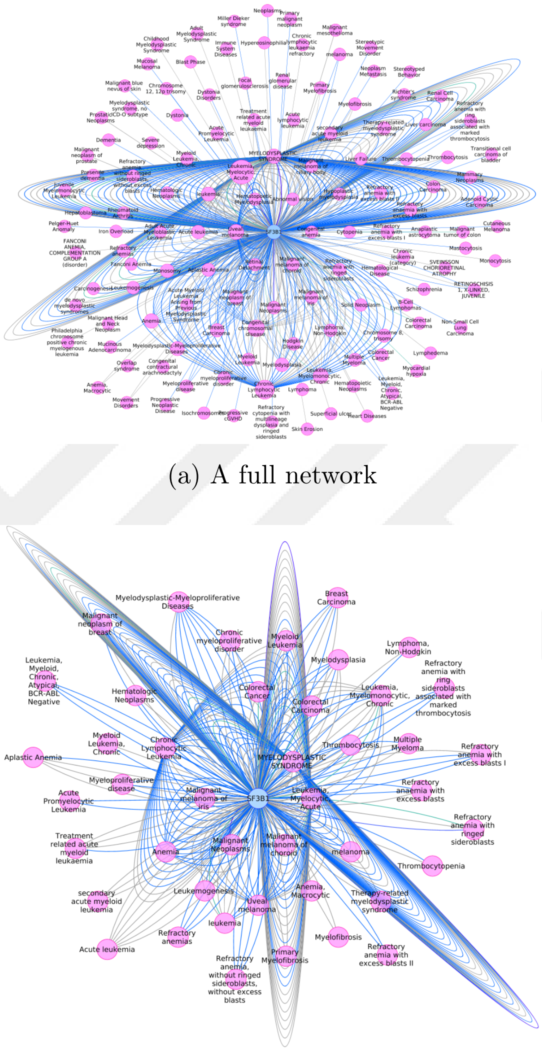
To confirm these outcomes, DisGeNET was used to identify the highly associated diseases. Two parameters were used to evaluate DisGeNET data:

1. Gene-diseases association (GDA) score which is computed by DisGeNET database [68] based on number and type of sources (level of curation, organisms), and number of studies reported that association. It ranges from 0 to 1.

2. Similarity, variant-diseases association (VDA) score was considered for variant node, but it is computed based on only numbers of sources and publications.

3. Disease's degree which is calculated by NetworkAnalyzer. Higher degree means higher number of edges connects gene or variant node to disease node, indicating higher number of evidences reporting the association between gene/variant and disease.

In the constructed gene-disease association network, 131 pink nodes (131 disease) are connected to SF3B1 by 472 edges based on: association type, number of associated SNPs and studies that support each association as shown in Figure 4.2.3a. For more relevant network representation, all diseases that have one edge, non-cancer or hemic, and lymphatic diseases were excluded as shown in Figure 4.2.3b. Then, the remaining list of diseases was sorted based on degree metric. MDS, AML, and CLL have the highest degree values as 84, 58, and 33 and their association scores equal to 0.7, 0.5, and 0.5, respectively. Then, uveal melanoma is connected by 15 edges and 0.6 association score, followed by six hemic and lymphatic diseases. After that, malignant neoplasm of breast and breast carcinoma are placed by 6 edges for each with an association score equals to 0.01. In addition, variants-diseases association network was constructed as shown in Figure 4.2.4. In comparison with COSMIC, some results are consistent, while others are different. K700E (dbSNP: rs559063155) is ranked first by 8 connected diseases including MDS, AML, and breast cancer. Its association score with AML equals 0.7, while it is 0.01 for MDS and breast. K666N (dbSNP: rs377023736) is connected to MDS and AML with 0.7 association score for each. Interestingly, K141 does not exist in DisGeNET database which may indicate that it is not a pathogenic mutation or not reported with clinical impact.


Neoplasms
Primary malignant neoplasm
Miller Dieker syndrome
Adult Myelodysplastic Syndrome
Childhood Myelodysplastic Syndrome
Immune System Diseases
Hypereosinophilia
Chronic lymphocytic leukaemia refractory
Malignant mesothelioma
Stereotypic Movement Disorder
melanoma
Mucosal Melanoma
Blast Phase
Focal glomerulosclerosis
Renal glomerular disease
Primary Myelofibrosis
Neoplasm Metastasis
Stereotyped Behavior
Malignant blue nevus of skin
Chromosome 12, 12p trisomy
Dystonia Disorders
Richter's syndrome
Renal Cell Carcinoma
Myelofibrosis
Refractory anemia with ring sideroblasts associated with marked thrombocytosis
Myelodysplastic syndrome, no CD-O subtype
Prostatic Neoplasms
Dystonia
Treatment related acute myeloid leukaemia
Acute lymphocytic leukemia
Therapy-related myelodysplastic syndrome
Liver carcinoma
secondary acute myeloid leukemia
Acute Promyelocytic Leukemia
Dementia
Severe depression
Malignant neoplasm of prostate
Myeloid Leukemia, Chronic
MYELODYSPLASTIC SYNDROME
Liver Failure
Thrombocytopenia
Thrombocytosis
Transitional cell carcinoma of bladder
Refractory anemia, without ringed sideroblasts, without excess blasts
Presenile dementia
Leukemia, Myelocytic, Acute
Malignant melanoma of ciliary body
Mammary Neoplasms
Juvenile Myelomonocytic Leukemia
Hematologic Neoplasms
leukemia
Hematopoetic Myelodysplasia
Abnormal vision
Hypoplastic myelodysplasia
Refractory anemia with excess blasts II
Colon Carcinoma
Adenoid Cystic Carcinoma
Refractory anaemia with excess blasts
Hepatoblastoma
Rheumatoid Arthritis
Pelger-Huet Anomaly
Iron Overload
Adult Acute Myeloblastic Leukemia
Acute leukemia
Uveal melanoma
SF3B1
Congenital anemia
Cytopenia
Refractory anemia with excess blasts I
Anaplastic astrocytoma
Malignant tumor of colon
Cutaneous Melanoma
Mastocytosis
FANCONI ANEMIA, COMPLEMENTATION GROUP A (disorder)
Refractory anemias
Fanconi Anemia
Monosomy
Aplastic Anemia
Retinal Detachment
Malignant melanoma of choroid
Refractory anemia with ringed sideroblasts
Hematological Disease
Chronic leukemia (category)
SVEINSSON CHORIORETINAL ATROPHY
Monocytosis
Carcinogenesis
Leukemogenesis
Acute Myeloid Leukemia Arising from Previous Myelodysplastic Syndrome
Malignant neoplasm of breast
Malignant Neoplasms
Malignant melanoma of iris
Solid Neoplasm
Schizophrenia
RETINOSCHISIS 1, X-LINKED, JUVENILE
de novo myelodysplastic syndromes
Malignant Head and Neck Neoplasm
Anemia
Breast Carcinoma
Congenital chromosomal disease
B-Cell Lymphomas
Colorectal Carcinoma
Non-Small Cell Lung Carcinoma
Philadelphia chromosome positive chronic myelogenous leukemia
Mucinous Adenocarcinoma
Myelodysplastic-Myeloproliferative Diseases
Hodgkin Disease
Lymphoma, Non-Hodgkin
Chromosome 8, trisomy
Lymphedema
Overlap syndrome
Congenital contractural arachnodactyly
Myeloid Leukemia
Myelodysplasia
Multiple Myeloma
Colorectal Cancer
Myocardial hypoxia
Anemia, Macrocytic
Myeloproliferative disease
Chronic myeloproliferative disorder
Chronic Lymphocytic Leukemia
Lymphoma
Leukemia, Myelomonocytic, Chronic
Hematopoietic Neoplasms
Leukemia, Myeloid, Chronic, Atypical, BCR-ABL Negative
Movement Disorders
Progressive Neoplastic Disease
Isochromosomes
Progressive cGVHD
Refractory cytopenia with multilineage dysplasia and ringed sideroblasts
Superficial ulcer
Heart Diseases
Skin Erosion
Myelodysplastic-Myeloproliferative Diseases
Breast Carcinoma
Malignant neoplasm of breast
Chronic myeloproliferative disorder
Myeloid Leukemia
Lymphoma, Non-Hodgkin
Myelodysplasia
Leukemia, Myeloid, Chronic, Atypical, BCR-ABL Negative
Hematologic Neoplasms
Refractory anemia with ring sideroblasts associated with marked thrombocytosis
Colorectal Cancer
Leukemia, Myelomonocytic, Chronic
Colorectal Carcinoma
Myeloid Leukemia, Chronic
Chronic Lymphocytic Leukemia
Thrombocytosis
Multiple Myeloma
Aplastic Anemia
MYELODYSPLASTIC SYNDROME
Refractory anemia with excess blasts I
Myeloproliferative disease
Malignant melanoma of iris
SF3B1
Leukemia, Myelocytic, Acute
Refractory anaemia with excess blasts
Acute Promyelocytic Leukemia
Refractory anemia with ringed sideroblasts
Treatment related acute myeloid leukaemia
Anemia
Malignant Neoplasms
Malignant melanoma of choroid
melanoma
Thrombocytopenia
Anemia, Macrocytic
Therapy-related myelodysplastic syndrome
secondary acute myeloid leukemia
Leukemogenesis
Uveal melanoma
leukemia
Myelofibrosis
Refractory anemias
Acute leukemia
Primary Myelofibrosis
Refractory anemia with excess blasts II
Refractory anemia, without ringed sideroblasts, without excess blasts

(a) A full network

(b) A filtered network

Figure 4.2.3: (a) The constructed network consists of 131 diseases are linked to SF3B1 by 472 edges, this list of diseases are reported from 1985 to 2018. (b) A filtered network including 44 diseases that are connected to SF3B1 by 349 edges.

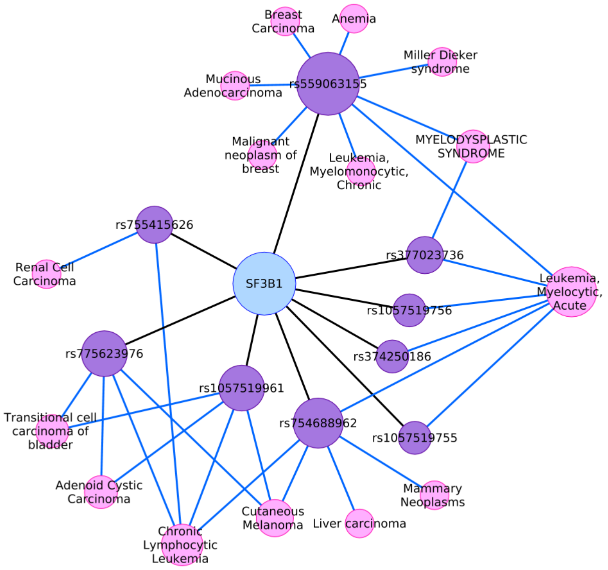


Figure 4.2.4: The constructed network shows one cyan node represents SF3B1 gene. It is connected to 9 purple nodes represent variants and each of them is connected to one or more pink nodes representing disease. Overall, the network has 25 noes and 37 edges.

4.3 Identifying Association of SF3B1 Families With Cancer

Each gene belongs to at least one family which contains group of genes produced from a common ancestral gene. Genes share same family usually have common structural properties and/or functions [69]. Hence, their relations with a specific disease may be similar. Accordingly, the association of SF3B1 families with hematologic vs. breast cancer was investigated, separately. Then, the obtained results were compared with those obtained from including only SF3B1. As reported in HGNC database, SF3B1 is a member of three families:

1. Armadillo-like helical domain containing (ARMH): 244 genes have a common superhelical structure, but they have different functions [70].

2. SF3b complex: consists of 7 genes, a member of “U2 small nuclear ribonucleoprotein” which is considered as a subgroup of five families. All of these subfamilies have “Major spliceosome” as a root family which contains 145 genes. Major spliceosome is a machinery complex splices introns from pre mature mRNA.

3. B-WICH chromatin-remodelling complex subunits (B-WICH): 8 genes positively regulate rRNA expression.

Two approaches were used to test the effect of including other genes from same family on the association with cancer. First, lists of genes for ARMH and B-WICH families were downloaded from HGNC. For major spliceosome, not all genes were included. As cryo-EM study showed that SF3b complex is disassociated after the late B^{act} state [2]. Spliceosomal C and spliceosomal P complexes formed after B^{act} state. In addition, spliceosomal E complex is a superfamily of “U1 small nuclear ribonucleoprotein". Therefore, they were not included in the input list. The final list included 244 ARMH genes, 8 genes from B-WICH complex, and 70 spliceosome components. Then each family list was inner joined with each of MDS, AML, CLL, and breast cancer data separately. The reported values of each family in 4.4 were counted after excluding SF3B1 entries. Overall, number of breast cancer patients is always the highest with respect to any family. Breast cancer patients carried the largest number of mutated genes as well. Comparing four cancer types with respect to same family pointed out that ARMH genes affect the most number of patients breast cancer, CLL, and AML except MDS which is associated with spliceosome families the most. For B-WICH family, it contributes the least to all of them as it is the smallest family containing 7 genes and SF3B1. Interestingly, only 6 mutated genes were detected in 242 MDS patient, while 46 mutated genes were detected in 248 AML patient. These data may reveal that few number of mutated genes are enough to develop MDS, while more mutated genes are needed to transform MDS into AML or develop AML without prior MDS.

Table 4.4: The association between each of the four diseases and SF3B1 or its family separately based on COSMIC database

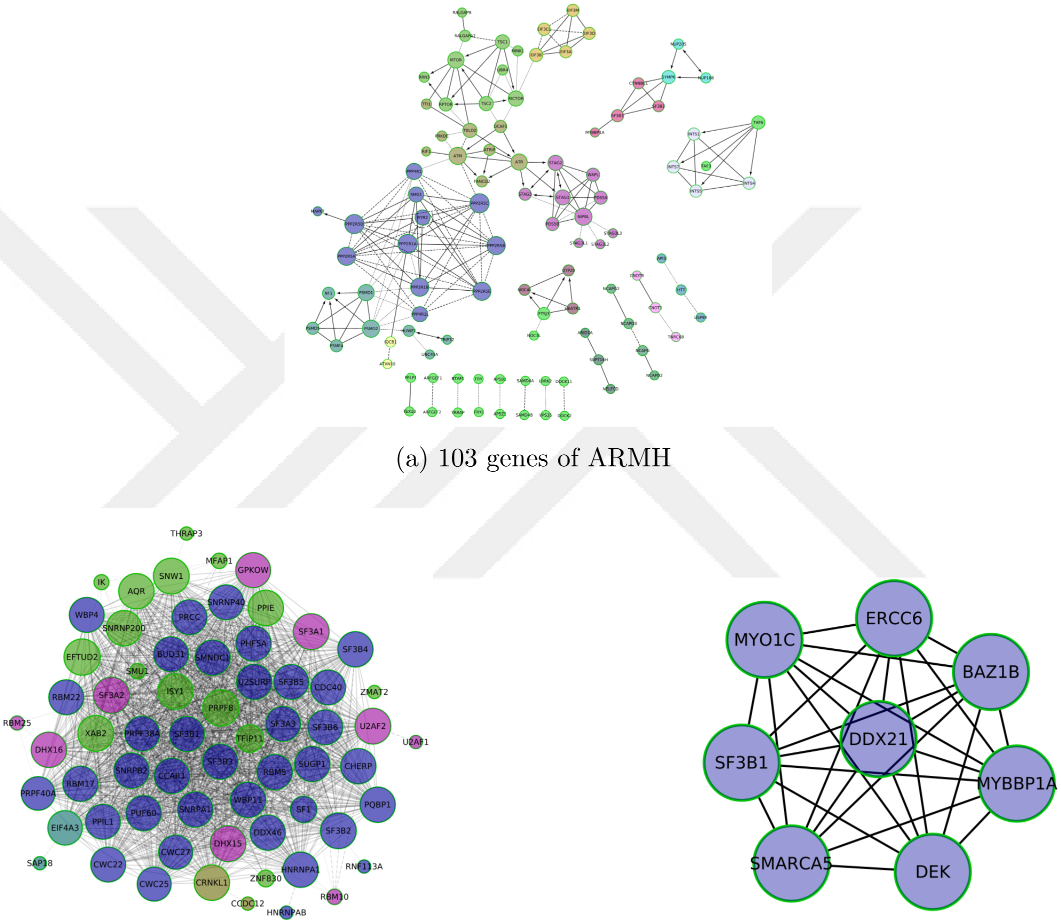
	Breast Cancer			
	SF3B1	**ARMH (243)**	**Subcomplex of Major spliceosome (69)**	**B-WICH (7)**
No of Genes	1	235	65	7
No of Mutations	51	49330	532	105
No of Samples	113	2534	821	232
	MDS			
No of Genes	1	38	6	1
No of Mutations	44	76	21	1
No of Samples	89	91	242	2
	AML			
No of Genes	1	218	46	7
No of Mutations	44	450	68	9
No of Samples	135	593	248	16
	CLL			
No of Genes	1	215	47	7
No of Mutations	57	500	34	5
No of Samples	378	597	108	18

For the other approach, Reactome FI app was used to derive a network containing genes of each family separately to find common pathways. Then, a gene-disease network was derived using DisGeNET app for each gene network. Based on number of associated genes and disease's degree parameters, impact of including genes from any of the three families on their association with cancer was tested. As explained in detail in the results section, adding genes have same origin affects the association with specific cancer types.

Interactions of SF3B1 with other genes from each of the three families are the following:

1. Armadillo-like helical domain containing: as shown in Figure 4.3.1a, SF3B1 has only 4 edges out of 188 edges of the whole network. It is linked to CTNNBL1, SYMPK, MYBBP1A, and SF3B2.
2. Subcomplex of Major spliceosome: It is a dense network in which SF3B1 has 50 edges and the maximum degree value is 53 as shown in Figure 4.3.1b.

3. B-WICH chromatin-remodelling complex subunits: SF3B1 has 7 edges as all other nodes as shown in Figure 4.3.1c.



(a) 103 genes of ARMH

(b) 65 genes of Subfamilies of Major spliceosome

(c) 8 genes of B-WICH

Figure 4.3.1: (a) The gene network contains 103 genes of ARMH family and 188 edges. They compose 16 components where the giant one has 56 genes and the second has 7 genes including SF3B1 are linked to 10 genes including SF3B1. (b) One component of 65 genes of some subfamilies of major spliceosome root family and 1321 edges. (c) One component of 8 genes of B-WICH family and 28 edges. All networks were built using Reactome FI plugin in Cytoscape.

DisGeNET analysis showed similar results to COSMIC as reported in Table 4.5. Breast carcinoma has the highest number of associated genes in each family's network. However, it is connected to highest number of edges only in ARMH network and lowest number in B-WICH and spliceosome networks. Numbers of edges that connect ARMH genes to breast cancer, CLL, and AML are the highest among other families. Unlike the mentioned diseases, MDS has the highest degree with spliceosome family. B-WICH family ranks second for breast cancer and AML based on number of edges, while ranks the lowest for MDS and CLL with 2 edges for each of them.

Table 4.5: The association between each of the four diseases and SF3B1 or its family separately based on DisGeNET and NetworkAnalyzer plugins

	Breast Cancer			
	SF3B1	**ARMH**	**Subcomplex of Major spliceosome**	**B-WICH**
No of associated genes	1	36	9	5
Degree	6	402	16	17
	MDS			
No of associated genes	1	9	3	1
Degree	84	21	37	2
	AML			
No of associated genes	1	15	5	2
Degree	58	55	18	29
	CLL			
No of associated genes	1	9	3	1
Degree	33	118	4	2

As it is mentioned before, “Major spliceosome" family has six subgroups, five of them have “U2 small nuclear ribonucleoprotein" as a subgroup that contains “SF3b complex". The following groups were included: spliceosomal A complex, spliceosomal B complex, spliceosomal Bact complex, U2 small nuclear ribonucleoprotein, and SF3b complex. For more detailed analysis, number of patients carried by each subfamily was reported and compared by SF3B1 individual among four diseases as shown in Figure 4.3.2. By comparing SF3B1 with other SF3b complex’s components, SF3B1 is the most frequently mutated gene in four diseases, however, SF3b complex is not the highest mutated subgroup in all diseases. For instance, spliceosome Bact complex is the most frequent subgroup associated with breast cancer, while spliceosome A complex is the most associated family with AML. These data pointed out which mutated genes are more crucial for each cancer type development.

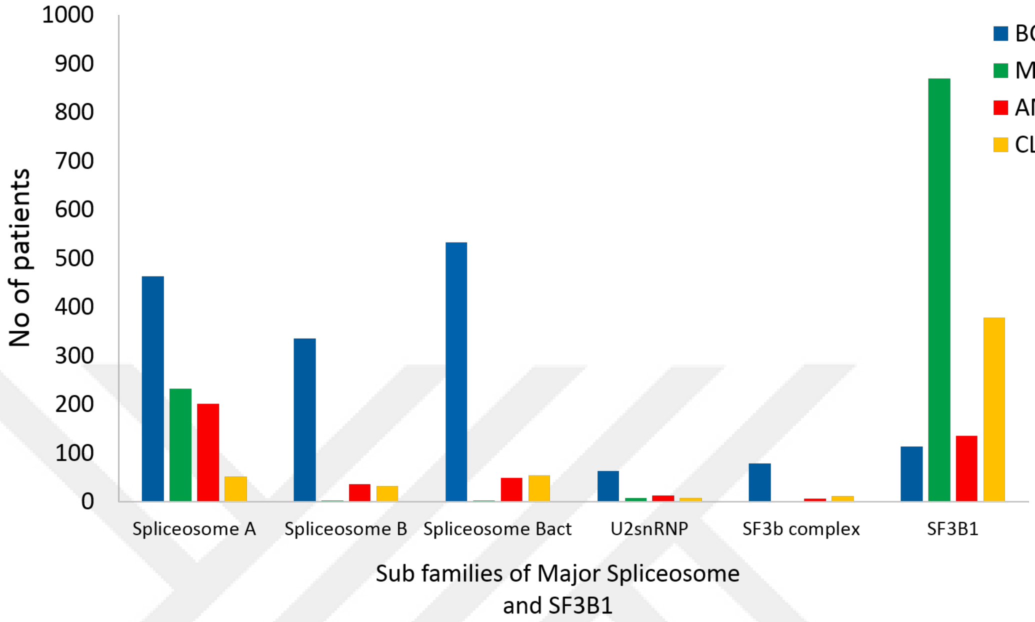


Figure 4.3.2: A clustered bar chart shows the number of patients who carried mutant genes from each subgroup of major spliceosome family. Breast cancer (BC), MDS, AML, and CLL that are colored in blue, green, red, and yellow, respectively.

4.4 Identifying Interactions Between SF3B1 and Different Pathways

Aberrant splicing has functional consequences in the cell by disturbing other vital pathways and leading to cancer development [36]. First, some queries were executed to find matching patients from any two/three/all cancers by using i) “INNER JOIN" clause to select the common “sample_name" attribute between two tables and ii) ‘WHERE" clause to specify that “gene_name" attribute is SF3B1 in one of the two tables. DB4S queries resulted in the following:

1. When SF3B1 is mutant in MDS:

   - There are 6 common patients with AML.
   - No common patients with CLL.

    - There are 9 common patients with breast cancer.

2. When SF3B1 is mutant in AML:
    - No common patients with MDS, CLL, or breast cancer.

3. When SF3B1 is mutant in CLL:
    - There are 2 common patients with MDS.
    - No common patients with AML.
    - There are 5 common patients with breast cancer.

4. When SF3B1 is mutant in breast cancer:
    - No common patients among MDS, AML, and CLL.

The returned numbers of samples from each query are low. However these results may indicate that aberrant splicing due to SF3B1 mutations may lead to developing i) MDS that could be followed by developing AML or breast cancer, ii) CLL that could be followed by developing breast cancer or MDS in extremely rare cases, iii) AML or breast cancer without prior cancer formation. These common patients had SF3B1 mutations in one disease and other mutant genes in the other disease. Accordingly, all other genes were collected with removing repeated ones and mapped to Cytoscape for pathway analysis to investigate the impact of aberrant splicing on other processes. As shown in Figure 4.4.1, other genes were shown to be involved in vital processes for the cell. For instance, 9 genes are involved in the process of negative and positive regulation of transcription by RNA polymerase II and others listed in Table 4.6.

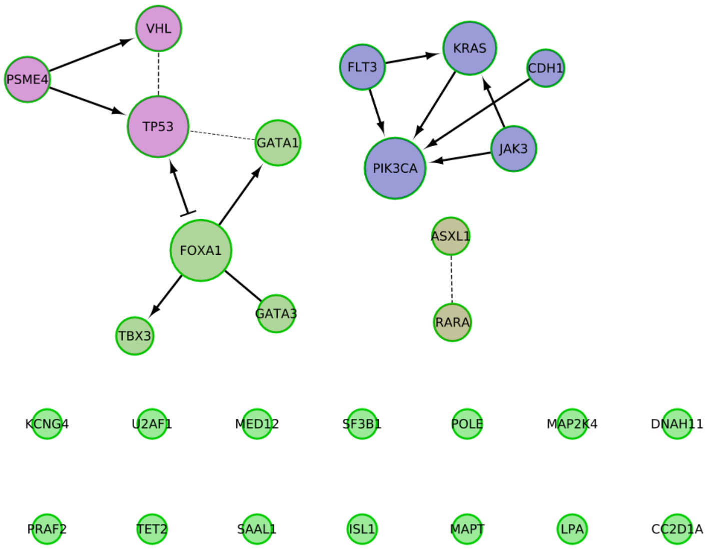


Figure 4.4.1: The constructed network by Reactome FI plugin in Cytoscape shows in total 28 nodes and 15 edges. The network consists of 3 connected components. The giant one contains 7 nodes and two small components consist of 5 and 2 nodes. There are 14 disconnected nodes. Edge's size represents FI score, while the shape represents FI direction. Node's size reflects its degree and node's color is based on FI clustering.

Table 4.6: Reactome FI analysis for the gene network displayed in Figure 4.4.1

Common Pathways	**No of genes**	**Genes from the network**
RNA Polymerase II Transcription(R)	8	FLT3,CDH1,VHL,JAK3, PIK3CA,RARA,KRAS,TP53
Signaling pathways regulating pluripotency of stem cells(K)	7	GATA3,GATA1,MED12, PSME4,RARA,KRAS,TP53
Human T-cell leukemia virus 1 infection(K)	5	JAK3,ISL1,TBX3,PIK3CA,KRAS
MAPK signaling pathway(K)	5	JAK3,MAP2K4,PIK3CA,KRAS,TP53
PI3K-Akt signaling pathway(K)	5	FLT3,MAP2K4,KRAS,MAPT,TP53
Common Biological Processes	**No of genes**	**Genes from the network**
Negative regulation of transcription by RNA polymerase II	9	FOXA1,GATA3,GATA1,VHL,ISL1, TBX3,RARA,TP53,CC2D1A
Positive regulation of transcription by RNA polymerase II	9	FOXA1,GATA3,GATA1,MED12, TET2,ISL1,ASXL1,RARA,TP53
Positive regulation of transcription, DNA-templated	8	GATA3,GATA1,MED12,CDH1, VHL,TBX3,RARA,TP53
Cytokine-mediated signaling pathway	6	FLT3,GATA3,JAK3,PIK3CA,KRAS,TP53
Negative regulation of cell proliferation/ trans	5	GATA3,GATA1,VHL,RARA,TP53
Negative regulation of apoptotic process	5	GATA1,VHL,TBX3,RARA,TP53
Positive regulation of cell proliferation	5	FLT3,ISL1,TBX3,RARA,KRAS

4.5 Relations Among Subtypes of Hematologic and Breast Cancers

Interactions within hematologic cancer or between any of them and breast cancer were investigated. Similar queries to the previous section were executed without specifying gene name. Inner join of each combination resulted in number of overlapping patients as shown in Figure 4.5.1. These values were relatively higher than the previous section because all genes were included with any conditions. Therefore, predicting relations among four cancers can be possible. For instance, MDS and AML has the most significant relation based on highest number of common patients (53). Also, relations between each of MDS/AML/CLL and breast cancer seem to be significant and among three of them as 27 patients were diagnosed by three cancer types. On the other hand, only one patient was common among three hematologic cancers and 3 patients were common among MDS, CLL, and breast. Regarding each two joined tables, number of common patients between CLL and AML/MDS were low, while it was relatively high for CLL and breast.

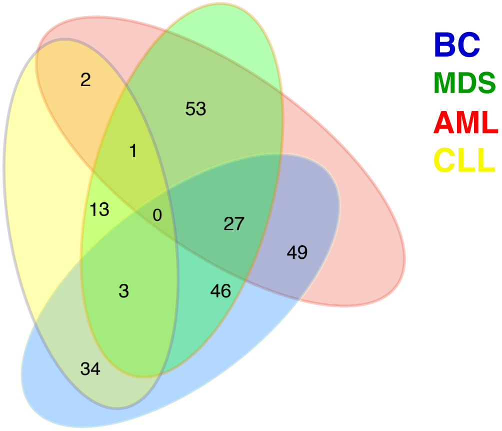


Figure 4.5.1: Venn diagram shows number of common patients among each possible combination among four cancer types. All data are reported except number of patients among AML, CLL, and breast cancer (BC) which equals to zero.

These patients carried a different set of mutant genes with respect to each disease. All of them were combined and filtered based on number of repetition to extract only frequently repeated gene. Then it was mapped into Cytoscape to identify common pathways if exists by using Reactome FI as shown in Figure 4.5.2. In Figure 4.4.1, which is constructed in the previous section, SF3B1 was a disconnected gene. In contrast to Figure 4.5.2, SF3B1 is connected to 3 genes which are connected to others, forming one connected component. Relations between SF3B1 and other genes involving in other pathways can be anticipated from the network shown in Figure 4.5.2.

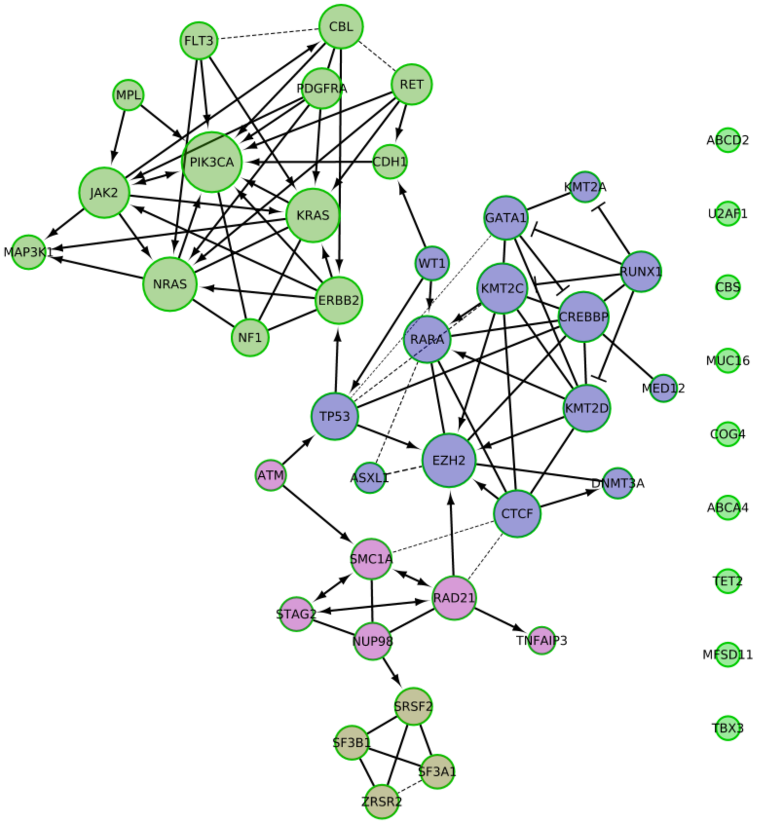


Figure 4.5.2: The constructed network by Reactome FI plugin in Cytoscape shows in total 42 nodes and 92 edges. The network consists of 10 components. The giant one contains 37 nodes and 9 components each of them is one node. Edge's size represents FI score, while the shape represents FI direction. Node's size reflects its degree and node's color is based on FI clustering.

Pathway enrichment analysis was conducted for this gene network by using Reactome FI plugin to obtain common pathways. The most common pathways, having highest number of genes, are i) RNA Polymerase II transcription, ii) microRNAs in cancer, and iii) central carbon metabolism in cancer as listed in Table 4.7.

Table 4.7: Reactome FI analysis for the gene network displayed in Figure 4.5.2

Common Pathways	**No of genes**	**Genes from the network**
Pathways in cancer(K)	14	RET,FLT3,CBL,NRAS,CDH1,ERBB2,JAK2,PDGFRA,CREBBP,RUNX1,PIK3CA,RARA,KRAS,TP53
RNA Polymerase II Transcription(R)	14	KMT2D,KMT2A,KMT2C,GATA1,MED12,ERBB2,CREBBP,RUNX1,RARA,SRSF2,KRAS,ATM,TP53,EZH2
MicroRNAs in cancer(K)	10	NRAS,ERBB2,PDGFRA,CREBBP,DNMT3A,PIK3CA,KRAS,ATM,TP53,EZH2
Central carbon metabolism in cancer(K)	8	RET,FLT3,NRAS,ERBB2,PDGFRA,PIK3CA,KRAS,TP53
SUMOylation(R)	8	RAD21,CREBBP,DNMT3A,SMC1A,STAG2,RARA,NUP98,TP53
RAF/MAP kinase cascade(R)	8	RET,FLT3,NRAS,ERBB2,JAK2,PDGFRA,NF1,KRAS
MAPK signaling pathway(K)	8	FLT3,NRAS,ERBB2,PDGFRA,MAP3K1,NF1,KRAS,TP53
PI3K-Akt signaling pathway(K)	8	FLT3,NRAS,ERBB2,JAK2,PDGFRA,PIK3CA,KRAS,TP53

4.6 Discussion

SF3B1 has a crucial role in splicing mechanism as it recognizes and selects the branch splice site along with other subunits of splicing machinery. SF3B1 is commonly mutated in many cancer types. Associated mutations cause aberrant splicing, production of abnormal transcripts, and cancer development [71–73]. This chapter demonstrated an analytical view about SF3B1 and its association with hematologic malignancies, represented by (MDS, AML, and CLL) vs. breast cancer. Although mutations of SF3B1 have crucial functional consequences in the cell, SF3B1 has recently received attention from researchers. The research's trend has peaked in 2013 for MDS and CLL articles, while AML and breast cancer lines peaked in 2016. In comparison with other associated genes with each disease, SF3B1 has well studied in MDS and CLL only. Therefore, a detailed analysis was performed to figure out if SF3B1 is worthy to be studied in breast cancer and AML patients also or not. Mutation analysis pointed out haematopoietic and lymphoid as the most frequently mutated tissue. In addition, top frequently mutated residues are located in SF3B1 heat repeats. These mutations may alter heat domain's conformation, affecting SF3B1 structurally and dynamically. This may be explain the reason behind not studying p.K141K although it was detected in 114 samples. As it is located in the N-terminal, it may not have structural/functional impact on SF3B1. These outcomes matched experimental findings [22, 33].

Network analysis of SF3B1 families showed that SF3B1 interacts with other genes from B-WICH and spliceosome complex families, due to functional similarities. On the other hand, SF3B1 is connected to only four out of 102 genes in ARMH gene network, because ARMH genes share a structural property with different function for each gene. Consequently, pathway enrichment analysis revealed “RNA Polymerase II Transcription" as the most common pathways by involving only 28 genes out of 103. It is followed by some signaling pathways shared by 11 genes such as i) the AMP-activated protein kinase (AMPK), ii) PIP3 activates AKT signaling, and iii) the phosphatidylinositol 3-kinase (PI3K)/protein kinase B (AKT) (PI3K-Akt). The reported data in Tables 4.4 & 4.5 indicate that mutations of SF3B1 can be enough to lead to MDS and CLL as the number of patients carried mutant SF3B1 is far away higher than number of patients carried other mutant splicing factors. On the other hand, contribution of other splicing factors moderately increase number of patients who were diagnosed by AML and dramatically increase for breast cancer patients. The least affected pathway in all diseases is RNA polymerase II transcription which is carried out by B-WICH complex genes followed by splicing mechanism in all diseases except MDS. Biological mechanisms that involve ARMH genes seem the most driven even for breast cancer, followed by CLL, AML, and MDS the least.

Common patients were found between breast cancer and MDS and/or AML. However, it was impossible to figure out which tumor was formed first. If MDS is firstly formed due to aberrant splicing which produce abnormal transcripts leading to AML or breast cancer development. The other possibility is that formation of breast tumor is prior to MDS or AML. As reported in literature, this case my occur due to chemo/radio therapies which are commonly used to treat breast cancer [74]. They may alter or damage DNA content, producing more mutant genes and consequently leading to MDS or AML development.

Last but not least, investigating the impact of aberrant splicing on other pathways showed that produced abnormal spliced transcripts encode to proteins involved in vital pathways; mitogen-activated protein kinase (MAPK) and PI3K-Akt signaling pathways. Associated biological processes are regulation of gene transcription, cell proliferation, and apoptosis. Misregulation of MAPK and/or PI3K-Akt signaling cause uncontrolled cell proliferation and resistance to apoptosis of tumor cells [75–77]. Additionally, positive and/or negative regulation of transcription by RNA polymerase II are also listed in Table 4.6, indicating the connections between mRNA splicing and transcription by polymerase II as reported before in an experimental study [78]. Without specifying SF3B1 mutations, more genes were retrieved from COSMIC and added to the network in Figure 4.5.2, more interactions among genes were formed. For instance, as shown in Figure 4.5.2, there are multiple paths connect SF3B1 to TP53 (top associated gene to CLL and breast cancer), that is connected to another cluster of genes containing JAK2 and FLT3, the top associated genes to MDS and AML, respectively. These complex interactions show the associations among different diseases and may explain how one disease can be developed and transformed to other multiple diseases.

4.7 Conclusion

In this chapter, data and network analysis shed light on SF3B1 and its interactions with other components involved in splicing or other vital processes in the cell. Most of the results were confirmed by two approaches, matched with experimental results as well. These findings may guide experimentalists and help them to save effort, time, and money for studying the relation between aberrant splicing and cancer development. Consequently, developing therapeutic targets for cancer can be improved. As explained in Section 4.2, the missense “K700E" mutation emerged as the most hotspot mutation among others. Accordingly, it was selected to be studied using molecular dynamics (MD) simulation in the next chapter. MD approach was used to assess the impact of K700E on SF3B1 structurally and dynamically which may provide insights into the linking among SF3B1, aberrant splicing, and cancer development.

Chapter 5

ASSESSMENT OF THE IMPACT OF SF3B1-K700E BY MEANS OF MOLECULAR DYNAMICS SIMULATIONS

In most cases, experimental research is time and money consuming. Therefore, scientists were directed to computational based studies and developed new algorithms and software tools, solving biological problems that cannot be solved by experimental techniques. For instance, molecular dynamics simulation was developed in the late 1950s. It is a tool that can be used to get insights into the dynamics of a system at the atomic and molecular level. In the 1970s, biophysicist started to apply MD on macromolecules such as proteins. Protein's structure is usually obtained from structural biology techniques such as x-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy or cryogenic electron microscopy (Cryo-EM) as in our case. Therefore, classical MD simulations were utilized to investigate the impact of a hotspot mutation of SF3B1 on its tertiary structure and dynamics. K700E mutation was selected for our simulations, as shown in Table 4.3, K700E (lysine ⇒ glutamic acid at position 700) is a common mutation in 14 cancers including breast cancer, myelodysplastic syndrome, acute myeloid leukemia, and chronic lymphocytic leukemia. It was selected based three attributes from COSMIC (March release): numbers of samples, tissues, and PubMed articles where it was reported. Therefore, it was identified to be such mutation whose impact was studied by using molecular dynamics (MD) simulations.

5.1 System Setup

Cryogenic electron microscopy (cryo-EM) structure of the human spliceosome (PDB ID:5Z56) at resolution 5.1 Å [2], which corresponded to B^{act} state, was retrieved from Protein Data Bank (PDB). The complex contains 52 proteins, three snRNAs, and synthetic pre-mRNA with a molecular weight of ca.1.8 megaDaltons. Due to the size of the system, not all SF3b complex's components were included, instead, five chains, namely, SF3B1, SF3B3, SF3B5, SF3B6, SF3B7, and pre-mRNA (See Figure 5.1.1) were used in MD simulations since they directly interact with SF3B1. SF3B2 and SF3B4 were excluded for the following reasons: i) they did not show any direct interactions with SF3B1, ii) for SF3B2: it lacks N & C termini with missing gaps in between, only 183 residues are available out of 895, iii) for SF3B2: it has 424 residues, but only 78 residues are available, and iv) both subunits interact together and with other splicing factors, not with SF3b complex components [79]. SF3B1 has 1304 amino acids but the obtained SF3B1 chain from cryo-EM structure was not complete. First 81 residues and 186 residues in between (90-489) are missing. Therefore, SWISS-MODEL webserver [80] was used to fill the missing gaps. However, the N-terminal of SF3B1 remained missing in the modeled structure (82-1304).

CHARMM-GUI [81] was exploited to prepare both wild type and the mutant systems. CHARMM36 force field [82] was used to model protein and RNA, whereas water molecules were represented by TIP3P water model [83]. After including water molecules, wild type system has 1089028 and the mutant protein has 1089281 atoms. The thickness of water layer was set to 16 Å in each direction to prevent interaction with the periodic images of the system. The system was neutralized with 0.15 M KCl.

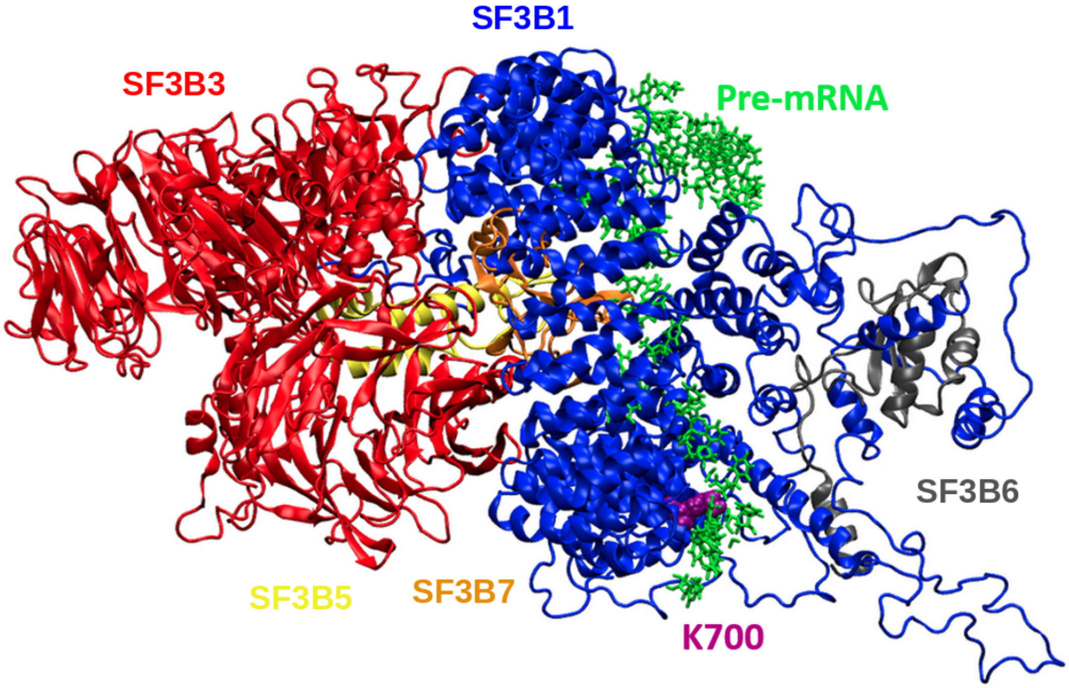


Figure 5.1.1: The depiction of the 3D structure of five domains of the SF3b complex included in this study. They were taken from the cryo-EM structure of human spliceosome complex (PDB:5z56) [2]. Each domain is shown in a different color. SF3B1, SF3B3, SF3B5, SF3B6, SF3B7, and pre-mRNA are shown in blue, red, yellow, grey, orange and green, respectively. The image was created using VMD [3].

5.2 Simulation Protocol

Molecular dynamics simulations were performed by using GROningen MAchine for Chemical Simulations (GROMACS) [84] package. Equilibration of the system was performed by using the NVT ensemble, whereas the NPT ensemble was used in the production step to maintain constant pressure and temperature throughout the simulation. Nose-Hoover and Parrinello-Rahman coupling algorithms were used with 1 ps and 5 ps coupling times to maintain the temperature at 310 °K and 1 bar, respectively. LINCS algorithm was used to constrain the bond lengths in hydrogen atoms [85]. The Particle Mesh Ewald (PME) method was used to compute long-range electrostatic interactions [86]. For van der Waals and short-range electrostatic interactions, the cut off value was set to (12 Å). The time step for integration was set to 2 femtoseconds. MD simulations were performed for 2 microseconds in total. Three replicates, each of which started with a different initial velocity, were used for both wild type and the mutant system to check the reliability of the results. The two replicates were run for 500 nanoseconds (ns), whereas the last one was run for 100 ns.

5.3 Simulation Analysis Tools

GROMACS package was used for MD trajectories analysis as it provides many tools and GRaphing and Advanced Computation and Exploration of data (Grace) tool for data plotting.

5.3.1 Root mean square deviation (RMSD)

Root mean square deviation is the most commonly used similarity measurement tool [87] which is given by the following equation:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2} \tag{5.1}$$

where, n represents pairs of equivalent atoms and d_i is the distance between the two atoms in the ith pair. Here, the RMSD is calculated by using the backbone atoms of the selected groups. Also, the initial structure in corresponding trajectories was used as reference [88].

5.3.2 Calculation of number of contacts

The number of contacts between pre-determined groups is computed by using “gmx mindist" module implemented in GROMACS. The cut-off value to calculate the number of contacts between the mutated residue and pre-mRNA was set to 0.3 nm, while 3 nm was used to calculate the number of contacts between side chain residues of p14 (residues 20-100) and pre-mRNA as p14 is defined in previous studies as a marker of branch point [22, 24].

5.3.3 Principal component analysis (PCA)

To determine the dominant motion of a structure, principal component analysis is computed by obtaining the covariance matrix of $C\alpha$ atoms with respect to the reference structure. Covariance matrix is computed using “gmx covar" module in GROMACS.

$$\boldsymbol{C}_{ij} = \langle M_{ij} \Delta r_i \Delta r_j \rangle \tag{5.2}$$

where $\boldsymbol{C}_{ij}$ corresponds to covariance matrix of $C\alpha$ atoms i and j. $M_{ij}\Delta r_i \Delta r_j$ refers to positional change from time-averaged structure for each coordinate of all $C\alpha$ atoms i and j. Diagonalization of covariance matrix is performed using “gmx covar" module in GROMACS. It results in a set of eigenvalues δ^2 and their corresponding eigenvectors ($\boldsymbol{v}$).

$$\boldsymbol{C}\boldsymbol{v} = \delta^2 \boldsymbol{v} \tag{5.3}$$

“gmx anaeig" module was used to analyze the first two eigenvectors that represent the directions and relative magnitudes [89] of more than 50% of the dominant motion of the system. Additionally, “rmsf" module was used to plot root mean square fluctuation (RMSF) per atom of the first two eigenvectors. Therefore, a particular residue that leads to fluctuations of the overall chain can be identified. RMSF is calculated as the averaging of all frames per each residue referenced to the initial frame [84]. RMSF is expressed as

$$RMSF = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(X_i(n) - \overline{X}_i)} \tag{5.4}$$

where N is the number of frames obtained from simulation time, $X_i(n)$ is the coordinates of backbone atom of $X_i$and $\overline{X}_i$ is the mean coordinate over simulation time.

5.3.4 Dynamic cross correlation map (DCCM)

Dynamic cross correlation (DCC) among $C\alpha$ of SF3B1 residues is computed using “bio3d" package for biological structural analysis written by R programming language [90]. “bio3d" library contains major functions needed for structural bioinformatics. For this study, “dccm.xyz" and “plot.dccm" modules were used to generate dynamical cross correlation matrices and to plot them, respectively. DCC is expressed as

$$DCC(i,j) = \frac{< \Delta r_i(t).\Delta r_j(t) >_t}{\sqrt{< \parallel \Delta r_i(t) \parallel^2 >_t}\sqrt{< \parallel \Delta r_j(t) \parallel^2 >_t}} \tag{5.5}$$

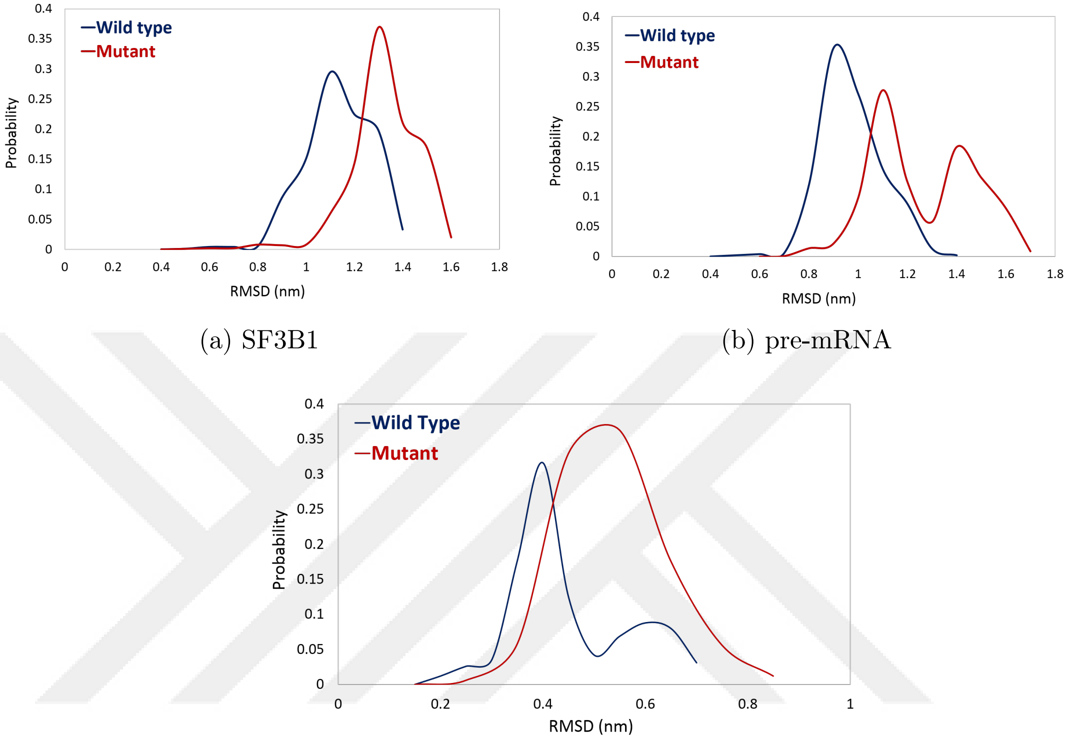
where $r_i(t)$ and $r_j(t)$ correspond to the coordinates of the i^{th} and j^{th} atoms as a function of time t, $<.>$ indicates the time ensemble average, $\Delta r_i(t) = r_i(t) - (< r_i(t) >)_t$ and $\Delta r_j(t) = r_j(t) - (< r_j(t) >)_t$ [91–93].

5.4 MD Trajectories Analysis

GROMACS [81] simulation generated 500 ns trajectories of two replicates from each system. The analysis was done for pre-mRNA and two components of SF3b complex; SF3B1 as it interacts directly with the pre-mRNA and other SF3b complex's components; SF3B6 (p14) as it is defined in previous studies as a marker of branch point and has interactions with SF3B1 [22, 24].

5.4.1 K700E mutation decreases the Stability of SF3B1, pre-mRNA and p14

As shown in 5.1.1, K700E is located in the heat domain of SF3B1 and interacts directly with pre-mRNA. Therefore, RMSD was computed for pre-mRNA and backbone atoms of SF3B1 to investigate the impact of the mutation on their stability. Then, the probability distributions were calculated for combined replicates for each component in wild type vs. mutant. As shown in both Figures 5.4.1a & 5.4.1b, the mutant system is right-skewed, indicating the reduction of stability. Same analysis was done for p14, which is shown to be in contact with pre-mRNA and SF3B1 as well. As shown in Figure 5.4.1c, the mutant system is also relatively right-skewed.



(a) SF3B1

(b) pre-mRNA

(c) p14

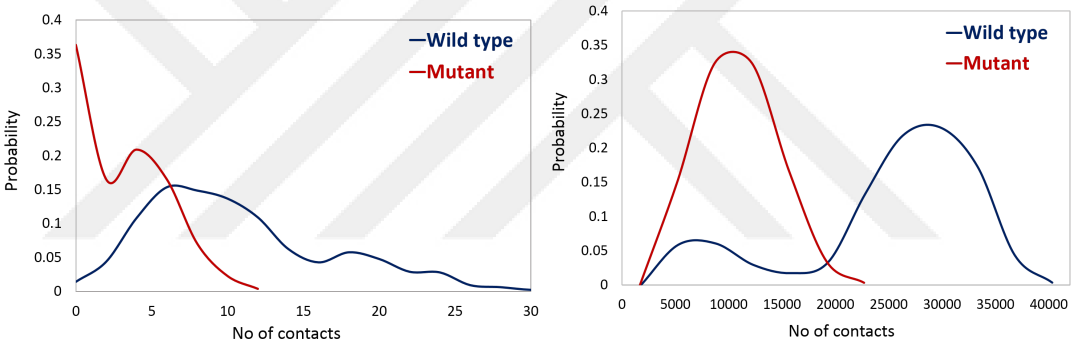
Figure 5.4.1: Probability plots of RMSD distributions of (a) SF3B1^{wt} in blue vs. SF3B1^{K700E} in red. (b) pre-mRNA in the wild type structure in blue vs. mutant structure in red. (c) p14 in the wild type structure in blue vs. mutant structure in red.

5.4.2 K700E mutation weakens the interactions formed with pre-mRNA

As shown in the above section, K700E decreases the stability of pre-mRNA and SF3B1, which is probably affects the interactions formed between them as well. Accordingly, the number of contacts was computed between K700E and pre-mRNA with cutoff distance $\leq$ 0.3 as the distance between them is (0.17 nm). Number of contacts was computed with respect to backbone and side chain of K700 residue. The results showed that backbone does not contribute to the interaction, while side chain forms bonds with pre-mRNA. As shown in Figure

5.4.2a, the mutant residue, Glutamic acid, has very low contact with pre-mRNA and reached zero contact after around half simulation time ($\sim$ 250ns). On the other hand, lysine as wild type highly interacts with pre-mRNA through the whole simulation time.

Additionally, the impact of K700E on the interactions formed between pre-mRNA and side chains of p14 residues ($20-100$) known as RNA recognition motif (RRM) [31] was investigated. As shown in Figure 5.4.2b, the number of contacts formed between p14^{RRM} and pre-mRNA drastically decreased.



(a) pre-mRNA & K700E

(b) pre-mRNA & p14

Figure 5.4.2: Probability plots of distributions of number of contacts formed between (a) pre-mRNA and side chain of amino acid K700 in wild type shown in blue vs. E700 mutant shown in red, with cutoff distance ≤ 0.3 nm. (b) pre-mRNA and side chains of p14 residues ($20 - 100$) in wild type shown in blue vs. mutant shown in red, with cutoff distance ≤ 3 nm.

5.4.3 K700E mutation affects the global dynamics of SF3B1

According to the obtained results from previous sections, the impact of the mutation on global dynamics of SF3B1 was investigated by performing essential dynamics analysis. The motion of both wild type and mutant SF3B1 was projected along with the first and the second eigenvectors of the mutant system as shown in Figure 5.4.3. The results showed that the second replicate of the mutant samples relatively wider conformational space than the wild type with respect to both eigenvectors. On the other hand, the first replicate of the mutant samples narrower space than the second replicate, but it is still in a different space than the wild type.

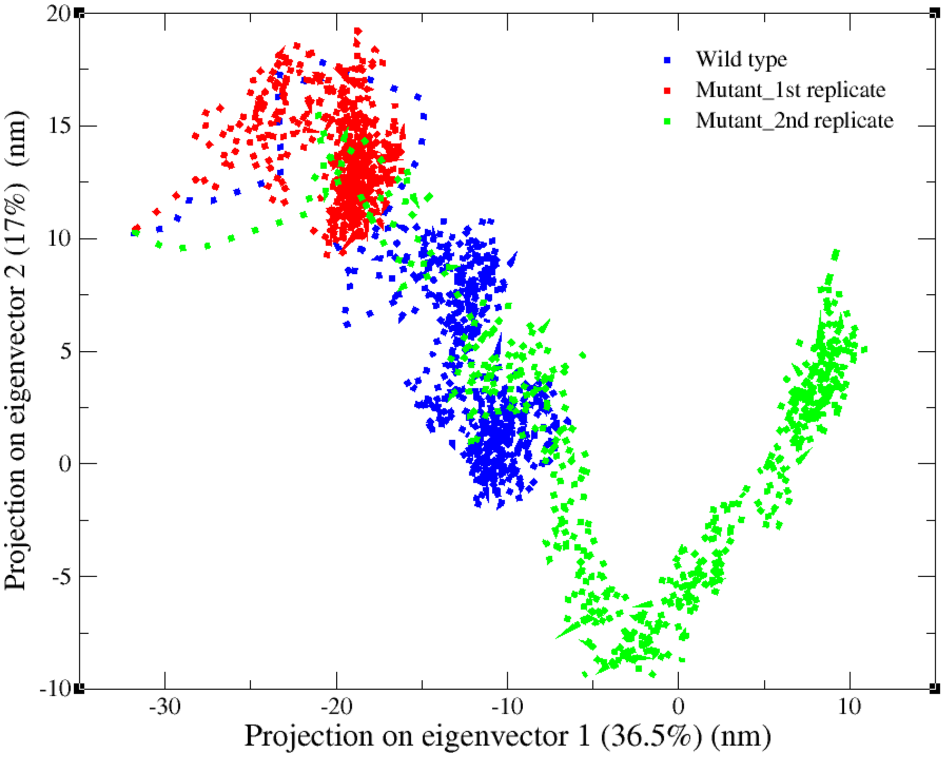


Figure 5.4.3: 2D projection of SF3B1 trajectories of wild type (blue) and two replicates of the mutant system (red and green) along the first two eigenvectors of SF3B1^{K700E} system.

Additionally, RMSF values per $C\alpha$ atoms of the first and the second eigenvectors of both wild type and mutant SF3B1 were calculated to highlight specific regions of SF3B1 fluctuate more than the others. The first two eigenvectors cumulatively account for more than 50% of total motion of the system. The results showed that the first and the second replicate of the mutant system displayed higher fluctuation for residues between 900-1125 which is revealed by the first eigenvector of systems as shown in Figure 5.4.4. Moreover, higher fluctuations for residues between 1125-1300 for the mutant system were observed as revealed by the second eigenvector.

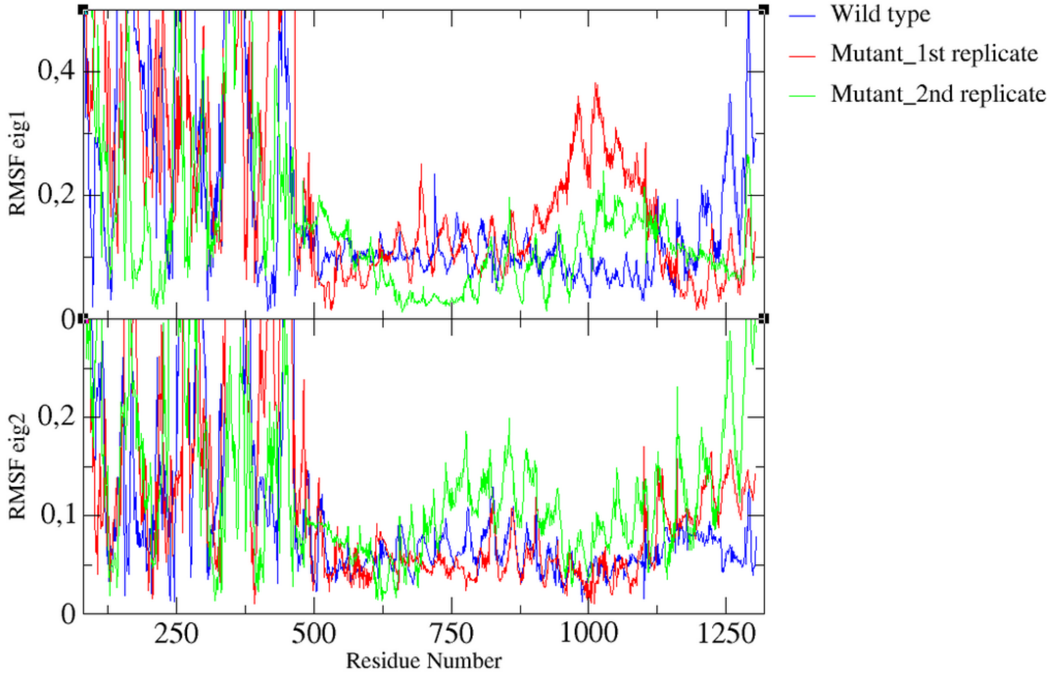
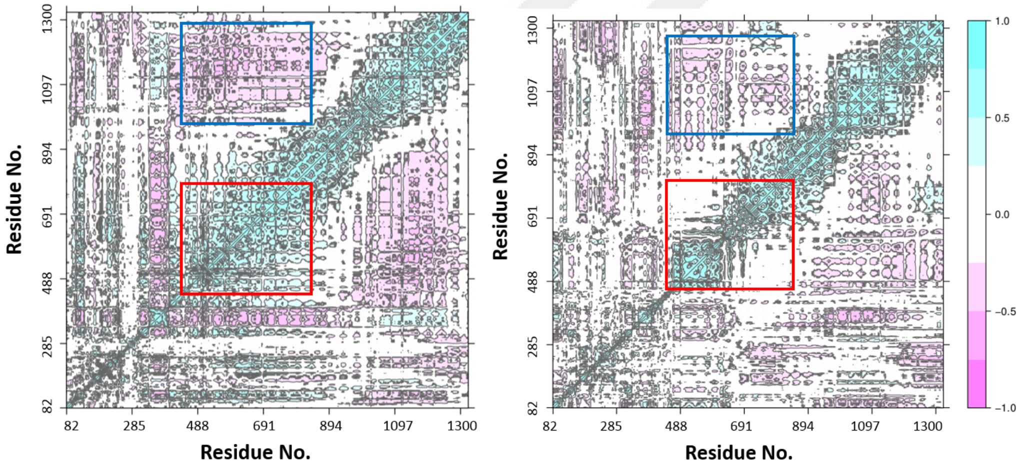


Figure 5.4.4: RMSF per $C\alpha$ atoms of the first two eigenvectors of SF3B1 obtained for wild type protein (blue) and two replicates of the mutant system (red and green)

5.4.4 K700E mutation distorts the communication among SF3B1 residues

In order to investigate the communication among SF3B1 residues in wild type vs. mutant, dynamic cross correlation was calculated for each system. Red square refer to the residues from 455 to 832 that show strong positive correlation with each other and strong negative correlation with another region which is from 932 to 1300 (shown as blue square), in the wild type protein. On the other hand, both positive and negative correlation patterns that are shown among residues $(455-832)$ and between $(455-832)$ and $(932-1300)$, respectively are drastically decreased in the mutant protein.

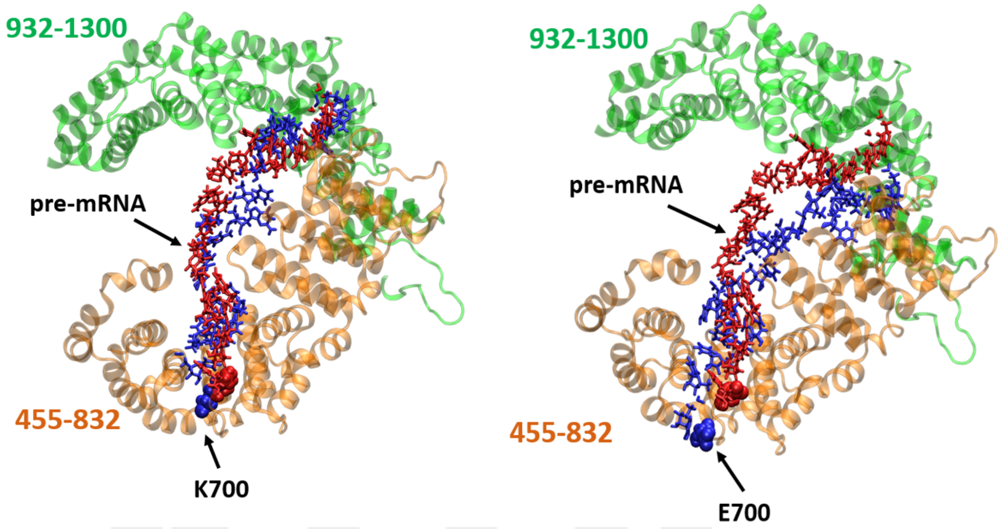
Mapping the two highlighted regions that are given in Figure 5.4.5a on the 3D structure of SF3B1 showed that the both highlighted region in the above section surround the pre-mRNA from both termini as shown in Figure 5.4.6a. Then, pre-mRNA and K700E residue were aligned with respect to these two regions in each system. The positioning of the pre-mRNA within the substructure was compared by presenting the first (red) and the last frame (blue) taken from the trajectories. In wild type system as shown in Figure 5.4.6a, both frames of pre-mRNA and K700 residue are almost superimposed on each other. Unlike the mutant system, their positions changed significantly with respect to the first frame, as shown in Figure 5.4.6b. This dynamic change reflects the impact of the mutation on dynamics and orientation of the pre-mRNA.



(a) Wild type

(b) Mutant

Figure 5.4.5: Dynamic cross-correlation map for SF3B1 is shown in (a) for the wild type system. As can be seen, there is a high positive correlation within the region 455 − 832, which is colored in red, and a high negative correlation between two highlighted regions sequences as 455 − 832 in red and 932 − 1300 in blue. (b) For the mutant system, the positive correlation decreased drastically within the region 455 − 832, which is colored in red, and low negative correlation between the two regions 455 − 832 in red and 932 − 1300 in blue.



(a) Wild type (b) Mutant

Figure 5.4.6: 3D structure of the two highlighted regions surrounding pre-mRNA in wild type system (a) and mutant system (b). α helical region $(455 - 832)$ in orange, α helical region $(932 - 1300)$ in green, both of pre-mRNA and K700E residue are shown in red representing the initial frame of simulation and blue color representing the last frame. The structural image was created using VMD [3].

5.5 Discussion

MD simulations analysis results showed that K700E mutation resulted in stability reduction of SF3b complex's components, particularly SF3B1 and those interact directly with mRNA. Thus, number of contacts between pre-mRNA and side chain of two groups, K700E and p14^{RRM}, were computed separately. Both plots show significant lower interaction in mutant structure. It was found that 700 residue interacts with mRNA from side chain group. In wild type, side chain of 700 residue is lysine which is a positive amino acid that causes attraction with mRNA. On the other hand, side chain of mutant residue is glutamic acid which is a negative amino acid that tends to repel the negatively charged mRNA, therefore, the distance between E700 and pre-mRNA increases leading to decreased interaction as shown in Figure 5.4.2a.

Experimental studies found that p14 interacts with SF3B1 and other spliceosome complex's components to act as a marker of branch point adenosine [24, 25]. Cryo-EM study showed that p14 is surrounded by N-terminal of SF3B1^{HD} [2] which starts at 463 residue. The distorted communication of SF3B1$^{455-832}$ in mutant system as shown in Figure 5.4.5b leads to a dynamical change of N-terminal of SF3B1^{HD}. Therefore, K700E mutation affects the positioning of p14 and its binding to SF3B1.

In addition, K700E mutation has a negative effect on the interaction between pre-mRNA and p14^{RRM} as shown in Figure5.4.2b. Consequently, the critical role of p14 in mRNA splicing mechanism may be disturbed, leading to misrecognition of branch point and aberrant splicing. The mutation K700E does not only affects the local dynamics of SF3B1, but also the global one. The high fluctuations seen from RMSF plot affect the conformation of SF3B1 and consequently distort the communication among SF3B1 residues as shown in DCCM. The combined sequence of the two highlighted regions in DCCM covers most of heat repeats as the 22 heat repeats begin from residue 463 until residue 1304 [22], locating pre-mRNA within its perimeter.

5.6 Conclusion

Comparative analysis of molecular dynamics trajectories of wild type and mutant systems showed that K700E mutation affects the structural and dynamical properties of SF3b complex's components as the following:

1. Increasing the motion's flexibility of SF3b complex's components such as SF3B1 and SF3B6 (p14) and pre-mRNA as well.
2. Interactions between pre-mRNA and p14^{RRM} and SF3B1^{K700} are declined dramatically in mutant system.
3. Changing the global dynamics of SF3B1, especially C-terminal residues showed higher fluctuations in mutant SF3B1.

4. Communication among SF3B1 residues is distorted in the mutant SF3B1

These adverse changes that caused by K700E are probably lead to alternative branch point selection, hence aberrant splicing of the pre-mRNA. Consequently, abnormal transcripts are translated into truncated proteins with altered functions that contribute to the development of the disease. Finally, the results of MD analysis provide molecular details that help to understand the functional consequences of K700E on the spliceosomal machinery.

Chapter 6

CONCLUSION & Future Work

In this thesis, publicly available databases, open source tools, and computational biophysics tools were utilized to analyze SF3B1 from different perspectives. Somatic mutations of SF3B1 cause aberrant splicing that lead to tumorigenesis [71–73]. They occur in many cancers such as hematologic cancer which is considered as the most associated type, uveal melanoma, skin, and breast cancer. Although this thesis started with breast cancer, hematologic malignancies, including MDS, AML, and CLL, were covered in the analyses. Data and network analyses findings help experimentalists conduct a more efficient and precise research to study the association among mutant SF3B1, aberrant splicing, and cancer development. It was concluded that SF3B1 can be solely studied as an associated splicing factor with blood cancer. On the other hand, investigating the association between splicing mechanism and breast cancer necessitate other splicing factors to be added in the study to obtain more significant relation between the mechanism and the disease.In contrast to spliceosome family, ARMH family has a common structural property instead of a common function. Therefore, constructing a network that contains ARMH genes including SF3B1 will give mechanistic insights into different pathways that are associated with breast cancer development. Additionally, functional consequences of aberrant splicing in the cell was were analyzed and showed that it promotes defects in other biological processes such as regulation of cell proliferation, apoptosis, and transcription by

RNA polymerase II. Besides, relations among different types of cancer (MDS, AML, CLL, and breast) were investigated. The significant relations were found between MDS & AML, each of MDS, AML, CLL, and breast separately, and MDS, AML, and breast, collectively. These results, though need experimental validations, may help developing i) new therapeutic approaches that target multiple pathways in cancerous cell and ii) prevention approaches for cancer to be transformed from one type to another.

Finally, atomistic MD simulations were performed to asses the molecular impact of SF3B1^{K700E} on pre-mRNA splicing mechanism. Comparative analysis of MD trajectories showed that the tertiary structure of SF3B1 is essential for its surface properties and its critical function in the selection of branch site in splicing mechanism. The point mutation K700E reduces the stability of SF3B1, distorts its structure and motion, decreases the interactions between pre-mRNA and both of K700E and p14^{RRM}. The conformational change due to K700E mutation distorts the structural integrity of SF3b complex. Consequently, selection of branch point selection is altered, and this leads to aberrant splicing of mRNA and translation of non-functional proteins. These adverse effects, considering the abundance of this mutation in cancer and SF3B1's critical function, may be contributing to the development and/or prognosis of cancer.

This thesis strongly proposes SF3B1 as a potential therapeutic target for cancer. Additionally, this combined in silico approach can be improved and employed in the characterization of impact for other cancer somatic mutations. Consequently, development of new diagnostic biomarkers and therapeutic targets in cancer can be improved. However, further studies are needed to validate this thesis's results as the following:

1. More patients data are needed to investigate associations between breast and blood cancers. Thus, it can be an effective step toward cancer drug repositioning.

2. Other cancer types can be included to improve the identification of more significant interactions that occur in the cancerous cell.

3. Identifying more complex interactions among genes and how they regulate different pathways will help developing new therapies that target multiple pathways.

4. MD simulations need experimental work to validate the impact of SF3B1-K700E mutation on mRNA splicing mechanism, leading to cancer.

Bibliography

[1] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, *et al.*, "STRING v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, no. D1, pp. D447–D452, 2014.

[2] X. Zhang, C. Yan, X. Zhan, L. Li, J. Lei, and Y. Shi, "Structure of the human activated spliceosome in three conformational states," *Cell Research*, vol. 28, no. 3, p. 307, 2018.

[3] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.

[4] G. K. Malhotra, X. Zhao, H. Band, and V. Band, "Histological, molecular and functional subtypes of breast cancers," *Cancer Biology & Therapy*, vol. 10, no. 10, pp. 955–960, 2010.

[5] Z. Zaidi and H. A. Dib, "The worldwide female breast cancer incidence and survival, 2018," *AACR; Cancer Research*, vol. 79, 2019.

[6] Z. Momenimovahed and H. Salehiniya, "Epidemiological characteristics of and risk factors for breast cancer in the world," *Breast Cancer: Targets and Therapy*, vol. 11, p. 151, 2019.

[7] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray, "Global cancer observatory: cancer today," *Lyon, France: International Agency for Research on Cancer*, 2018.

[8] G. N. Hortobagyi, J. de la Garza Salazar, K. Pritchard, D. Amadori, R. Haidinger, C. A. Hudis, H. Khaled, M.-C. Liu, M. Martin, M. Namer, *et al.*, “The global breast cancer burden: variations in epidemiology and survival,” *Clinical Breast Cancer*, vol. 6, no. 5, pp. 391–401, 2005.

[9] M. J. Ellis and C. M. Perou, “The genomic landscape of breast cancer as a therapeutic roadmap,” *Cancer Discovery*, vol. 3, no. 1, pp. 27–34, 2013.

[10] O. L. Griffith, N. C. Spies, M. Anurag, M. Griffith, J. Luo, D. Tu, B. Yeo, J. Kunisaki, C. A. Miller, K. Krysiak, *et al.*, “The prognostic effects of somatic mutations in ER-positive breast cancer,” *Nature Communications*, vol. 9, no. 1, p. 3476, 2018.

[11] S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, *et al.*, “Cosmic: somatic cancer genetics at high-resolution,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D777–D783, 2016.

[12] J. Wu, T. K. K. Mamidi, L. Zhang, and C. Hicks, “Integrating germline and somatic mutation information for the discovery of biomarkers in triple-negative breast cancer,” *International Journal of Environmental Research and Public Health*, vol. 16, no. 6, p. 1055, 2019.

[13] C. T. Gomillion *et al.*, “Assessing the potential of chitosan/polylactide nanoparticles for delivery of therapeutics for triple-negative breast cancer treatment,” *Regenerative Engineering and Translational Medicine*, vol. 5, no. 1, pp. 61–73, 2019.

[14] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, *et al.*, “Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis,” *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262–272, 2006.

[15] C. Desmedt, B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi, M. Delorenzi, M. Piccart, and C. Sotiriou, “Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes,” *Clinical Cancer Research*, vol. 14, no. 16, pp. 5158–5165, 2008.

[16] N. Bonifaci, A. Berenguer, J. Díez, O. Reina, I. Medina, J. Dopazo, V. Moreno, and M. A. Pujana, “Biological processes, properties and molecular wiring diagrams of candidate low-penetrance breast cancer susceptibility genes,” *BMC Medical Genomics*, vol. 1, no. 1, p. 62, 2008.

[17] S. Stamm, S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T. Thanaraj, and H. Soreq, “Function of alternative splicing,” *Gene*, vol. 344, pp. 1–20, 2005.

[18] A. Fabregat, K. Sidiropoulos, G. Viteri, O. Forner, P. Marin-Garcia, V. Arnau, P. D’Eustachio, L. Stein, and H. Hermjakob, “Reactome pathway analysis: a high-performance in-memory approach,” *BMC Bioinformatics*, vol. 18, no. 1, p. 142, 2017.

[19] S. M. Berget, C. Moore, and P. A. Sharp, “Spliced segments at the 5’ terminus of adenovirus 2 late mRNA,” *Proceedings of the National Academy of Sciences*, vol. 74, no. 8, pp. 3171–3175, 1977.

[20] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts, “An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA,” *Cell*, vol. 12, no. 1, pp. 1–8, 1977.

[21] M. C. Wahl, C. L. Will, and R. Lührmann, “The spliceosome: design principles of a dynamic RNP machine,” *Cell*, vol. 136, no. 4, pp. 701–718, 2009.

[22] C. Cretu, J. Schmitzová, A. Ponce-Salvatierra, O. Dybkov, E. I. De Laurentiis, K. Sharma, C. L. Will, H. Urlaub, R. Lührmann, and V. Pena, “Molecular architecture of SF3b and structural consequences of its cancer-related mutations,” *Molecular Cell*, vol. 64, no. 2, pp. 307–319, 2016.

[23] O. Gozani, R. Feld, and R. Reed, “Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is

required for assembly of spliceosomal complex A.," *Genes & Development*, vol. 10, no. 2, pp. 233–243, 1996.

[24] C. L. Will, C. Schneider, A. M. MacMillan, N. F. Katopodis, G. Neubauer, M. Wilm, R. Lührmann, and C. C. Query, "A novel u2 and u11/u12 snrnp protein that associates with the pre-mrna branch site," *The EMBO journal*, vol. 20, no. 16, pp. 4536–4546, 2001.

[25] C. C. Query, P. S. McCaw, and P. A. Sharp, "A minimal spliceosomal complex a recognizes the branch site and polypyrimidine tract.," *Molecular and Cellular Biology*, vol. 17, no. 5, pp. 2944–2953, 1997.

[26] O. Gozani, J. Potashkin, and R. Reed, "A potential role for u2af-sap 155 interactions in recruiting u2 snrnp to the branch site," *Molecular and Cellular Biology*, vol. 18, no. 8, pp. 4752–4760, 1998.

[27] D. M. Cass and J. A. Berglund, "The SF3b155 N-terminal domain is a scaffold important for splicing," *Biochemistry*, vol. 45, no. 33, pp. 10092–10101, 2006.

[28] C. Wang, K. Chua, W. Seghezzi, E. Lees, O. Gozani, and R. Reed, "Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis," *Genes & Development*, vol. 12, no. 10, pp. 1409–1414, 1998.

[29] M. A. Andrade and P. Bork, "Heat repeats in the huntington's disease protein," *Nature Genetics*, vol. 11, no. 2, p. 115, 1995.

[30] M. J. Schellenberg, R. A. Edwards, D. B. Ritchie, O. A. Kent, M. M. Golas, H. Stark, R. Lührmann, J. M. Glover, and A. M. MacMillan, "Crystal structure of a core spliceosomal protein interface," *Proceedings of the National Academy of Sciences*, vol. 103, no. 5, pp. 1266–1271, 2006.

[31] R. Spadaccini, U. Reidt, O. Dybkov, C. Will, R. Frank, G. Stier, L. Corsini, M. C. Wahl, R. Lührmann, and M. Sattler, "Biochemical and NMR analyses of an SF3b155–p14–U2AF-RNA interaction network involved in branch point definition during pre-mRNA splicing," *RNA*, vol. 12, no. 3, pp. 410–425, 2006.

[32] M. A. Andrade, C. Perez-Iratxeta, and C. P. Ponting, “Protein repeats: structures, functions, and evolution,” *Journal of Structural Biology*, vol. 134, no. 2-3, pp. 117–131, 2001.

[33] K. Yoshida and S. Ogawa, “Splicing factor mutations and cancer,” *Wiley Interdisciplinary Reviews: RNA*, vol. 5, no. 4, pp. 445–459, 2014.

[34] J. W. Harbour, E. D. Roberson, H. Anbunathan, M. D. Onken, L. A. Worley, and A. M. Bowcock, “Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma,” *Nature Genetics*, vol. 45, no. 2, p. 133, 2013.

[35] M. J. Ellis, L. Ding, D. Shen, J. Luo, V. J. Suman, J. W. Wallis, B. A. Van Tine, J. Hoog, R. J. Goiffon, T. C. Goldstein, *et al.*, “Whole-genome analysis informs breast cancer response to aromatase inhibition,” *Nature*, vol. 486, no. 7403, p. 353, 2012.

[36] J. D. Fackenthal and L. A. Godley, “Aberrant RNA splicing and its functional consequences in cancer cells,” *Disease Models & Mechanisms*, vol. 1, no. 1, pp. 37–42, 2008.

[37] H. Dolatshad, A. Pellagatti, F. G. Liberante, M. Llorian, E. Repapi, V. Steeples, S. Roy, L. Scifo, R. N. Armstrong, J. Shaw, *et al.*, “Cryptic splicing events in the iron transporter ABCB7 and other key target genes in SF3B1-mutant myelodysplastic syndromes,” *Leukemia*, vol. 30, no. 12, p. 2322, 2016.

[38] P. Papasaikas, J. R. Tejedor, L. Vigevani, and J. Valcárcel, “Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery,” *Molecular cell*, vol. 57, no. 1, pp. 7–22, 2015.

[39] J. Zhang, A. M. Ali, Y. K. Lieu, Z. Liu, J. Gao, R. Rabadan, A. Raza, S. Mukherjee, and J. L. Manley, “Disease-causing mutations in SF3B1 alter splicing by disrupting interaction with sugp1,” *Molecular Cell*, vol. 76, no. 1, pp. 82–95, 2019.

[40] J. M. Jiménez-Vacas, V. Herrero-Aguayo, E. Gómez-Gómez, A. J. León-González, P. Sáez-Martínez, E. Alors-Pérez, A. C. Fuentes-Fayos,

A. Martínez-López, R. Sánchez-Sánchez, T. González-Serrano, *et al.*, "Spliceosome component SF3B1 as novel prognostic biomarker and therapeutic target for prostate cancer," *Translational Research*, vol. 212, pp. 89–103, 2019.

[41] J. M. Jimenez-Vacas, J. L. Lopez-Canovas, M. C. Vazquez-Borrego, S. Pedraza-Arevalo, V. Herrero-Aguayo, P. Saez-Martinez, A. J. Montero-Hidalgo, R. B. Encinas, A. Lara-Lopez, J. M. Perez-Gomez, *et al.*, "SF3B1 as novel target for the treatment of multiple endocrine-related cancers," in *21st European Congress of Endocrinology*, vol. 63, BioScientifica, 2019.

[42] B. Liu, M. Ki, O. Abdel-Wahab, and S. Chandarlapaty, "Abstract 4654: Hotspot mutations in the core spliceosomal protein SF3B1 promote breast tumorigenesis," *Cancer Research*, vol. 79, no. 13 Supplement, pp. 4654–4654, 2019.

[43] J. Borišek, A. Saltalamacchia, A. Gallì, G. Palermo, E. Molteni, L. Malcovati, and A. Magistrato, "Disclosing the impact of carcinogenic SF3b mutations on pre-mrna recognition via all-atom simulations," *Biomolecules*, vol. 9, no. 10, p. 633, 2019.

[44] J. Borišek, A. Saltalamacchia, A. Spinello, and A. Magistrato, "Exploiting Cryo-EM structural information and all-atom simulations to decrypt the molecular mechanism of splicing modulators," *Journal of Chemical Information and Modeling*, 2019.

[45] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, *et al.*, "Cosmic: the catalogue of somatic mutations in cancer," *Nucleic Acids Research*, vol. 47, no. D1, pp. D941–D947, 2018.

[46] B. Braschi, P. Denny, K. Gray, T. Jones, R. Seal, S. Tweedie, B. Yates, and E. Bruford, "Genenames.org: the HGNC and VGNC resources in 2019," *Nucleic Acids Research*, vol. 47, no. D1, pp. D786–D792, 2018.

[47] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, *et al.*, "The reactome pathway

knowledgebase," *Nucleic Acids Research*, vol. 42, no. D1, pp. D472–D477, 2013.

[48] M. Piacentini, P. Morgan, J. Miltner, R. Ravanini, R. Peinthor, M. Kleusberg, J. Clift, and J. Haller, "DB browser for SQLite," 2015.

[49] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[50] G. Wu, X. Feng, and L. Stein, "A human functional protein interaction network and its application to cancer data analysis," *Genome Biology*, vol. 11, no. 5, p. R53, 2010.

[51] A. Bauer-Mehren, M. Rautschka, F. Sanz, and L. I. Furlong, "Disgenet: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks," *Bioinformatics*, vol. 26, no. 22, pp. 2924–2926, 2010.

[52] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282–284, 2007.

[53] C. Di, Q. Zhang, Y. Chen, Y. Wang, X. Zhang, Y. Liu, C. Sun, H. Zhang, J. D. Hoheisel, *et al.*, "Function, clinical application, and strategies of Pre-mRNA splicing in cancer," *Cell Death & Differentiation*, vol. 26, no. 7, pp. 1181–1194, 2019.

[54] B.-D. Wang and N. H. Lee, "Aberrant RNA splicing in cancer and drug resistance," *Cancers*, vol. 10, no. 11, p. 458, 2018.

[55] J. Makki, "Diversity of breast carcinoma: histological subtypes and clinical relevance," *Clinical Medicine Insights: Pathology*, vol. 8, pp. CPath–S31563, 2015.

[56] A. M. MacMillan, C. C. Query, C. R. Allerson, S. Chen, G. L. Verdine, and P. A. Sharp, "Dynamic association of proteins with the pre-mRNA branch region.," *Genes & Development*, vol. 8, no. 24, pp. 3008–3020, 1994.

[57] J. E. Arenas and J. N. Abelson, “Prp43: An RNA helicase-like factor involved in spliceosome disassembly,” *Proceedings of the National Academy of Sciences*, vol. 94, no. 22, pp. 11798–11802, 1997.

[58] J.-B. Fourmann, J. Schmitzová, H. Christian, H. Urlaub, R. Ficner, K.-L. Boon, P. Fabrizio, and R. Lührmann, “Dissection of the factor requirements for spliceosome disassembly and the elucidation of its dissociation products using a purified splicing system,” *Genes & Development*, vol. 27, no. 4, pp. 413–428, 2013.

[59] R.-T. Tsai, R.-H. Fu, F.-L. Yeh, C.-K. Tseng, Y.-C. Lin, Y.-h. Huang, and S.-C. Cheng, “Spliceosome disassembly catalyzed by Prp43 and its associated components Ntr1 and Ntr2,” *Genes & Development*, vol. 19, no. 24, pp. 2991–3003, 2005.

[60] J. I. Murray, R. B. Voelker, K. L. Henscheid, M. B. Warf, and J. A. Berglund, “Identification of motifs that function in the splicing of non-canonical introns,” *Genome Biology*, vol. 9, no. 6, p. R97, 2008.

[61] M. D. Ohi, C. W. Vander Kooi, J. A. Rosenberg, L. Ren, J. P. Hirsch, W. J. Chazin, T. Walz, and K. L. Gould, “Structural and functional analysis of essential pre-mRNA splicing factor Prp19p,” *Molecular and Cellular Biology*, vol. 25, no. 1, pp. 451–460, 2005.

[62] R. Hogg, J. C. McGrail, and R. T. O’Keefe, “The function of the nineteen complex (ntc) in regulating spliceosome conformations and fidelity during pre-mrna splicing,” 2010.

[63] S. Das and A. R. Krainer, “Emerging functions of SRSF1, splicing factor and oncoprotein, in RNA metabolism and cancer,” *Molecular Cancer Research*, vol. 12, no. 9, pp. 1195–1204, 2014.

[64] Y. Huang, R. Gattoni, J. Stévenin, and J. A. Steitz, “SR splicing factors serve as adapter proteins for TAP-dependent mRNA export,” *Molecular Cell*, vol. 11, no. 3, pp. 837–843, 2003.

[65] J. N. Ihle and D. G. Gilliland, “Jak2: normal function and role in hematopoietic disorders,” *Current Opinion in Genetics & Development*, vol. 17, no. 1, pp. 8–14, 2007.

[66] N. Daver, R. F. Schlenk, N. H. Russell, and M. J. Levis, “Targeting FLT3 mutations in AML: review of current knowledge and evidence,” *Leukemia*, vol. 33, no. 2, pp. 299–312, 2019.

[67] T. Chen, Z. Wang, W. Zhou, Z. Chong, F. Meric-Bernstam, G. B. Mills, and K. Chen, “Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types,” *BMC Genomics*, vol. 17, no. 2, p. 394, 2016.

[68] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, “The disgenet knowledge platform for disease genomics: 2019 update,” *Nucleic Acids Research*, 2019.

[69] R. E. Pyeritz, B. R. Korf, and W. W. Grody, *Emery and Rimoin’s Principles and Practice of Medical Genetics and Genomics: Foundations*. Academic Press, 2018.

[70] M. R. Groves and D. Barford, “Topological characteristics of helical repeat protein,” *Current Opinion in Structural Biology*, vol. 9, no. 3, pp. 383–389, 1999.

[71] S. Alsafadi, A. Houy, A. Battistella, T. Popova, M. Wassef, E. Henry, F. Tirode, A. Constantinou, S. Piperno-Neumann, S. Roman-Roman, *et al.*, “Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage,” *Nature Communications*, vol. 7, p. 10615, 2016.

[72] R. B. Darman, M. Seiler, A. A. Agrawal, K. H. Lim, S. Peng, D. Aird, S. L. Bailey, E. B. Bhavsar, B. Chan, S. Colla, *et al.*, “Cancer-associated SF3B1 hotspot mutations induce cryptic 3’ splice site selection through use of a different branch point,” *Cell Reports*, vol. 13, no. 5, pp. 1033–1045, 2015.

[73] C. DeBoever, E. M. Ghia, P. J. Shepard, L. Rassenti, C. L. Barrett, K. Jepsen, C. H. Jamieson, D. Carson, T. J. Kipps, and K. A. Frazer, "Transcriptome sequencing reveals potential mechanism of cryptic 3'splice site selection in SF3B1-mutated cancers," *PLoS Computational Biology*, vol. 11, no. 3, p. e1004105, 2015.

[74] R. E. Smith, "Risk for the development of treatment-related acute myelocytic leukemia and myelodysplastic syndrome among patients with breast cancer: review of the literature and the national surgical adjuvant breast and bowel project experience," *Clinical Breast Cancer*, vol. 4, no. 4, pp. 273–279, 2003.

[75] L. Santarpia, S. M. Lippman, and A. K. El-Naggar, "Targeting the MAPK–RAS–RAF signaling pathway in cancer therapy," *Expert Opinion on Therapeutic Targets*, vol. 16, no. 1, pp. 103–119, 2012.

[76] C. Porta, C. Paglino, and A. Mosca, "Targeting PI3K/Akt/mTOR signaling in cancer," *Frontiers in Oncology*, vol. 4, p. 64, 2014.

[77] W.-C. Huang and M.-C. Hung, "Induction of Akt activity by chemotherapy confers acquired resistance," *Journal of the Formosan Medical Association*, vol. 108, no. 3, pp. 180–194, 2009.

[78] M. J. Hicks, C.-R. Yang, M. V. Kotlajich, and K. J. Hertel, "Linking splicing to Pol II transcription stabilizes pre-mrnas and influences splicing patterns," *PLoS Biology*, vol. 4, no. 6, p. e147, 2006.

[79] Y. Yang, A. Hadjikyriacou, Z. Xia, S. Gayatri, D. Kim, C. Zurita-Lopez, R. Kelly, A. Guo, W. Li, S. G. Clarke, *et al.*, "PRMT9 is a type ii methyltransferase that methylates the splicing factor SAP145," *Nature Communications*, vol. 6, p. 6428, 2015.

[80] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. de Beer, C. Rempfer, L. Bordoli, *et al.*, "Swissmodel: homology modelling of protein structures and complexes," *Nucleic Acids Research*, vol. 46, no. W1, pp. W296–W303, 2018.

[81] J. Lee, X. Cheng, J. M. Swails, M. S. Yeom, P. K. Eastman, J. A. Lemkul, S. Wei, J. Buckner, J. C. Jeong, Y. Qi, *et al.*, “CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field,” *Journal of Chemical Theory and Computation*, vol. 12, no. 1, pp. 405–413, 2015.

[82] J. B. Klauda, R. M. Venable, J. A. Freites, J. W. O’Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell Jr, and R. W. Pastor, “Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types,” *The Journal of Physical Chemistry B*, vol. 114, no. 23, pp. 7830–7843, 2010.

[83] P. Mark and L. Nilsson, “Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K,” *The Journal of Physical Chemistry A*, vol. 105, no. 43, pp. 9954–9960, 2001.

[84] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers,” *SoftwareX*, vol. 1, pp. 19–25, 2015.

[85] L. Wang, Y. Yuan, X. Chen, J. Chen, Y. Guo, M. Li, C. Li, and X. Pu, “Probing the cooperative mechanism of the μ–δ opioid receptor heterodimer by multiscale simulation,” *Physical Chemistry Chemical Physics*, vol. 20, no. 47, pp. 29969–29982, 2018.

[86] S. Yuan, S. Filipek, K. Palczewski, and H. Vogel, “Activation of g-protein-coupled receptors correlates with the formation of a continuous internal water pathway,” *Nature Communications*, vol. 5, p. 4733, 2014.

[87] I. Kufareva and R. Abagyan, “Methods of protein structure comparison,” in *Homology Modeling*, pp. 231–257, Springer, 2011.

[88] I. Aier, P. K. Varadwaj, and U. Raj, “Structural insights into conformational stability of both wild-type and mutant EZH2 receptor,” *Scientific Reports*, vol. 6, p. 34984, 2016.

[89] L. Yang, G. Song, A. Carriquiry, and R. L. Jernigan, “Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes,” *Structure*, vol. 16, no. 2, pp. 321–330, 2008.

[90] B. J. Grant, A. P. Rodrigues, K. M. ElSawy, J. A. McCammon, and L. S. Caves, “Bio3d: an R package for the comparative analysis of protein structures,” *Bioinformatics*, vol. 22, no. 21, pp. 2695–2696, 2006.

[91] S. K. Mishra and R. L. Jernigan, “Protein dynamic communities from elastic network models align closely to the communities defined by molecular dynamics,” *PLOS One*, vol. 13, no. 6, p. e0199225, 2018.

[92] K. Kasahara, I. Fukuda, and H. Nakamura, “A novel approach of dynamic cross correlation analysis on molecular dynamics simulations and its application to Ets1 dimer–DNA complex,” *PLOS One*, vol. 9, no. 11, p. e112419, 2014.

[93] J. McCammon, “Protein dynamics,” *Reports on Progress in Physics*, vol. 47, no. 1, p. 1, 1984.

COMBINED NETWORK ANALYSIS AND MOLECULAR DYNAMICS SIMULATIONS STUDY FOR CHARACTERIZATION OF PREVALENT SOMATIC MUTATIONS IN BREAST CANCER: SF3B1 CASE STUDY

ORIGINALITY REPORT

9%	5%	5%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Nashville State Community College Student Paper	1%
2	d-nb.info Internet Source	<1%
3	www.nature.com Internet Source	<1%
4	res.mdpi.com Internet Source	<1%
5	journals.plos.org Internet Source	<1%
6	Submitted to University of Leicester Student Paper	<1%
7	Submitted to Jawaharlal Nehru University (JNU) Student Paper	<1%

Juan M. Jiménez-Vacas, Vicente Herrero-