FUNCTIONAL ENRICHMENT METHODOLOGY FOR ANALYZING OMICS DATA TO STUDY THE AETIOLOGY OF RARE DISEASES

by

Ceren Saygı

B.Sc., Biological Sciences and Bioengineering, Sabancı University, 2008

M.Sc., Biological Sciences and Bioengineering, Sabancı University, 2010

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Molecular Biology and Genetics

Boğaziçi University

2018

To my family and

Prof. Uğur Sezerman,

ACKNOWLEDGEMENTS

First of all, I am grateful to my thesis advisor Prof. Nesrin Özören for her support for the success of this collaborative study. She is an advisor who knows the importance of working in a project that one loves and hence allowed me to work on a topic that I love.

I am deeply grateful to my thesis co-advisor Prof. Uğur Sezerman. I met him fifteen years ago, the very first day of my Sabancı University undergraduate years. He is the one who helped me to continue in science while I was feeling exhausted and depressed. He is also the one who introduced me to the invaluable scientists and treated me extraordinarily tolerant and supportive. I consider him as a family member, not just an advisor. I believe that our paths will always intersect for a lifetime.

I would like to show my gratitude to my thesis committee members Prof. Esra Battaloğlu, Prof. Uğur Özbek, Prof. Yasemin Alanay and Prof. Necla Birgül for their valuable criticism.

I would particularly like to thank Prof. Yasemin Alanay. She is one of the major role players to me in succeeding this thesis since she came up with an idea of sending me to Hamburg to learn NGS data analysis and also guide me while writing projects and articles.

I would like to express the most profound appreciation to Prof. Nurten Akarsu, whenever I think about her, my eyes are brimming with tears. She gave me courage that I cannot even imagine. I had a chance to work with her for a few days, it was like a dream. I am working harder to get a chance to work with her in the future.

I thank you so very much to Prof. Aslıhan Tolun to encourage me in difficult times, give me the morale to move on and share her invaluable life experiences with me. I am going to come to visit her many years later, she is an academician I appreciate so much.

I would like to offer my special thanks to Prof. Kerstin Kutsche and Malik Alawi for accepting me in their labs. Prof. Kerstin Kutsche taught me very essential details regarding the variant prioritization, she has great contributions to my scientific way of thinking. When I face with the computer terminal for the first time years after the course of C ++ I took in my undergraduate years, Malik gave me a lot of effort to teach Linux and Bash scripting language and most importantly he didn't complain about this process at all. He is not just a colleague for me, but also the one who let me have good times in Hamburg.

I am grateful to all academic staff of the Boğaziçi University Molecular Biology and Genetics department for their contribution to my improvement and the department secretary Ümit Bayraktar for her limitless kindness and help. I would thank to Aslı Gündoğdu and Burçak Özeş for just being there for the hard times and letting me feel that I can always trust them. Also, I would like to thank Aslı Uğurlu, Elif Eren, and Nazmiye Özkan for their friendship. My sincere thanks also goes to Neşe Coşkun for being my heart sister and always empowering me since the day we met at Sabancı University.

I also have a few words to my lovely lab members that I received generous support: Thank you Okan to improve my Linux knowledge and correct simple mistakes of my scripts in the middle of the night without demoralizing me, and also for his talks when I feel a little under the weather. Thank you Arda for his great help in expanding my expertise in data analysis. Thank you Ege to answer my countless questions without any hesitation. If I have anything to ask, I come near him first all the time. Thank you Begüm to be there when things were not going as well as I expected and also thank you to comment on little annoying stuff with me without any hesitation. Thank you Nogay to tolerate me when I talk incessantly and narrate many stories almost every afternoon while he drives me to home. Thank you Aslı to give me the basic knowledge of molecular dynamics simulations and to be an example of being always strong, independent from what life brings. Thank you Rüçhan to be a very kind person and always push me to study more. Thank you Zeynep, Melis, Umut, Baran, Narod, Tuğçe, and Oğulcan to motivate me all the time.

And lastly, I thank my parents and my dearest brother who made the basis of my self-confidence, encouraged me to ct justly and supported me in my difficult times.

ABSTRACT

FUNCTIONAL ENRICHMENT METHODOLOGY FOR ANALYZING OMICS DATA TO STUDY AETIOLOGY OF RARE DISEASES

Rare diseases (RDs) are a large and diverse group of disorders and defined by low prevalence, in other words, it is any disease that affects a small percentage of the population. According to OMIM and Orphanet, ~7000 different RDs have been estimated, but the number of phenotypes that remain to be defined could be considerably higher. The difficulty in obtaining the correct diagnosis is the most dramatic problem to be solved for the patients, about 30% still lack a diagnostic definition. The patients living with rare diseases visit an average of 7.3 physicians before receiving an accurate diagnosis and the mean length of time from symptom onset to accurate diagnosis is 4.8 years. Late diagnoses delay specific treatments and may have severe and life-threatening consequences. Molecular diagnosis is the most prominent way to facilitate earlier and accurate diagnosis, and hence an effective treatment for rare undiagnosed cases. In this dissertation project, a novel bioinformatics workflow is constructed for whole-exome/genome sequencing data analysis, variant prioritization and pathogenicity prediction from a cascade of different tools shading light into different aspects of the diagnostic process. The pathogenicity mechanisms of mutations are elucidated via molecular dynamics (MD) simulations. The newly developed pipeline is planned to be used for diagnosis of undiagnosed patients with a suspected genetic disorder, where other testing modalities have been inconclusive or non-informative. The workflow was tested on several undiagnosed clinical cases with their family members and achieved high success rates by identifying the causative variant. For two of these families, the pathogenicity mechanisms of mutations were described via MD simulations, and these findings have been submitted to two different SCI journals and passed the editorial approval. The diagnosis of one of these families was Periventricular Nodular Heterotopia, while the other was Nail Dysplasia-10. Both of the diseases are extremely rare that is seen in one in a million cases. In conclusion, we developed a unique workflow for molecular diagnosis of rare undiagnosed diseases. Our pipeline contributes to the already existing knowledge through the combination of population frequency, pathogenicity prediction tools, gene intolerance scores, and MD simulations for the first time.

ÖZET

ÖZGÜN BİR OMİK VERİ ANALİZ YÖNTEMİ İLE NADİR HASTALIKLARIN ETİYOLOJİSİNİN TAYİNİ

Nadir hastalıklar, toplumda görülme sıklığı son derece düşük olan ve literature girmiş yaklaşık 7000 hastalığı başlığı altında toplayan bir hastalık grubudur. Günümüzde, nadir hastalıklarla yaşayan hastaların yaklaşık %30'una tanı konulamamaktadır, tanı alabilen hastalar için ise semptom başlangıcından kesin tanıya kadar geçen ortalama süre 4.8 yıldır ve hastalar kesin tanı almadan önce ortalama 7.3 doktoru ziyaret etmektedirler. Bu durum, hem genetik tanının diğer tanı yaklaşımları arasında daha erken önceliklendirilmesini hem de nadir hastalıklarla ilişkili yeni genlerin tanımlanmasını zorunlu kılmaktadır. Bu çalışmada, tüm ekzom/genom dizileme verisinin analizi, varyant önceliklendirme ve varyant patojenisite mekanizmasının açıklanması için farklı yöntemleri bir araya getirerek ve birkaç nadir hastalığı model olarak kullanarak, yeni bir analiz yöntemi geliştirdik. Bu analiz yöntemi tanı konmamış nadir hastalıklardan muzdarip birey ya da bireyler içeren aileler üzerinde test edilmiş, vakalarda patojenik varyantı tanımlayarak ve hastalığı teşhis ederek yüksek bir başarı oranı elde edilmiştir. Bu ailelerden ikisinde moleküler dinamik simülasyon yöntemi ile mutasyonların patojenite mekanizmaları açıklanmış ve bu bulgular rapor edilmek üzere SCI kapsamındaki dergilere gönderilmiş, yayına kabul edilmiştir. Bu ailelerden birinin teşhisi Periventriküler Nodüler Heterotopi olurken diğerininki ise Tırnak Displazisi-10 olmuştur. Bu hastalıkların ikisi de milyonda bir görülen son derece nadir hastalıklar sınıfına girmektedirler. Genleri hastalık ile ilişkilendirmek karmaşık, çok aşamalı bir süreçtir. Günümüzün büyük veri analiz aktivitelerinin çoğuna benzemekle birlikte, klinik doğası gereği daha da karmaşıklaşmaktadır. Sonuç olarak, şüphelenilen genetik bozukluğu olan ancak tanı konamamış hastaların tanısında kullanılmak üzere bir yöntem geliştirdik. Üç ana basamaktan oluşan yöntemimiz; varyantların populasyonda görülme sıklığını, birçok patojenite tahmin aracının kombinasyonunu ve gen intoleransı puanlarını moleküler dinamik simülasyonları ile birleştirerek ve nadir hastalıklara uygulayarak halihazırda varolan literatüre büyük katkı sağlamaktadır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS/ABBREVIATIONS

RD	Rare Disease
OMIM	Online Mendelian Inheritance of Men
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
NGS	Next-Generation Sequencing
SNV	Single Nucleotide Polymorphisms
GATK	Genome Analysis Toolkit
BWA	Burrows-Wheeler Alignment Tool
VarAFT	Variant Annotation and Filter Tool
VAAST	Variant Annotation, Analysis and Search Tool
TransVar	Trans-level variant annotator for precision genomics
MAGI	Bayesian-like Method for Metabolite, Annotation, and Gene Integration
UTR	Untranslated Region
ExAC	Exome Aggregation Consortium
GnomAD	Genome Aggregation Database
MGI	Mouse Genome Informatics
CADD	Combined Annotation Dependent Depletion
REVEL	Rare Exome Variant Ensemble Learner
M-CAP	The Mendelian Clinically Applicable Pathogenicity
SIFT	Sorting Intolerant From Tolerant
PolyPhen	Polymorphism Phenotyping
PDB	Protein Data Bank
NMR	Nuclear Magnetic Resonance
MD	Molecular Dynamics
VMD	Visual Molecular Dynamics
BQSR	Base Quality Score Recalibration
PCR	Polymerase Chain Reaction
VQSR	Variant Quality Score Recalibration
IGV	Integrative Genomics Viewer
MAF	Minor Allele Frequency

RVIS	Genic Intolerance Residual Variation Intolerance Score
pLI	Loss-of-Function Intolerance
z	Synonymous/Non-synonymous Intolerance
CNV	Copy Number Variation
BLAST	Basic Local Alignment Search Tool
I-TASSER	Iterative Threading Assembly Refinement
NpT	Constant Pressure and Temperature
NvT	Constant Volume and Temperature
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Flexibility
NAMD	Nanoscale Molecular Dynamics
YASARA	Yet Another Scientific Artificial Reality Application
GNM	Gaussian Network Model
ANM	Anisotropic Network Model
PME	Particle-Mesh Ewald
NH	Neuronal Heterotopia
PNH	Periventricular Nodular Heterotopia
NCDC 1-10	Nail Disorder, Nonsyndromic Congenital 1-10
CRD	Cysteine-Rich Domains
AA	Arachidonic Acid
DVL	Dishevelled
GFP	Green Fluorescent Protein

1. INTRODUCTION

1.1. Rare Diseases

Rare diseases (RDs) are a diverse group of diseases and defined on the basis of low prevalence, in other words, any disease which affects “small" percentage of a population can be added to the rare disease list (Boykott *et al.*, 2013). However, the threshold for defining “small” changes from region to region. In the European population, RD is classified as a disease which affects not more than five individuals per 10,000. In the USA, RD is defined as “any disease or condition that affects less than 200,000 people in the United States”. In Japan, the definition changes to a condition that affects less than 50,000 people in the country. About 80% of RDs have genetic origins and mostly still unknown. RDs also called orphan diseases because drug companies are not interested in adopting them to develop new treatments.

Patients with severe illnesses see physicians to be diagnosed and be provided the information needed. However, the life of people with rare diseases is not that easy, due to the low prevalence of their diseases and hence the lack of knowledge of their health care providers, they often face challenges. Among those challenges, the most persisting ones are delayed diagnosis and misdiagnosis. Despite the fact that early and accurate diagnosis is very critical, about 30% of patients still wait for diagnosis; moreover, the rest can get an accurate diagnosis quite late. The faster the diagnosis is made, the better for the patients; patients with a longer delay in diagnosis have a more severe disease in comparison to the time of initial diagnosis. Moreover, early diagnosis allows better management and facilitates primary preventive measures.

Unfortunately, in rare diseases, different conditions mostly produce overlapping symptoms, instead of characteristically manifesting themselves as repeating collections of stereotypical symptoms which together define a disease status. This is the main factor of why some patients remain undiagnosed despite undergoing an exhaustive workup. For instance, when patients see a clinician with symptoms of epilepsy and mental retardation, there are more than thirty rare diseases where these symptoms may be related. Periventricular Nodular Heterotopia, Pseudo-Torch Syndrome, Tranebjaerg Svejgaard

syndrome. It is very challenging to distinguish most rare diseases from each other, as symptoms can be very mild to distinguish one condition from another.

Even though RDs are individually infrequent, they are common in cumulative; there are approximately 7,000 described RDs. The majority (50-75%) affects children; they are collectively responsible for 35% of deaths in the first year of life, one-third of children born with a rare disease will not live to see their fifth birthday (Eurodis, 2005). The importance of new methods to help early and accurate diagnosis is extremely crucial, and awareness in this regard is increasing day by day. Molecular testing is the most prominent way to enable accurate diagnosis and potentially pave the way for appropriate treatment of rare undiagnosed diseases. It should be added to the routine testing approach for diagnosis of undiagnosed patients with a suspected genetic disorder. Furthermore, the addition of molecular diagnosis to the routine testing approaches is not enough, it must be prioritized earlier than other diagnostic approaches.

Turkey is a country where consanguineous marriage and having children at young ages are common. This cultural phenomenon raises the importance of rare disease studies in the country. In such a society where rare diseases are seen at higher percentages in comparison to the other societies, speed up and strengthen the diagnosis phase and the birth follow-up programs to prevent the increase in the number of patients is very critical for both the quality of life and the country's economy.

This section will give information about the two diseases which were diagnosed, in the scope of this thesis:

1.1.1. Periventricular Nodular Heterotopia

Neuronal heterotopia (NH) is defined by the presence of normal neurons such as periventricular, subcortical and leptomeningeal glioneuronal in an inappropriate location due to primary failure of neuronal migration (Barkovich *et al*., 2001, 2005, 2012; Guerrini and Barba, 2010). Periventricular nodular heterotopia (PNH, OMIM: 300049) is the most widespread type of neuronal heterotopia and defined by groups of normal neurons line the ventricular walls. It is the most significant reason of drug-resistant epilepsy, characterized

mainly by seizures; some affected individuals have been reported with additional symptoms like mild intellectual disability, movement problems and dyslexia (Fallil *et al*., 2015).

The genetic basis of PNH is X-linked dominant and was first found to be associated with F-actin-binding protein A (*FLNA*) gene in 1998 (Fox *et al*., 1998). In a study of PNH patients in 2001, *FLNA* mutations were detected in 83% of familial and 19% of sporadic cases (Sheen *et al*., 2001). This form of PNH is also associated with cardiac malformations, mainly in females (Bardon-Cancho *et al*., 2014; Parrini *et al*., 2006; Scherer *et al*., 2005). Other types of PNH have also been known; such as an autosomal recessive PNH due to mutations in the *ARFGEF2* gene (OMIM 608097), PNH with chromosome 5p duplications (OMIM 608098) (Sheen *et al*., 2004).

As more than one gene can cause the same disease, defects in one gene can also lead to more than one disease. This can happen for a variety of reasons: The mutations may fall in different locations of a protein, they may differ in their magnitudes of effect on function or have different functional effects on a protein. Tissue specificity can also cause different diseases for mutations in different parts of a gene. Moreover, modifier genes may cause more severe, less severe phenotypes or novel phenotypes. And finally, different mutations on the protein-protein interaction interface region of a protein may exhibit different etiology depending on which region of interfaces they are present in. *FLNA* is an example of such a single gene that causes more than one disease since it causes both neurological and non-neurological diseases. It recruits F-actin into extended networks, binds many cellular components other than F-actin; such as transcription factors, membrane receptors, enzymes and signaling intermediates (Stossel *et al*., 2001; Feng and Walsh, 2004; Popowicz *et al*., 2006).

FLNA consists of 48 exons and encodes a large (280-kD) cytoplasmic actin-binding phosphoprotein that connects membrane receptors to the actin cytoskeleton (Carroll *et al*., 1982; Chen *et al*., 1989; Patrosso *et al*., 1994). As we described in Figure 4.1, the protein is composed of three main functional domains: An actin-binding domain at the N-terminus, a rod domain with 23 repeats divided by two hinge regions and a dimerization and binding domain at the C-terminus (Noegel *et al*., 1989; Gorlin *et al*., 1990; Hock *et al*., 1990).

The majority of individuals with PNH with *FLNA* mutations are female with no mental retardation and partial epilepsy-(Kamuro and Tenokuchi, 1993; Dobyns *et al*., 1996, Poussaint *et al*., 2000). In contrast, liveborn males with *FLNA* mutations are very rare. The mutations are mostly lethal for males, as suggested by the common occurrence of miscarriages and premature male deaths of affected mothers and skewed sex-ratio in the families (Kamuro and Tenokuchi, 1993; Huttenlocher *et al*., 1994; Jardine *et al*., 1996; Moro *et al*., 2002). The fetal viability of male patients seems to depend on the severity of the *FLNA* mutation. Mild to moderate variants in surviving males; either missense, splice site or truncations near the C-terminus, mostly manifest milder clinical phenotypes in females and thus avoid detection of the disease. (Sheen *et al*., 2001; Moro *et al*., 2002). In other words, splicing or severe truncations presumed loss of function of the FLNA lead to male lethality and only partial-loss-of-function variants are found in surviving males. These observations in males point out the obligatory presence of FLNA during human embryonic development (Robertson, *et al*., 2005).

Since 1998, more than 60 mutations of *FLNA* have been reported in patients with PNH, and these mutations are distributed homogeneously on the protein (Robertson *et al*., 2003; Hidalgo-Bravo *et al*., 2005; Stefanova *et al*., 2005). However, if we consider male patients with 14 different *FLNA* mutations the picture is quite different. Instead of a homogeneous distribution, mutations seen in viable males are grouped into the beginning or end of the protein (Figure 1.1).

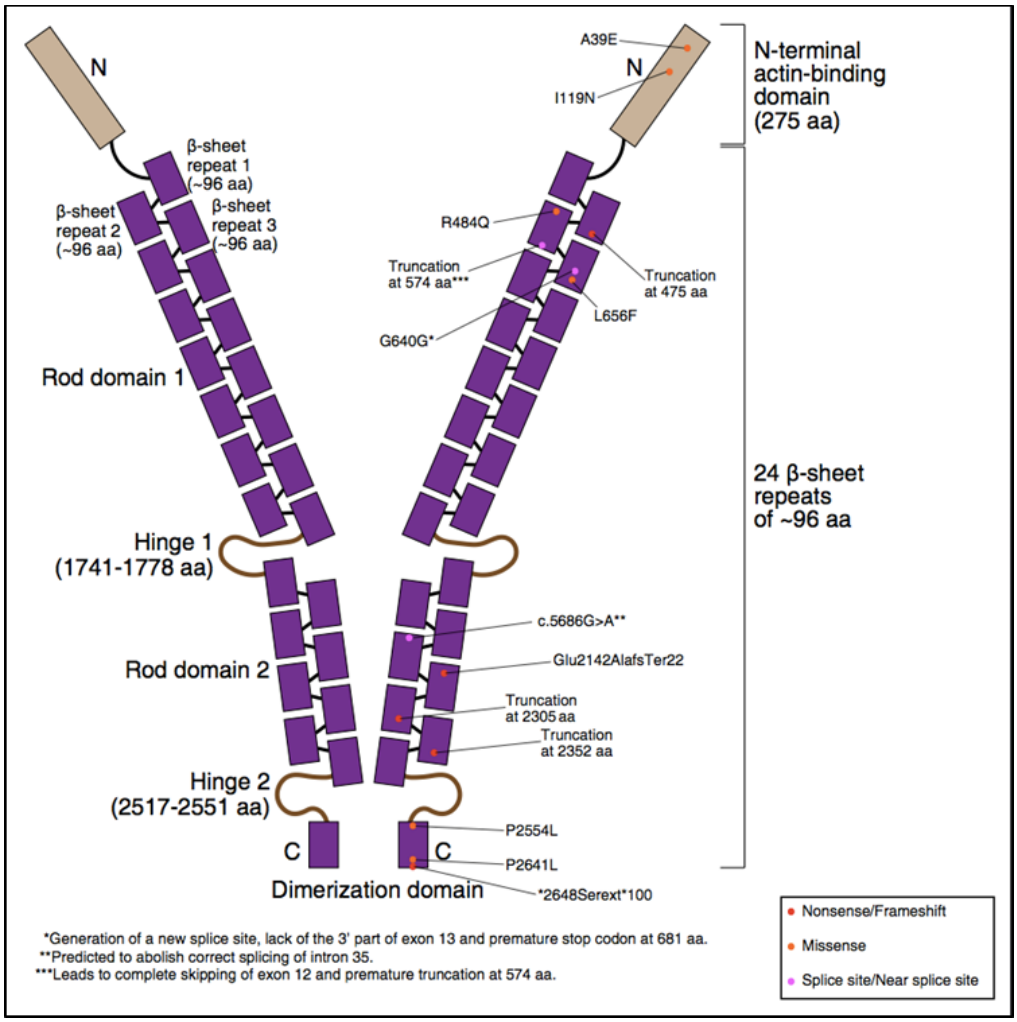


Figure 1.1. The schematic representation of reported male *FLNA* mutations in the literature

The first living male with *FLNA* mutation reported the literature in 2001 (Sheen *et al*., 2001). Since that day, only 19 male patients with *FLNA* mutations have been published, including the two male cases of this study (Sheen *et al*., 2001; Parrini *et al*., 2004; Gerard-Blanluet *et al*., 2006; Hehr *et al*., 2006; Kasper *et al*., 2012; Fergerot *et al*., 2012; Oegema *et al*., 2013; Oda *et al*., 2015; Lange *et al*., 2015; Liu *et al*., 2017). The type/location of mutations and the age of onset of male patients can be found in Table 1.1.

Table 1.1. *FLNA* mutations reported in viable males

Number of patients	AO	Amino acid Alteration	Nucleotide Alteration	Type of Mutation	Exon	Parental Consanguinity	Ref
1	-	Truncation at 2305aa	C to G at 6915bp	Nonsense	40	Sporadic	Sheen *et al*., 2001
1	-	Leu to Phe at 656aa	C to T at 1966bp	Missense	12	Sporadic	Sheen *et al*., 2001
1	15	Protein truncation at 574aa	A>G substitution	Splice Site	intron 11 splice site	Sporadic	Parrini *et al.*, 2004
Dizygotic twin boys	At birth	Pro2641Leu	7922C>T	Missense	48	Familial	Gerard-Blanluet *et al*., 2006
1	At birth	G640G, premature stop codon at 681aa	1923C>T	Splice Site	12	Not known	Hehr *et al*., 2006
1	15	abolish correct splicing of intron 35	5686G> A	Splice site	36	Familial	Kasper *et al.*, 2012
1	_	Ile119Asn	356T > A	Missense	2	Familial	Fergelot *et al*, 2012
1	20	Ala39Glu	116C > A	Missense	2	Familial	Fergelot *et al*, 2012
3	At birth	*2648Serext*100	7941_7942delCT	No-stop Frameshift	48	Familial	Oegema *et al*., 2013
2	At birth	Glu2142AlafsTer22	6425_6428delAGAG	Frameshift	40	Familial	Oda *et al*., 2015
1		Ser2352*	7055_7070delCTTTTGCAGTCAGCCT	Nonsense	42	Sporadic	Lange *et al.*, 2015
2	27 5	Pro2554Leu Gly475*	7661C>T 1425C>A	Missense Nonsense	46 10	Sporadic Sporadic	Liu *et al*., 2017
2	11 16	R484Q	G1451A	Missense	10	Familial	This study

1.1.2. Nail disorder, nonsyndromic congenital, 10

Nails grow over the nail bed throughout life as a result of matrix epithelial cell differentiation. The development of human nails starts around the ninth week of gestation and is completed during the fifth month of pregnancy (Baran *et al*., 2012). Human hereditary nail disorders are divided into 10 different subtypes (Nail disorder, nonsyndromic congenital 1-10; NCDC 1-10; OMIM 161050, 149300, 151600, 206800, 164800, 107000, 605779, 607523, 614149, 614157). They constitute a rare and heterogeneous group of ectodermal dysplasia and occur as isolated and/or syndromic ectodermal conditions, where other ectodermal appendages are also involved. Nail dysplasia can also be associated with skeletal dysplasia phenotypes. Five genes have been found to be associated thus far; namely *HPGD, RSPO4, PLCD1, COL7A1* and *FZD6* (Khan *et al*., 2015). Even though considerable advances were achieved in the diagnosis and

management of nail disorders, the knowledge of the molecular pathways of nail growth and morphogenesis is still quite limited.

In 2011, Fröjmark *et al*. were the first to identify the mutations in *FZD6* gene as a cause of autosomal recessive nail dysplasia (NCDC10, OMIM 614157). They reported two consanguineous Pakistani families with 11 members affected by isolated nail dysplasia (Khan *et al*., 2015; Fröjmark *et al*., 2011). According to the study of Fröjmark *et al*., the homozygous *FZD6* mutations (p.Glu584* and p.Arg511Cys) result in dysfunctional FZD6 and the loss of FZD6 followed by misregulation of several FZD6-mediated pathways needed for the formation and regeneration of nails in a proper manner. The action of FZD6 protein at the molecular level was studied by Cui *et al*. with the study of claw development in mice, and their findings pointed out a regulatory role for FZD6-mediated Wnt signaling in the differentiation process of claw/nail formation (Cui *et al*., 2013). To date, seven different mutations have been reported in eleven families, including two missense, two nonsense, two frameshifts and one compound heterozygous mutation (Table 1.2). Five of these seven mutations are clustered in the C-terminus which suggests that the C-terminal region could be a mutation hotspot. The discovered variants include: amino acid substitutions in highly conserved residues and nonsense/frameshift variants leading to signaling disruption in the C-terminal cytoplasmic domain (Fröjmark *et al*., 2011; Naz *et al*., 2012; Raza *et al*., 2013; Wilson *et al*., 2013; Kasparis *et al*., 2016; Mohammadi-asl *et al*., 2017).

Table 1.2. *FZD6* mutations that are found to be associated with NCDC10

Amino acid Change	**Mutation Type**	**Mode of Inheritance**	**Location**	**Ref**
Gly422Asp	Missense	Homozygous	6th transmembrane domain	Raza *et al.,* 2012
Arg509Ter	Nonsense	Homozygous	C-terminus	Wilson *et al.,* 2013
Arg511Cys	Missense	Homozygous	C-terminus	Fröjmark *et al.,* 2011
Gly559Aspfs*16	Frameshift	Homozygous	C-terminus	Kasparis *et al.,* 2016
Glu584Ter	Nonsense	Homozygous	C-terminus	Fröjmark *et al.,* 2011
Ser620Cysfs*75	Frameshift	Homozygous	C-terminus	Mohammadi-asl *et al.,* 2017
Arg96Cys/ Glu438Lys	Missense	Compound Heterozygous	N-terminus/3rd extracellular loop	Wilson *et al.,* 2013

As we described in Figure 1.2, FZD6 is composed of seven transmembrane domains (amino acids 202–222, 234–254, 284–305, 325–345, 371–391, 417–437, 474–494) and seven topological domains (amino acids 19–201, 223–233, 255–284, 306–324, 346–370, 392–416, 438–473) (Figure 1.2). The crystal structure of FZD6 has not yet been deposited to Protein Data Bank (PDB). It belongs to the frizzled family and, in general, frizzled family proteins expose their N-terminus on the extracellular side that contains a cysteine-rich domain (CRD) that binds the receptor's ligands (Yang-Snyder *et al*., 1996; Nusse *et al*., 2003). All known interaction partners bind the extracellular cysteine-rich domains (CRD) of FZD proteins (Rodriguez *et al*., 2005; Smallwood *et al*., 2007; Bafico *et al*., 1999; Nam *et al*., 2006; Mercurio *et al*., 2004; Dann *et al*., 2001). Even though this domain is necessary for ligand binding, it is not known to be necessary for signal transduction (Povelones *et al*., 2005). Mutagenesis studies have shown that there are several residues in the intracellular loops and the C-terminus that are very critical for signaling (Cong *et al*., 2004).

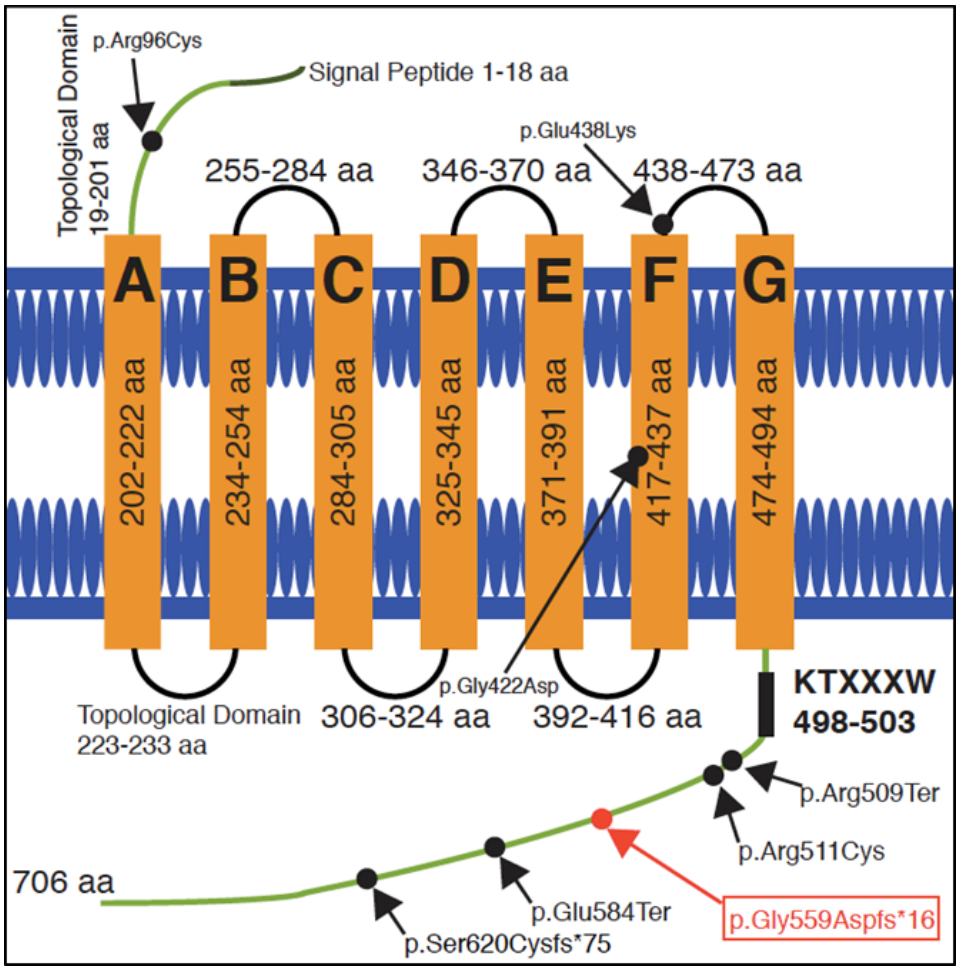


Figure 1.2. The schematic representation of reported *FZD6* mutations in the literature

1.2. Whole Exome/Genome Sequencing

Mendelian diseases are described based on the assumption that one variant is responsible for the disease of an individual, and the pattern of inheritance is consistent with the transfer of alleles at a single locus. The identification of the genetic defects in Mendelian diseases traditionally conducted by focusing on the regions that are inherited with the disease and linkage analysis is a predominant statistical method used for more than seventy years (Morton *et al*., 1955; Ott *et al*., 1999; Teare *et al*., 2005). Specifically, parametric linkage analysis is used for traits with a Mendelian form of inheritance. In this technique, polymorphic markers are utilized to follow the co-segregation of variations with the phenotype in the family (Teare *et al*., 2005). The LOD score, developed by Newton Morton, compares the probability of the two loci are indeed linked, to the probability of

observing this by chance. Positive LOD (logarithm (base 10) of odds) scores approve the existence of linkage, and negative LOD scores indicate that linkage is less probable. This technique can quantify the evidence of the involvement of variants in disease predisposition, but it cannot find the causative gene. Identifying the real causative variants responsible for linkage signals has challenges due to the difficulties in sequencing large regions highlighted by linkage peaks. Moreover, errors on the level of inbreeding have an enormous influence on the LOD scores (Miano *et al*., 2000). Some prefer to arbitrarily put first or second cousin consanguinity into the pedigrees if there is evidence for inbreeding without known details which makes the absolute value of the LOD score meaningless (Hildebrandt *et al*., 2009).

Homozygosity mapping ensures a rapid mapping of autosomal recessively inherited genes in consanguineous families via the identification of chromosomal regions with homozygous segments (ROH). ROH defines the genomic regions that occur if two copies of an ancestral haplotype came together in an individual. In consanguineous families, the same genomic segments are inherited from both parents, hence the members of these families have much more homozygous stretches on their genomes. This situation leads to a higher prevalence of recessive diseases in these families; homozygosity mapping is based on this observation. To date, ROH was found to be associated with an increased risk of schizophrenia (Lencz *et al*., 2007; Keller *et al*., 2012), Alzheimer's disease (Nalls *et al*., 2009; Ghani *et al*., 2015), autism (Chahrour *et al*., 2012; Lin *et al*., 2013), intellectual disabilities (Gamsiz *et al*., 2013), lung (Orloff *et al*., 2012), breast (Thomsen *et al*., 2015) and thyroid cancer (Thomsen *et al*., 2016). Moreover, ROH was also found to have an effect on inbreeding depression, bone mineral density (Yang *et al*., 2015), height (Joshi *et al*., 2015), cognitive ability. However, no effect of ROH was seen on several complex disorders, namely bipolar disorder, colorectal cancer, breast cancer, prostate cancer, and childhood acute lymphoblastic leukemia.

Even though the alternatives of traditional linkage analysis haven't become the common choice for homozygosity mappings, several methods have been developed. There are two main algorithms to identify ROH; sliding-window algorithms and Hidden-Markow Model (HMM) algorithms. Sliding-window algorithms scan each chromosome by moving a fixed size window throughout the genome in search of stretches of sequential

homozygous SNPs. This approach is applied in PLINK (Purcell *et al*., 2007), Homozygosity-Mapper (Seelow and Schuelke, 2012) and HomSI (Gormez *et al*., 2014). HMM algorithms account for background levels of linkage disequilibrium, like the one implemented in H3M2 (Magi *et al*., 2014) and BCFtools/RoH (Narasimhan *et al*., 2016).
As a result of new technical improvements, direct identification of the causative variants from WES/WGS data of affected individuals plus their healthy family members and omitting the predominant linkage step has apparently gained popularity. Moreover, it is also preferred to combine WES/WGS with linkage studies/homozygosity mapping to identify causal variants, since some rare variants are quite common in the general population. But, sample size with sufficient statistical power is very significant for the success of the linkage analysis. For instance, when only one affected offspring is present in a sibship, statistical power reduces dramatically (Wong *et al*., 1986). Hence, this combination is often successful for large families due to the noticeable decrease in statistical significance of the peaks for the trio analysis.

More than hundreds of publications about application of WES on rare diseases, ~25% report mutations from known disease-causing genes that match the symptoms of the patient being investigated and one can surely predict that the number of unpublished studies is much higher. As the number of known disease-causing genes grows, de facto conversion of WES from a research tool to a diagnostic tool becomes inevitable (Bamshad *et al*., 2011). Till now, most of the medical treatments were designed according to the one-size-fits-all approach; hence treatments are successful for some but not for others. Precision Medicine takes individual differences in patients' genes into account. Ivacaftor which was developed for cystic fibrosis patients can be an excellent example for this. Cystic fibrosis is an autosomal recessive disease that affects almost 70.000 people around the world and found to be associated with the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Genetic understanding of cystic fibrosis results in the categorization of the disease into subgroups. The channel arrives the cell surface, but channel activity is not sufficient in some subgroups, the channel stays in the cell cytoplasm in another subgroup. The Ivacaftor was designed to improve the opening time of activated CFTR channels at the cell surface. So, for patients whose channels do not reach to the cell surface, it might only have some minimal effect. In patients with adequately transported

channels effect could be dramatic, this is the situation for 5% of patients with the G551D mutation (Ramsey *et al*., 2011; Brodlie *et al*., 2015).

Since the first human genome was sequenced in 2001 at the cost of around US$3 billion, the price of sequencing an entire genome remained expensive to be used for routine medical practice. However, NGS approaches that entered the research setting in 2008 resulted with a significant decline in sequencing prices. Now, genomes can be sequenced for almost US$500. As a result, newly developed genomic applications pave the way not only for precision medicine but also for diagnosis of rare disorders, where conventional techniques have failed.

One of the obstacles for understanding the potential of WES in personalized medicine is the bioinformatics analysis which mostly requires a strong computer power. Analysis of WES data with publicly or commercially available algorithms requires a proper computational infrastructure. As a second, many publicly available algorithms focus on a single aspect and do not provide a workflow from start to finish which is required for the construction of a bioinformatics pipeline. Thirdly, there are no gold standards for translating WES into clinical knowledge, since different diseases may need different strategies for the data analysis.

Multiple methods have been developed to analyze data with respect to the different kinds of variants. For calling single nucleotide polymorphisms (SNVs) and short indels, software including the Genome Analysis Toolkit (GATK) (McKenna *et al*., 2010), FreeBayes (Garrison *et al*., 2012) and SAMtools (Li *et al*., 2009) can be applied in combination with a short-read aligner like Burrows-Wheeler Alignment Tool (BWA) (Li& Durbin, 2009) or Bowtie2 (Langmead *et al*., 2009). It has been shown for WES, that a combination of different variant callers and short-read aligners outperforms any single method. Sensitivity can be substantially increased with neglectable impact and specificity (Bao *et al*, 2014). A combination of tools is also required to target the different types of variants. An efficient analysis strategy would, on the one hand, need to integrate multiple methods for each type of variants, and on the other hand, it would also need to integrate results for the different types of variants into a single comprehensive solution.

Once the variants are obtained, there are a variety of tools that can be used for the annotation step; such as VarAFT, VAAST, TransVar, MAGI, SNPnexus, VarMatch and Annovar. Among them, Annovar is preferred more than the others since it is easily updated if new information is available for the annotation and has several functionalities that are very beneficial for the ones working on rare diseases. It is a command-line Perl program and can be used where standard Perl modules are installed. It annotates the functional effects of variants and can compare the frequency of variants in known variation databases, such as dbSNP, the 1000 Genomes Project.

In the present, our ability to sequence the genome is much greater than our ability to interpret an enormous number of resulting variants. Not only methods used for alignment, variant calling, and filtering can considerably influence the variant detection; but also differentiating disease-causing variants from an enormous number of candidate variants is a multidimensional task. A considerable proportion of cases still remain undiagnosed even after WES was performed.

In the previous paragraphs of this part of the introduction, the significance of the tools used for bioinformatic analysis was pointed out. Not only the tools used to obtain the variants are still changing from the lab to the lab, but also the variant prioritization strategy is also quite unique to each lab itself.

Variant interpretation means the determination of which of the potentially functional variants found in patients' genome actually contributes to their disease. Assumptions such as, if a variant creates a premature stop codon, it is more damaging than a missense change or if a variant creates a non-synonymous change, it is more damaging than a synonymous change are mostly ill-advised. A human carries lots of loss of function alleles in heterozygous and homozygous states on average. A stop codon in a less conserved gene could be more tolerated than a missense variant in a very conserved gene. Moreover, synonymous variations are associated with human diseases by affecting splicing (Sheikh *et al*., 2013) mRNA stability (Nackley *et al*., 2006) and altering protein conformation (Kimchi-Sarfaty *et al*., 2007). One must also keep in mind that a variant can be damaging but not disease-causing, so candidate variants filtered according to population frequency and classified by pathogenicity and intolerance tools must be evaluated very

carefully. Pathogenicity scores should never be considered as a hundred percent correct. They are useful start¬ing points, but they are only starting points, the strengths and weaknesses of those tools vary. In addition, over-filtering (excluding causative variants) and under-filtering (ending up with an enormous number of variants) of variants must be avoided. As a result of its complexity and impact on patient diagnosis, the prioritization process is based largely on expert interpretations and literature review.

The databases and tools that have been used in this dissertation project are listed below:

1.2.1. Exome Aggregation Consortium (ExAC) Database

Exome data for 60,706 individuals of various ancestries were assembled by the Exome Aggregation Consortium. 10,195,872 candidate variants were identified with a subset of 7,404,909 high-quality variants, including 317,381 insertions or deletions (indels). This data was used to create objective metrics of pathogenicity and to identify which genes are subjected to strong selection against several mutations (Karczewski *et al*., 2016).

They deviation of variations from expectation is quantified with a Z score (Samocha *et al*., 2014), it is zero for synonymous variants, but significantly shifted to higher values for missense and truncating variants. They also developed an expectation maximization algorithm by using the observed and expected truncating variant counts within each gene and separated each gene as a loss-of-function (LoF) intolerant (pLI) (pLI ≥ 0.9) or LoF tolerant (pLI ≤ 0.1) (Lek *et al*., 2016).

Using data from 60,706 exomes in Exome Aggregation Consortium, Cassa *et al*. also analyzed the genome-wide distribution of protein-truncating variants and developed a scoring system (shet) to predict the mode of inheritance probabilities and fitness loss due to the truncation mutations for each gene. They have also predicted the phenotypic severity, the age of onset, penetrance in a set of high-confidence haploinsufficient disease-associated genes (Cassa *et al*., 2017).

1.2.2. Genome Aggregation Database (GnomAD)

GnomAD is the next release version of ExAC and contains both 123,136 whole exome and 15,496 genome sequencing data from unrelated individuals. Individuals who were affected by severe pediatric diseases and their first-degree relatives were removed.

1.2.3. Genic Intolerance

Petrovski *et al*. developed a system to score gene intolerance, the system measures if genes have relatively less or more functional variation than expected based on neutral variations on the gene; using the data from the NHLBI Exome Sequencing Project. They support their system with the fact that the genes responsible for Mendelian diseases are less tolerant of genetic variation in comparison to the genes that associate with no disease. Basically, the lower the residual variation intolerance score as percentiles (RVIS%), the higher the intolerance of the gene.

The defined threshold to divide common and rare variants is r=0.1% MAF in the combined ESP6500 population. Y is the missense and truncating SNVs with MAF > r and X is the total number of protein-coding variants observed in a gene. They then regress Y on X and take the studentized residual as RVIS (Petrovski *et al*., 2013).

1.2.4. PrimatAI

PrimatAI trained deep neural network algorithm by using hundreds of thousands of common variants from the sequencing of six non-human primate species and identified pathogenic mutations in rare disease patients (Sundaram *et al*., 2018).

Even though assaying common variation across diverse human populations is an effective strategy for detecting benign variants (Lek *et al*., 2016), the total amount of common variation in humans is limited due to bottleneck events which a large fraction of ancestral diversity was lost (Mallick *et al*., 2016). Population sizes of databases are still not large enough, hence from out of more than 70 million potentially protein altering missense

substitutions in the reference genome, only around 1/1,000 are present at greater than 0.1% MAF (Lek *et al*., 2016, Liu *et al*., 2011).

Outside of modern human populations, chimpanzees comprise the next closest extant species, since we share 99.4% amino acid sequence identity (Chimpanzee Sequencing Analysis Consortium, 2005). If polymorphisms that are identical-by-state similarly affect fitness in the two species, the presence of a variant at high allele frequencies in chimpanzee populations should indicate benign consequence in human.

PrimateAI threshold of > 0.8 is for likely pathogenic classification, < 0.6 is for likely benign, and 0.6–0.8 is as intermediate in genes with dominant modes of inheritance, and a threshold of > 0.7 is for likely pathogenic and < 0.5 for likely benign in genes with recessive modes of inheritance (Sundaram *et al*., 2018).

1.2.5. Mouse Genome Informatics (MGI)

Model organisms are vital to reveal the mechanisms of human diseases. MGI is the international database resource to investigate the genetic basis of human diseases by translating information from mouse phenotypes and disease models. It provides phenotypes for over 50,000 mutant alleles in mice and provides experimental model descriptions for over 1500 human diseases. Curated data from scientific publications are integrated with those from high-throughput phenotyping and gene expression centers. Data are standardized using defined, hierarchical vocabularies such as the Mammalian Phenotype (MP) Ontology, Mouse Developmental Anatomy and the Gene Ontologies (GO) (Law *et al*., 2018).

1.2.6. Combined Annotation–Dependent Depletion (CADD)

CADD, REVEL, and M-CAP are designed to solve the same problem, and all three have different strengths and weaknesses. Neutral rare variants are more difficult to be distinguishable from disease-causing variants, and most tools tend to classify them as damaging (Li *et al*., 2013; Hodgkinson *et al*., 2013). Many of the tools predict the pathogenicity of missense variants based on the conservation of amino acid or nucleotide,

biochemistry of the amino acid substitutions and population frequency (Kumar *et al*., 2009; Li *et al*., 2009; Chun *et al*., 2009; Adzhubei *et al*., 2010; Schwarz *et al*., 2010; Reva *et al*., 2011; Choi *et al*., 2012; Shihab *et al*., 2013; Carter *et al*., 2013; Kircher *et al*., 2014; Quang *et al*., 2015; Niroula *et al*., 2015). However, individual tools often disagree since different predictive features are taken into consideration and combining the results of multiple predictors can improve performance (Gonzalez-Perez *et al*., 2011; Crockett *et al*., 2012; Lopes *et al*., 2012; Olatubosun *et al*., 2012; Li *et al*., 2012; Li *et al*., 2013; Frousios *et al*., 2013; Capriotti *et al*., 2013; Bendl *et al*., 2014; Dong *et al*., 2015).

The American College of Medical Genetics (ACMG) suggests that any variant with an allele frequency higher than 5% in control population can be classified as benign (Richards *et al*., 2015), and it is usual to lower the threshold to reduce the number of variants to a manageable number (Taylor *et al*., 2015; Lek *et al*., 2016). Even after such filtering, between 200-500 missense and truncating variants that are not present in any database of control individuals are remained (Taylor *et al*., 2015; Lek *et al*., 2016; 1000 Genomes Project Consortium, 2011). So, pathogenicity prediction tools that can differentiate between neutral and pathogenic rare variants are urgent needs.

CADD integrates information from several annotations of genetic variation into one score by contrasting variants that survived natural selection with simulated mutations (Paten *et al*., 2008). For mutation simulation, they used an empirical model of sequence evolution with CpG dinucleotide-specific rates. To generate annotations, they used the Ensembl Variant Effect Predictor (VEP) (McLaren *et al*., 2010), data from the ENCODE Project (ENCODE Project Consortium, 2012) and information from UCSC Genome Browser tracks (Meyer *et al*., 2013). Annotations include conservation metrics such as GERP (Cooper *et al*., 2005), phastCons (Siepel *et al*., 2005) and phyloP (Pollard *et al*., 2010); regulatory information (ENCODE Project Consortium, 2012) such as genomic regions of DNase I hypersensitivity (Boyle *et al*., 2008) and transcription factor binding (Johnson *et al*., 2007); transcript information such as distance to exon-intron boundaries or expression levels in commonly studied cell lines (ENCODE Project Consortium, 2012); and protein-level scores such as those generated with Grantham (Grantham *et al*., 1974), SIFT (Ng *et al*., 2003) and PolyPhen (Adzhubei *et al*., 2010). C scores are highest for

potential nonsense variants and then for missense and canonical splice-site variants, moreover intergenic variants has with the lowest C scores.

1.2.7. Rare Exome Variant Ensemble Learner (REVEL)

Many of the meta-predictors performed superior performance in comparison to different algorithms for the prediction of benign or pathogenic variants (Katsonis *et al*., 2014). REVEL and M-CAP are two well-known meta-predictors; they collect multiple tools into a single prediction output.

To develop REVEL scores, the team trained a random forest on the set of variants from the Human Gene Mutation Database (HGMD), Exome Sequencing Project (ESP) (Tennessen *et al*., 2012) and Atherosclerosis Risk in Communities (ARIC) study (The ARIC investigators, 1989) European-American and African-American populations, the 1000 Genomes Project (KGP) (Abecasis *et al*., 2012) European, Yoruban, and Asian populations by using the R ‘‘randomForest’’ package (Liaw *et al*., 2002) with 1,000 binary classification trees (Breiman *et al*., 2001; Hastie *et al*., 2009). They incorporated 18 pathogenicity prediction scores from 13 tools as predictive features: MutPred, FATHMM, VEST, Poly-Phen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, and phastCons (Ioannidis *et al*., 2016).

1.2.8. The Mendelian Clinically Applicable Pathogenicity (M-CAP)

The features M-CAP uses for classification are based on existing pathogenicity scores, measures of evolutionary conservation and the cross-species analog to the frequency within the human population (Hastie *et al*., 2003; Ogutu *et al*., 2011). To evaluate the pathogenicity of missense variants as a machine learning task, M-CAP uses pre-existing and new features. It uses nine established pathogenicity programs: SIFT13, PolyPhen-2, CADD15, MutationTaster, MutationAssessor, FATHMM22, LRT23, MetaLR16 and MetaSVM16. It also joints seven established measures of base pair, amino acid, genomic region, and gene conservation: RVIS24, PhyloP25, PhastCons26, PAM250, BLOSUM62, SIPHY28, and GERP29. Moreover, M-CAP brings 298 new features derived

from the multiple-sequence alignment of 99 primates, mammalian and vertebrate genomes to the human genome (Kuhn *et al*., 2013).

1.3. Computational Protein Structure and Functional Impact Prediction

Despite all of the advances, referring disease causation to prioritized variants still remains an inexact process. The most critical point to keep in mind is that: If a variant of a protein is found to be associated with a certain disease, this does not mean that the variant is pathogenic. Low population frequency, pathogenicity scores, intolerance scores, and more can only inform about how that variant might damage a gene; but since reporting it is a very big step that affects the patient's life, it is very essential to have additional proofs to show the pathogenicity. The importance of molecular dynamic (MD) simulations appeared at this stage.

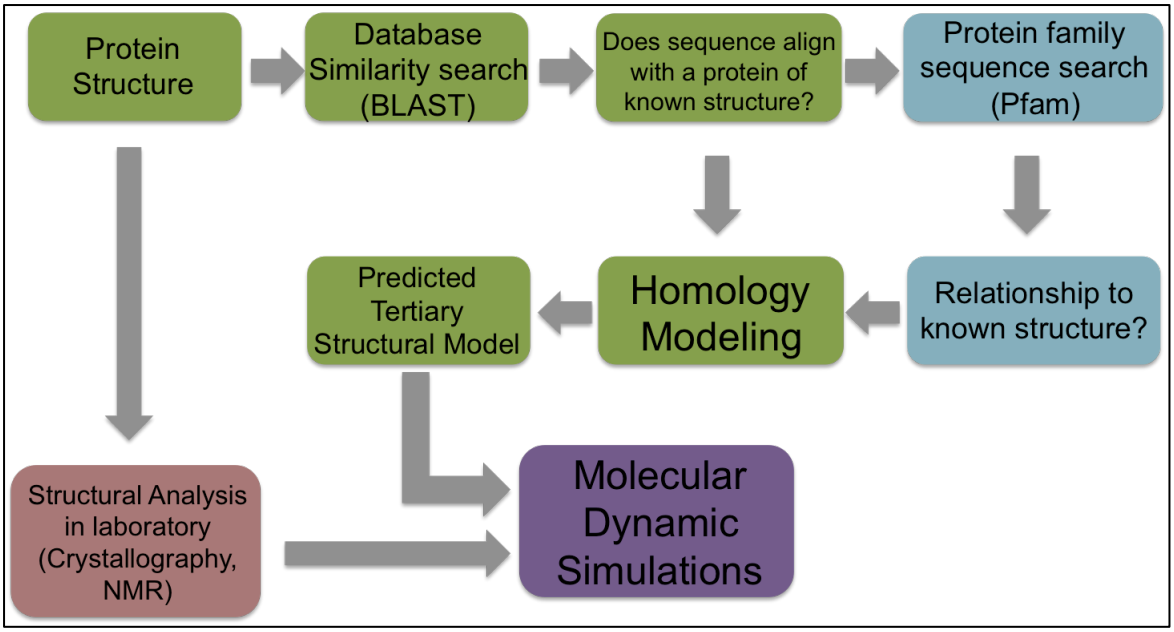


Figure 1.3. Workflow for computational protein structure determination

Above diagram shows the basic workflow to decide the most reliable way for protein structure prediction (Figure 1.3). The study of the macromolecular structure is a key point in the understanding of biology. Protein data bank (PDB) holds around 100,000 proteins crystal or NMR structure which is quite low when we consider all proteins in the human. However, obtaining a high-resolution structure of a protein requires a time-

consuming experimental process; due to the difficulties in the nature of crystallization and obtaining a sufficient amount of protein. In addition, proteins are flexible, and dynamics can play a significant role in their functionality, they also undergo conformational changes while performing their function and crystallographic studies cannot observe proteins in motion. So, both comparative and ab initio methods for the prediction of protein structure gains much more interest year by year; due to their speed in obtaining the results and their increase in reliability and consistency. These computational methods can be used not only to predict the structure of proteins but also to determine the effect of mutations on the protein.

In order to predict the structure of proteins or determine the effect of mutations on protein structures with the MD simulation, one must either have a crystal/NMR structure of a protein or obtain a predictive structure by comparative/homology modeling. Homology modeling means modeling a protein's structure by using a known experimental structure of another homologous protein. MD is a computer simulation method that measures the physical movements several hundreds of atoms and molecules. The interactions of atoms and molecules for a fixed period of time are allowed to give an idea about the dynamic evolution of the system. It was initially developed for theoretical physics in the late 1950s, but now it is also applied in materials science, chemical physics and the modeling of biomolecules.

2. PURPOSE

Rare diseases are a type of diseases which affect a small number of people in comparison to the general population. A remarkable number of patients remain undiagnosed despite undergoing an exhaustive workup. Early and accurate diagnosis is very critical since patients who present with longer diagnostic delays have more advanced disease compared to the time of initial diagnosis. Moreover, the sooner the diagnosis is made, the less the expenses will be.

The goal of this dissertation is to create a novel pipeline for analyzing WES/WGS datasets from undiagnosed patients with suspected rare Mendelian disorders. An effective approach is created for data analysis, variant prioritization and pathogenicity prediction from a cascade of different tools shading light into different aspects of the diagnostic process and using several undiagnosed rare diseases as a model. Linux operating system and Bash shell scripting are used for the bioinformatics analysis. The statistical analysis is carried out in R. For variant discovery, GATK, and for annotating SNP/indel calls, the software Annovar is used. Performing WES/WGS data analysis allows us to determine the quality of the variants obtained and also organize the resulting output format with the content appropriate for our variant prioritization strategy. We choose which tools to take into account and these tools automatically assigned their scores during bioinformatics analysis; the bioinformatics part involves prioritization part as well. And finally, molecular dynamics simulations are done to model both wild-type and mutant proteins and elucidate the pathogenicity mechanisms of mutations.

The WES datasets used to establish the bioinformatics methodologies are generated from several index cases with an undiagnosed disorder, plus their family members. Our novel approach achieved a high success rate by identifying the causative variant and providing the diagnosis. The pipeline contributes to the already existing knowledge through the combination of population frequency, pathogenicity prediction tools, gene intolerance scores and MD simulations for the first time.

3. MATERIALS AND METHODS

3.1. Tools, databases, cases, and primers

Table 3.1. Resources that must be downloaded (compiled) into the server before starting to the analysis

Tools	Databases
Trimmomatic version 0.36	Reference Assembly (hg19/GRCh37)
BWA	Indexes of the reference
Picard	HapMap 3.3 sites (hapmap.vcf)
Samtools	Omni 2.5 sites (omni.vcf)
GATK	1000G high-confidence sites (1000G.vcf)
BedTools	dbSNP (dbsnp.vcf)
R and R libraries ggplot2 and gsalib	Mills & 1000G Gold Standard Indels (mills.vcf)
ANNOVAR	MD5Sums
Linux	Annovar Database
VMD Software	ExAC database
CHARMM36 Software	GnomAD database
The FoldX Suite Software	1000 Genomes database
	OMIM database

A minimum of 4 to 8 Gb of computer memory is needed for exome, 12 Gb is needed for whole-genome.

3.1.1. Case I

The index patient is a sixteen year old male. There is no known parental consanguinity. His clinical symptoms are occipital lobe epilepsy and epileptic status in sleep. Pedigree analysis demonstrated one additional eleven year-old affected brother and healthy parents.

3.1.2. Case II

The index patient is a thirty-three year old female. Her parents are first cousins. She was diagnosed with congenital nail dysplasia at birth. She was referred for genetic counselling during her first pregnancy. She had thickened, hard, shiny, hyperplastic and hyperpigmented, claw-shaped (onycholysis) nails on the hands and feet. All nails in all four extremities were affected. Intermitently she loses her nails and the newly grown ones are similarly affected. They become hard, thickened and claw-shaped in time. Pedigree analysis demonstrated two additional affected sisters and a healthy brother with parental consanguinity (Figure 4.13). The patient and her sisters did not give consent for publishing of images. Her sisters were also recognized to have nail dysplasia from birth. Photographic images were examined and they suggested an identical phenotype. Both sisters were married with unrelated partners and had healthy children. The patient's was not related to her partner. Analysis of the pedigree suggested autosomal recessive inheritance pattern. Therefore, recurrence of the phenotype was considered low. She was counseled accordingly. She later gave birth to a healthy boy. Postpartum she was diagnosed with uveitis and treated for ocular tuberculosis without pulmonary involvement. Her treatment regimen included nine months of anti-tuberculosis agents. Corticosteroids were also used for the first two months.

Primers

- *FLNA*

  5'to 3' F: GAGGCAAGGGAGGGGTC
  5' to 3' R: CATCATCAGGTGGGGAGG

- *FZD6*

  5'to 3' F: CCAATCAGTGAAAGTCGAAGAGTAC
  5' to 3' R: TTCACTCCGCGCACTTTCA

3.2. Whole Exome/Genome Data Analysis Pipeline

In this part of the dissertation, WES/WGS data analysis will be explained in detail. The workflow we applied for WES/WGS data analysis composed of the following steps:

- Quality Control
- Preprocessing: Trimming the Ends and Adaptors
- Mapping the Reads to the Reference Genome: Preparation of input files and actual alignment
- Post-Alignment Processing
  Sorting and Conversion of .sam file
  Marking and Removing PCR Duplicates
  Indexing .bam file
  Base Quality Score Recalibration (BQSR)
- Germline Variant Calling (Haplotyping) and Joint Genotyping
  Haplotypingand Joint Genotyping
  Variant Quality Score Recalibration
- Variant Annotation
  Run Annovar
  Visualization of the Variants by IGV

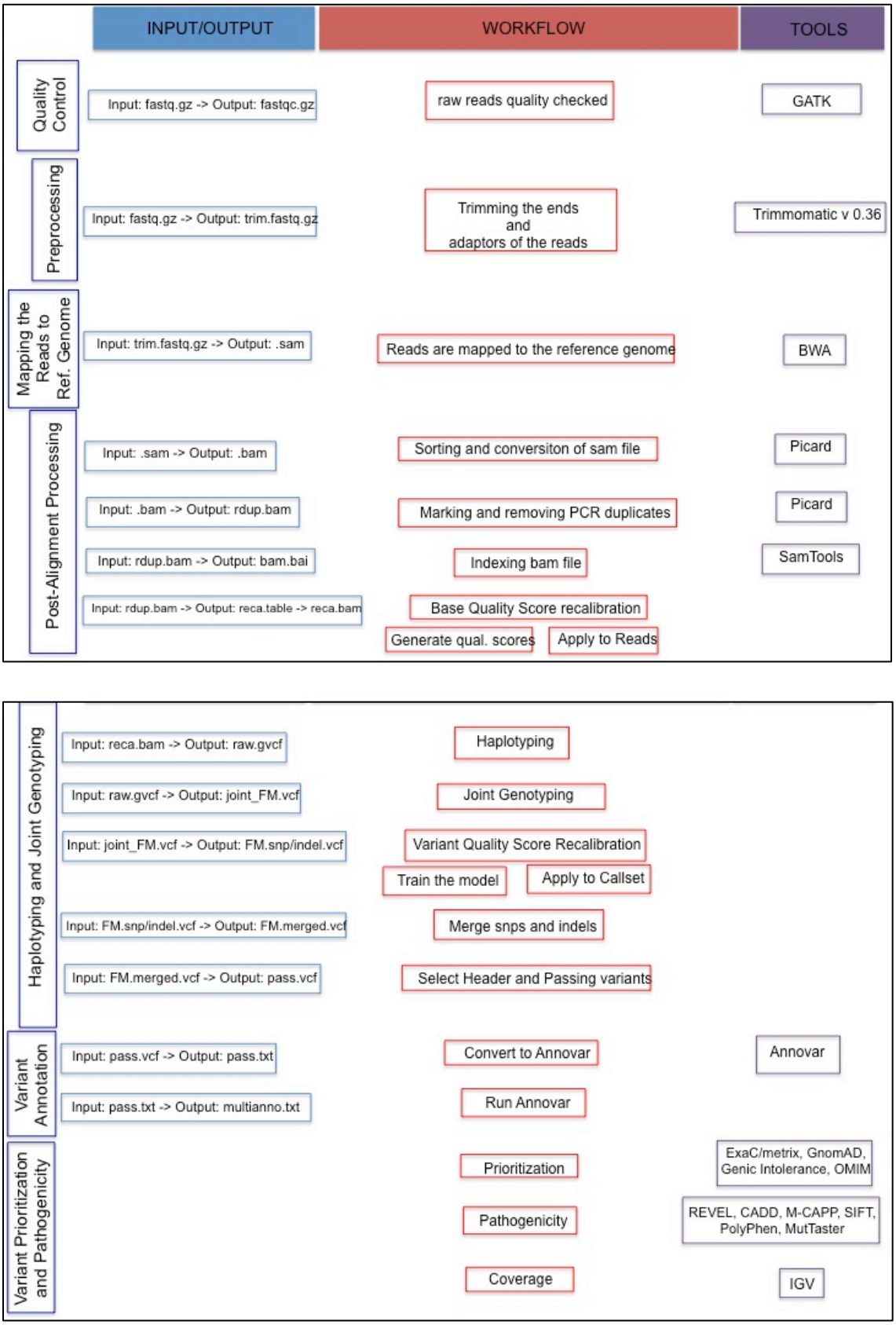


Figure 3.1. Scheme of the input&output file formats, workflow, and the tools used

3.2.1. Quality Control

The raw data of sequencing machine is called as FastQ files, it includes a sequence information like fasta files and also an information about the quality of the sequence. A fastq file typically consists of four lines for a single read (Figure 3.2).

```
@EAS100R:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCAG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Figure 3.2. Typical fastq file

The first line begins with @ and represents the name of the read and information about the position on the flow-cell for Illumina sequences (Figure 3.3).

```
@<instrument name>:<run-id>:<flow-cell-id>:<flowcell lane>:<tile within the
flow-cell lane>:<x-coordinate of the detected cluster>:<y-Position of the
detected cluster> <member of a pair>:<Y if the read is filtered, N otherwise>:<0
when control bits are on, even number otherwise>:<multiplex index>
```

Figure 3.3. First line of fastq files for Illumina sequences

The sequence of the read is found on the second line. A,C,G,T and N characters can be seen. The third line has either '+' sign or the read name after the '@' sign. The last line is the ASCII, it encodes of the Phred Scaled quality of the base two lines above.

Exome data we used in this dissertation are generated by a paired end sequencing. Sequence data has two files, all the forward sequences and all the reverse sequences in the same order. The molecules are greater than twice the reading length of an average Illumina experiment, do not overlap and have a gap between them. This approach helps to detect PCR duplicates.

Analysis of the quality of the FastQ files is the first step to do before starting the analysis.

fastqc *.fastq.gz -t 8

FastQC report is used which outputs graphics such as read-length plots, read-quality plots, sequence-duplication levels. Among all of the graphics, most crucial ones are per base sequence quality, per base sequence content and adaptor content. On the below, quality scores of R1 for one of the members of trios are seen and the most critical points that must be taken into consideration from each is explained briefly.

Measure	Value
Filename	62346540_S64_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	41076766
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	47

Figure 3.4. Basic statistics of fastq files

Sequence length and GC content is optimal, but total amount of sequence read is lower. This is one of the signs of low coverage (Figure 3.4).

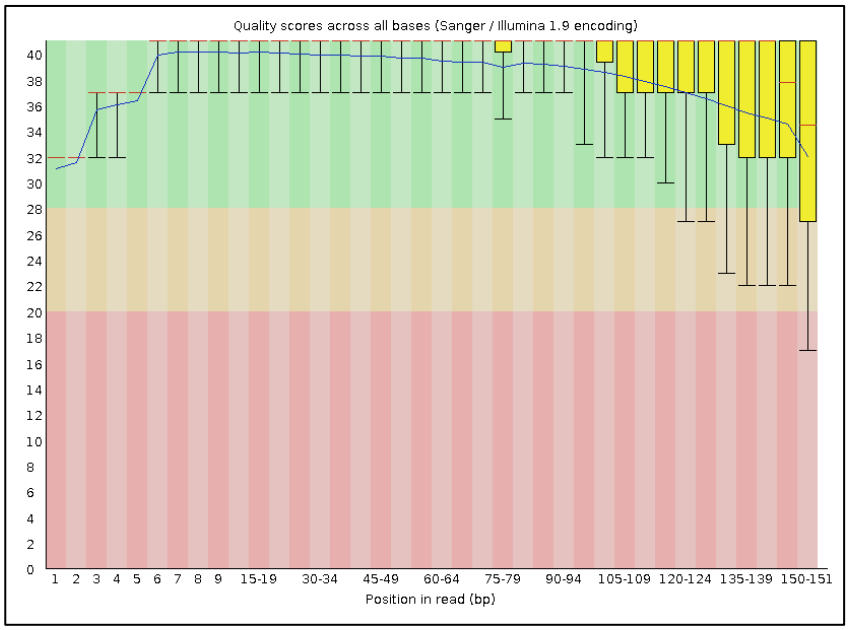


Figure 3.5. Per base sequence quality

The median value is the red line, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent 10% and 90% points and the mean quality is the blue line. This data looks consistent, quality is high along the reads (Figure 3.5).

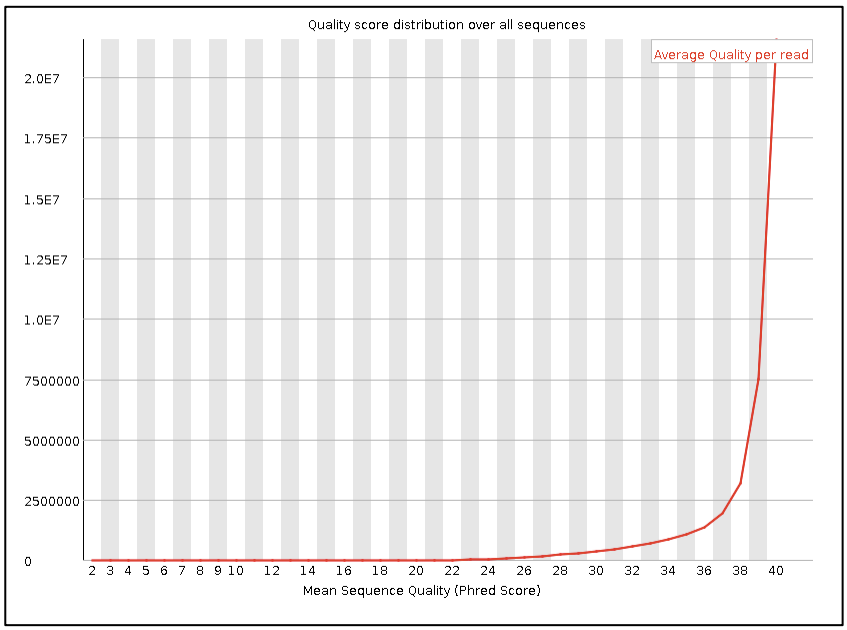


Figure 3.6. Per sequence quality scores

Quality scores are uniformly distributed, most are high quality sequences (Figure 3.6).

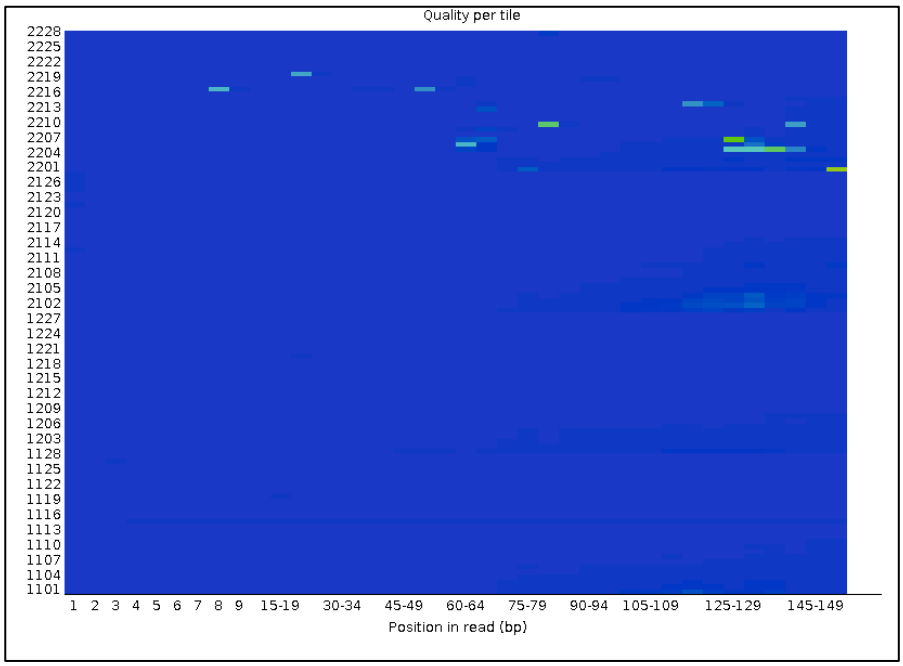


Figure 3.7. Per tile sequence quality

The plot shows the deviation from the average quality for each tile. The colours are on a cold to hot scale, colder colours mean the quality is at or above the average for that base, and hotter colours mean a tile had worse qualities than other tiles for that base. A good plot should be all blue if there is no loss in quality associated with only one part of the flowcell and this plot's quality looks quite high (Figure 3.7).

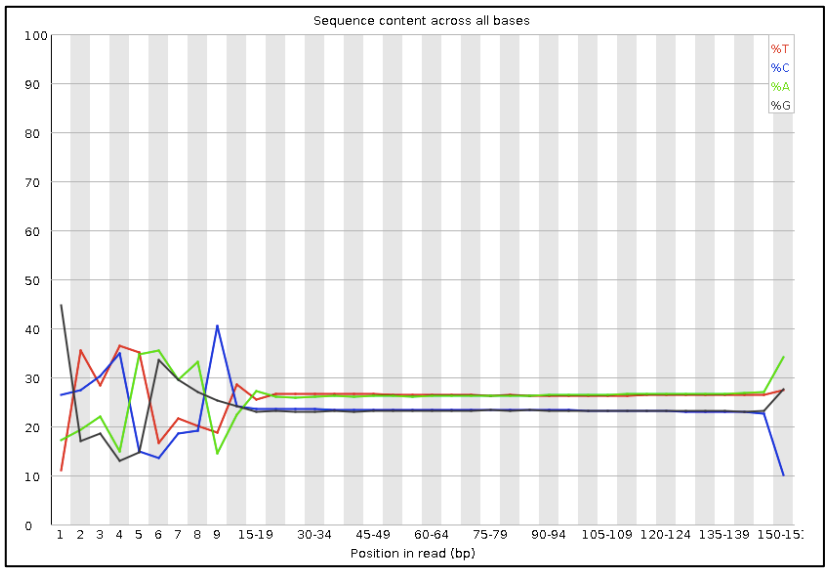


Figure 3.8. Per base sequence content

Assuming the data is a random sample from the sequence space, the base content at each position the contribution should be identical. Thus, straight lines are expected to be seen. The first few bases might indeed show some erratic behavior, which could be due to noncompletely random primers. For this case, the ends have to be trimmed (Figure 3.8).

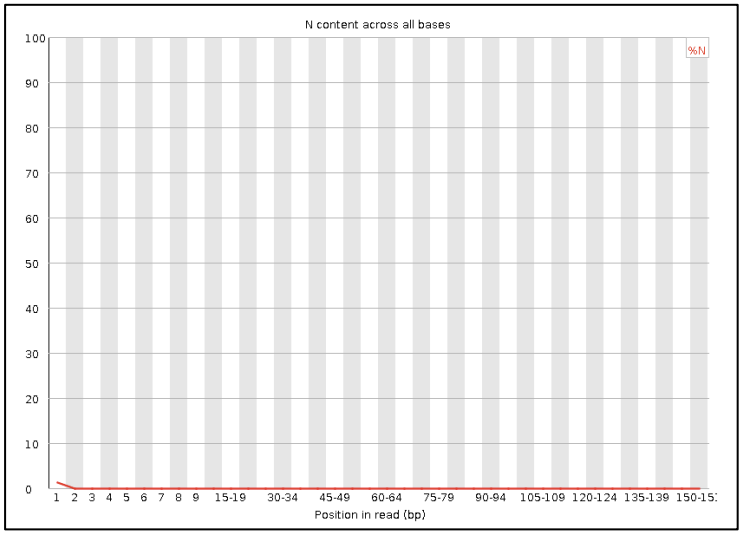


Figure 3.9. Per base N content

If a sequencing machine cannot make a confident base call then it will represented by N. For this case, unknown base is not seen (Figure 3.9).

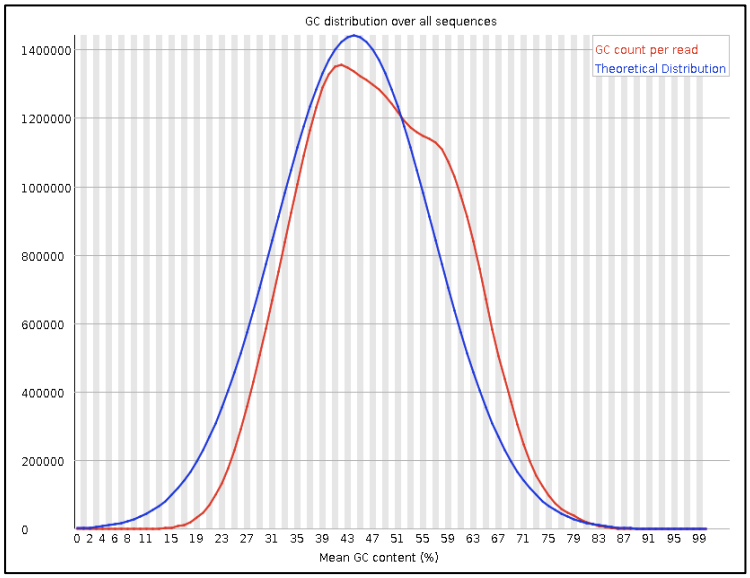


Figure 3.10. Per sequence GC content

The GC content of each sequence is measured and compared with a modelled normal distribution of GC content. The above result fits with expected (Figure 3.10).

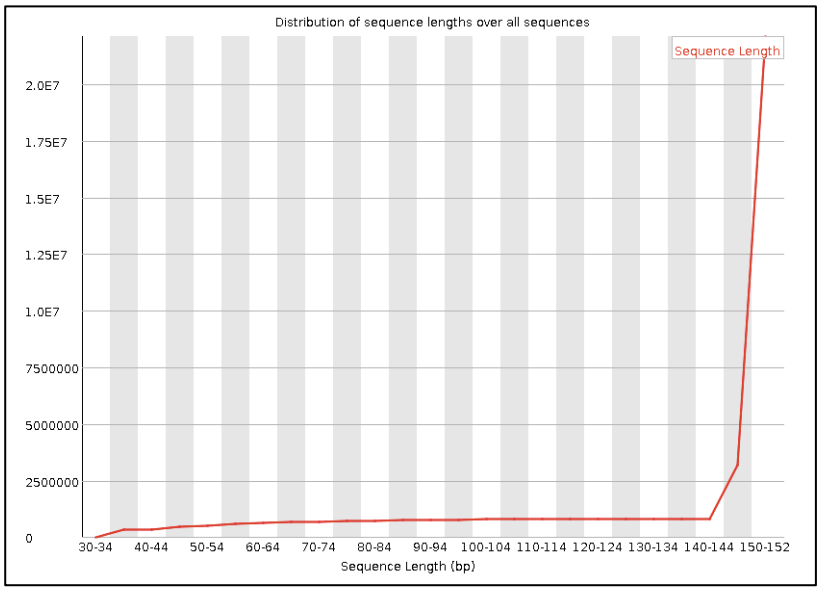


Figure 3.11. Sequence length distribution

This graph shows the distribution of fragment sizes in the data. The reads form our data are mostly around 150bp, as expected (Figure 3.11).

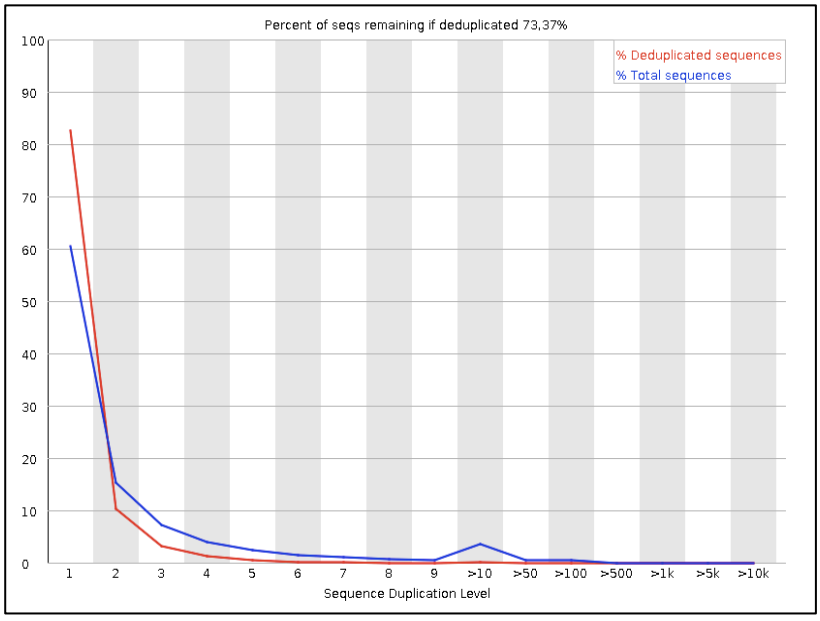


Figure 3.12. Sequence duplication levels

The degree of duplication is counted for every sequence in a library and the plot shows the relative number of sequences with different degrees of duplication. For this case, the duplication level is very low (Figure 3.12). A low level of duplication may indicates a very high level of coverage of the target sequence and the lesser the PCR artifacts seen, whereas a high level of duplication may indicates enrichment bias such as PCR over amplification.

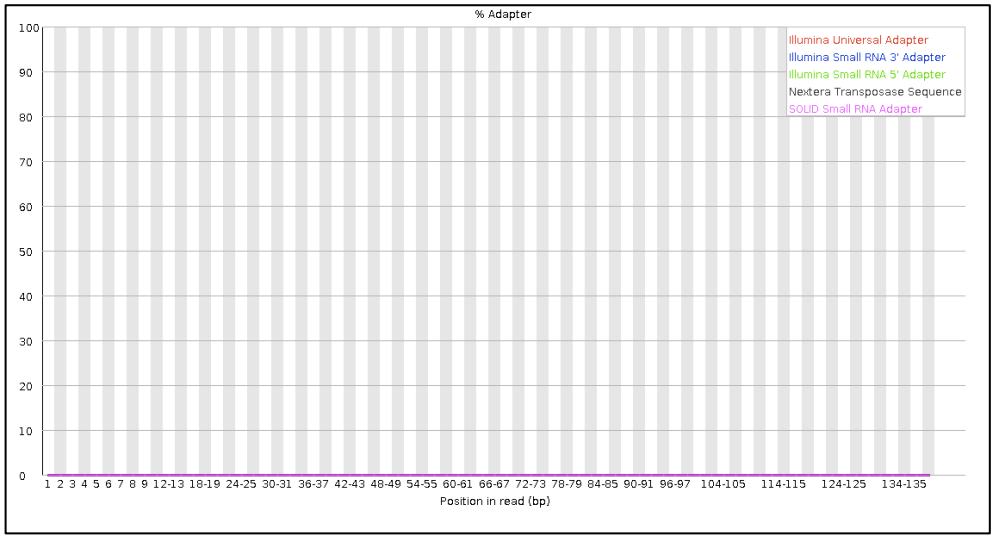


Figure 3.13. Adapter content

Adaptor contents are one kind of overrepresented sequences in the library, they have to be trimmed not to negatively affect the mapping process. For this case, adaptor contamination is not detected (Figure 3.13).

3.2.2. Preprocessing

When one receives sequence data from a sequencing provider, the data is typically in a raw state, in other words, it is not immediately usable for variant discovery analysis. Raw data quality check and trimmimg the adaptors/ends of the reads need to be done before starting to map the sequence reads to the reference genome.

Adapters must be ligated to the 5’ and 3’ ends of each single DNA molecule after fragmentation step for Illumina short read sequencing. These adapter sequences hold

barcoding sequences, forward/reverse primers and the binding sequences to immobilize the fragments to the flow cell and allow bridge-amplification.

Since the adapter sequences are synthetic and they are not seen in the genomic sequence, adapter contamination leads to NGS alignment errors and an increased number of unaligned reads. Hence, they need to be trimmed before starting to mapping. Not only adaptors but also end bases are needed to be clipped in order to cope with lower quality bases. Typically they are seen in the 3' end. For both ends and adaptors trimming, Trimmomatic version 0.36 is used.

INPUT: fastq.gz -> OUTPUT: trim.fastq.gz

```
for x in *_R1_*fastq.gz; do java -jar /opt/bioinf/Trimmomatic-0.36/trimmomatic-0.36.jar
PE $x ${x/_R1_/_R2_} ${x/_001.fastq.gz/_trim.fastq.gz}
${x/_001.fastq.gz/_unpaired.fastq.gz} ${x/R1_001.fastq.gz/R2_trim.fastq.gz}
${x/R1_001.fastq.gz/R2_unpaired.fastq.gz} -threads 24 CROP:150 TRAILING:10
MINLEN:40; done
```

3.2.3. Mapping the Reads to the Reference Genome

For humans the most well-known options for reference sequence are UCSC and Ensembl. We have used the UCSC release of the human genome hg19.

Necessary Resources

- Software

  BWA

  SAMtools

  Picard Tools

  Genome Analysis Toolkit (GATK)

  R and R libraries ggplot2 and gsalib
- Hardware

  A minimum of 2 Gb of memory is recommended for WES and 4 to 8 Gb is for WGS. The software can run on Linux or MacOS X with Java 1.7 installed.

- Files

  Sequence trimmed reads in fastq format (trimmed_reads.fq)

  The human reference genome in fasta format (reference.fa)

A database of known variants in vcf format (dbsnp137.vcf) If it is compressed, it must be unpacked.

Gunzip *gz

The GATK uses two files to access the reference file: a dictionary of the contig names and sizes, an index file to allow random access to the reference bases. These files must be generated to use the reference file in fasta format with GATK tools.

i. The BWA index is generated by running this BWA command:
   bwa index -a bwtsw /opt/storage/GATK_hg19/ucsc.hg19.fasta
ii. The fasta file index is generated by running this SAMtools command:
   samtools faidx /opt/storage/GATK_hg19/ucsc.hg19.fasta
iii. The sequence dictionary is generated by running the this Picard command:
   java -jar /tools/picard/picard.jar CreateSequenceDictionary \
   REFERENCE=/opt/storage/GATK_hg19/GATK_hg19/ucsc.hg19.fasta \
   OUTPUT=/opt/storage/GATK_hg19/GATK_hg19/ucsc.hg19.dict

Burrows-Wheeler Aligner (BWA2) is used to map the reads to the human genome. It works very well with Illumina data and has the ability to run using several threads.

INPUT: trim.fastq.gz -> OUTPUT: .sam

```
for x in *_1_paired.fastq.gz; do bwa mem -t 12 -M -R
"@RG\tID:"${x/_1_paired.fastq.gz/}"\tSM:"${x/_1_paired.fastq.gz/}"\tPL:illumina\tLB:st
andard" /opt/storage/GATK_hg19/ucsc.hg19.fasta $x ${x/_1_/_2_} >
${x/_1_paired.fastq.gz/.sam}; done
```

BWA uses many different options to adjust the mapping. The -t commands BWA the number of threads to use, the –M marks shorter split hits as secondary and –R STR

completes read group header line. BWA outputs the alignments in .sam format; summarizing position, quality, and structure for each read (Figure 3.14).

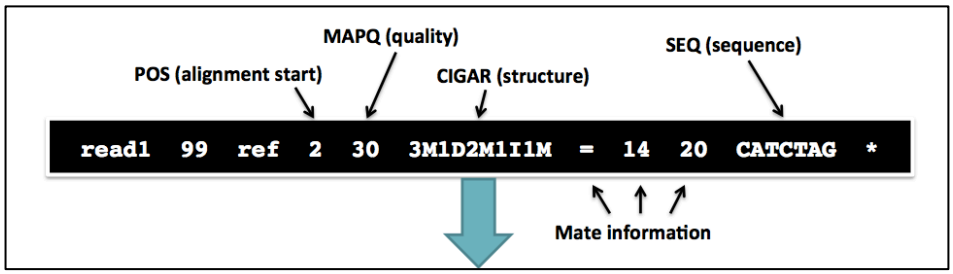


Figure 3.14. .sam file format

3.2.4. Post-Alignment Processing

The .sam file is the starting point to obtain the binary alignment/map format (.bam). For converting the .sam file to .bam Picard is used. It compresses and indexes the .sam file, so portions of the file can be accessed without the need to load the whole file.

INPUT: .sam -> OUTPUT: .bam

(for x in *.sam; do java -Xmx4g -jar /tools/picard/picard.jar SortSam 'INPUT='$x 'OUTPUT='/mnt/data/${x/.sam/.bam} SORT_ORDER=coordinate; done) |& tee run.log

Before doing any kind of manipulation or analysis using a reference sequence, the files usually has to be indexed. .bam index file is created by using Samtools as .bam.bai.

INPUT: rdup.bam -> OUTPUT: .bam.bai

for x in *rdup.bam; do samtools index $x; done

Remove Duplicates. PCR duplicates arise at the step of PCR amplification of fragments. Since they share the same sequence and the same alignment position, they can lead to problems in SNP calling since some allele may be overrepresented due to

amplification biases. Again, Picard is used to mark PCR duplicates in the .bam with a certain tag and then remove those marked ones.

INPUT: .bam -> OUTPUT: rdup.bam

```
(for x in *bam; do echo $x; java -Xmx3g -jar /tools/picard/picard.jar MarkDuplicates
'INPUT='$x 'OUTPUT='${x/.bam/.rdup.bam} METRICS_FILE=metrics.txt
REMOVE_DUPLICATES=true; done) |& tee run.log
```

Base Quality Score Recalibration (BQSR). DNA sequencing machines provide a rate for the quality of each base that they read, which is called Phred score. A Phred score of 10 represents 90% accuracy, 20 equals 99%, 30 equals 99.9% and so on. The produced scores are subject to technical errors, leading to over- or under-estimated base quality scores. BQSR is a machine learning process to model these errors empirically and adjust the quality scores accordingly. More accurate base qualities are obtained and this increases the accuracy of the variant calls.

The base recalibration process involves two steps: Generating quality scores and applying the scores to bases. The program builds a model of covariation based on the data and a set of known variants, then it adjusts the base quality scores in the data based on the model.

Generating Quality Scores. INPUT: rdup.bam -> OUTPUT: reca.table

```
for x in *.rdup.bam; do java -Xmx12g -jar /opt/bioinf/gatk/GenomeAnalysisTK.jar --
num_cpu_threads_per_data_thread 12 -T BaseRecalibrator -R
/ref/Homo_sapiens/gatk_hg19/ucsc.hg19.fasta -I $x -knownSites
/ref/Homo_sapiens/gatk_hg19/dbsnp_138.hg19.vcf -knownSites
/ref/Homo_sapiens/gatk_hg19/Mills_and_1000G_gold_standard.indels.hg19.vcf -
knownSites /ref/Homo_sapiens/gatk_hg19/1000G_phase1.indels.hg19.vcf -o
${x/.bam/.reca.table}; done
```

Applying the Scores to Bases. INPUT: reca.table -> OUTPUT: reca.bam

```
for x in *.rdup.bam; do java -Xmx12g -jar /opt/bioinf/gatk/GenomeAnalysisTK.jar --
num_cpu_threads_per_data_thread 12 -T PrintReads -R
/ref/Homo_sapiens/gatk_hg19/ucsc.hg19.fasta -I $x -BQSR ${x/.bam/.reca.table} -o
${x/.bam/.reca.bam}; done
```

Optional Before/After Plotting for Chromosome 20 only. There is also an optional step for building a second model and generating before/after plots to see the effects of the recalibration process. This tool performs the first step described above: it builds the model of covariation and produces the recalibration table.

```
for x in 62357526_S10.rdup.bam; do java -Xmx12g -jar
/opt/bioinf/gatk/GenomeAnalysisTK.jar --num_cpu_threads_per_data_thread 12 -T
BaseRecalibrator -R /ref/Homo_sapiens/gatk_hg19/ucsc.hg19.fasta -I $x -knownSites
/ref/Homo_sapiens/gatk_hg19/dbsnp_138.hg19.vcf -knownSites
/ref/Homo_sapiens/gatk_hg19/Mills_and_1000G_gold_standard.indels.hg19.vcf -
knownSites /ref/Homo_sapiens/gatk_hg19/1000G_phase1.indels.hg19.vcf -o
${x/.bam/.pre.table} -L chr20; done

for x in 62357526_S10.rdup.bam; do java -Xmx12g -jar
/opt/bioinf/gatk/GenomeAnalysisTK.jar --num_cpu_threads_per_data_thread 12 -T
BaseRecalibrator -R /ref/Homo_sapiens/gatk_hg19/ucsc.hg19.fasta -I $x -knownSites
/ref/Homo_sapiens/gatk_hg19/dbsnp_138.hg19.vcf -knownSites
/ref/Homo_sapiens/gatk_hg19/Mills_and_1000G_gold_standard.indels.hg19.vcf -
knownSites /ref/Homo_sapiens/gatk_hg19/1000G_phase1.indels.hg19.vcf -o
${x/.bam/.post.table} -BQSR ${x/.bam/.pre.table} -L chr20; done

for x in 62357526_S10.rdup.bam; do java -Xmx12g -jar
/opt/bioinf/gatk/GenomeAnalysisTK.jar -T AnalyzeCovariates -R
/ref/Homo_sapiens/gatk_hg19/ucsc.hg19.fasta -L chr20 -before ${x/.bam/.pre.table} -after
${x/.bam/.post.table} -plots recalibration_plots.pdf; done
```

3.2.5. Haplotyping and Joint Genotyping

Once the pre-processed data are obtained, everything is ready for getting SNP calls. The next step is identifying the sites where data displays variation relative to the reference genome taking into consideration that some variations can be caused by mapping and sequencing artifacts.

The workflow involves running Haplotype Caller on each sample separately in gvcf mode to get an intermediate file format. The gvcfs of multiple samples are then run through a joint genotyping to get a multi-sample vcf callset, which can then be filtered to balance sensitivity and specificity as desired.

Necessary Resources

- Software
  GATK
- Hardware
  A minimum of 2 Gb of memory is recommended for WES and 4 to 8 Gb is for WGS. The software can run on Linux or MacOS X with Java 1.7 installed.

- Files
  The processed sequence data in bam format.
  The human reference genome in fasta format (reference.fa)

The Haplotype Caller program can call SNPs and indels via local *de novo* assembly if it encounters an active region, a region that shows many variations. It discards the current mapping information for that region and reassembles the reads. Haplotype Caller algorithm:

- Define active regions: The program detects regions that show significant variation and then applies *de novo* assembly to those regions.
- Determine haplotypes of the active region: For each active region, the program assembles a De Bruijn-like graph for the reassembly and identification of

haplotypes and then to identify potentially variant sites it realigns each haplotype against the reference haplotype using the Smith-Waterman algorithm.

- Determine likelihoods of the haplotypes: For each active region, PairHMM algorithm is used to perform a pairwise alignment of each read against each haplotype and a matrix of likelihoods of haplotypes are produced. These probabilities are then marginalized to obtain the likelihoods of alleles for each variant site.
- Assign sample genotypes: For each variant site, the program performs Bayes' rule, then the most likely genotype is assigned to the sample.

INPUT: reca.bam -> OUTPUT: raw.g.vcf

```
for x in *.reca.bam; do java -Xmx36g -jar /tools/GATK/GenomeAnalysisTK.jar --
num_cpu_threads_per_data_thread 12 -T HaplotypeCaller -R
/opt/storage/GATK_hg19/ucsc.hg19.fasta -I $x --emitRefConfidence GVCF -o
${x/.reca.bam/.raw.g.vcf}; done
```

GenotypeGVCFs merges gVCF records that were produced. At each position of the input gVCFs, this tool combines all records, adjust genotype likelihoods, re-genotype the recently combined record and re-annotate it.

INPUT: raw.g.vcf -> OUTPUT: joint_FM.vcf

```
java -Xmx32g -jar /opt/bioinf/gatk/GenomeAnalysisTK.jar -T GenotypeGVCFs -R
/ref/Homo_sapiens/gatk_hg19/ucsc.hg19.fasta --variant 62304381_S42.rdup.raw.g.vcf --
variant 62304392_S43.rdup.raw.g.vcf --variant 62304396_S44.rdup.raw.g.vcf --variant
62304409_S45.rdup.raw.g.vcf -o 62304381_62304392_62304396_62304409_joint.vcf
```

3.2.6. Variant Quality Score Recalibration (VQSR)

Variant quality score recalibration (VQSR) uses machine learning to identify annotation profiles of variants that are likely to be real. It requires a large callset (minimum 30 exomes, more than one whole genome if possible) and highly curated sets of known variants. The aim is to assign a well-calibrated probability to each variant call to create accurate call sets by filtering.

Necessary Resources

- Software

  GATK

  RStudio, R libraries ggplot2 and gsalib
- Hardware

  A minimum of 2 Gb of memory is recommended for WES and 4 to 8 Gb is for WGS. The software can run on Linux or MacOS X with Java 1.7 installed.
- Files

  Call set in vcf format (raw_variants.vcf)

  The human reference genome in fasta format (reference.fa)

  Sets of known/true variants in vcf format for training the model:

  HapMap 3.3 sites (hapmap.vcf) (International HapMap 3 Consortium *et al*., 2010)

  Omni 2.5 sites (omni.vcf) (Durbin *et al*., 2010)

  1000G high-confidence sites (1000G.vcf) (Durbin *et al*., 2010)

  dbSNP (dbsnp.vcf) (Sherry *et al*., 2001)

  Mills & 1000G Gold Standard Indels (mills.vcf)

The approach is composed of two steps:

Train Model Using HapMap (Separate models for SNPs and Indels). This model is based on known sites and Gaussian mixture model is used to determine different parameters. Aim is to develop an estimate of the relationship between SNP call annotations and the likelihood that a SNP is a true genetic variant, not a sequencing or data analysis artifact.

Before continuing with training of snps and indels; annotations, the desired truth sensitivity and model parameters must be specified for the program to evaluate the likelihood of SNPs being real. Annotations (coverage, quality by depth, FisherStrand, MappingQualityRankSumTest, ReadPosRankSumTest) are mostly included in the input file. Model parameters are used to define the percentage of the worst scoring variants and the minimum number of worst scoring variants to use when building the model of bad variants.

The desired truth sensitivity values are used by the program to generate tranches. 100.0, 99.9, 99.0 and 90.0 are first to fourth tranche thresholds used by default. The threshold values mean the sensitivity that can be obtained when we apply them to the call sets to train the model. The lowest tranche is highly specific but less sensitive and each subsequent tranche introduces additional true positive calls and also false positive calls. This allows to filter variants based on how delicate one needs the call set to be, as opposed to applying hard filters.

```
#VQSR SNPs
java -Xmx24g -jar /tools/GATK/GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R /reference/hg19/bwa/ucsc.hg19.fasta --num_threads 6 \
-input /mnt/30_epigenetics_10_centogene_joint.vcf \
-input /mnt/1003308-1003310-1003311/S1_S4_S8_S2_S9_S3_joint.vcf \
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
/db/vcf/hg19/hapmap_3.3.hg19.sites.vcf \
-resource:omni,known=false,training=true,truth=false,prior=12.0
/db/vcf/hg19/1000G_omni2.5.hg19.sites.vcf \
-resource:1000G,known=false,training=true,truth=false,prior=10.0
/db/vcf/hg19/1000G_phase1.snps.high_confidence.hg19.sites.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
/db/vcf/hg19/dbsnp_138.hg19.vcf \
-an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an SOR -an DP \
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
-mode SNP \
--recal_file snp.recal \
--tranches_file snp.tranches \
--rscript_file snp.plots.R
```

```
#VQSR INDELs
java -Xmx24g -jar /tools/GATK/GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R /reference/hg19/bwa/ucsc.hg19.fasta --num_threads 6 \
-input /mnt/30_epigenetics_10_centogene_joint.vcf \
-input /mnt/1003308-1003310-1003311/S1_S4_S8_S2_S9_S3_joint.vcf \
-tranche 100.0 -tranche 99.9 -tranche 99.0 -tranche 90.0 \
-mode INDEL \
--recal_file indel.recal \
--tranches_file indel.tranches \
--rscript_file indel.plots.R \
--maxGaussians 4 \
-resource:mills,known=false,training=true,truth=true,prior=12.0
/db/vcf/hg19/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
/db/vcf/hg19/dbsnp_138.hg19.vcf \
-an QD -an DP -an FS -an SOR -an ReadPosRankSum -an MQRankSum
```

Apply the Model to the Callset. The trained Gaussian mixture model is applied to both known and novel variations to assess the likelihood of being real. The score added as an info to each variant is called as VQSLOD, it is the log odds of being a real variant versus being a false variant.

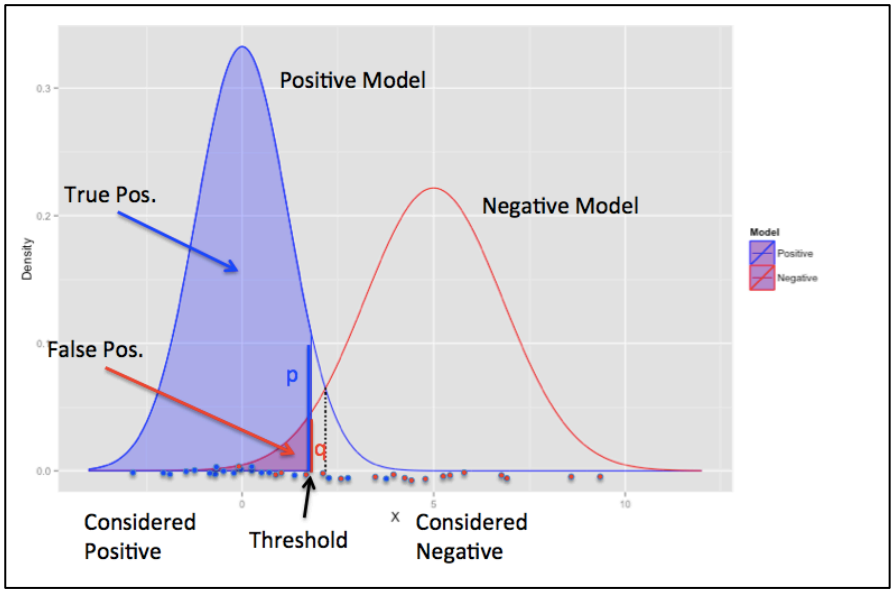


Figure 3.15. Threshold model of VQSR

INPUT: joint_FM.vcf -> OUTPUT: FM.snp.vcf & FM.indels.vcf

```
java -jar /tools/GATK/GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R /reference/hg19/bwa/ucsc.hg19.fasta \
-input /mnt/1003308-1003310-1003311/S1_S4_S8_S2_S9_S3_joint.vcf \
--ts_filter_level 99.9 \
-tranchesFile /mnt/1003308-1003310-1003311/snp.tranches \
-recalFile /mnt/1003308-1003310-1003311/snp.recal \
-mode SNP \
-o /mnt/1003308-1003310-1003311/snp_S1_S4_S8_S2_S9_S3_joint.vcf

java -jar /tools/GATK/GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R /reference/hg19/bwa/ucsc.hg19.fasta \
-input /mnt/1003308-1003310-1003311/S1_S4_S8_S2_S9_S3_joint.vcf \
--ts_filter_level 99.9 \
-tranchesFile /mnt/1003308-1003310-1003311/indel.tranches \
-recalFile /mnt/1003308-1003310-1003311/indel.recal \
-mode INDEL \
-o /mnt/1003308-1003310-1003311/indel_S1_S4_S8_S2_S9_S3_joint.vcf
```

Merge SNPs and Indels. INPUT: FM.snp.vcf & FM.indels.vcf -> OUTPUT: FM.merged.vcf

```
java -jar /tools/picard/picard.jar MergeVcfs INPUT=snp.YNS_EPIX_32333435.vcf
INPUT=indel.YNS_EPIX_32333435.vcf OUTPUT=YNS_EPIX_32333435.merged.vcf

for x in *.snp.vcf; do java -jar /mnt/tools/picard/picard.jar MergeVcfs 'INPUT='$x
'INPUT='${x/.snp./.indel.} 'OUTPUT='${x/.snp./.merged.}; done (WITH LOOP -
OPTIONAL)
```

Select Header and Passing Variants. INPUT: FM.merged.vcf -> OUTPUT: pass.vcf

```
grep -P "^#|PASS" YNS_EPIX_32333435.merged.vcf | uniq > YNS_EPIX_32333435.pass.vcf
for x in *.merged.vcf; do grep -P "^#|PASS" $x | uniq > ${x/merged/pass}; done
```

3.2.7. Variant Annotation

Convert to Annovar. For annotating SNP calls the software Annovar is used. ANNOVAR takes a simple format that includes chr, start, end, ref, alt, plus optional fields, as an input. To use Annovar, one must convert .vcf file format to the Annovar input file format. The convert2annovar.pl script can convert other "genotype calling" format into ANNOVAR format.

INPUT: pass.vcf -> OUTPUT: pass.txt

```
for x in *pass.vcf; do perl /tools/annovar/convert2annovar.pl -format vcf4 $x -allsample -withfreq -include > ${x/vcf/txt}& done
```

Run Annovar. Annovar annotates functional effects of variants; the location of each variant with respect to genes; exonic, intronic, intergenic, splice site, 5'/3'-UTR and so on. The option preferred in this project is to gene-based annotation.

INPUT: pass.txt -> OUTPUT: multianno.txt

```
(perl /mnt/tools/annovar/table_annovar.pl /mnt/YNS_EPIX_32333435.pass.txt /mnt/tools/annovar/humandb/ -buildver hg19 -protocol refGene,avsnp147,snp138NonFlagged,clinvar_20170130,popfreq_max_20150413,exac03nontcga,revel,dbnsfp33a -operation g,f,f,f,f,f,f,f --argument '--splicing_threshold 40',,,,,,, -nastring " --otherinfo;) |& tee run.log
```

3.2.8. Visualization and Interpretation of NGS Read Alignments via Integrative Genomics Viewer (IGV)

IGV is a visualization tool for aligned read data and read coverage. Paired ends sequencing produce reads from both ends of genomic fragments of known size. IGV color-codes paired ends if the insert size is larger than expected, falls on different chromosomes, or has unexpected pair orientations (Figure 3.16). Several file formats can be used as an input of IGV, but for sequence alignment data, only .sam or .bam files can be used; we prefer to use .bam files. IGV also requires bam index files and index file must be named by appending .bai. For instance, the index file for abc.bam must be named abc.bam.bai or abc.bai (Robinson *et al.*, 2011).

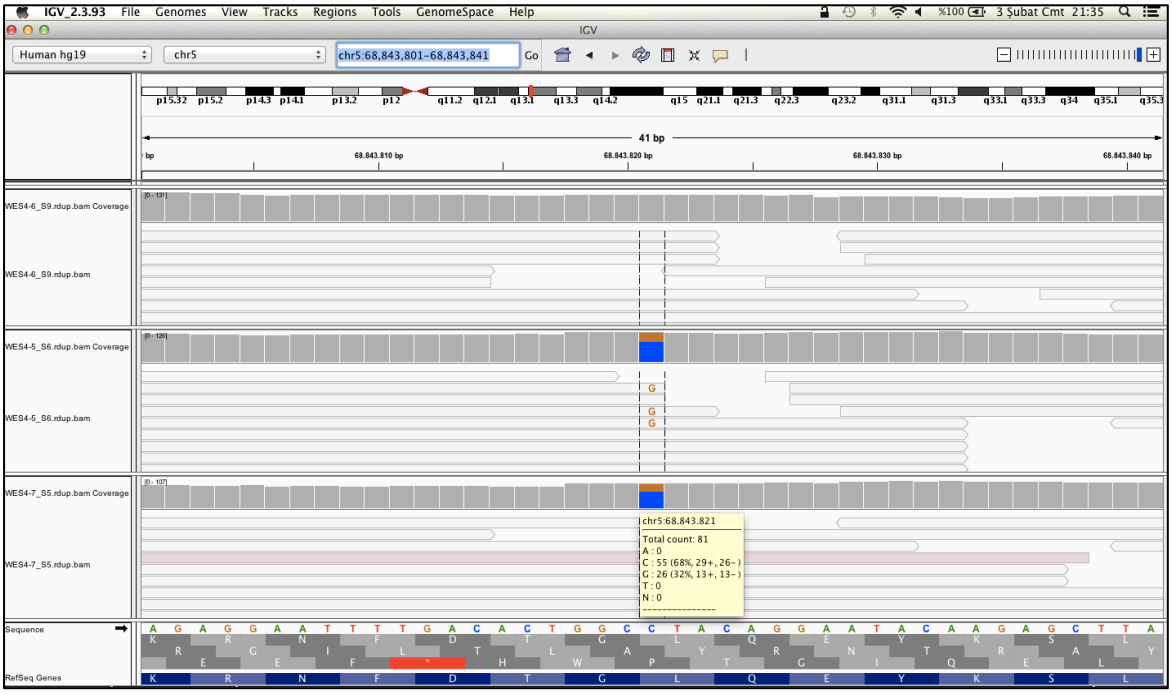


Figure 3.16. The IGV application window

Visualization of the datasets aids to check the average coverage, enrichment specificity and a number of random errors in the dataset. Moreover, clear segregation mistakes are easily detected by quickly looking at specific genes or regions and comparing multiple datasets detection.

3.3. Variant Filtration ad Prioritization Strategy

We preferred a trio-based analysis instead of a proband-only. The samples from the child (affected sister/brother) and both of their biological parents were analyzed. This enabled us to identify *de novo* variants that are present only in the child, to filter out rare benign familial variants, and to establish the phase of variants in recessive or imprinted disorders by inheritance.

Our pipeline outputs three different variant lists namely *de novo* heterozygous, homozygous and compound heterozygous variants. By using the affected and healthy family members as an input, it reports the *de novo* heterozygous, homozygous and compound heterozygous variant lists that are seen in only the affected members. Independent from the pedigree information, the first thing we do is figuring out all *de novo* heterozygous, homozygous and compound heterozygous variant scenarios separately (Figure 3.17).

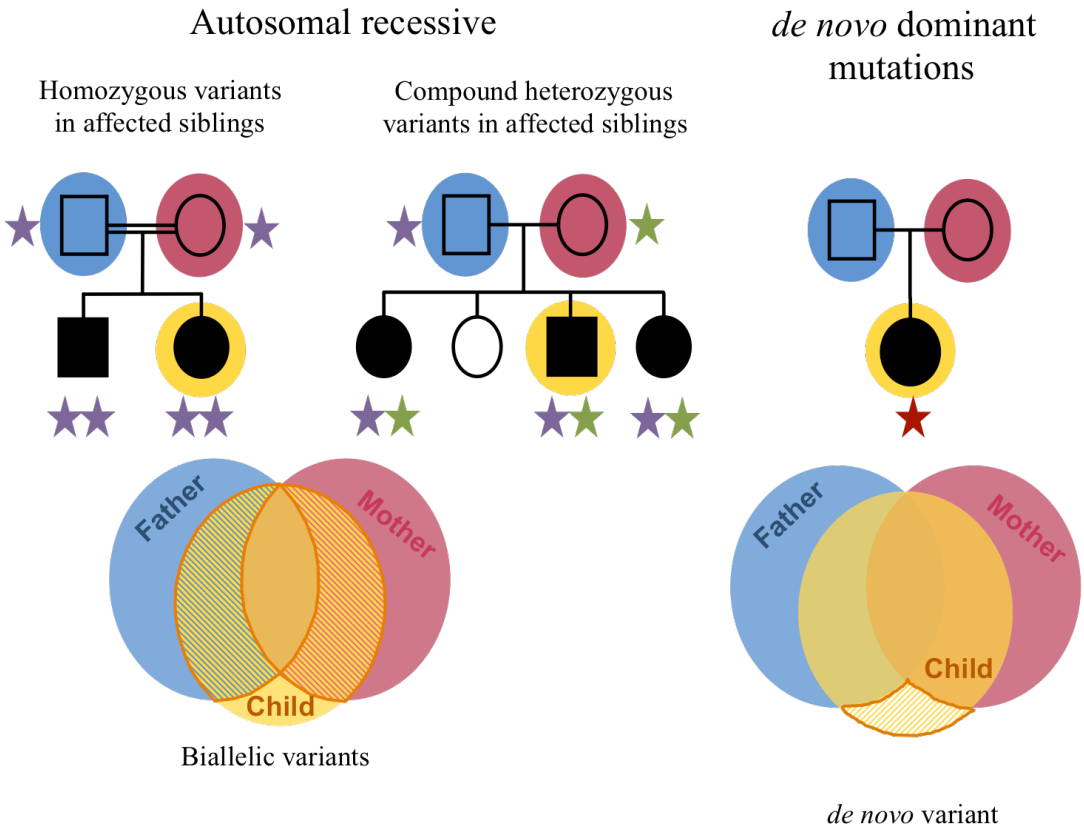


Figure 3.17. Disease gene identification strategies

One of the clearest proof to consider a variant as benign is its high allele frequency in the human population, being too high for causation of a disease (Lek *et al.*, 2016; Whiffen *et al.*, 2017). Hence, from variants with a MAF of <0.1% and without homozygous carriers in public databases, the ones predicted to affect protein coding were analyzed. If no prominent variant is found, the threshold is increased to <0.5%. For the intronic alterations, the ones at exon-intron boundaries from -10 bases to +10 bases are retained. If no prominent variant is found, the threshold is increased to -40 to +40. Then, we prioritize variant lists and start with the list that matches the most expected segregation pattern. Symptoms of affected individuals and family history were reviewed to prioritize variants with the highest degree of symptom match. Evidence from various sources; population databases, computational assessments, PubMed, OMIM and MGI were gathered. Seven different tools are being considered for pathogenicity estimations; namely CADD, REVEL, M-CAP, PrimatAI, SIFT, Polyphen, and MutationTaster. We also investigated the effect of the splice site mutations via four different splice site prediction programs; Human Splicing Finder, NetGene2Server, Berkeley Drosophila Genome Project-Splice Site Prediction by Neural Network and Oriel SpliceView. Instead of expecting the support for the disease-causing effect of the variant from all of the *in silico* tools, the information obtained from each tool is taken into account; since each tool has several varying strengths and weaknesses.

Table 3.2. Classes of evidences to prioritize the variants in rare diseases

Evidence Level	Evidence	Aim and Examples
Gene Level	Genetic	***the gene shows statistically low number of variants***
		Exome Aggregation Consortium Database (ExAC) Genome Aggregation Database (GnomAD) Genic Intolerance
	Experimental	***Model systems: Animal models with mutated/knock-out gene present a phenotype that has overlaps with the human disease***
		Mouse Genome Informatics (MGI)
		Protein Interactions: The product of the gene interacts with proteins which found to be related with the disease of interest
		Protein Interactions: STRING
	Literature	***Biochemical function: The product of the gene has a function consistent with the phenotype***
		Deep literature search
Variant Level	Genetic	***the variant is found in databases with a very low frequency or not found in any databases of healthy population cohorts***
		the variant is co-inherited with the disease in affected families
		PopFreq Exome Aggregation Consortium Database (ExAC) Genome Aggregation Database (GnomAD)
	Informatic	***Conservation: the variant show evolutionary conservation***
		Predicted affect on function: the variant is found on the gene predicted to cause functional effect
		Combined Annotation–Dependent Depletion (CADD) Rare Exome Variant Ensemble Learner (REVEL) The Mendelian Clinically Applicable Pathogenicity (M-CAP) PrimatAI Human Splicing Finder (HSF) NetGene2 Server Berkeley Drosophila Genome Project Splice Site Prediction by Neural Network Oriel SpliceView SIFT PolyPhen MutTaster Homozygosity Mapper BCFtools/ROH
	Simulation	***Comparing the mutated and native protein by studying the physical movements several hundreds of atoms in solution with explicit solvent representations***
		Homology Modeling Molecular Dynamics Simulations (MD)
	Literature	***Whether the variation is located at the functional domains/motifs or the mutational hotspots of the protein***
		Deep literature search

As mentioned, variant prioritization is based on several factors including population allele frequency, conservation, penetrance, prevalence and mode of inheritance of the variant and patient symptoms.

3.3.1. Population Allele Frequency

Population-level minor allele frequency (MAF) is critical since causative alleles for most Mendelian disorders are expected to be rare, because of their deleterious effects on reproductive fitness. Large-scale genome/exome sequencing efforts have cataloged protein-coding variations observed more than 250,000 individuals and, among all, ExAC (Lek *et al.*, 2016) and GnomAD are the largest datasets of variant allele frequency. Typically, causative alleles are less likely to be found in these databases. If they are found, then they are not present with high allele frequencies. In any global population, >5% MAF is considered benign for most of the Mendelian disorders, except for the well-known founder alleles (Richards *et al.*, 2015).

Using the ExAC dataset, Kobayashi and colleagues pointed out that 97.3% of pathogenic variants in genes that are found to be associated with disorders had MAF < 0.1% (Kobayashi *et al.*, 2017). Hence, we consider MAF threshold of 0.1% as a suitable starting point. If no prominent variant is found, the threshold is increased to 0.5%.

3.3.2. Gene Constraints

Several groups have developed statistical methods that predict the tolerance of a gene to synonymous, missense and loss-of-function changes by using the measurements of allele frequencies in populations.

The Genic Intolerance Residual Variation Intolerance Score (RVIS) uses more than 6,500 exomes from the NHLBI-Exome-Sequencing-Project. The linear model compares the number of common functional variants and the number of total variants observed in the gene. Constrained genes have less common functional variation than expected and have lower RVIS % score, genes with more common functional variants have higher RVIS % score (Petrovski *et al.*, 2013).

Moreover, ExAC database is used 60,706 exomes to calculate the probability of loss-of-function intolerance (pLI) and synonymous/non-synonymous intolerance (z) for each gene in the human genome. For the loss-of-function intolerance (pLI), they model the expected number of *de novo* mutations per gene, compares the observed and expected numbers of loss-of-function variants to derive a probability score. The closer the pLI is to 1, the less tolerant to variation the gene is and pLI >0.9 is an important sign for pathogenicity. For the synonymous/non-synonymous intolerance (z), they again model the expected number of mutations per gene, compares the observed and expected numbers of variants to derive the probability score. The higher the z, the less tolerant to variation the gene is predicted to be and z >3 is an important sign for pathogenicity.

3.3.3. Splice Variant Location

We have searched the locations of pathogenic splice site mutations in the literature and noticed that the majority were on the region of ± 10, with a few pathogenic mutations around ± 30. Moreover, Bergant *et al.* reanalyzed 1,059 unsolved cases of WES specifically for copy number variations (CNVs), splice site variants, breakpoints, and mtDNAs. Among all pathogenic splice-site mutations, majority nonconsensus splicing variants were seen at positions +4 and +5 (63.6%), ± 1 and ±2 contributing to 1.2%. Two splice defects seen at positions −3, −12 and two synonymous variants were predicted to affect splicing (Bergant *et al.*, 2017). Hence, we decided to take ± 10 as an initial threshold and ± 40 as the second threshold for splice site mutations.

3.3.4. Conservation Status

The degree of conservation is obtained by aligning human protein sequences to homologous protein sequences from other organisms. The less conserved the column, the more likely it will be tolerated.

3.3.5. Region of Homozygosity

To check whether our prominent gene from homozygous variant list is within a ROH or in the case that there are genes in homozygous filtered variants that are close to each other less than 10Mb to check whether these genes are within a ROH, WES data was used to perform homozygosity mapping by using two programs; namely Homozygosity-Mapper (Seelow and Schuelke, 2012) and BCFtools/ROH (Narasimhan *et al.,* 2016). Homozygosity-Mapper identifies ROH with a sliding-window approach and also proposed that the program always identify the same genomic regions as conventional linkage analyses. The graphical visualization tool, easy-to-use and very fast to obtain a qualitative measure of homozygosity around the locus of interest. The .vcf file upload into the databases as inputs and analysis of projects are done within a few minutes. Independent from parameters like family structure or allele frequencies, the score is measured from the observed homozygosity. In addition to Homozygosity-Mapper, BCFtools/ROH is also used in this study. It is an extension to the BCFtools software package, applies HMM to simulated data and real data from the 1000-Genomes-Project. However, to keep in mind, all these ROH programs are more useful if you have more than one family with the same disease.

3.3.6. Phenotypic Evaluation

Mendelian diseases show themselves as combinations of stereotypical symptoms that together define a disease. However, in the case of rare diseases, many different conditions produce overlapping symptoms.

If the prioritized genes of the patient are associated with any rare disease, text-mining the literature must be done to answer if the disease has similar symptoms with the patient's phenotype. Considering the overlapping symptoms of rare diseases very carefully, the main symptoms and exceptional symptoms should be noted and, if necessary, the patient should be re-examined for certain conditions.

3.3.7. Threshold for *in silico* Pathogenicity Prediction Tools

CADD. Based on their observation on training sets, they set up at several thresholds. C-score >10 means the 10% most deleterious substitutions and C-score >20 indicates the 1% most deleterious substitutions.

REVEL. The score for each variant can range from zero to one, reflecting the proportion of trees in the random forest that classified the variant as pathogenic. Pathogenicity threshold can be set either as 0.50 or 0.75. 75.4% of disease mutations, 10.9% of neutral variants have a score above 0.5; whereas 52.1% of disease mutations, 3.3% of neutral variants, and 4.1% of all missense variants have a score above 0.75.

M-CAP. The threshold to define a variant as pathogenic is > 0.025 for M-CAP.

PrimatAI. The recommended threshold is > 0.8 for likely pathogenic classification, <0.6 for likely benign, and 0.6–0.8 as intermediate in genes with dominant modes of inheritance, on the basis of the enrichment of de novo variants in cases as compared to controls, and a threshold of >0.7 for likely pathogenic and <0.5 for likely benign in genes with recessive modes of inheritance.

3.4. Validation via Sanger Sequencing

3.4.1. DNA Extraction from Peripheral Blood

200 µl of blood sample from the tube containing K_2EDTA was transferred to sterile 1.5ml centrifuge tubes, 20 µl proteinase K and 200 µl red blood cell lysis buffer were added. The sample was vortexed for approximately 15 sec. and then incubated at 56 °C for 10 min. After incubation, 200 µl absolute ethanol was added to the tube and vortexed again for approximately 15 sec. Then, the mixture was applied to a QIAamp spin column and centrifuged at 6000 x g (8000 rpm) for 1 min. The filtrate was discarded, and the column was washed with 500 µl wash buffer-1 by centrifuging at 6000 x g (8000 rpm) for 1 min. The filtrate was discarded, and the column was washed with the same amount of wash buffer-2 by centrifuging at full speed for 3 min. Discard the filtrate. Finally, 200 µl water

was added on top of the filtrate, and the sample was incubated at room temperature for 1 min. and centrifuged at 6000 x g (8000 rpm) again for 1 min. to elute the genomic DNA.

3.4.2. Quantitative Analysis of Extracted DNA

Qubit fluorometer was used to determine the concentration of the extracted DNA samples. The measurement was done at 260 nm to determine the nucleic acid concentration. To monitor the purity of DNA samples, 230 and 280 nm absorbances were also measured. A260/A280 ratio between 1.8-2 and A260/A230 ration higher than 2 were considered as a sign of purity. For WES, minimum 1 μg of a sample was used.

3.4.3. Polymerase Chain Reaction (PCR)

The primers for the target regions were designed via Primer3 and whether the primer pairs amplify only one product with the right position and size is checked via UCSC in-silico PCR.

PCR reactions included 100 ng DNA, 10 μl 5X polymerase buffer, 3 μl $MgCl_2$, 1 μl dNTPs, 0,2 μl of each primer with a concentration of 100 pmol and 1U of *Taq* polymerase in a volume of 50 μl. The program was as follows: An initial denaturation step at 96 °C for 2 min., followed by 35 cycles of 30 sec. at 94 °C, 30 sec. at annealing temperature, 40 sec. at 72 °C and a final elongation for 5 min. at 72 °C. PCR products were loaded on 1% agarose gels to verify the amplicon size with respect to the appropriate DNA ladder.

3.4.4. Sanger Sequencing

Due to the Mendelian patterns of inheritance, segregation in family members is strong evidence for variant pathogenicity. In accordance with this, lack of segregation is accepted as strong evidence that a variant is not pathogenic.

When a condition is inherited in a dominant manner, co-segregation in affected family members is strong supporting evidence for pathogenicity. When the *de novo* occurrence of a variant is suspected, the absence of past family history of the disease

considered strong evidence of pathogenicity. When a condition is inherited in a recessive manner, the maternal and paternal copies of damaging variant and homozygosity in the affected individual is strong supporting evidence for pathogenicity. For the compound heterozygosity, the damaging variants are at different positions in the maternal and paternal copies of the same gene.

Sequence validation and segregation analysis were performed by Sanger-sequencing. Sequence electropherograms were analyzed using the FinchTV (Geospiza, USA). Mutation nomenclature refers to GenBank mRNA reference sequence.

3.5. Homology Modeling

Basically, the steps required in homology modeling, also called comparative modeling are the following:

- Template identification and initial alignment: The percentage identity between the sequence of interest and a possible template must be high enough to be detected with basic sequence alignment programs such as BLAST (Altschul *et al*., 1990).
- Alignment correction: More sophisticated methods are used for better alignment, for instance, CLUSTAL-W (Thompson, Higgins, and Gibson, 1994).
- Backbone generation
- Loop modeling: The alignment between model and template sequence usually contains deletions and insertions. All insertions or deletions in the alignment are modeled as loops and turns
- Side-chain modeling: The side-chain conformations of residues in structurally similar proteins often have similar the torsion angles. So, it makes sense to copy conserved residues from the template to the model and achieve higher accuracy than by simply copying the backbone and repredicting the side chains.
- Overall model optimization and verification: Modeling programs either restrain the atom positions or apply hundreds of steps of energy minimization to

optimize the model. The most straightforward approach to model optimization is running a molecular dynamics simulation of the model.

I-TASSER (Iterative Threading Assembly Refinement) is used for homology modeling. It is an automated modeling and structure-based function annotation server (Yang *et al*., 2015).

3.6. Molecular Dynamics Simulations

VMD (Visual Molecular Dynamics) is written in C++; it is a molecular graphics program with three-dimensional graphics rendering and coloring options to display and analyze biopolymers. The molecules are displayed as one or more representations, and the atoms in each representation are chosen using atom selection syntax. VMD provides a graphical user interface for program control and also a text interface using the Tcl embeddable parser to allow for complex scripts with variable substitution and function calls. It can be used to display MD simulation trajectories. New molecules from MD simulations can be uploaded into VMD from a set of structure files that include static information about the system, such as bond connectivity and atomic mass and charge values; and also coordinate files that contain the positions of all the atoms that make up the molecule. The read data from formats other than PBD or PSF are converted into those formats and then can be used as an input.

The main steps of molecular dynamics simulation can be summarized as follows:

- If the protein we are interested in has a known crystal structure or an NMR structure in the Protein Data Bank (PDB), this structure can be used for molecular dynamics simulations. If not, one should start with the homology modeling as described in the previous section and use the predicted structure.
- dH_2O and NaCl should be added to the structure at the nearest 5 Å. The addition of NaCl is necessary to neutralize the charge balance and to calculate the electrostatic balance accurately. The distance should be at least 5 Å not to interfere with the protein itself.

- The energy of the structure should be minimized either by NpT (constant pressure and temperature) or NvT (constant volume and temperature). Then, the simulation can be started. We simulated the motion of proteins (mutated protein and wild-type protein) on a femtosecond (10−15 s) timescale and mimicked the true folding process by calculating the electrostatic charges via CHARMM36 for 20 ns. Two different temperature conditions are preferred, 310 °K and 450 °K. In the case of high temperature, more energy is given to the system, the system progress faster. Hence, we are also able to observe the movements of proteins further.
- At the end of the analysis, MD simulations provide four main information:

  RMSD (Root Mean Square Deviation): Average distance of the backbone atoms. This calculates how close the protein backbone atoms are to each other.

  RMSF (Root Mean Square Flexibility): Calculating the displacement distance of each amino acid.

  Salt-Bridge Analysis: Salt-bridges are bonds between oppositely charged and sufficiently close to each other residues to have an electrostatic attraction. They have a role in protein structure and specificity of the interaction of proteins with other biomolecules. Salt-bridge analysis calculate whether any new salt-bridge dissociations or association is seen in the protein as a result of the mutation.

  FoldX ($\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt}$)*:* Calculates the difference in the structural stability of the protein.

3.6.1. FLNA Protein

The crystal structure of the N-terminus unit of flaming protein (PDB ID: 4M9P) was captured as monomer (Light *et al.*, 2012). The identified mutant (R484Q) as described in bioinformatics section was constructed as in-silico with VMD (Humphrey *et al.*, 1996). Native and mutant complexes were solvated in water boxes and neutralized with NaCl using the cutoff distance for solute and protein of 5 Å. The overall charge of all systems was adjusted to zero. Native and mutant systems were composed of ~125,000 atoms.

The monomer and dimer systems were simulated with the NAMD program by using CHARMM 36 all-atom force fields parameters including the correction maps (MacKerell *et al.*, 1998; MacKerell *et al.*, 2004; Brooks *et al.,* 2009; Philips *et al.*, 2005). Along with all simulations, water molecules were treated explicitly using TIP3P model (Jorgensen *et al.*, 1983). Following a 10,000-step minimization with the Greedy algorithm, native and mutant systems were equilibrated at 298 °K at 1 atm for 1 ns. The production simulations were performed along 20 ns trajectory at 310 °K collected as NpT ensemble. To reveal the impacts of high temperature on the system within shorter MD trajectories, MD simulations were also run at 450 °K along 20 ns as NvT ensemble. The pressure and temperature controls of systems were done with Langevin pressure and temperature coupling. All MD simulations were performed with an integration velocity as 2 fs, and the long-range Coulomb interactions were calculated with particle-mash Ewald (PME) method in (x,y,z) dimensions (Essman *et al.*, 1995; Darden *et al.*, 1993). To test the reproducibility of simulations, they were repeated at 310 °K NpT ensemble and at 450 °K NvT ensemble, which different velocity seeds were used.

VMD was used for both visualization and analysis of MD trajectories (Humphrey *et al*., 1996). For all system, RMSD, RMSF, and salt-bridge analysis were performed. In addition to that, the initial and final coordinates of 20 ns MD simulation trajectories, the mutant and native configurations were saved in every 5 ns along trajectories, and the effects of mutations on destabilization tendency of enzyme were calculated as $\Delta\Delta G$ (kcal/mole) FoldX analysis used as YASARA plug-in (Schymkowitz *et al.*, 2005; Krieger *et al.,* 2002).

Hinge motion prediction

HingeProt is a tool developed to predict hinge motions of a protein by employing Elastic Network Models which are Gaussian Network Model (GNM) to predict the motion in the slowest mode and Anisotropic Network Model (ANM) to predict the direction of this motion (Emekli *et al.*, 2008). HingeProt was used for the minimized and equilibrated wild-type and mutant (R484Q) structures to observe the effect of the mutation on hinge motion of FLNA. Moreover, cut-off distances for GNM and ANM were chosen as a default of HingeProt, which were 10 Å and 18 Å, respectively.

3.6.2. FZD6 Protein

Based on the NGS sequencing results, the protein sequence containing our mutant (p.Gly559Aspfs*16; c.1676_1683delGAACCAGC.), called as FZD6 mutant, were translated via Ensembl Tool. Since the crystal structure of FZD6 protein is not known yet, through I-TASSER, the structure of FZD6 mutant was predicted, and the best model among generated several ones was selected based on C-score (Zhang *et al.*, 2008; Yang *et al.*, 2015). All MD simulations were performed with NAMD program, where CHARMM 36 all-atom force fields including the correction maps were used (MacKerell *et al.*, 1998; MacKerell *et al.*, 2004; Brooks *et al.*, 2009; Philips *et al.*, 2005). Water molecule was described as TIP3P model (Jorgensen *et al.*, 1983). Before production simulations, FZD6 mutant was subjected to 40,000-step minimization performed with Greedy algorithm, and it was followed by 2 ns equilibration, collected as NpT ensemble at 298 °K. During simulations, periodic boundary conditions were applied at all dimensions, the systems were run with 2 fs velocity, and Langevin pressure coupling was used to keep pressure constant at 1 atm. Particle-mesh Ewald (PME) method was used to calculate the electrostatic interactions (Essman *et al.,* 1995; Darden *et al.*, 1993). Production simulations of FZD6 and FZD6 mutant were collected at 310 °K as 20 ns followed by the minimization of ionized systems with the Greedy algorithm and NpT equilibration performed at 298 °K with 2 fs step velocity. All MD simulations were run twice by changing their initial coordinates to avoid any artifact. The native and mutant systems were prepared to MD simulation, e.g. solvation and ionization, with VMD. Also, the visualization and analysis of systems, e.g. RMSD and RMSF calculations, and salt-bridge interactions were performed with VMD (Humphrey *et al.,* 1996).

4. RESULTS

4.1. Rare Disease Cohort

Molecular testing is the most prominent way to prevent delayed diagnosis of undiagnosed rare disorders. Its contribution to diagnosis is still numerically not yet calculated. We built up a unique workflow that contributes to the already existing knowledge through the combination of selected threshold for population frequency, pathogenicity prediction tools, gene intolerance scores, and MD simulations for the first time. The workflow created in this dissertation project was tested on families with members of undiagnosed diseases and achieved a high success rate by identifying the causative variant. Since only two of these families have undergone molecular dynamics simulation to explain pathogenicity mechanisms of the mutations, these two families were written to the result part of the dissertation.

4.2. Case I

The family was ascertained in the Medipol Hospital, Istanbul and enrolled for a molecular genetic study at Acıbadem University as part of a whole exome/genome sequencing project investigating undiagnosed disorders in Turkish families. Written approved informed consent was obtained from all participants. We immediately pursued WES analysis. At the end of the WES data analysis, we had 666 variants for de novo heterozygous, 755 variants for homozygous and just 152 variants for compound heterozygous scenario. After the population frequency and splice site filtering, quantity decreased to 21, 19, 2 for de novo heterozygous, homozygous, compound heterozygous respectively. After recruiting data for gene intolerance (ExAC, GnomAD and Genic Intolerance DB), mouse phenotype (MGI), pathogenicity scores (CADD, REVEL, M-CAP, PrimatAI, SIFT, Polyphen, and MutationTaster); five genes were highlighted namely *SULT1C3, SLAIN1, ZFHX3, TKTL1, FLNA* (Table 4.1 and Table 4.2). We have also checked the top shet score gene list of Cassa *et al.* and only *FLNA* gene is found in top list with a score of 0,33. However, PrimatAI score for *FLNA* is also 0,54; which is on the area of likely benign.

Table 4.1. The most prominent homozygous variants for case I

Gene	Variant_Type	Change	OMIM	MGI
SULT1C3			_	_
SLAIN1	frameshift insertion	NM_001242868:exon1:c.219_220insGG:p.A73fs	_	_
ZFHX3	nonframeshift deletion	NM_001164766:exon8:c.2436_2450del:p.812_817del	Prostate cancer, susceptibility	Mice homozygous for a knock-out allele exhibit prenatal lethality. Mice heterozygous for the same allele exhibit partial postnatal lethality, decreased body size and prolonged conception time.
TKTL1	nonsynonymous SNV	NM_001145934:exon5:c.G628A:p.D210N	_	Mice homozygous for a knock-out allele exhibit increased susceptibility to DSS-induced colitis.
FLNA	nonsynonymous SNV	NM_001110556:exon10:c.G1451A:p.R484Q		"Females heterozygous for an X-linked, ENU-induced mutation exhibit dilated pupils and milder cardiac, sternum, and palate defects than males. Hemizygous males are inviable and exhibit incomplete septation of the outflow tract, septal defects, cleft palate and incomplete fusion of the sternum."

Table 4.2. More information regarding the most prominent homozygous variants for case I

Gene	ExAc	GnomAD	ExAc Metrix	Genic Intolerance	SIFT PolyPhen MutTaster	REVEL M-CAP CADD
SULT1C3	_	_	_	94.95%	_ _ _	_ _ _
SLAIN1	_	1.000 26376/26376 HOM:13188	_	46.12%	_ _ _	_ _ _
ZFHX3	_	_	_	0.26%	_ _ _	_ _ _
TKTL1	_	_	_	3.11%	D B D	0.158 0.010 20,5
FLNA	0,04887 4/81858 HOM:0 HEMI:3	0.00007887 14/177518 HOM:0 HEMI:6	z = 4.95 806.2/519 pLI = 1.00 54.4/1	0.44%	D D D	0.574 0.635 26,8

All the articles in the literature related to these five genes were compiled at this stage. Known functions, domains, motifs of each protein; distribution of mutations on the

gene, associated diseases, their inheritance patterns, exceptions regarding the inheritance pattern, overlapping and differentiating symptoms in patient phenotypes were all collected, and hypothesis created. Among all, the *FLNA* gene was the most significant gene that can be associated with the disease due to the findings in the literature and results of MD simulations.

The available clinical symptoms for the evaluation of the WES results can be listed as occipital lobe epilepsy and epileptic status in sleep. After the analysis steps that described in the method section and deep literature search, the genetic diagnosis we proposed for the patient was Periventricular Nodular Heterotopia. This diagnosis was confirmed by the clinician as the patient re-examined afterward for additional evaluation to determine the clinical fit.

Here, we reported the transmission of PNH from a clinically asymptomatic mother to two sons, in a fully penetrant classical X-linked dominant manner. We identified a novel c. 1451G>A, p.R484Q change in *FLNA* exon 10. Mutation nomenclature refers to GenBank mRNA reference sequence NM_001110556. Using whole-exome sequencing, the index case and his affected brother are found to be hemizygous for the missense mutation (Figure 4.1). This mutation leads to the substitution of a very conserved amino acid and not previously reported in the literature.

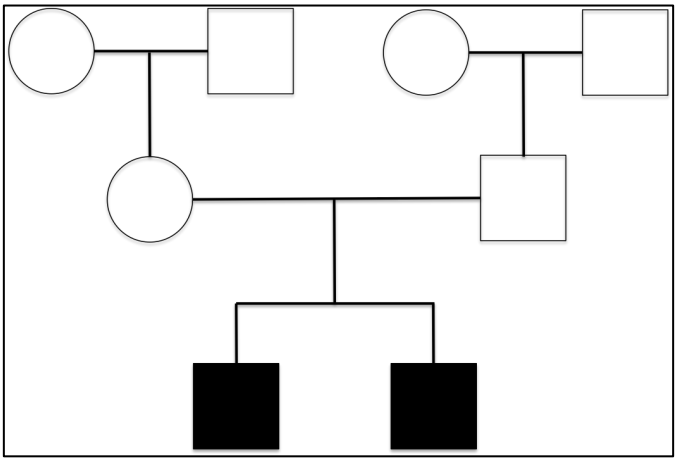


Figure 4.1. Pedigree of a non-consanguineous Turkish family segregating X-linked dominant PNH. Circles and squares represent females and males, respectively. Clear symbols represent unaffected while filled symbols represent affected individuals

The mutation was confirmed via Sanger sequencing. The healthy mother is a heterozygous carrier for this mutation, and the healthy father is carrying the reference allele while the index case and his affected brother are hemizygous as expected (Figure 4.2).

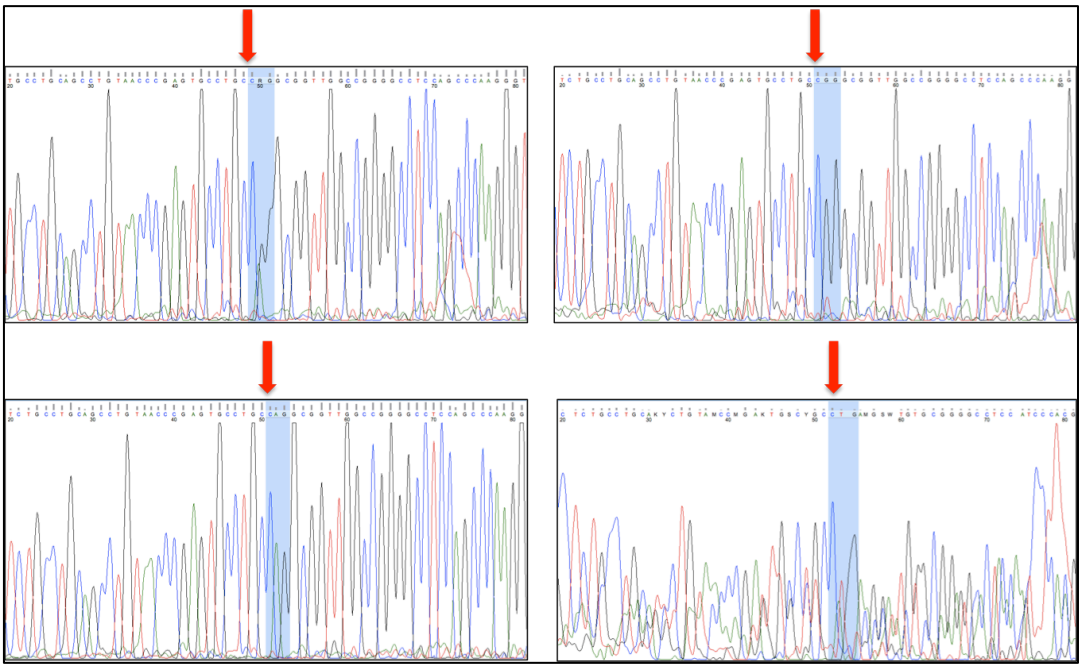


Figure 4.2. The electropherograms of the mother showing her heterozygous state while the the father showing his wild-type state in the upper panel. The index case and his affected brothers are hemizygous for c. 1451G>A. The arrow designates the position of the variant

Evolutionary conservation of the missense amino acid in other FLNA orthologues was examined by Clustal Omega (Sievers *et al*., 2014). The functional importance of the missense mutation is strengthened by the fact that the region is quite conserved in several species including human, rat, and mouse. Species abbreviations and accession numbers are as follows: Cf, Castor fiber, APD32923.1; Hs, Homo sapiens, NP_001447.2; Mm,Mus musculus,NP_034357.2; Rn, Rattus norvegicus, NP_001128071.1; Oh, Ophiophagus hannah, ETE70682.1; Bt, Bos taurus, NP_001193443.1; Mn, Macaca nemestrina, XP_011716088.1; Ml, Myotis lucifugus, XP_023608786.1 (Figure 4.3).

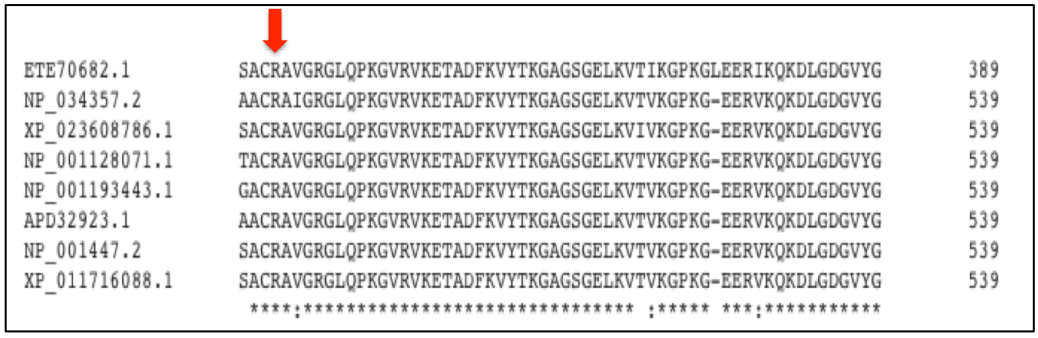


Figure 4.3. Partial amino acid sequence of human FLNA protein with orthologues from other species. The p. R484Q is indicated by an arrow

4.2.1. MD Simulations

Along 20 ns of MD trajectories, RMSD values are fitted against the crystal structures of native and mutant (R484Q) FLNA proteins. Specifically, there is a dramatic change in RMSD patterns of the mutant protein as it is jumping from ~2.4 Å to ~3.4 Å within 0.08 ns around ~17.6 ns of 20 ns MD trajectories whereas RMSD pattern of native protein remains stable along with a whole trajectory (Figure 4.4.a). For further explanation of jump in RMSD pattern of the protein upon R to Q mutation, we measure the flexibilities of residues in mutant FLNA protein between 17.58th ns and 17.66th ns (Figure 4.4) and categorize RMSF values from highest to smallest ones (Table 4.3).

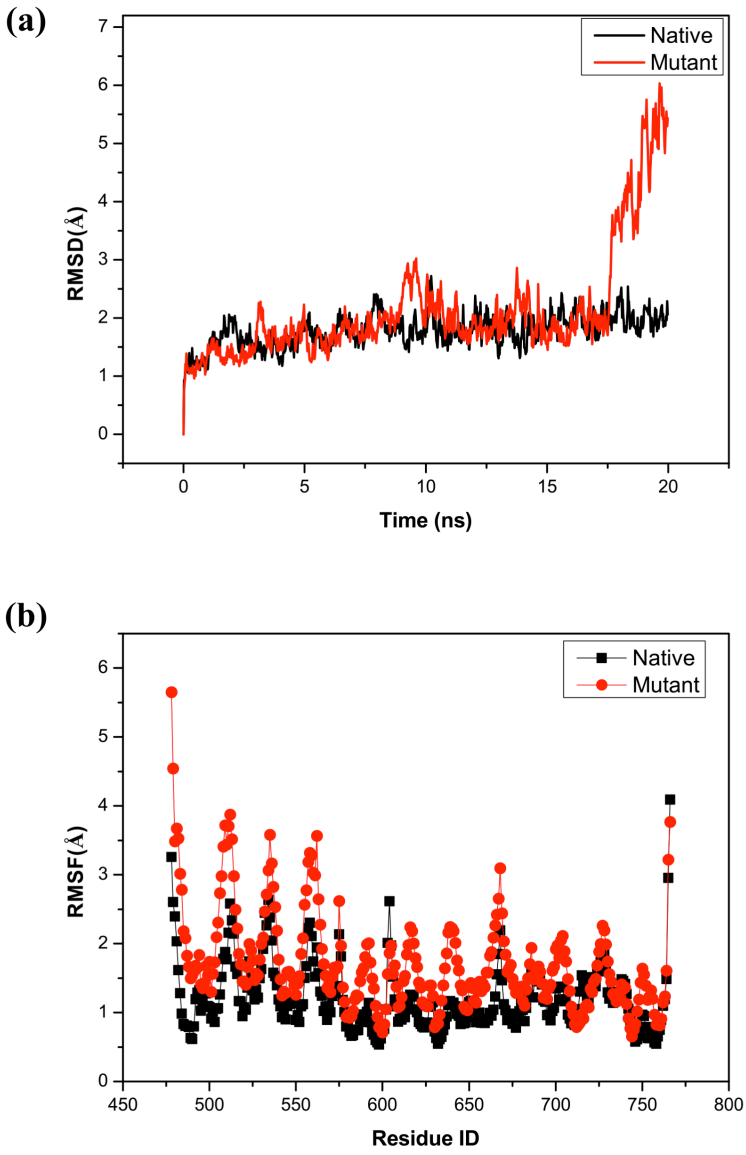


Figure 4.4. a) RMSD and b) RMSF results of modeled FLNA proteins, native and mutant, along 20 ns

28 out of 290 residues experienced change in RMSF values more than 1 Å. Pro480, Ser481 and Ala482 residues are included in these 28 residues out of 290 in mutant FLNA protein as displayed in Table 4.3.

Table 4.3. The changes in RMSF values in mutant FLNA protein upon R484Q replacement at 310 °K

Residue ID	ΔRMSF (Å)	Residue ID	ΔRMSF (Å)	Residue ID	ΔRMSF(Å)	Residue ID	ΔRMSF (Å)
766	1,60	478	1,31	701	1,20	557	1,11
535	1,60	532	1,30	531	1,18	703	1,09
512	1,58	533	1,30	509	1,17	537	1,08
668	1,57	514	1,25	558	1,15	482	1,07
511	1,57	667	1,21	575	1,15	536	1,07
534	1,51	576	1,21	481	1,14	479	1,04
513	1,33	669	1,20	607	1,14	480	1,04

Yet; these particular differences in RMSF values of residues in mutant FLNA protein are not enough to explain this particular jump in RMSD pattern (Figure 4.4.a). Hence, the snapshots are taken from MD trajectories from 16th ns to 20th ns in every 1 ns. As displayed in Figure 4.5, the gap between domain 3 and 4 begins to open from 17th ns of trajectory. This particular gap becomes more apparent at 18th ns and remains through the end of the trajectory.

In addition to differences in RMSD patterns of native and mutant FLNA proteins, it is also worthy to note that residues in mutant FLNA are more flexible with respect to residues in native one. Specifically, we focus on the RMSF pattern around Arg484 to reveal the impacts of R484Q mutant and conclude that there is an increase in flexibilities of residues around Arg484 as in the range of ~1.4-fold to ~2.8-fold, resid displayed in Figure 4.5. Upon R484Q mutation, there is ~2.8-fold increase in flexibility of 484^{th} residue, see Figure 4.6.

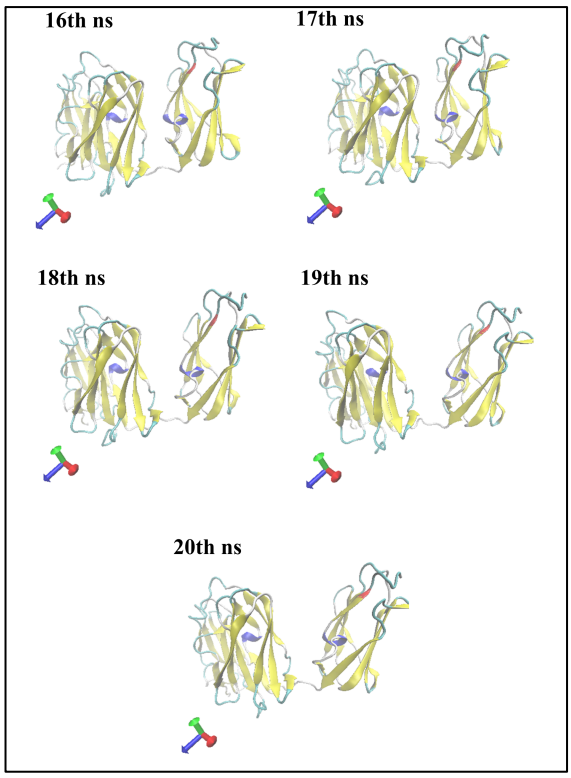


Figure 4.5. The configurations of mutant FLNA protein between 16th ns and 20th ns at 310 °K. Here, the secondary structures of FLNA protein are displayed in 'Cartoon' format by indicating R484Q residue as a red

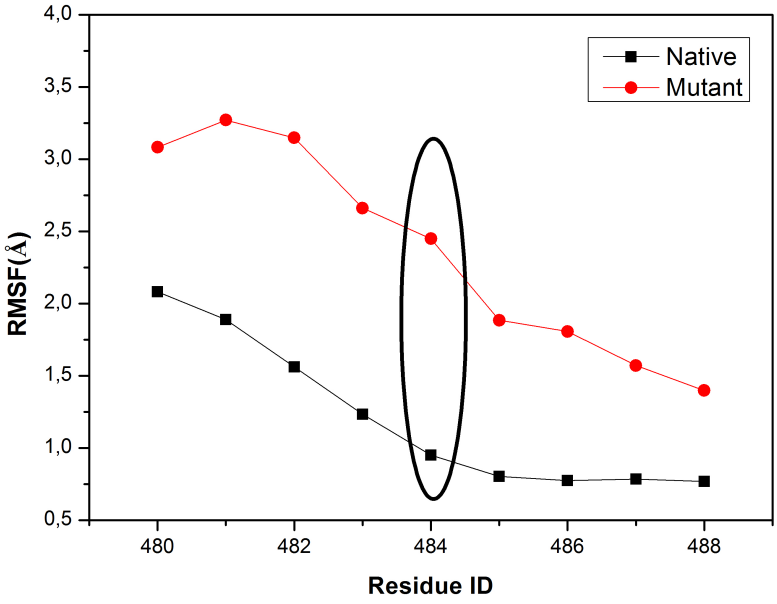


Figure 4.6. RMSF patterns of residues around 484th amino acid in native and mutant FLNA protein at 310 °K

After revealing the structural differences in backbone motion and flexibilities of residues involved in domain 3, 4 and 5 of FLNA protein, we focus on how R484Q mutation alters the intra-molecular interactions within the protein. Specifically, Arg484 residue is involved in salt-bridge interaction with Glu642 in native FLNA protein (Figure 4.7) with ~3.5 Å average distance, stabilized after 9th ns through the end of the trajectory at 310 °K (Figure 4.9). With Arg to Gln replacement at 484th position, this particular interaction is lost, and it results in the distancing of domain 3 from domain 4 by disrupting the stabilization of interface between domain 3 and 4, Figure 4.5.

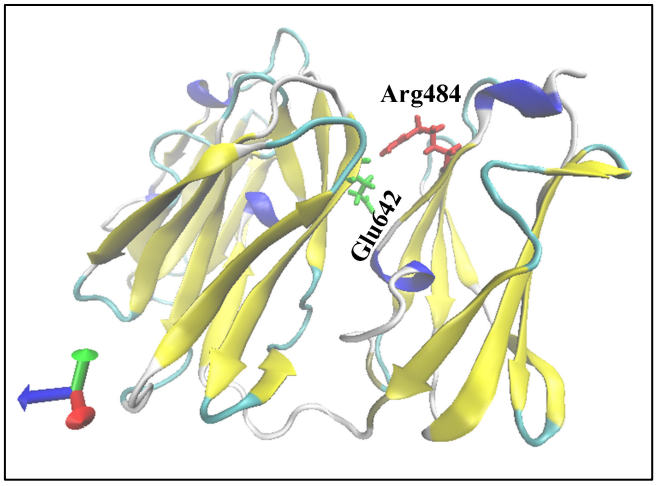


Figure 4.7. Arg484:Glu642 salt-bridge interaction in FLNA protein

Further, the evolutions of other salt-bridge formations, named as Glu499:Arg488, Asp653:Lys493, Glu614:Lys493, Asp653:Arg496 and Glu652:Arg496, along interaction interphase between domain 3 and 4 are examined along 20 ns MD trajectories at 310 °K in order to reveal whether R484Q mutant affects their strength/ their existence or not. We conclude that just Glu499:Arg488 salt-bridge interaction out of five is affected with Arg to Gln replacement at the 484th position. As displayed in Figure 4.8, this interaction becomes weakened after 12th ns and its strength is lost especially after 17th ns with average distance more than ~7 Å. Herein, it is worthy to note that this weakness time period of Glu499:Arg488, 17th ns, corresponds to time period that domain 3 is going away from domain 4 as displayed in Figure 4.5. Together the loss of Arg484:Glu642 interaction with the weakness of Glu499:Arg488 interaction, we conclude that Arg484:Glu642 interaction in FLNA protein has a role to stabilize the interaction between domain 3 and domain 4.

Even other salt-bridge interactions existed in interaction interphase of domain 3 and domain 4 are conserved; it is not enough to prevent a distancing of domain 3 from domain 4. Hence, we consider Arg484 as the most crucial residue along with the interaction interphase of domain 3 and 4 for keeping all three domain of FLNA protein together.

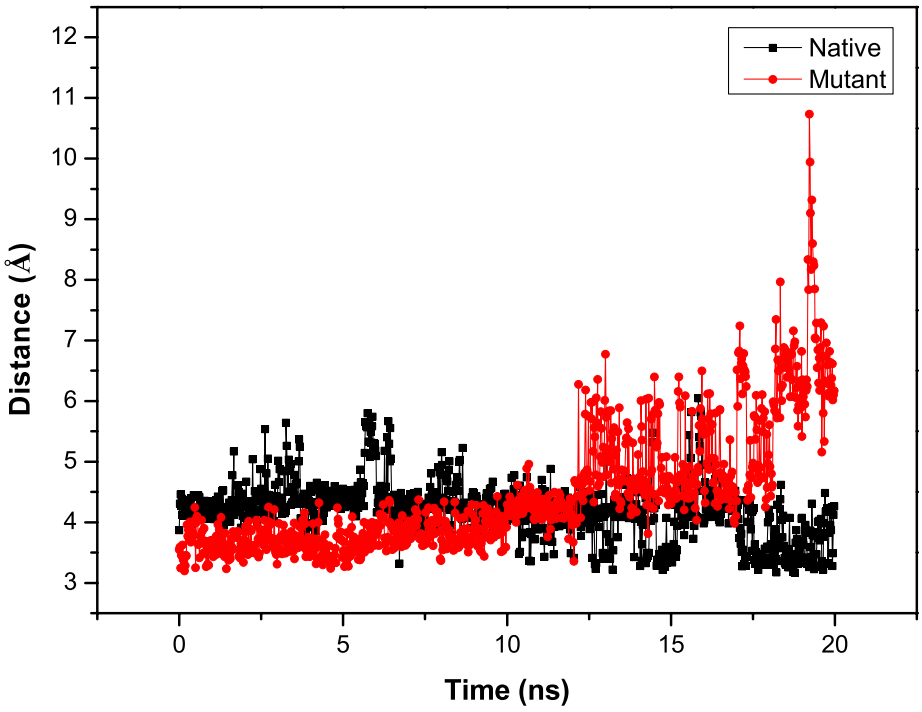


Figure 4.8. The evolution of Glu499:Arg488 salt-bridge interaction in FLNA native and mutant proteins along 20 ns MD trajectory at 310 °K

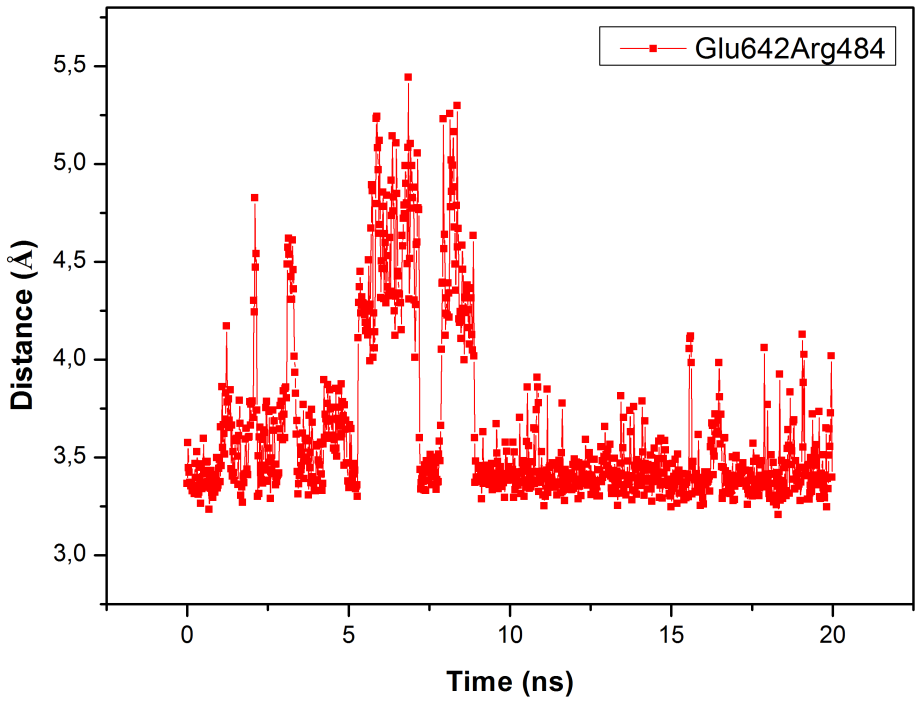


Figure 4.9. The evolution of salt-bridge interaction between Arg484 and Glu642 in native FLNA protein along 20 ns MD at 310 °K

Moreover, we aim to reveal the impact of R484Q replacement on destabilization tendency of FLNA protein. Through 20 ns of MD trajectory at 310 °K, the configurations of mutant and native FLNA protein are saved in every 5 ns, and the destabilization tendencies of overall mutant and native FLNA proteins are documented by calculating ΔΔG (kcal/mole) with FoldX (Schymkowitz *et al*., 2005). As displayed in Figure 4.10, we do not get noticeable difference in ΔΔG (kcal/mole) values of native and mutant FLNA. It means that R484Q mutant doesn't lead to the full dissociation of whole protein by leading to higher ΔΔG value.

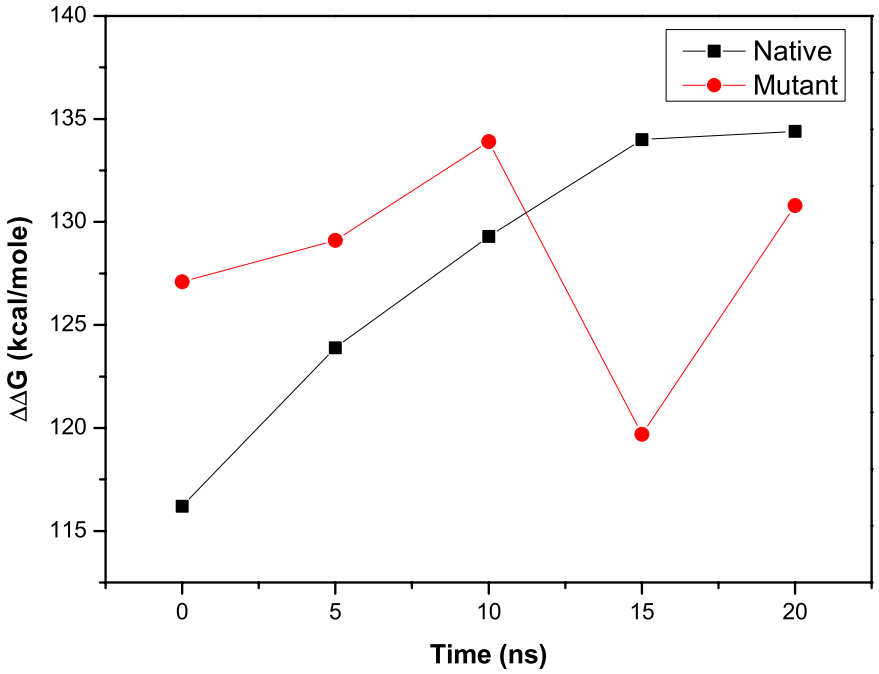


Figure 4.10. FoldX results of native and mutant FLNA protein at 310 K along 20 ns MD trajectories

The slowest models of HingeProt were chosen as prediction results. There were no differences between the predictions for hinge motions of the minimized and equilibrated wild-type and mutant structures. For both structures, the residue 576 was predicted as the hinge residue by separating the protein into two rigid parts. Therefore, the same hinge motions were expected to observe for the wild-type and mutant FLNA structures in an adequate time interval. Furthermore, the predicted motion was observed in the 20 ns MD trajectory of mutant FLNA protein between 16th ns and 20th ns at 310 °K as can be seen in Figure 4.5 whereas not observed in the wild-type trajectories. The reason for this observation could be the weakened interaction at the 484th residue by reduced bulkiness with Gln and disrupted salt-bridge interaction between Arg484 and Glu642 in the R484Q mutant. These weakened interactions were probably brought about the observation of the predicted hinge motion much earlier in the mutant than expected in the wild-type structure.

4.3. Case II

The family was ascertained in the Acıbadem Hospital, Istanbul and enrolled for a molecular genetic study at Acıbadem University as part of a whole exome/genome sequencing project investigating undiagnosed disorders in Turkish families. Written approved informed consent was provided from all participants. We immediately pursued WES analysis. At the end of the WES data analysis, we had 96 variants for de novo heterozygous, 421 variants for homozygous and 185 variants for compound heterozygous scenario. After the population frequency and splice site filtering, quantity decreased to 19, 46, 3 for de novo heterozygous, homozygous, compound heterozygous respectively. After recruiting data for gene intolerance (ExAC, GnomAD, and Genic Intolerance databases), mouse phenotype (MGI), pathogenicity scores (CADD, REVEL, M-CAP, PrimatAI, SIFT, Polyphen, and MutationTaster); four genes were highlighted namely *IGSF9, FZD6, EPYC, DIO2* (Table 4.4). We have also checked the top shet score gene list of Cassa *et al*. but none of the gene is found in the top list.

Table 4.4. The most prominent homozygous variants for case II

Gene	Variant_Type	Change	OMIM	MGI
IGSF9	splicing	NM_020789:exon 20:c.3183-6T	–	"Mice homozygous are viable and fertile but show abnormal miniature inhibitory postsynaptic currents and increased susceptibility to pharmacologically induced seizures."
FZD6	exonic	FZD6:NM_00131 7796:exon5:c.761 _768del:p.G254fs	Nail disorder, nonsyndromic congenital, 10	"Homozygous mice for one mutation display abnormal hair follicle orientation. Another mutation of this gene does not appear to result in a phenotype."
EPYC	splicing	NM_004950:exon 3:c.166-6T>-	–	"Mice homozygous for a knock-out exhibit short femurs and borderline osteoarthritis at 9 months of age."
DIO2	splicing	NM_013989:exon 2:c.223-9->T>T	–	"Mice homozygous for a disruption in this gene display elevated thyroxine (T4) and thyroid-stimulating hormone levels and changes in the metabolism and excretion of iodothyronines."

Table 4.4. The most prominent homozygous variants for case II (cont.)

Gene	ExAc	GnomAD	ExAc Metrix	Genic Intolerance	SIFT PolyPhen MutTaster	REVEL M-CAP CADD
IGSF9		–	z = 0.92 438.5/399 pLI = 0.57 41.7/9	10.44%	– – –	– – –
FZD6	–	–	z = 0.71 209.9/189 pLI = 0.00 20.2/14	20.07%	– – –	– – –
EPYC	–	–	z = -1.71 94.1/128 pLI = 0.00 11.2/8	59.92%	– – –	– – –
DIO2	–	–	z = 0.44 97.9/89 pLI = 0.22 6.8/2	45.71%	– – –	– – –

Text-mining the literature for each remained variant is the first thing to do at this stage. Known functions, domains, motifs of each protein; distribution of mutations on the gene, associated diseases, their inheritance patterns, exceptions regarding the inheritance pattern, overlapping and differentiating symptoms in patient phenotypes were all collected, and hypothesis created. Among all, the *FZD6* gene was the most significant gene that can be associated with the disease even though the mutation does not have high intolerance or pathogenicity prediction scores, due to its known association with NCDC10 in the literature and the results of MD simulations.

The available clinical symptoms for the evaluation of the WES results can be listed as thickened, hard, shiny, hyperplastic and hyperpigmented, claw-shaped (onycholysis) nails on the hands and feet. After the analysis steps that described in the method section and deep literature search, the genetic diagnosis we proposed for the patient was Nail Dysplasia 10. This diagnosis was confirmed by the clinician as the patient re-examined afterward for additional evaluation to determine the clinical fit.

Here, we report a consanguineous Turkish family with three affected individuals with homozygous 8 bp deletion mutation, p.Gly559Aspfs*16; c.1676_1683delGAACCAGC. The unaffected parents of each child are heterozygous carriers for this mutation. This amino acid change creates a premature stop codon at position 16 of the new reading frame where 133 amino acid is lost in C-terminus compared to native protein. Mutation nomenclature refers to GenBank mRNA reference sequence NM_001317796. So far, only two reports have described a small frameshift deletion in the *FZD6* gene (Kasparis *et al*., 2016; Mohammadi-asl *et al*., 2017). Previously, our mutation was reported in two other Turkish families with NDNC10, and therefore, this is the third Turkish family with the same mutation, indicating that all three families have a common ancestor.

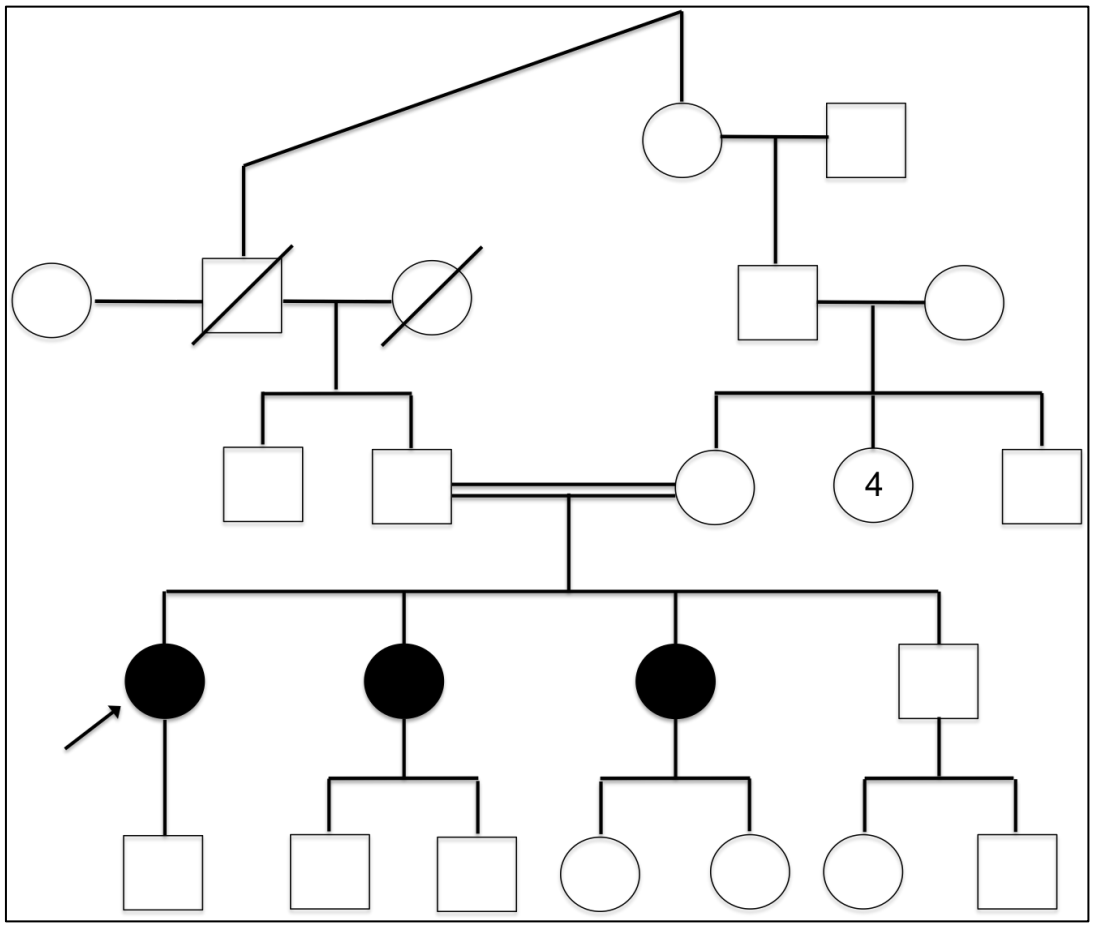
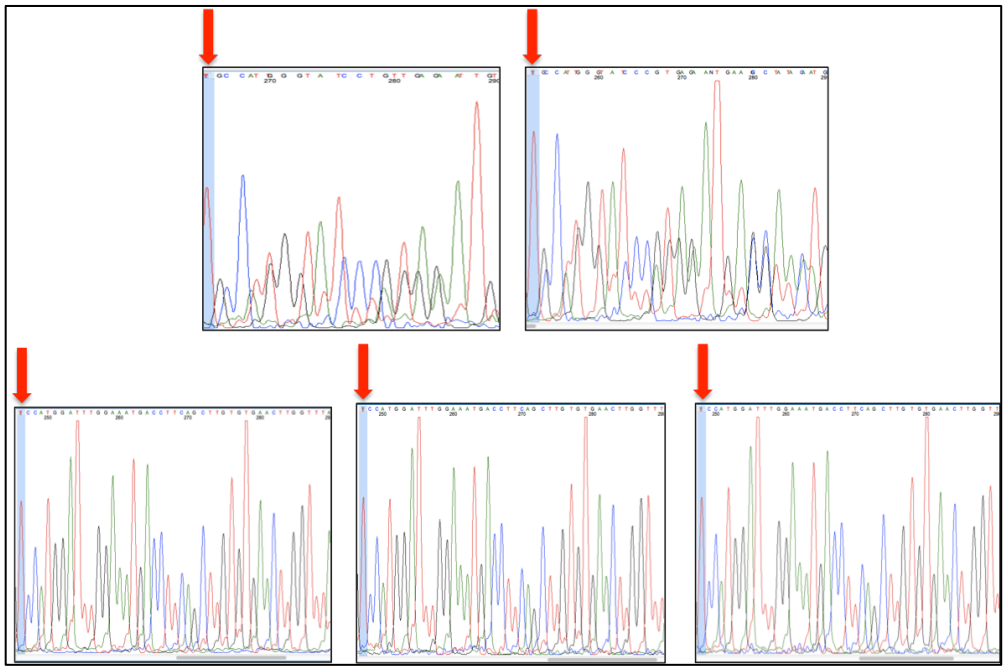


Figure 4.11. Pedigree of a consanguineous Turkish family segregating NCDC10. Circles and squares represent females and males, respectively. Clear symbols represent unaffected individuals while filled symbols represent affected individuals

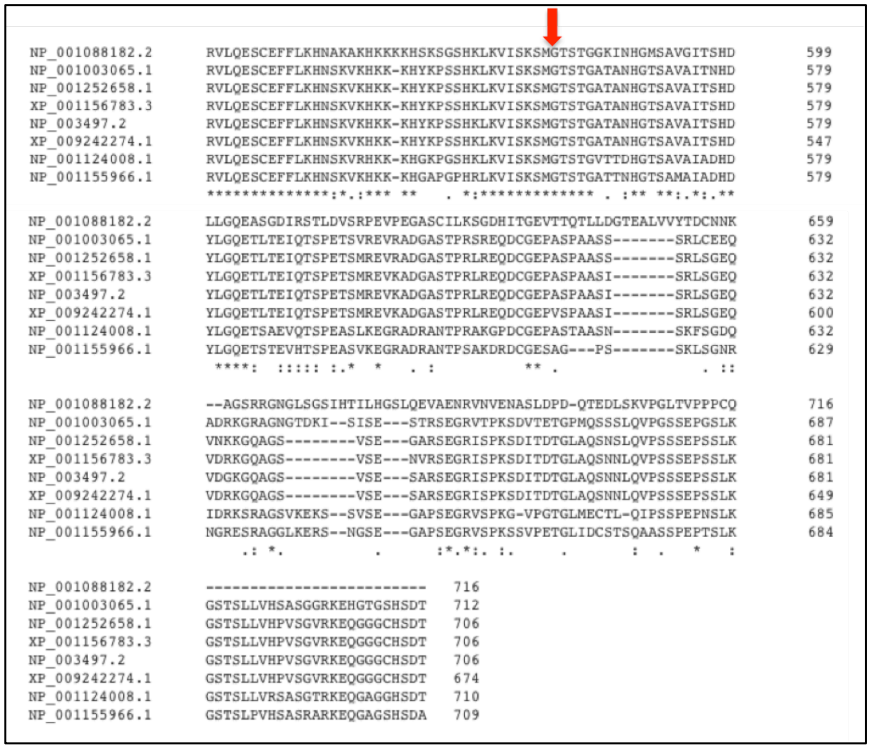
The mutation was confirmed via Sanger sequencing. The electropherograms of the father and mother show their heterozygous state for deletion in the upper panel while the index case and her two affected sisters are homozygous for c.1859delC. The arrow designates the position of the variant (Figure 4.12.a).

Evolutionary conservation of the disappeared amino acid region in other FZD6 orthologues was examined using Clustal Omega (Sievers *et al*., 2014). The potential functional significance of the frameshift mutation is supported by the fact that the region is quite conserved in several species including human, rat, and mouse. Species abbreviations and accesstion numbers are as follows: Species abbreviations are as follows: Hs, Homo sapiens; Pt, Pan troglodytes; Ma m, Macaca mulatta; Pa, Pongo abelii; Bt, Bos taurus, Cf, Canis lupus familiaris; Rn, Rattus norvegicus; Mm, Mus musculus; Xl Xenopus laevis. The accession numbers for the respective proteins are as follows: Hs, NP_003497.2; Pt, XP_001156717.1; Mm, NP_032082.2; Pa, XP_009242274.1 ; Cf, NP_001003065.1; Rn, NP_001124008.1; Mm, NP_032082.2; Xl, NP_001088182.2 (Figure 4.12.b). The Clustal Omega results of the paralogs of FZD6 suggest that both the N- and C-terminal regions are highly variable compared to the transmembrane domains. In the light of this fact, we deduce that there is an alteration in the interactions partners and thus of variations of functions in FZD family proteins.

(a)



(b)



```
NP_001088182.2      RVLQESCEFFLKHNAKAKHKKKHKSKSGSHKLKVISKSMGTSTGGKINHGMSAVGITSHD       599
NP_001003065.1      RVLQESCEFFLKHNSKVKHKK-KHYKPSSHKLKVISKSMGTSTGATANHGTSAVAITNHD       579
NP_001252658.1      RVLQESCEFFLKHNSKVKHKK-KHYKPSSHKLKVISKSMGTSTGATANHGTSAVAITSHD       579
XP_001156783.3      RVLQESCEFFLKHNSKVKHKK-KHYKPSSHKLKVISKSMGTSTGATANHGTSAVAITSHD       579
NP_003497.2         RVLQESCEFFLKHNSKVKHKK-KHYKPSSHKLKVISKSMGTSTGATANHGTSAVAITSHD       579
XP_009242274.1      RVLQESCEFFLKHNSKVKHKK-KHYKPSSHKLKVISKSMGTSTGATANHGTSAVAITSHD       547
NP_001124008.1      RVLQESCEFFLKHNSKVRHKK-KHGKPGSHKLKVISKSMGTSTGVTTDHGTSAVAIADHD       579
NP_001155966.1      RVLQESCEFFLKHNSKVKHKK-KHGAPGPHRLKVISKSMGTSTGATTNHGTSAMAIADHD       579
                    **************:*.:*** **    . *:************** . :** **:.*:.**

NP_001088182.2      LLGQEASGDIRSTLDVSRPEVPEGASCILKSGDHITGEVTTQTLLDGTEALVVYTDCNNK       659
NP_001003065.1      YLGQETLTEIQTSPETSVREVRADGASTPRSREQDCGEPASPAASS-------SRLCEEQ       632
NP_001252658.1      YLGQETLTEIQTSPETSMREVRADGASTPRLREQDCGEPASPAASS-------SRLSGEQ       632
XP_001156783.3      YLGQETLTEIQTSPETSMREVKADGASTPRLREQDCGEPASPAASI-------SRLSGEQ       632
NP_003497.2         YLGQETLTEIQTSPETSMREVKADGASTPRLREQDCGEPASPAASI-------SRLSGEQ       632
XP_009242274.1      YLGQETLTEIQTSPETSMREVKADGASTPRLREQDCGEPVSPAASI-------SRLSGEQ       600
NP_001124008.1      YLGQETSAEVQTSPEASLKEGRADRANTPRAKGPDCGEPASTAASN-------SKFSGDQ       632
NP_001155966.1      YLGQETSTEVHTSPEASVKEGRADRANTPSAKDRDCGESAG---PS-------SKLSGNR       629
                     ****:  ::::: :.*  *   . :          ** .               . ::

NP_001088182.2      --AGSRRGNGLSGSIHTILHGSLQEVAENRVNVENASLDPD-QTEDLSKVPGLTVPPPCQ       716
NP_001003065.1      ADRKGRAGNGTDKI--SISE---STRSEGRVTPKSDVTETGPMQSSSLQVPGSSEPGSLK       687
NP_001252658.1      VNKKGQAGS--------VSE---GARSEGRISPKSDITDTGLAQSNSLQVPSSSEPSSLK       681
XP_001156783.3      VDRKGQAGS--------VSE---NVRSEGRISPKSDITDTGLAQSNNLQVPSSSEPSSLK       681
NP_003497.2         VDGKGQAGS--------VSE---SARSEGRISPKSDITDTGLAQSNNLQVPSSSEPSSLK       681
XP_009242274.1      VDRKGQAGS--------VSE---SARSEGRISPKSDITDTGLAQSNNLQVPSSSEPSSLK       649
NP_001124008.1      IDRKSRAGSVKEKS--SVSE---GAPSEGRVSPKG-VPGTGLMECTL-QIPSSPEPNSLK       685
NP_001155966.1      NGRESRAGGLKERS--NGSE---GAPSEGRVSPKSSVPETGLIDCSTSQAASSPEPTSLK       684
                        .: *.          .      :*.*:. :.     .       :  .   *   :

NP_001088182.2      -------------------------   716
NP_001003065.1      GSTSLLVHSASGGRKEHGTGSHSDT   712
NP_001252658.1      GSTSLLVHPVSGVRKEQGGGCHSDT   706
XP_001156783.3      GSTSLLVHPVSGVRKEQGGGCHSDT   706
NP_003497.2         GSTSLLVHPVSGVRKEQGGGCHSDT   706
XP_009242274.1      GSTSLLVHPVSGVRKEQGGGCHSDT   674
NP_001124008.1      GSTSLLVRSASGTRKEQGAGGHSDT   710
NP_001155966.1      GSTSLPVHSASRARKEQGAGSHSDA   709
```

Figure 4.12. (a) The electropherograms of the family. (b) Partial amino acid sequence of the human FZD6 protein in comparison with orthologues from other species. The mutation point of c.1676_1683delGAACCAGC frameshift deletion is indicated by an arrow

4.3.1. MD Simulations

The crystal structure of FZD6 protein has not yet deposited to Protein Data Bank (PDB). Thus, 3D structures of FZD6 protein, including native form and p.Gly559Aspfs*16 mutation, were modeled via I-TASSER before performing 20 ns MD simulations. As displayed in Figure 4.14, mutant protein (FZD6 mutant) displays higher backbone motion compared to native one. From beginning of 4th ns through the end of MD trajectory, the tendency for increased RMSD is clearly observed and the particular differences in backbone responses of native and mutant ones become more apparent within last 5 ns. As in line with RMSD patterns, higher RMSF values are observed in FZD6 mutant. Especially, Leu253-Cys282, Ala329-Phe380 and His549-Ser571 regions display higher RMSF values then the native protein (Fig 4.13). The increased flexibilities in Leu253-Cys282 and Ala329-Phe380 regions of FZD6 mutant could be explained by the crosstalk between these regions from loop structures, closely located to each others (Fig 4.14).

(a)

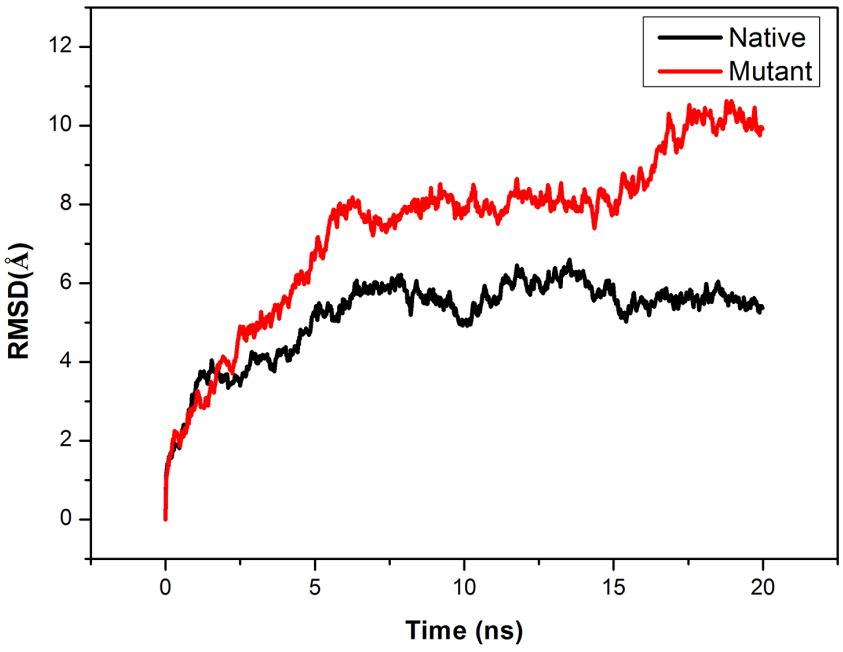


Figure 4.13. RMSD (a) and RMSF (b) results of modeled FZD6 proteins, native and mutant, along 20 ns

(b)

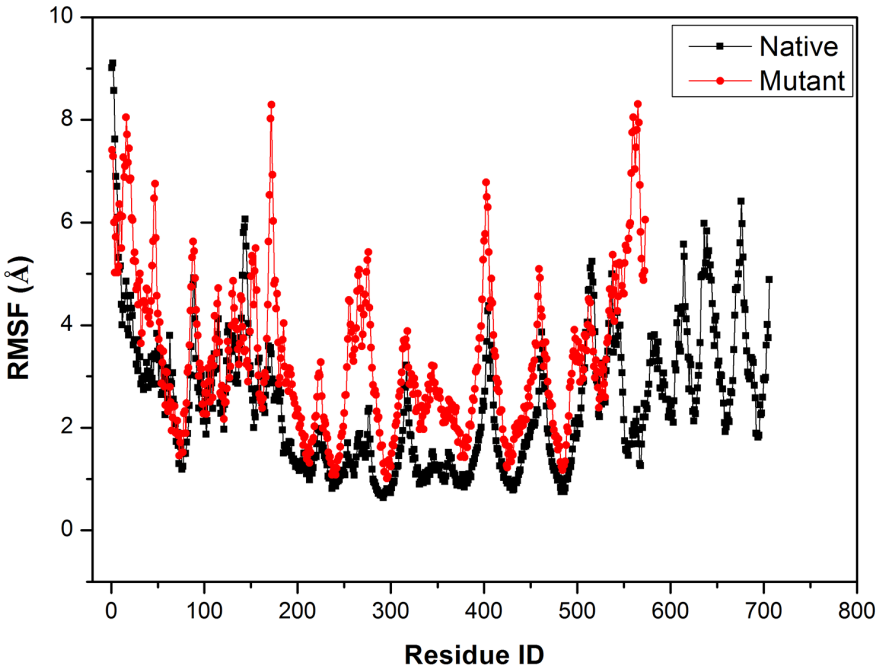


Figure 4.13. RMSD (a) and RMSF (b) results of modeled FZD6 proteins, native and mutant, along 20 ns (cont.)

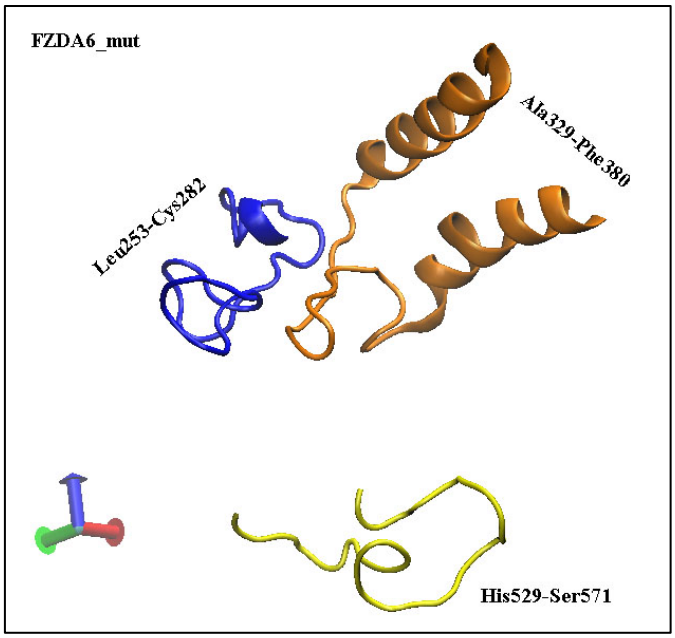


Figure 4.14. The secondary structure formations in FZD6 mutant displaying higher RMSF values compared to native protein

It is also important to notice that KTxxxW motif, considered as significant for FZD signaling, is not closely located to these listed regions. Specifically, we focus on the flexibilities of residues in KTxxxW motif. As displayed in Figure 4.15, the increase in flexibilities of KTxxxW motif has been captured in mutant protein. This particular increase in KTxxxW motif could be considered as a part of general trend observed in mutant protein with respect to native one.

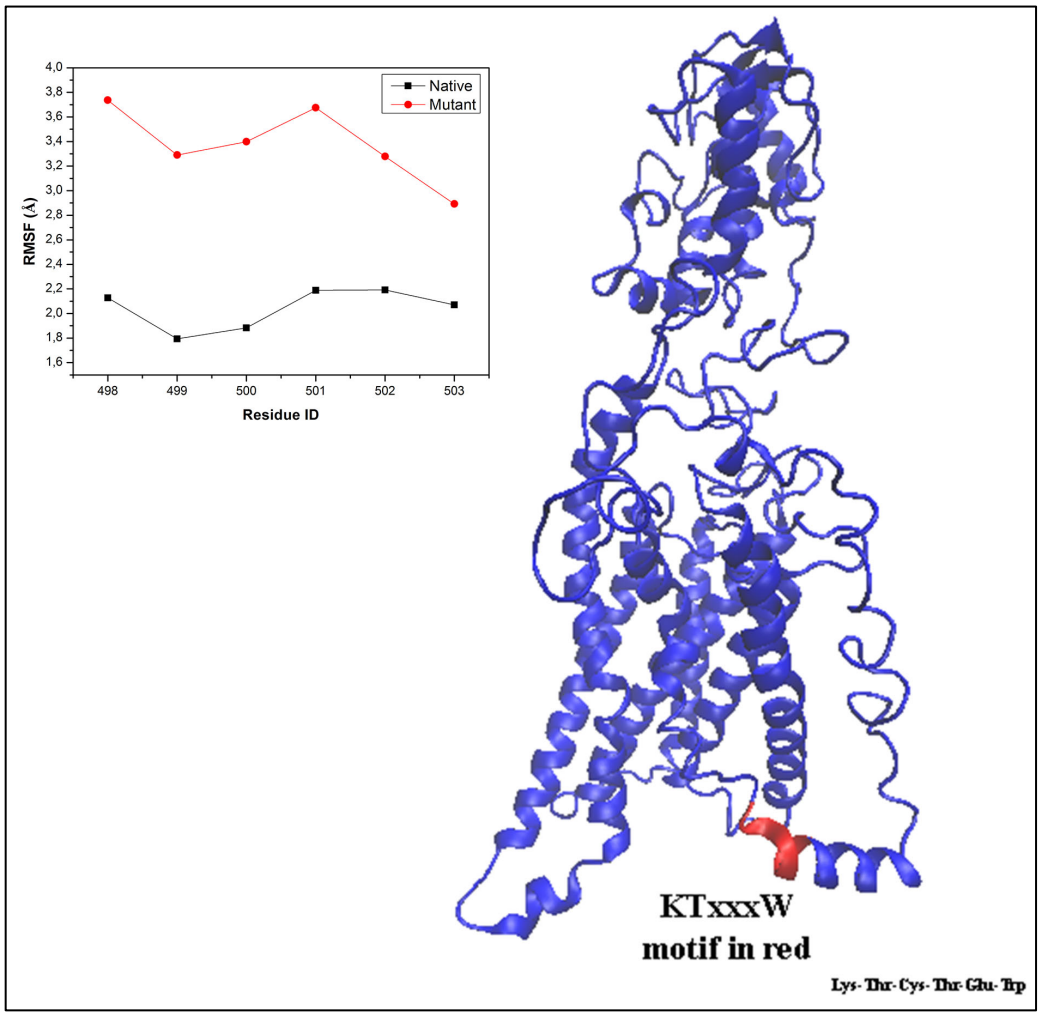


Figure 4.15. Comparison of the flexibilities of residues in KTxxxW motif

Secondly, the significant salt-bridge interactions for native and mutant proteins are evaluated along 20 ns MD trajectory. Due to the existence of p.Gly559Aspfs*16 mutation, 28 salt-bridge interactions among 89 ones are not existed in the mutant proteins. This number corresponds to almost 30% of all salt-bridge interactions, and the loss of this portion would be one of reasons behind increased RMSD pattern (Figure 4.13a). Among

these 28 salt-bridge interactions, 7 ones are established between C-terminal structure and β–sheet structures, being considered as a part of the seven trans membrane-spanning receptor (Figure 1.2). Thus, the loss of these particular interactions results in the weakening of intramolecular interactions in an obvious way. As displayed in Figure 4.16, these salt-bridge interactions are mostly strong, except Glu697-Lys552, to contribute the stability of protein in positive manner.

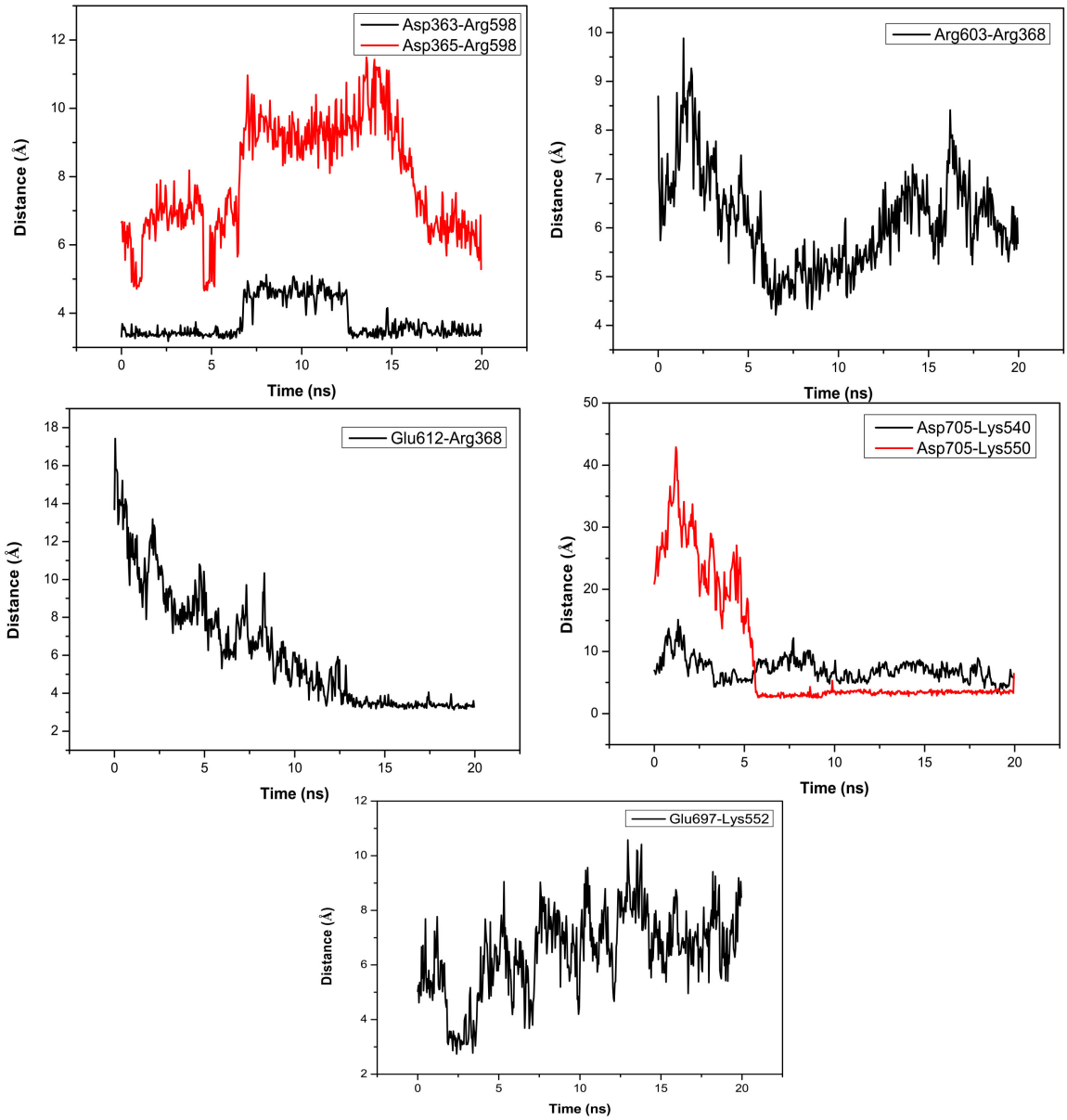


Figure 4.16. The salt-bridge interactions loss in FZD6 mutant upon mutation

5. DISCUSSION

5.1. Overview

Rare diseases (RDs) are any kind of diseases that affect a small percentage of the population. About 30% are still lacking a diagnostic definition, and the rest can get an accurate diagnosis quite late. Molecular diagnosis is the most prominent way to facilitate accurate diagnosis and an effective and appropriate treatment for rare undiagnosed cases. WES/WGS provides a great potential to develop high-throughput and low-cost platforms for clinical diagnostics. Technical hurdles to clinical applications of these methods are mainly downstream bioinformatics analysis and causative variant detection.

This dissertation aims to construct a bioinformatics workflow to diagnose undiagnosed patients with a suspected genetic disorder, where other testing modalities have been inconclusive or noninformative. More specifically, the objectives were to 1) develop a bioinformatics workflow for detection of SNVs/indels 2) develop a variant prioritization workflow aiming not to miss the causative variant and also end up with manageable number of variants through the selected threshold for population frequency, excellent combination of pathogenicity prediction tools and gene intolerance scores 3) explain the pathogenicity mechanisms of mutations via molecular dynamics simulations. Since the prices of WGS are falling rapidly and turnaround time including data analysis can be reduced to a few days, WGS is now started to be considered as an alternative to WES. It is important in this regard that our pipeline is also appropriate for WGS.

The bioinformatics workflows created from the very first step, the WES paired-end library files. Trimmomatic was used to remove adapters and low quality (Phred quality score <5) bases (Bolger, Lohse, & Usadel, 2014). Further processing was performed following the GATK best practice recommendations. Briefly, BWA mem v0.7.12 is used to map the trimmed reads to the human reference genome (UCSC GRCh37/hg19) (Li & Durbin, 2009) and then Picard tools (v1.141) were used to mark and remove the duplicate reads. GATK (v3.4) was used for indel realignment, BQSR, calling variants, joint genotyping and VQSR (McKenna *et al*., 2010). Annovar (v2015-03-22) was used to

annotate and filter variations against public databases (dbSNP138, 1000 Genomes Project, and ExAC Browser) (Wang, Li, and Hakonarson; 2010).

Our pipeline outputs three different variant lists with the information of the affected and control cases as an input. It outputs three different variant lists; de novo heterozygous, homozygous and compound heterozygous variants. Independent from the pedigree information, we figure out all variants that are inherited in an autosomal dominant/recessive or X-chromosome/Y-chromosome manner.

The first strategy that we consider to filter the benign variants is their existence in the human population at an allele frequency higher than >0.1%. So, variants with a MAF of <0.1%, without homozygous carriers in public databases and predicted to affect protein coding were taken into consideration. For the intronic alterations, the ones at exon-intron boundaries ranging from -10 to + 10 are retained. If no prominent variant is found, then the threshold is increased to -40 to +40 (Figure 5.1). To prioritize those rare variants, data collected from various sources: ExAC and GnomAD for allele frequencies; MGI for mouse phenotypes; CADD, REVEL, M-CAP, PrimatAI, SIFT, Polyphen, and MutationTaster for pathogenicity predictions and Pubmed for the literature search. For the splice site mutations, four splice site prediction programs were used; Human Splicing Finder, NetGene2Server, Berkeley Drosophila Genome Project-Splice Site Prediction by Neural Network and Oriel SpliceView. Instead of expecting the support for the disease-causing effect from all of the tools, the information obtained from each tool is taken into account.

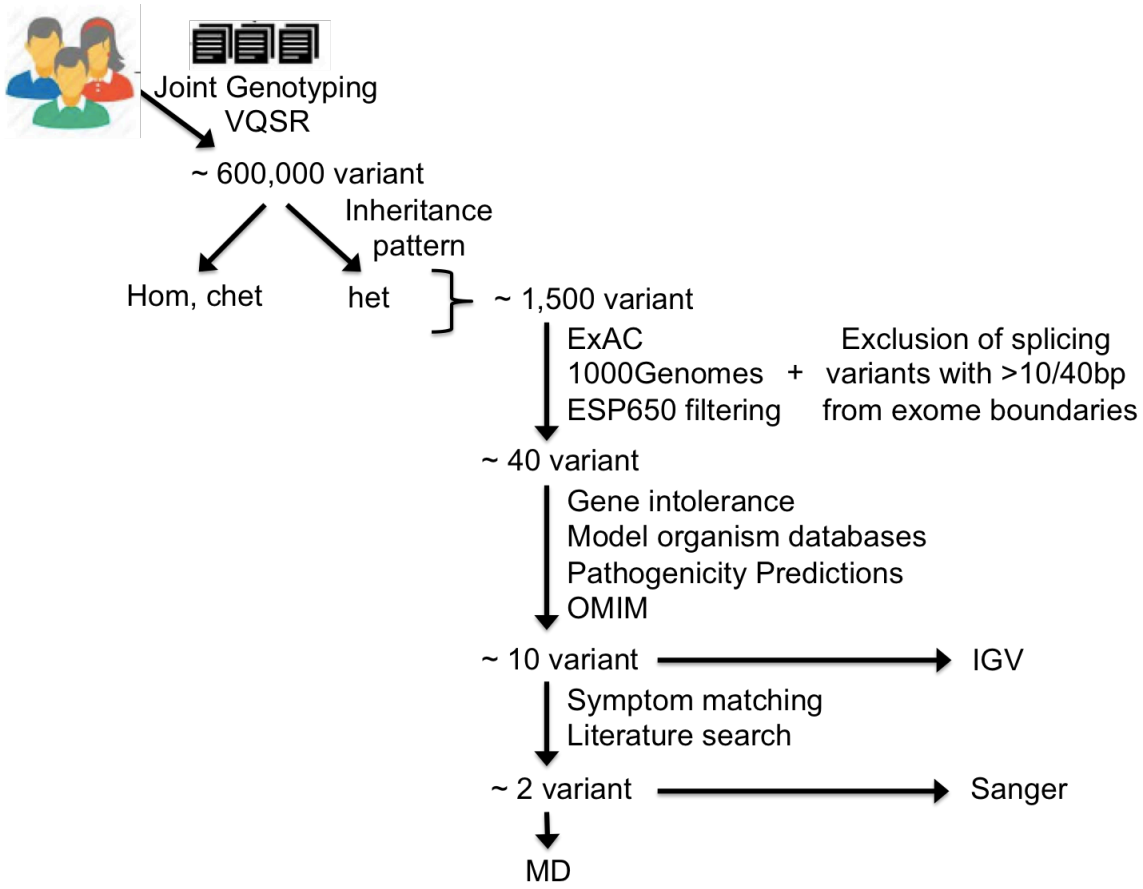


Figure 5.1. The workflow for filtering large genomic datasets generated from rare Mendelian diseases

After prioritization of the variants via the criteria listed above, we continue with text-mining the literature for each remained variant. Finally, we have done MD simulations for the most prominent one to compare the wild-type and mutated proteins and predict the pathogenicity mechanism of the causative variant. MD is a computer simulation method that mimics and monitors the physical movements of several hundreds of atoms and molecules in solution with explicit solvent representations on a femtosecond timescale by calculating the electrostatic charges. It is a very suitable method for routine clinical practice and should be added to the molecular diagnosis pipeline since the time of each simulation takes will decrease as computational power in cpu/gpu develops in the near future (Teo *et al*., 2014) and the information obtained with this technique provides very significant contributions to the variant pathogenicity as long as it is interpreted by experienced scientists.

Finally, the workflow was applied to several undiagnosed cases and their family members and successfully prioritized variants and provided the diagnosis. For two of these families, the pathogenicity mechanisms of mutations were described via molecular dynamics simulations.

5.2. Bioinformatic Analysis, Diagnosis and Functional Impact Prediction of Case I

The available clinical symptoms for the index case are occipital lobe epilepsy and epileptic status in sleep. After the analysis steps that described in the method section and deep literature search, the genetic diagnosis we proposed for the patient was *FLNA*-Related Periventricular Nodular Heterotopia. This diagnosis was confirmed by the clinician as the patient re-examined afterward.

The human FLNA protein is a 280 kDa elongated protein composed of an N-terminus followed by a long rod region (Fig. 1.1). This rod region consists of 24 domains (each containing about 96 amino acids) that are separated by a flexible, non-conserved, 25-residue-long hinge 1 segment; rod 1 (domains 1-15) and rod 2 (containing domains 1-15). The other hinge segment is 35-residue-long and stays between the domains 23 and 24 (Gorlin *et al*. 1990). FLNA binds to actin through its actin-binding domain located at the N-terminus, whereas the C-terminal domain contains the site of homo-dimerization and binds to membrane glycoproteins.

FLNA expression is essential not only for mammalian development (Feng *et al*., 2006; Ferland *et al*., 2006; Hart *et al*., 2006), but also for development of other organisms, such as Dictyostelium (Khaire *et al*., 2007; Annesley *et al*., 2007), Drosophila and C. elegans (Kovacevic *et al*., 2010). *FLNA*-associated PNH is predominantly seen in women with difficult to treat seizures. In contrast, hemizygous *FLNA* mutations in males are mostly lethal, and it is assumed that the loss of function mutations in males results in a more severe phenotype (Fox *et al*., 1998; Sheen *et al*., 2003; Guerrini and Parrini, 2010). So far, only nineteen viable males have been reported in the literature (Sheen *et al*., 2001; Parrini *et al*., 2004; Gerard-Blanluet *et al*., 2006; Hehr *et al*., 2006; Kasper *et al*., 2012; Fergelot *et al*., 2012; Oegema *et al*., 2013; Oda *et al*., 2015; Lange *et al*., 2015; Liu *et al*., 2017; Saygı *et al*., 2018). There are two different scenarios to explain the liveborn males

with *FLNA* mutation: Mosaicism and partial loss of function of the gene. There are several reports that revealed the mosaicism with *FLNA* in males with a classical PNH phenotype (Guerrini *et al*., 2004; Parrini *et al*., 2004). In addition, mutations in *FLNA* consistent with residual function were reported to cause PNH in males, with a less severe outcome; often missense changes, or alleles that only truncate the extreme C-terminus (Parrini *et al*., 2006; Sheen *et al*., 2001; Sole *et al*., 2009). Hehr *et al*. reported a male with splice site mutation in *FLNA*. They showed that the splice site resulted in both normal and aberrant mRNA transcripts and this can result by retaining some normal FLNA function (Hehr *et al*., 2006). Oda *et al*. also reported two boys with a milder phenotype who carry in-frame exome skipping of a 4 bp deletion in *FLNA* (Oda *et al*., 2016).

Mutations in *FLNA* cause several allelic X-linked disorders, not only PNH, but also skeletal syndromes like Melnick–Needles syndrome, cardiac valvular dystrophy, chronic idiopathic intestinal pseudo-obstruction, FG syndrome, and terminal osseous dysplasia (Robertson, *et al*., 2005; Unger *et al*., 2007; Sun *et al*., 2010). Since numerous binding partners interact with FLNs, the pathological mechanisms of the diseases are most likely attributed to the loss of partner binding or aberrant interactions caused by mutations (Robertson *et al*., 2005). This supports its diverse interactions with many different protein networks. The long list of FLNA-interacting proteins supports this point of view (Feng and Walsh, 2004).

Based on the crystal structure of FLNA protein containing domain 3, 4, and 5, it has been suggested that the interfaces between domain 3, 4 and 5 are highly conserved, and these interfaces are interacting through β-sheet formations, located side by side with each other (Sethi *et al*., 2014). The interaction between domain 3 and 4 is mediated via the edges of β-sheets whereas domain 4 and domain 5 interact along three β-sheet formations. Compared to the interaction between domain 4 and 5, tighter one is established between domain 3 and 4 with polar residues (Sethi *et al*., 2014). In FLNA protein, Trp582 is playing a crucial role to establish the proper interactions between domain 4 and 5. This particular domain-domain interaction reported for the crystal structure of FLNA protein, domain 3, 4 and 5, is a unique property of the entire Ig superfamily.

Our mutant, R484Q, is located to interaction interface between domain 3 and 4. Upon R to Q mutation, the characteristic of the interface as being polar is not altered. Based on previous studies, it is concluded that mutation proline or valine to glutamine or aspartate has to lead the substantial changes even being in a hydrophobic core by destabilizing the individual β-sheet formations.

Molecular dynamics simulations of FLNA protein including domain 3, 4 and 5, is firstly carried out by Sethi *et al*. along 50 ns MD trajectory, collected at 300 °K as NpT ensemble. This theoretical calculation is performed to validate SAXS data suggesting that domain 3 is changing its orientation with respect to domain 4 and domain 5 whose interactions remained intact (Sethi *et al*., 2014). Thermal stability assay performed with individual domain has suggested that the presence of domain 5 stabilizes domain 4 whereas domain 3 does have an additional role to stabilize domain 3 (Sethi *et al*., 2014).

In summary, we diagnosed a family PNH in an X-linked dominant transmission from a clinically asymptomatic mother to two sons. We detected a novel c. 1451G>A change in *FLNA* exon 10, leading to the substitution of a very conserved amino acid (p.R484Q). Arg484 residue is involved in salt-bridge interaction with Glu642 in native FLNA protein. With Arg to Gln replacement at 484th position, this particular interaction is lost, and it results in the distancing of domain 3 from domain 4 by disrupting the stabilization of interface between domain 3 and 4. Together the loss of Arg484:Glu642 interaction, the Glu499:Arg488 interaction is also weakened; we conclude that Arg484:Glu642 interaction in FLNA protein has a role to stabilize the interaction between domain 3 and domain 4. Even other salt-bridge interactions existed in interaction interphase of domain 3 and domain 4 are conserved; it is not enough to prevent a distancing of domain 3 from domain 4. Hence, we consider Arg484 as the most crucial residue along with the interaction interphase of domain 3 and 4 for keeping all three domain of FLNA protein together.

5.3. Bioinformatic Analysis, Diagnosis and Functional Impact Prediction of Case II

The available clinical symptoms for the index case are thickened, hard, shiny, hyperplastic and hyperpigmented, claw-shaped (onycholysis) nails on the hands and feet. After the analysis steps that described in the method section and deep literature search, the genetic diagnosis we proposed for the patient was Nail Dysplasia 10. This diagnosis was confirmed by the clinician as the patient re-examined afterward.

So far, ten types of NDNC were recorded in the literature, six of which are inherited in an autosomal dominant manner. The genes associated with human hereditary nail disorders are listed as *HPGD, RSPO4, PLCD1, COL7A1* and *FZD6* (Khan *et al*., 2015). *HPGD* gene is found to be associated with isolated congenital nail clubbing (OMIM 119900) and is responsible for the metabolism of prostaglandins. Following irritation or injury, arachidonic acid (AA) is released and oxygenated by calcium-dependent enzyme systems leading to the formation of prostaglandins. Specifically, prostaglandin E2 is readily detectable in equine acute inflammatory exudates. Moreover, both the influx of extracellular calcium and mobilization of intracellular calcium are very critical for the process of prostaglandin formation (Taylor *et al*., 1990). Another gene is *RSPO4*, linked to nail disorder, nonsyndromic congenital (NCDC4; OMIM 206800); encodes a secreted protein R-spondin 4 with a known role in embryonic development and homeostatic self-renewal in adult tissues; besides its role in Wnt signaling which has both anti-inflammatory and pro-inflammatory functions. *PLCD1* is linked to NDNC3 (OMIM 151600); a member of the phospholipase C family that regulates homeostasis of the immune system in the skin. The lack of PLCD1 protein induces skin inflammation; since the skin of *PLCD1* -/- mice shows typical inflammatory phenotypes, including increased dermal cellularity, leukocyte infiltration and expression of pro-inflammatory cytokines. Moreover, exogenously expressed PLCD1 attenuates LPS-induced expression of IL-1b (Ichinohe *et al*., 2007). Another gene related to nail disorders is *COL7A1*, which the alpha chain of type VII collagen that is associated with NDNC8 (OMIM 607523). Mutations in COL7A1 induce lifelong severe skin and mucosal blistering followed by scarring, caused by loss of adhesion between the epidermis and the dermis. *COL7A1* -/- mice also display blisters and erosion at sites of trauma, subepidermal blistering, and high postnatal lethality. Finally, *FZD6* gene is known to be associated with NDNC10 (OMIM 614157). FZD6 is a member

of G-protein coupled receptor Class 6. It is the largest member of the FZD family and C-terminus of FZD3 and FZD6 are longer than the other FZDs. The gene function is defined as a negative regulator of the canonical Wnt/beta-catenin signaling. FZD6 signaling was shown to activate beta-catenin in a study of patients affected by nail dysplasia. The same study reported that Wnt3a signaling causes beta-catenin accumulation in healthy, but not *FZD6*-mutant fibroblasts, indicating a canonical role of FZD6 in this context (Fröjmark *et al*., 2011).

The common intersection point of all known genes is their association with the immune system, specifically innate immunity. Some NCDC patients show inflammation problems and all known NDCD associated genes play roles in inflammation (Gattinoni *et al*., 2010; Kuehl & Egan, 1980). The truncation mutation we found was previously reported in two other Turkish families, indicating founder effect. The phenotype of the affected individuals in our family is very similar to the other two families; except the uveitis in the index patient. The diagnosis of uveitis and possible ocular tuberculosis in the index patient is noteworthy. Ocular follow-up of our patient will probably help differentiate between an autoinflammatory granulomatous process and ocular tuberculosis.

Limited information is available regarding the interaction of WNT–FZD protein families. However, Kilander *et al*., showed via fluorescence recovery after photo-bleaching (FRAP) that recombinant WNT-1, -2, 3A, -4, -5A, -7A, -9B, and -10B affect FZD6 surface mobility and thus directly act on FZD6 (Kilander *et al*., 2014). The loss of interaction partners we proposed due to our truncation mutation could mainly be WNT family proteins. WNT pathway and innate immunity are also shown to be interrelated. There is an interaction among the WNT signaling network, inflammatory cytokines, and innate immune signaling pathways (Gatica-andrades *et al*., 2017). Individual WNT proteins was shown to have pro- or anti-inflammatory functions. WNT ligands and WNT/β-catenin signaling was found to positively regulate LPS-induced pro-inflammatory cytokines. The WNT signaling pathway plays a major role in regulating tolerance versus immunity, particularly in DCs, and more (Swafford *et al*., 2015). Therefore, it is not unexpected that immune-related problems are seen in NCDC patients.

Fröjmark *et al*. were the first to link the mutations in *FZD6* gene to autosomal recessive nail dysplasia (Fröjmark *et al*., 2011). They identified two different mutations in two large consanguineous Pakistani families. Affected individuals were homozygous for missense mutation p.Arg511Cys for one family and homozygous for nonsense mutation p.Glu584Ter in the other family. Later, several reports of other patients from Pakistan, Iran, and Turkey were reported. Naz *et al*. reported two more Pakistani families where affected individuals were also homozygous for mutation p.Glu584Ter, indicating a common ancestor (Naz *et al*., 2012). Raza *et al*. also reported another Pakistani family with a homozygous p.Gly422Asp; c.1265G>A mutation (Raza *et al*., 2013). At the same year, two other families with new mutations were reported; in one family affected individuals were homozygous for missense mutation p.Arg509Ter and in the second family affected individuals were compound heterozygous for mutations p.Arg96Cys/p.Glu438Lys (Wilson *et al*., 2013). Moreover, in 2016, homozygous mutation for an 8 bp deletion, p.Gly559Aspfs*16; c.1676_1683delGAACCAGC was detected in two Turkish families (Kasparis *et al*., 2016). In 2017, a homozygous 1bp deletion variant, c.1859delC (p.Ser620Cysfs*75) was seen in an Iranian family (Mohammadi-asl *et al*., 2017). To date, seven different mutations have been reported in eleven families, including two missense, two nonsense, two frameshifts, and one compound heterozygous (Fröjmark *et al*., 2011; Naz *et al*., 2012; Raza *et al*., 2013; Wilson *et al*., 2013; Kasparis *et al*., 2016; Mohammadi-asl *et al*., 2017). Five out of mutations are clustered in the C-terminus, which suggests that the C-terminal region could be a mutation hotspot.

Currently very little is known regarding the structure and function of Frizzled receptors. In general, the link between FZDs and heterotrimeric G proteins is a matter of discussion in the field. Through mutagenesis studies, it has been revealed that several residues in the intracellular loops and the C-terminus of FZDs play critical roles for signaling. Specifically, the mutation of the highly conserved internal KTxxxW motif between 498th and 503rd positions in the C-terminus or single amino acid changes in the first (R340A) or the third (L524A) intracellular loops of, another protein from human FZD family, FZD5 completely abolished FZD signaling. The same mutations completely ablated the binding of the phosphoprotein DVL and its membrane recruitment by FZD (Cong *et al*., 2004) which is a central player in FZD-induced signal transduction and functionally necessary for all FZD signaling pathways (Wallingford *et al*., 2005; Malbon *et*

al., 2006). PDZ domain of DVL directly binds the KTxxxW motif of FZD (Wong *et al*., 2003).

Also, in general, agonists binding to GPCRs were shown to induce changes in C-tail conformation that is necessary for activating heterotrimeric G protein (Nie *et al*., 2001). Prolonged agonist stimulation catalyzes the phosphorylation of the C-tail, promoting arrestin binding, desensitization, and GPCR internalization (Drake *et al*., 2006). Studies utilizing the peptides encoding the C-tail of FZD also suggest alpha-helicity in C-terminus of the Frizzleds seems a must for efficient protein-protein interaction with DVL and other downstream signaling elements (Punchihewa *et al*., 2009; Lemma *et al*., 2013). Moreover, shortening the C-tail beyond C507 of, another protein from human FZD family, FZD5 impaired normal DVL recruitment and the ability of Wnt activator to activate Lef/ Tcf-dependent transcription (Cong *et al*., 2004; Tauriello *et al*., 2012).

In terms of experimental studies, the existing knowledge about FZD6 is still limited. Fröjmark *et al*. expressed wild-type and mutant (p.Glu584X and p.Arg511Cys) variants of FZD6 fused to green fluorescent protein (GFP) in HEK293T cells. While the missense mutation has no or little effect on total FZD6 levels, no expression was detected from the nonsense one. FZD6 was shown to be critical for the morphogenesis of hair follicles in Drosophila and mice. *FZD6* -/- mice are viable and fertile; but among more than 100 *FZD6* -/- mice examined, all have abnormal macroscopic hair whorls (Guo *et al*., 2004). Besides, *FZD3* -/- and *FZD6* -/- double-mutant mice die within minutes of birth and have a mis-oriented pattern of inner-ear sensory hair cells, this points out the role for FZD6 in planar-cell polarity. According to the same study of Wang *et al*, *FZD6* gene is expressed in all sensory hair cells and in many nonsensory epithelial cells in the inner ear (Wang *et al*., 2006). Moreover, Fröjmark *et al*. reinvestigated the *FZD6* -/- mouse model and about 50% of male knock-out mice, but none of the female mice had absent or abnormal claws compared to wild-type mice. To link the expression of FZD6 to early nail development they also checked FZD6 expression in mouse embryos at several embryonic days and revealed that at E16.5 there was an expression of FZD6 in the epidermis of the digital tip in the region corresponding to the developing nail bed and ventral part of the digit (Fröjmark *et al*., 2011). In addition to that, Naz *et al*. reported a strong expression level of FZD6 in the ventral nail matrix and some FZD6 staining in the nail bed (Wilson *et al*., 2013).

Due to the lack of crystal structure of FZD6, computer-based analyzes of FZD6 is also very limited. The first attempt made by Mohammadi-asl, *et al*. as predicting the formation of multiple helical secondary structures in the distal cytoplasmic region of the p.Ser620Cysfs*75 mutant protein which does not exist in the normal protein via I-TASSER (Roy *et al*., 2010). Moreover, they used NtePhos 3.1 server and revealed pathogenic consequence of the mutation by disturbing the cytoplasmic domain structure and signaling through loss of phosphorylation residues (Mohammadi-asl *et al*., 2017). They concluded that the nonsense mutation causes the loss of the distal end of the topological domain (amino acids 495-706). This domain mediates Wnt/beta-catenin signaling by relocalization and phosphorylation of disheveled proteins. Even though, KTXXXW is present, loss of phosphorylation residues and formation of unusual helical secondary structures can result in lack of proper response to WNT-3A and WNT-5A activation consistent with previous studies (Fröjmark *et al*., 2011; Naz *et al*., 2012; Mohammadi-asl *et al*., 2017).

For the same reason, we performed the homology modeling of native and mutant forms of FZD6 protein with I-TASSER. To gain more insight about the impacts of mutation on the structure of the protein, we performed 20 ns MD simulations and concluded that FZD6 mutant displayed higher RMSD pattern compared to native. This result suggests us that the introduction of stop codon to C-terminus, associated with the translation of new 15 amino acids upon the frame-shift, results in increased tendency for unfolding than native structure with higher backbone motion at 310 °K. This result is also supported with RMSF pattern that Leu253-Cys282, Ala329-Phe380, and His549-Ser571 regions (mutant numbering) are more flexible in FZD6 mutant compared to native one. Specifically, His549-Ser571 region (mutant numbering) displays almost ~8-fold more flexibility compared to native. Hence, the loss of C-terminus, even being in partial, would disrupt the intramolecular interactions and we could end up with unstable protein. It is also crucial to notice that the conservation of KTxxxW motif in C-terminus, previously suggested as essential for FZD signaling in FZD5 protein, is not enough for FZD6 mutant. This fact suggests that the problem in our case can be the loss of structural integrity in addition to the loss of signaling region.

Along 20 ns MD trajectory, we also consider the impacts of mutations in terms of intramolecular interactions such as salt-bridge formations. Upon this particular mutation, almost 30% of salt-bridge interactions are lost. Even explaining protein stability is a complex issue, there is a well-known fact in literature that the intramolecular interactions, such as salt-bridge formations, are crucial elements for the stability of proteins and their positions on 3D structure of protein contribute to its stability in different manners, e.g., the salt-bridge formations on protein surface contribute to enzyme stability less than 1 kcal/mole (Marti *et al*., 2003) while those buried and positioned on hydrophobic core contribute more than 4 kcal/mole (Anderson *et al*., 1990). The loss of these seven salt-bridge formations, established with β–sheet structures, considered as a part of the seven transmembrane-spanning receptor (Fig 1.1), would adversely affect the protein stability and result in its non-functionality. Except for Glu697-Lys552 interactions, these particular salt-bridge interactions are strong enough to contribute to the stability of the protein in a positive manner. As a well-known fact, entropy is a crucial element of thermodynamic of macromolecules to create a favorable environment for protein or substrate binding, happened in the signaling pathway. Upon the alterations in entropy of protein with negative manner caused by the loss of these salt-bridge interactions, the expected interaction(s) of protein would be either lacked or disrupted in FZD6 mutant and this non-functionality happens.

In summary, we diagnosed a consanguineous Turkish family with NCDC10 via the identification of a homozygous frameshift mutation, p.Gly559Aspfs*16, in *FZD6* gene. Published functional data of FZD family proteins convincingly demonstrate the importance of C-terminus on signal transduction of FZDs. The 8 bp deletion reported p.Gly559Aspfs*16 leads to loss of 133 amino acid in C-terminus of FZD6 compared to native protein. We propose that the pathogenicity of this frameshift mutation is caused by disturbing the C-terminal domain structure and hence interaction partners of FZD6.

6. CONCLUSION

The sequencing technology has become much faster and cheaper since the first human genome was sequenced in 2001 at the cost of around US$3 billion. Now, genomes can be sequenced for around US$500. As a result, WES/WGS has entered the medical practice; they will pave the way not only for precision medicine but also for diagnosis of rare disorders, where conventional techniques have failed. WES/WGS was proven to be a cost-effective method to detect disease-causing somatic or germline variants, but, it still cannot dominate the clinical field. Even though, there are many publicly available algorithms for data analysis; they tend to focus on a single aspect and do not provide an extensive workflow from start to finish. Also, there are no gold standards for translating WES/WGS into clinical knowledge; however, different diseases may require different strategies. These are the most critical factors that prevent their widespread usage in the clinical field.

With this study, we have shown that with an effective methodology including bioinformatics analysis, variant prioritization and the elucidation of pathogenicity mechanisms, patients can reach a rapid and reliable diagnosis. This study carried the genetic findings one step further and report the effect of the mutations on protein structure and hence the pathogenicity mechanisms. Our pipeline adds to the already existing knowledge through the selected threshold for population frequency, excellent combination of pathogenicity prediction tools, gene intolerance scores, and MD simulations. MD simulation is a very appropriate method that should be added to the workflow due to its reliability and ability to be implemented in a shorter time in the near future due to the developing computational power in cpu/gpu (Teo *et al*., 2014). It should be routinely performed for molecular diagnoses in addition to WES and WGS. The WES datasets used to help establish the bioinformatics methodologies was tested on undiagnosed index cases and their family members. Our novel approach achieved a high success rate by identifying the causative variant and providing the diagnosis. For two of these families, the pathogenicity mechanisms of mutations were described via MD simulations, and these findings have been submitted to two different SCI journals and passed the editorial approval.

REFERENCES

Abecasis GR, Auton A, Brooks LD, *et al*. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

Anderson DE, Becktel WJ, Dahlquist FW. pH-induced denaturation of proteins: a single salt-bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. Biochemistry. 1990;29(9):2403-8.

Annesley SJ, Bandala-sanchez E, Ahmed AU, Fisher PR. Filamin repeat segments required for photosensory signalling in Dictyostelium discoideum. BMC Cell Biol. 2007;8:48.

An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.

Bafico A, Gazit A, Pramila T, Finch PW, Yaniv A, Aaronson SA. Interaction of frizzled related protein (FRP) with Wnt ligands and the frizzled receptor suggests alternative mechanisms for FRP inhibition of Wnt signaling. J Biol Chem. 1999;274(23):16180-7.

Bamshad MJ, Ng SB, Bigham AW, *et al*. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745-55.

Baran R, de Berker D, Holzberg M, Thomas L. Baran and Dawber's Diseases of the Nails and Their Management, 4th edn. Oxford: Wiley- Blackwell; 2012.

Bardón-cancho EJ, Muñoz-jiménez L, Vázquez-lópez M, Ruíz-martín Y, García-morín M, Barredo-valderrama E. Periventricular nodular heterotopia and dystonia due to an ARFGEF2 mutation. Pediatr Neurol. 2014;51(3):461-4.

Barkovich AJ, Kuzniecky RI, Jackson GD, Guerrini R, Dobyns WB. Classification system for malformations of cortical development: update 2001. Neurology. 2001;57(12):2168-78.

Barkovich AJ, Kuzniecky RI, Jackson GD, Guerrini R, Dobyns WB. A developmental and genetic classification for malformations of cortical development. Neurology. 2005;65(12):1873-87.

Barkovich AJ, Guerrini R, Kuzniecky RI, Jackson GD, Dobyns WB. A developmental and genetic classification for malformations of cortical development: update 2012. Brain. 2012;135(Pt 5):1348-69.

Benayoun L, Spiegel R, Auslender N, *et al*. Genetic heterogeneity in two consanguineous families segregating early onset retinal degeneration: the pitfalls of homozygosity mapping. Am J Med Genet A. 2009;149A(4):650-6.

Bendl J, Stourac J, Salanda O, *et al*. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. 2014;10(1):e1003440.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

Boycott KM, Vanstone MR, Bulman DE, Mackenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet. 2013;14(10):681-91.

Brodlie M, Haq IJ, Roberts K, Elborn JS. Targeted therapies to improve CFTR function in cystic fibrosis. Genome Med. 2015;7:101.

Brooks BR, Brooks CL, Mackerell AD, *et al*. CHARMM: the biomolecular simulation program. J Comput Chem. 2009;30(10):1545-614.

Boyle AP, Davis S, Shulha HP, *et al*. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132(2):311-22.

Burrows NP, Jones RR. Yellow nail syndrome in association with carcinoma of the gall bladder. Clin Exp Dermatol. 1991;16(6):471-3.

Carroll RC, Gerrard JM. Phosphorylation of platelet actin-binding protein during platelet activation. Blood. 1982;59(3):466-71.

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3:S3.

Cassa CA, Weghorn D, Balick DJ, *et al*. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nat Genet. 2017;49(5):806-810.

Capriotti E, Altman RB, Bromberg Y. Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics. 2013;14 Suppl 3:S2.

Chahrour MH, Yu TW, Lim ET, *et al*. Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. PLoS Genet. 2012;8(4):e1002635.

Chen M, Stracher A. In situ phosphorylation of platelet actin-binding protein by cAMP-dependent protein kinase stabilizes it against proteolysis by calpain. J Biol Chem. 1989;264(24):14282-9.

Cheng WC, Chung IF, Tsai CF, *et al*. YM500v2: a small RNA sequencing (smRNA-seq) database for human cancer miRNome research. Nucleic Acids Res. 2015;43(Database issue):D862-7.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE. 2012;7(10):e46688.

Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-61.

Cong F, Schweizer L, Varmus H. Wnt signals across the plasma membrane to activate the beta-catenin pathway by forming oligomers containing its receptors, Frizzled and LRP. Development. 2004;131(20):5103-15.

Cooper GM, Stone EA, Asimenos G, *et al*,. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901-13.

Crockett DK, Ridge PG, Wilson AR, *et al*. Consensus: a framework for evaluation of uncertain gene variants in laboratory test reporting. Genome Med. 2012;4(5):48.

Cui CY, Klar J, Georgii-heming P, *et al*. Frizzled6 deficiency disrupts the differentiation process of nail development. J Invest Dermatol. 2013;133(8):1990-7.

Dann CE, Hsieh JC, Rattner A, Sharma D, Nathans J, Leahy DJ. Insights into Wnt binding and signalling from the structures of two Frizzled cysteine-rich domains. Nature. 2001;412(6842):86-90.

Darden T., Y.D., Pedersen L. Particle mesh Ewald: An N(log) method for Ewald sums in large system. Journal of Chemical Physics B. 1993;98(12): 10089-10092.

Dawn teare M, Barrett JH. Genetic linkage studies. Lancet. 2005;366(9490):1036-44.

Dobyns WB, Guerrini R, Czapansky-beilman DK, *et al*. Bilateral periventricular nodular heterotopia with mental retardation and syndactyly in boys: a new X-linked mental retardation syndrome. Neurology. 1997;49(4):1042-7.

Dong C, Wei P, Jian X, *et al*. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-37.

Doward W, Perveen R, Lloyd IC, Ridgway AE, Wilson L, Black GC. A mutation in the RIEG1 gene associated with Peters' anomaly. J Med Genet. 1999;36(2):152-5.

Drake MT, Shenoy SK, Lefkowitz RJ. Trafficking of G protein-coupled receptors. Circ Res. 2006;99(6):570-82.

Drmanac R, Sparks AB, Callow MJ, *et al*. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327(5961):78-81.

Ekşioğlu YZ, Scheffer IE, Cardenas P, *et al*. Periventricular heterotopia: an X-linked dominant epilepsy locus causing aberrant cerebral cortical development. Neuron. 1996;16(1):77-87.

Emekli U, Schneidman-duhovny D, Wolfson HJ, Nussinov R, Haliloglu T. HingeProt: automated prediction of hinges in protein structures. Proteins. 2008;70(4):1219-27.

Essman, U., Perera, L., Berkowitz, M., Darden, T., Lee, H., and Pedersen, G.G. A smooth particle mesh Ewald method. Journal of Chemical Physics. 1995;(19):8577-8593.

European Organisation for Rare Diseases. Rare Diseases: Understanding this Public Health Priority. (Eurodis, 2005).

Fallil Z, Pardoe H, Bachman R, *et al*. Phenotypic and imaging features of FLNA-negative patients with bilateral periventricular nodular heterotopia and epilepsy. Epilepsy Behav. 2015;51:321-7.

Feng Y, Walsh CA. The many faces of filamin: a versatile molecular scaffold for cell motility and signalling. Nat Cell Biol. 2004;6(11):1034-8.

Fergelot P, Coupry I, Rooryck C, *et al*. Atypical male and female presentations of FLNA-related periventricular nodular heterotopia. Eur J Med Genet. 2012;55(5):313-8.

Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?. Nat Rev Genet. 2008;9(2):102-14.

Finegold DN, Kimak MA, Lawrence EC, *et al*. Truncating mutations in FOXC2 cause multiple lymphedema syndromes. Hum Mol Genet. 2001;10(11):1185-9.

Friedländer MR, Lizano E, Houben AJ, *et al*. Evidence for the biogenesis of more than 1,000 novel human microRNAs. Genome Biol. 2014;15(4):R57.

Fox JW, Lamperti ED, Ekşioğlu YZ, *et al*. Mutations in filamin 1 prevent migration of cerebral cortical neurons in human periventricular heterotopia. Neuron. 1998;21(6):1315-25.

Fox JW, Walsh CA. Periventricular heterotopia and the genetics of neuronal migration in the cerebral cortex. Am J Hum Genet. 1999;65(1):19-24.

Frousios K, Iliopoulos CS, Schlitt T, Simpson MA. Predicting the functional consequences of non-synonymous DNA sequence variants--evaluation of bioinformatics tools and development of a consensus strategy. Genomics. 2013;102(4):223-8.

Frydman M, Weinstock AL, Cohen HA, Savir H, Varsano I. Autosomal recessive Peters anomaly, typical facial appearance, failure to thrive, hydrocephalus, and other anomalies: further delineation of the Krause-Kivlin syndrome. Am J Med Genet. 1991;40(1):34-40.

Frojmark AS, Schuster J, Sobol M *et al*. Mutations in Frizzled 6 cause isolated autosomal-recessive nail dysplasia. Am J Hum Genet 2011; 88: 852–60.

Gamsiz ED, Viscidi EW, Frederick AM, *et al*. Intellectual disability is associated with increased runs of homozygosity in simplex autism. Am J Hum Genet. 2013;93(1):103-9.

Gatica-andrades M, Vagenas D, Kling J, *et al*. WNT ligands contribute to the immune response during septic shock and amplify endotoxemia-driven inflammation in mice. Blood Adv. 2017;1(16):1274-1286.

Gattinoni L, Ji Y, Restifo NP. Wnt/beta-catenin signaling in T-cell immunity and cancer immunotherapy. Clin Cancer Res. 2010;16(19):4695-701.

Gérard-blanluet M, Sheen V, Machinis K, *et al*. Bilateral periventricular heterotopias in an X-linked dominant transmission in a family with two affected males. Am J Med Genet A. 2006;140(10):1041-6.

Ghani M, Reitz C, Cheng R, *et al*. Association of Long Runs of Homozygosity With Alzheimer Disease Among African American Individuals. JAMA Neurol. 2015;72(11):1313-23.

Gorlin JB, Yamin R, Egan S, *et al*. Human endothelial actin-binding protein (ABP-280, nonmuscle filamin): a molecular leaf spring. J Cell Biol. 1990;111(3):1089-105.

Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974;185(4154):862-4.

González-pérez A, López-bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88(4):440-9.

Gormez Z, Bakir-gungor B, Sagiroglu MS. HomSI: a homozygous stretch identifier from next-generation sequencing data. Bioinformatics. 2014;30(3):445-7.

Guo N, Hawkins C, Nathans J. Frizzled6 controls hair patterning in mice. Proc Natl Acad Sci USA. 2004;101(25):9277-81.

Gupta S, Samra D, Yel L, Agrawal S. T and B cell deficiency associated with yellow nail syndrome. Scand J Immunol. 2012;75(3):329-35.

Guerrini R, Barba C. Malformations of cortical development and aberrant cortical networks: epileptogenesis and functional organization. J Clin Neurophysiol. 2010;27(6):372-9.

Guerrini R, Mei D, Sisodiya S, *et al*. Germline and mosaic mutations of FLN1 in men with periventricular heterotopia. Neurology. 2004;63(1):51-6.

Hanson IM, Fletcher JM, Jordan T, *et al*. Mutations at the PAX6 locus are found in heterogeneous anterior segment malformations including Peters' anomaly. Nat Genet. 1994;6(2):168-73.

Hastie, T., Tibshirani, R., and Friedman, J.H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Manhattan, New York: Springer; 2009.

Hehr U, Hehr A, Uyanik G, Phelan E, Winkler J, Reardon W. A filamin A splice mutation resulting in a syndrome of facial dysmorphism, periventricular nodular heterotopia, and severe constipation reminiscent of cerebro-fronto-facial syndrome. J Med Genet. 2006;43(6):541-4.

Hidalgo-bravo A, Pompa-mera EN, Kofman-alfaro S, Gonzalez-bonilla CR, Zenteno JC. A novel filamin A D203Y mutation in a female patient with otopalatodigital type 1 syndrome and extremely skewed X chromosome inactivation. Am J Med Genet A. 2005;136(2):190-3.

Hildebrandt F, Heeringa SF, Rüschendorf F, *et al*. A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS Genet. 2009;5(1):e1000353.

Hock RS, Davis G, Speicher DW. Purification of human smooth muscle filamin and characterization of structural domains and functional sites. Biochemistry. 1990;29(40):9441-51.

Hodgkinson A, Casals F, Idaghdour Y, Grenier JC, Hernandez RD, Awadalla P. Selective constraint, background selection, and mutation accumulation variability within and between human populations. BMC Genomics. 2013;14:495.

Hoque SR, Mansour S, Mortimer PS. Yellow nail syndrome: not a genetic disorder? Eleven new cases and a review of the literature. Br J Dermatol. 2007;156(6):1230-4.

Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph. 1996;14(1):33-8, 27-8.

Huttenlocher PR, Taravath S, Mojtahedi S. Periventricular heterotopia and epilepsy. Neurology. 1994;44(1):51-5.

Ichikawa Y, Shimizu H, Arimori S. "Yellow nail syndrome" and rheumatoid arthritis. Tokai J Exp Clin Med. 1991;16(5-6):203-9.

Ichinohe M, Nakamura Y, Sai K, Nakahara M, Yamaguchi H, Fukami K. Lack of phospholipase C-delta1 induces skin inflammation. Biochem Biophys Res Commun. 2007;356(4):912-8.

Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437(7055):69-87.

Jamieson RV, Perveen R, Kerr B, *et al*. Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. Hum Mol Genet. 2002;11(1):33-42.

Jardine PE, Clarke MA, Super M. Familial bilateral periventricular nodular heterotopia mimics tuberous sclerosis. Arch Dis Child. 1996;74(3):244-6.

Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497-502.

Jorgensen WL, C.J., Madura JD, Impey RW, Klein ML, Comparison of Simple Potential Functions for Simulating Liquid Water. J Chem Phys. 1983;79(2):926-76.

Joshi PK, Esko T, Mattsson H, *et al*. Directional dominance on stature and cognition in diverse human populations. Nature. 2015;523(7561):459-462.

Kamatani M, Rai A, Hen H, *et al*. Yellow nail syndrome associated with mental retardation in two siblings. Br J Dermatol. 1978;99(3):329-33.

Kamuro K, Tenokuchi Y. Familial periventricular nodular heterotopia. Brain Dev. 1993;15(3):237-41.

Karczewski KJ, Weisburd B, Thomas B, *et al*. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45(D1):D840-D845.

Kasper BS, Kurzbuch K, Chang BS, *et al*. Paternal inheritance of classic X-linked bilateral periventricular nodular heterotopia. Am J Med Genet A. 2013;161A(6):1323-8.

Kasparis C, Reid D, Wilson NJ, *et al*. Isolated recessive nail dysplasia caused by FZD6 mutations: report of three families and review of the literature. Clin Exp Dermatol. 2016;41(8):884-889.

Katsonis P, Lichtarge O. A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. Genome Res. 2014;24(12):2050-8.

Keller MC, Simonson MA, Ripke S, *et al*. Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. PLoS Genet. 2012;8(4):e1002656.

Khaire N, Müller R, Blau-wasser R, *et al*. Filamin-regulated F-actin assembly is essential for morphogenesis and controls phototaxis in Dictyostelium. J Biol Chem. 2007;282(3):1948-55.

Khan S, Basit S, Habib R, Kamal A, Muhammad N, Ahmad W. Genetics of human isolated hereditary nail disorders. Br J Dermatol. 2015;173(4):922-9.

Kilander MB, Dahlström J, Schulte G. Assessment of Frizzled 6 membrane mobility by FRAP supports G protein coupling and reveals WNT-Frizzled selectivity. Cell Signal. 2014;26(9):1943-9.

Kimchi-sarfaty C, Oh JM, Kim IW, *et al*. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science. 2007;315(5811):525-8.

Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-5.

Kovacevic I, Cram EJ. FLN-1/filamin is required for maintenance of actin and exit of fertilized oocytes from the spermatheca in C. elegans. Dev Biol. 2010;347(2):247-57.

Krieger E, Koraimann G, Vriend G. Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. Proteins. 2002;47(3):393-402.

Kuehl FA, Egan RW. Prostaglandins, arachidonic acid, and inflammation. Science. 1980;210(4473):978-84.

Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinformatics. 2013;14(2):144-61.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.

Lambert EM, Dziura J, Kauls L, Mercurio M, Antaya RJ. Yellow nail syndrome in three siblings: a randomized double-blind trial of topical vitamin E. Pediatr Dermatol. 2006;23(4):390-5.

Lange M, Kasper B, Bohring A, *et al*. 47 patients with FLNA associated periventricular nodular heterotopia. Orphanet J Rare Dis. 2015;10:134.

Law M, Shaw DR. Mouse Genome Informatics (MGI) Is the International Resource for Information on the Laboratory Mouse. Methods Mol Biol. 2018;1757:141-161.

Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843-54.

Lencz T, Lambert C, Derosse P, *et al*. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci USA. 2007;104(50):19942-7.

Lek M, Karczewski KJ, Minikel EV, *et al*. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.

Lemma V, D'agostino M, Caporaso MG, *et al*. A disorder-to-order structural transition in the COOH-tail of Fz4 determines misfolding of the L501fsX533-Fz4 mutant. Sci Rep. 2013;3:2659.

Li MX, Kwan JS, Bao SY, *et al*. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genet. 2013;9(1):e1003143.

Li B, Krishnan VG, Mort ME, *et al*. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744-50.

Lin PI, Kuo PH, Chen CH, *et al*. Runs of homozygosity associated with speech delay in autism in a taiwanese han population: evidence for the recessive model. PLoS ONE. 2013;8(8):e72056.

Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011;32(8):894-9.

Liu W, Yan B, An D, Xiao J, Hu F, Zhou D. Sporadic periventricular nodular heterotopia: Classification, phenotype and correlation with Filamin A mutations. Epilepsy Res. 2017;133:33-40.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res. 2012;40(7):e53.

Londin E, Loher P, Telonis AG, *et al*. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. Proc Natl Acad Sci USA. 2015;112(10):E1106-15.

Lopes MC, Joyce C, Ritchie GR, *et al*. A combined functional annotation score for non-synonymous variants. Hum Hered. 2012;73(1):47-51.

Mackerell AD, Bashford D, Bellott M, *et al*. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B. 1998;102(18):3586-616.

Mackerell AD, Feig M, Brooks CL. Improved treatment of the protein backbone in empirical force fields. J Am Chem Soc. 2004;126(3):698-9.

Magi A, Tattini L, Palombo F, *et al*. H3M2: detection of runs of homozygosity from whole-exome sequencing data. Bioinformatics. 2014;30(20):2852-9.

Malbon CC, Wang HY. Dishevelled: a mobile scaffold catalyzing development. Curr Top Dev Biol. 2006;72:153-66.

Maldonado F, Tazelaar HD, Wang CW, Ryu JH. Yellow nail syndrome: analysis of 41 consecutive patients. Chest. 2008;134(2):375-381.

Mallick S, Li H, Lipson M, *et al*. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201-206.

Marti DN, Bosshard HR. Electrostatic interactions in leucine zippers: thermodynamic analysis of the contributions of Glu and His residues and the effect of mutating salt-bridges. J Mol Biol. 2003;330(3):621-37.

Mckenna A, Hanna M, Banks E, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.

Mclaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-70.

Mercurio S, Latinkic B, Itasaki N, Krumlauf R, Smith JC. Connective-tissue growth factor modulates WNT signalling and interacts with the WNT receptor complex. Development. 2004;131(9):2137-47.

Meyer LR, Zweig AS, Hinrichs AS, *et al*. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 2013;41(Database issue):D64-9.

Mohammadi-asl J, Pourreza MR, Mohammadi A, Eskandari A, Mozafar-jalali S, Tabatabaiefar MA. A novel pathogenic variant in the FZD6 gene causes recessive nail dysplasia in a large Iranian kindred. J Dermatol Sci. 2017;88(1):134-138.

Moro F, Carrozzo R, Veggiotti P, *et al*. Familial periventricular heterotopia: missense and distal truncating mutations of the FLN1 gene. Neurology. 2002;58(6):916-21.

Morton NE. Sequential tests for the detection of linkage. Am J Hum Genet. 1955;7(3):277-318.

Naz G, Pasternack SM, Perrin C, *et al*. FZD6 encoding the Wnt receptor frizzled 6 is mutated in autosomal-recessive nail dysplasia. Br J Dermatol. 2012;166(5):1088-94.

Nackley AG, Shabalina SA, Tchivileva IE, *et al*. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science. 2006;314(5807):1930-3.

Nalls MA, Guerreiro RJ, Simon-sanchez J, *et al*. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. Neurogenetics. 2009;10(3):183-90.

Nam JS, Turcotte TJ, Smith PF, Choi S, Yoon JK. Mouse cristin/R-spondin family proteins are novel ligands for the Frizzled 8 and LRP6 receptors and activate beta-catenin-dependent gene expression. J Biol Chem. 2006;281(19):13247-57.

Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics. 2016;32(11):1749-51.

Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4.

Ng SB, Turner EH, Robertson PD, *et al*. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461(7261):272-6.

Ng SB, Buckingham KJ, Lee C, *et al*. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010;42(1):30-5.

Nie J, Lewis DL. The proximal and distal C-terminal tail domains of the CB1 cannabinoid receptor mediate G protein coupling. Neuroscience. 2001;107(1):161-7.

Niroula A, Urolagin S, Vihinen M. PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS ONE. 2015;10(2):e0117380.

Nishimura DY, Searby CC, Alward WL, *et al*. A spectrum of FOXC1 mutations suggests gene dosage as a mechanism for developmental defects of the anterior chamber of the eye. Am J Hum Genet. 2001;68(2):364-72.

Noegel AA, Leiting B, Witke W, *et al*. Biological roles of actin-binding proteins in Dictyostelium discoideum examined using genetic techniques. Cell Motil Cytoskeleton. 1989;14(1):69-74.

Nusse R. Wnts and Hedgehogs: lipid-modified proteins and similarities in signaling mechanisms at the cell surface. Development. 2003;130(22):5297-305.

Oda H, Sato T, Kunishima S, *et al*. Exon skipping causes atypical phenotypes associated with a loss-of-function mutation in FLNA by restoring its protein function. Eur J Hum Genet. 2016;24(3):408-14.

Oegema R, Hulst JM, Theuns-valks SD, *et al*. Novel no-stop FLNA mutation causes multi-organ involvement in males. Am J Med Genet A. 2013;161A(9):2376-84.

Ogutu JO, Piepho HP, Schulz-streeck T. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. 2011;5 Suppl 3:S11.

Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M. PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat. 2012;33(8):1166-74.

Orloff MS, Zhang L, Bebek G, Eng C. Integrative genomic analysis reveals extended germline homozygosity with lung cancer risk in the PLCO cohort. PLoS ONE. 2012;7(2):e31975.

Ott J. Analysis of human genetic linkage, Vol. Vol. xxiii. Baltimore: Johns Hopkins University Press; 1999.

Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761-3.

Pang SM. Yellow nail syndrome resolution following treatment of pulmonary tuberculosis. Int J Dermatol. 1993;32(8):605-6.

Parrini E, Ramazzotti A, Dobyns WB, *et al*. Periventricular heterotopia: phenotypic heterogeneity and correlation with Filamin A mutations. Brain. 2006;129(Pt 7):1892-906.

Paten B, Herrero J, Fitzgerald S, *et al*. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 2008;18(11):1829-43.

Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008;18(11):1814-28.

Patrosso MC, Repetto M, Villa A, *et al*. The exon-intron organization of the human X-linked gene (FLN1) encoding actin-binding protein 280. Genomics. 1994;21(1):71-6.

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013;9(8):e1003709.

Phillips JC, Braun R, Wang W, *et al*. Scalable molecular dynamics with NAMD. J Comput Chem. 2005;26(16):1781-802.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110-21.

Popowicz GM, Schleicher M, Noegel AA, Holak TA. Filamins: promiscuous organizers of the cytoskeleton. Trends Biochem Sci. 2006;31(7):411-9.

Poussaint TY, Fox JW, Dobyns WB, *et al*. Periventricular nodular heterotopia in patients with filamin-1 gene mutations: neuroimaging findings. Pediatr Radiol. 2000;30(11):748-55.

Povelones M, Nusse R. The role of the cysteine-rich domain of Frizzled in Wingless-Armadillo signaling. EMBO J. 2005;24(19):3493-503.

Purcell S, Neale B, Todd-brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.

Punchihewa C, Ferreira AM, Cassell R, Rodrigues P, Fujii N. Sequence requirement and subtype specificity in the high-affinity interaction between human frizzled and dishevelled proteins. Protein Sci. 2009;18(5):994-1002.

Ramsey BW, Davies J, Mcelvaney NG, *et al*. A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. N Engl J Med. 2011;365(18):1663-72.

Raza SI, Muhammad N, Khan S, Ahmad W. A novel missense mutation in the gene FZD6 underlies autosomal recessive nail dysplasia. Br J Dermatol 2013; 168:422–5.

Razi E. Familial yellow nail syndrome. Dermatol Online J. 2006;12(2):15.

Rehm HL, Bale SJ, Bayrak-toydemir P, *et al*. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013;15(9):733-47.

Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.

Richards S, Aziz N, Bale S, *et al*. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.

Robertson SP. Filamin A: phenotypic diversity. Curr Opin Genet Dev. 2005;15(3):301-7.

Rodriguez J, Esteve P, Weinl C, *et al*. SFRP1 regulates the growth of retinal ganglion cell axons through the Fz2 receptor. Nat Neurosci. 2005;8(10):1301-9.

Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010;5(4):725-38.

Samman PD, White WF. The "Yellow Nail" Syndrome. Br J Dermatol. 1964;76:153-7.

Samocha KE, Robinson EB, Sanders SJ, *et al*. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46(9):944-50.

Scherer C, Schuele S, Minotti L, Chabardes S, Hoffmann D, Kahane P. Intrinsic epileptogenicity of an isolated periventricular nodular heterotopia. Neurology. 2005;65(3):495-6.

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575-6.

Schulte G, Bryja V. The Frizzled family of unconventional G-protein-coupled receptors. Trends Pharmacol Sci. 2007;28(10):518-25.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005;33(Web Server issue):W382-8.

Seelow D, Schuelke M, Hildebrandt F, Nürnberg P. HomozygosityMapper--an interactive approach to homozygosity mapping. Nucleic Acids Res. 2009;37(Web Server issue):W593-9.

Sethi R, Seppälä J, Tossavainen H, *et al*. A novel structural unit in the N-terminal region of filamins. J Biol Chem. 2014;289(12):8588-98.

Shihab HA, Gough J, Cooper DN, *et al*. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65.

Sheen VL, Dixon PH, Fox JW, *et al*. Mutations in the X-linked filamin 1 gene cause periventricular nodular heterotopia in males as well as in females. Hum Mol Genet. 2001;10(17):1775-83.

Sheen VL, Jansen A, Chen MH, *et al*. Filamin A mutations cause periventricular heterotopia with Ehlers-Danlos syndrome. Neurology. 2005;64(2):254-62.

Sheikh TI, Mittal K, Willis MJ, Vincent JB. A synonymous change, p.Gly16Gly in MECP2 Exon 1, causes a cryptic splice event in a Rett syndrome patient. Orphanet J Rare Dis. 2013;8:108.

Siepel A, Bejerano G, Pedersen JS, *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-50.

Sievers F, Higgins DG. Clustal omega. Curr Protoc Bioinformatics. 2014;48:3.13.1-16.

Simpson MA, Irving MD, Asilmaz E, *et al*. Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. Nat Genet. 2011;43(4):303-5.

Smallwood PM, Williams J, Xu Q, Leahy DJ, Nathans J. Mutational analysis of Norrin-Frizzled4 recognition. J Biol Chem. 2007;282(6):4057-68.

Stefanova M, Meinecke P, Gal A, Bolz H. A novel 9 bp deletion in the filamin a gene causes an otopalatodigital-spectrum disorder with a variable, intermediate phenotype. Am J Med Genet A. 2005;132A(4):386-90.

Stossel TP, Condeelis J, Cooley L, *et al*. Filamins as integrators of cell mechanics and signalling. Nat Rev Mol Cell Biol. 2001;2(2):138-45.

Sundaram L, Gao H, Padigepati SR, *et al*. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161-1170.

Sun Y, Almomani R, Aten E, *et al*. Terminal osseous dysplasia is caused by a single recurrent mutation in the FLNA gene. Am J Hum Genet. 2010;87(1):146-53.

Swafford D, Manicassamy S. Wnt signaling in dendritic cells: its role in regulation of immunity and tolerance. Discov Med. 2015;19(105):303-10.

Tauriello DV, Jordens I, Kirchner K, *et al*. Wnt/β-catenin signaling requires interaction of the Dishevelled DEP domain and C-terminus with a discontinuous motif in Frizzled. Proc Natl Acad Sci USA. 2012;109(14):E812-20.

Taylor SM, Laegreid WW, Englen MD, *et al*. Influence of extracellular calcium on the metabolism of arachidonic acid in alveolar macrophages. J Leukoc Biol. 1990;48(6):502-11.

Taylor JC, Martin HC, Lise S, *et al*. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47(7):717-726.

Tennessen JA, Bigham AW, O'connor TD, *et al*. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64-9.

The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. Am J Epidemiol. 1989;129(4):687-702.

Thomsen H, Filho MI, Woltmann A, *et al*. Inbreeding and homozygosity in breast cancer survival. Sci Rep. 2015;5:16467.

Thomsen H, Chen B, Figlioli G, *et al*. Runs of homozygosity and inbreeding in thyroid cancer. BMC Cancer. 2016;16:227.

Unger S, Mainberger A, Spitz C, *et al*. Filamin A mutation is one cause of FG syndrome. Am J Med Genet A. 2007;143A(16):1876-9.

Valleix S, Niel F, Nedelec B, *et al*. Homozygous nonsense mutation in the FOXE3 gene as a cause of congenital primary aphakia in humans. Am J Hum Genet. 2006;79(2):358-64.

Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. Science. 2007;318(5858):1931-4.

Vincent A, Billingsley G, Priston M, *et al*. Phenotypic heterogeneity of CYP1B1: mutations in a patient with Peters' anomaly. J Med Genet. 2001;38(5):324-6.

Wang Y, Guo N, Nathans J. The role of Frizzled3 and Frizzled6 in neural tube closure and in the planar polarity of inner-ear sensory hair cells. J Neurosci. 2006;26(8):2147-56.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

Wallingford JB, Habas R. The developmental biology of Dishevelled: an enigmatic protein governing cell fate and cell polarity. Development. 2005;132(20):4421-36.

Whiffin N, Minikel E, Walsh R, *et al*. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med. 2017;19(10):1151-1158.

Wilson NJ, Hansen CD, Azkur D *et al*. Recessive mutations in the gene encoding frizzled 6 cause twenty nail dystrophy – expanding the differential diagnosis for pachyonychia congenita. J Dermatol Sci 2013; 70:58–60.

Withers SJ, Gole GA, Summers KM. Autosomal dominant cataracts and Peters anomaly in a large Australian family. Clin Genet. 1999;55(4):240-7.

Wong HC, Bourdelas A, Krauss A, *et al*. Direct binding of the PDZ domain of Dishevelled to a conserved internal sequence in the C-terminal region of Frizzled. Mol Cell. 2003;12(5):1251-60.

Yamamoto A, Nagano T, Takehara S, Hibi M, Aizawa S. Shisa promotes head formation through the inhibition of receptor protein maturation for the caudalizing factors, Wnt and FGF. Cell. 2005;120(2):223-35.

Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Res. 2015;43(W1):W174-81.

Yang TL, Guo Y, Zhang JG, Xu C, Tian Q, Deng HW. Genome-wide Survey of Runs of Homozygosity Identifies Recessive Loci for Bone Mineral Density in Caucasian and Chinese Populations. J Bone Miner Res. 2015;30(11):2119-26.

Yang-snyder J, Miller JR, Brown JD, Lai CJ, Moon RT. A frizzled homolog functions in a vertebrate Wnt signaling pathway. Curr Biol. 1996;6(10):1302-6.

Zaidman GW, Flanagan JK, Furey CC. Long-term visual prognosis in children after corneal transplant surgery for Peters anomaly type I. Am J Ophthalmol. 2007;144(1):104-108.

Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9:40.