REPUBLIC OF TURKEY

ACIBADEM MEHMET ALİ AYDINLAR UNIVERSITY

INSTITUTE OF HEALTH SCIENCES

INVESTIGATION OF RARE DISEASE-CAUSING VARIANTS USING A PRIORITIZATION STRATEGY

TUĞÇE BOZKURT

MASTER THESIS

DEPARTMENT of BIOSTATISTICS and BIOINFORMATICS

SUPERVISOR

Prof. Dr. Uğur Sezerman

ISTANBUL-2020

DECLARATION

I declare that this thesis work is my own work, I had no unethical behavior at any stages from the planning to the writing of the thesis, I obtained all the information in this thesis in accordance with academic and ethical rules, I cited all the information and comments that were not obtained with this thesis work, and I provided resources in the list of references. I also declare that there was no violation of any patents and copyrights during the study and writing of this thesis.

22.07.2020

Tuğçe Bozkurt

TABLE OF CONTENTS

LIST OF ABBREVATIONS AND SYMBOLS

1KGP	1000 Genome Project
3D	Three-dimensional
ACMG	American College of Medical Genetics and Genomics
BAM	Binary alignment map
BQSR	Base Quality Score Recalibration
BWA	Burrows-Wheeler Aligner
CADD	Combined annotation– dependent depletion
COE	Collier/Olf/Ebf
CSF	Cerebrospinal fluid
DBD	DNA-Binding Domain
EBF	Early B-cell Factor
EEG	Electroencephalography
ESHG	European Society of Human Genetics
ExAC	Exome Aggregation Consortium
GATK	Genome Analysis Toolkit
GLUT1DS1	GLUT1-deficiency syndrome 1
gnomAD	The Genome Aggregation Database

GXD	Mouse Gene Expression Database
HADDS	Hypotonia, Ataxia, And Delayed Development Syndrome
HGNC	Human Gene Nomenclature Committee
HLH	Helix-loop-helix
HPO	Human Phenotype Ontology
HTML	HyperText Markup Language
IGV	Integrative Genomics Viewer
IPT	Ig-like/plexins/transcription Factors
LP	Lumbar puncture
M-CAP	The Mendelian clinically applicable pathogenicity
MAF	Minor Allele Frequency
MD	Molecular dynamic
MGD	Mouse Genome Database
MGI	Mouse Genome Informatics
MRI	Magnetic Resonance Imaging
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
NHLBI	The National Heart, Lung, and Blood Institute
ns	Nanosecond

o/e	Observed/expected
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase Chain Reaction
PKA	Protein Kinase A
pLI	Probability of being Loss-of-function Intolerant
PRKAR1A	Protein Kinase CAMP-dependent, Regulatory Type I Alpha
PSI-BLAST	Position-Specific Iterated BLAST
RD	Rare disease
REVEL	Rare exome variant ensemble learner
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuations
RVIS	Residual Variation Intolerance Score
SAM	Sequence alignment map
SLC2A1	Solute carrier family 2 facilitated glucose transporter member 1
SNV	Single Nucleotide Variation
TAD	Transactivation Domain
TCL	Tool Command Language
UCSC	The University of California at Santa Cruz
UTR	Untranslated Region

VCF	Variant Call Format
VMD	Visual Molecular Dynamics
VUS	Variants of uncertain significance
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
WHO	World Health Organization

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I would like to express the deepest gratitude to my thesis advisor Prof. Dr. Ugur Sezerman, for enabling me to work on a topic that I want. He is not only a supervisor that makes this study possible but also the ones that treated me constructive and supportive in every sense. I will always be proud to be a student of him.

I express sincere thanks to my jury members, Prof. Dr. Yasemin Alanay and Asst. Prof. Dr. Burcu Bakir-Gungor, for their valuable evaluation. I also offer my special thanks to my substitute jury members, Prof. Dr. Cengiz Yakicier and Prof. Dr. Sibel Aylin Ugur-Iseri for their precious review.

I would like to thank intimately for Doc. Dr. Emel Timucin for encouraging me when the time was difficult to decide and sharing her invaluable life experiences with me. I will always admire her personality.

I would like to offer my special thanks to Dr. Ceren Saygi for teaching the general concepts and answering my questions regardless of the kilometers between us.

I have friends in Sezerman Lab to thank for supporting me in every minute of this journey. Thank you Okan for helping me in Linux scripting and correcting my simple errors even in the middle of the night. Thank you Berk E. for answering my countless questions about any computational stuff without complaint. Thank you Umut for his collaboration to the first case I analyzed. Thank you Ege for helping me in the implementation of the workflow and replying me whenever I stuck. Thank you Narod for being my sister and making easier my life in the hard times. Thank you Baris for all the things he shared with me especially at the time of Corona. Thank you Eray for being there when things were not going as well as I expected. Thank you Fatma for

being my sister and being there whenever I need. Thank you Nogay for being a witty friend and helping me to transfer the raw data. Thank you Berk G. for keeping the ACU solidarity and contributing to my R scripts. Thank you Tayyip for being a gentle friend and providing peanut support. Thank you Melis for giving advising to me about scientific and non-scientific topics. Thank you Hazal, Sila, Ruchan, Baran, Gokhan, Ece, Orhan, Burcin, Metin for motivating me all the time.

I convey my sincere thanks to Cagla for being my heart sister and always standing by me since the day we met in undergraduate years.

Finally, I am grateful to my parents for encouraging and supporting me all the time.

SUMMARY

Rare diseases are a group of diseases that affect approximately 400 million people globally, although they are individually rare in a population. The low prevalence, heterogeneity between phenotype and genotype relationship, and incomplete knowledge about them make the precise diagnosis difficult. In the post-genomic era, NGS-based approaches have become an accurate way for diagnosis. WES is one of these approaches based on the enrichment of protein-coding regions of human genes. WES enables the generation of a vast amount of genomic variation data. Identifying the potential effects of these variations and their association with over 6,000 unique phenotypes is the main challenge in RDs. Several bioinformatics tools are developed to discover the causative mutations among hundreds of variations. However, these tools generally center upon a particular aspect of the entire discovery and offer no gold standard. This thesis aimed to build a comprehensive WES workflow, including a data analysis pipeline and variant filtration strategy, for the prioritization of variants underlying RDs. The clinical impact of the workflow was tested on three previously unsolved WES cases. Two proband-based and one trio-based cases demonstrated that the workflow could be successfully applied to both approaches for diagnostics. The workflow revealed the novel variants in the SLC2A1 and PRKAR1A in proband-based cases diagnosed with GLUT1DS1 and Acrodysostosis-1, respectively. In the trio case diagnosed with HADDS, the disease-causing variant in the EBF3 was detected. In conclusion, the workflow has the potential to accelerate the diagnostic odyssey of patients with RDs and to contribute to the literature.

Keywords: Rare Diseases, Variant Prioritization, Whole Exome Sequencing

ÖZET

Önceliklendirme Stratejisi Kullanarak Nadir Hastalığa Neden Olan Varyantların Araştırılması

Nadir hastalıklar, bir popülasyonda bireysel olarak nadir olmalarına rağmen, dünya genelinde yaklaşık 400 milyon insanı etkileyen bir hastalık grubudur. Düşük prevalans, fenotip ile genotip ilişkisi arasındaki heterojenite ve bunlar hakkındaki eksik bilgi, kesin tanıyı zorlaştırmaktadır. Post-genomik dönemde, NGS temelli yaklaşımlar teşhis için kesin bir yol haline gelmiştir. WES, insan genlerinin protein kodlayan bölgelerinin zenginleştirilmesine dayanan bu yaklaşımlardan biridir. WES çok miktarda genomik varyasyon verisi oluşturulmasını sağlar. Bu varyasyonların potansiyel etkilerini ve bunların 6.000'den fazla benzersiz fenotip ile ilişkisinin belirlenmesi, nadir hastalıkların ana zorluktur. Yüzlerce varyasyon arasındaki nedensel mutasyonları keşfetmek için çeşitli biyoinformatik araçlar geliştirilmiştir. Bununla birlikte, bu araçlar genellikle tüm keşfin belirli bir yönüne odaklanır ve altın standardı sunmaz. Bu tez projesi, nadir Mendel hastalıklarının altında yatan varyantların önceliklendirilmesi için bir veri analizi yöntemi ve varyant filtrasyon stratejisini içeren kapsamlı bir WES iş akışı oluşturulmasını amaçlamıştır. İş akışının klinik etkisi, önceden çözülmemiş üç WES vakası üzerinde test edilmiştir. İki proband tabanlı ve bir trio tabanlı vaka, iş akışının teşhis için her iki yaklaşıma da başarıyla uygulanabileceğini göstermiştir. İş akışı, sırasıyla GLUT1DS1 ve Acrodysostosis-1 teşhisi konulan proband bazlı vakalarda SLC2A1 ve PRKAR1A'daki yeni varyantları ortaya çıkardı. HADDS tanısı konulan trio vakada, EBF3'te hastalığa neden olan varyant tespit edildi. Sonuç olarak, iş akışı, nadir hastalıkları olan hastaların tanısal yolculuklarını hızlandırma ve literatüre katkıda bulunma potansiyeline sahiptir.

Anahtar kelimeler: Nadir Hastalıklar, Tüm Ekzom Dizileme, Varyant Önceliklendirme

1. BACKGROUND AND AIM OF THE STUDY

Rare diseases (RDs) are health conditions that affect a small proportion of individuals in a population (1). Although the frequency definition differs from population to population, RDs affect 400 million people worldwide (2,3). The prevalence statistics about RDs suggest that despite individually rare in a particular community, they are collectively common health issues with 6,000 - 8,000 unique phenotypes (4,5).

It is still challenging to precisely identify these rare phenotypes whose majority caused by a genetic origin (5). The challenges mainly arise from the heterogenic nature between the genotype and phenotype correlation and incomplete knowledge about them. These difficulties may result in taking a precise diagnosis after years from the onset of symptoms and seeing many physicians during this time for patients (6). Furthermore, the delay in diagnosis also affects the search for treatment options. Despite the limited number of orphan drugs, long-term complications can be alleviated or reprieved by targeted therapies for some RDs if diagnosed early (7). Therefore, identifying of the genetic cause of RDs is vital for not only an accurate diagnosis but also the treatment opportunity.

In recent years, NGS technology has become an unbiased and accurate way for the comprehensive characterization of RDs (8). NGS is performed by commercially available platforms based on massively parallel sequencing of DNA molecules (9). WES is one of the NGS approaches based on the enrichment of protein-coding regions of known human genes (10). Since most of the variants underlying rare Mendelian disorders impair protein-coding regions, WES is considered as a powerful approach for identifying them (11). After the initial reports about the usage of WES in the clinical area, the method has increasingly become common for RD diagnostics (12–

14). This advancement in the clinical utility of WES is mostly driven by bioinformatics analysis approaches (15).

WES is performed as a series of biochemical and computational procedures (16). As a result of the procedures, a vast amount of data is generated (17). However, analysis and interpretation of the data are complex processes and require highly specific computational power and expertise. A variety of open-source tools and software exist to enable the analysis of WES data. These tools and software generally center upon a particular aspect of the entire process, and they are implemented with different analysis pipeline by each laboratory group.

Despite the advancement in the analysis tools and software, WES has a 25-30% diagnostic yield (14,18,19). It is reported that the causative variants which belong to some patients in the undiagnosed percentage may not be identified in the first analysis, although they already in the sequenced exome (20–22). There are various reasons proposed that cause to remain a pathogenic variant to be unidentified in the initial exome. These reasons may arise from the processes of bioinformatic analysis workflow and phenotype examination such that when the phenotype is reported in detail, this increases the probability that a pathogenic mutation would be prioritized in the bioinformatic analysis (23). Recent studies reported that undiagnosed patients could get a precise diagnosis by reanalysis of the same WES data, using different workflows (20,23,22).

Considering the diversity in available tools and software and the yield of re-analyzes, building a unique analytical workflow that improves the clinical utility of WES cases is essential. In this dissertation, it is aimed to develop workflow and to test it on unsolved cases. The workflow introduced in this dissertation can efficiently integrate WES data and improve the diagnostic yield of WES. The results revealed from the implementation of the workflow will also contribute to RD literature.

2. INTRODUCTION

2.1. Rare Diseases

RD is generally defined as a health condition that affects a small proportion of individuals in a population. While in the European population, when a disease affects fewer than 1 in 2,000 people (24), it is described as rare; in the United States, it is considered rare if it affects fewer than 200,000 people (25). The reason why these diseases remain rare between individuals is that they usually negatively affect reproductive fitness (26). But, according to the World Health Organization (WHO) report, it is estimated that nearly 400 million people suffer from RDs on the global scale (3). These statistics suggest that despite individually rare in a particular population, RDs are collectively common health issues with between 6,000 - 8,000 unique phenotypes cataloged to date (4,5). The majority of these phenotypes, over 70%, are caused by a genetic origin (5). However, identifying an RD is still challenging due to the genetic and phenotypic heterogeneity in these diseases and incomplete knowledge. A precise diagnosis takes nearly 4,8 years from the onset of symptoms, and the patients see about 7,3 physicians until that time (6).

Moreover, these challenges in diagnosis lead to delay in searching for treatment options. Although there are a limited number of orphan drugs, which are the compounds used for treating rare diseases, dietary therapy or enzyme replacement therapy can significantly improve the quality of life for some patients with rare metabolic disorders (27). For instance, Solute carrier family 2 facilitated glucose transporter member 1 (SLC2A1) encoding a glucose transporter across the blood-brain barrier is the causative gene for GLUT1-deficiency syndrome 1 (GLUT1DS1) (28) characterized by developmental delay, infantile seizures, acquired microcephaly, spasticity and ataxia (29). The ketogenic diet supplies an

alternative fuel for the brain instead of glucose and results in a marked recovery in symptoms (30). Therefore, precise diagnostics approaches open the door identifying the disease and finding therapy options.

Next-generation sequencing (NGS) based technologies are precise and unbiased ways to detect disease-causing variation (8). Whole-exome sequencing (WES) is one of the sequencing methods that supposed to cover the protein-coding regions in the genome (10). These regions constitute only approximately 1–2% of the entire human genome but, 85% of all DNA variations that have an impact on human disease are located in exomes (31,32). Additionally, most alleles that are known to cause rare Mendelian disorders disrupt exonic parts of the genome; therefore, WES is considered as a feasible approach for identifying variants underlying these disorders. (11). The initial usage of WES in RD research introduced with the identification of genes caused Freeman–Sheldon syndrome (12), Miller syndrome (13), and Schinzel–Giedion syndrome (33). After these initial reports, WES has increasingly become available on both research and clinical areas, as NGS costs have fallen (34).

2.2. Fundamentals of WES

In 1977, an article describing a new method for DNA sequencing was reported by Allan Maxam and Walter Gilbert. In this study, it was shown that the terminally labeled bases were treated with base-specific chemical cleavage, and the products of DNA molecules were separated in the electrophoresis system (35). As an alternative to this method, an approach including the usage of chain-terminating dideoxynucleotide analogs that cause to finish the DNA synthesis was described by Sanger et al (36). The Sanger method has been the commonly used approach for DNA sequencing throughout nearly 30 years (9). In 2005, sequencing technology was evolved by releasing the first massively parallel pyrosequencing

platform, which was the new era of high-throughput genomic analysis (37). While the automated Sanger method is recognized as a “first-generation”, newer high-throughput approaches are known as “next-generation sequencing” (38).

Commercially available platforms perform NGS based on massively parallel sequencing of clonally amplified or single DNA molecules that are separated in a flow cell (9). Although several companies develop platforms differing in sequencing chemistries and technical details, all NGS platforms utilize a similar strategy (39). First of all, the sample nucleic acids are broken into smaller pieces by several enzymatic or mechanical methods. Companies develop their bases, namely adapters, and use them in the samples' sequencing by ligating them to the small sample fragments during library preparation. If the system requires amplification of the sample to enable the creation of signals at a level which the system can detect, the Polymerase Chain Reaction (PCR)-mediated procedure is generally employed. Even though high-fidelity DNA polymerases are used in this procedure, a small number of nucleotide changes may occur, resulting in lower accuracy than 100% depending on the system type. In order to pool several samples in a single sequencing run, companies represent different types of library preparation strategies. In order to illustrate, an example can be given with Illumina, which dominates the NGS market. In its procedure, nucleotides which are conjugated to a fluorescent marker are quenched by a present chemical. Each of them is added to the growing strand once in every fragment population (40). Another common strategy to increase accuracy can be obtained through sequencing both DNA strands, and producing forward and reverse reads, called paired-end sequencing.

WES is one of the NGS approaches based on the enrichment of protein-coding regions of known human genes. WES is applied as a series of experimental and computational procedures. The experimental process of WES generally follows a

similar procedure, without regard to the capture method and NGS platform used (16).

The capture allows the investigation of the protein-coding regions in the genome. For that purpose, several types of DNA or RNA-based sequences are designed and synthesized compatible with the targeted exonic regions. Although the capture probes' content shows the difference between commercial reagent sets, they focus on coding regions and their proximal intronic flanking sequence. They may also include as targets 5' and 3' untranslated region (UTR) sequences (41).

The procedure typically starts with the NGS library preparation from genomic DNA. Genomic DNA is randomly fragmented, and several micrograms are used to prepare a library. The fragment library is then hybridized with oligonucleotide capture probes with blocking bases complementary to the adaptors. The captured fragments are purified and then amplified. The barcode sequences, which allow sample indexing in the sequencing process could be introduced during post-capture amplification if they were not presented during the initial library preparation. Then, massively parallel sequencing is applied to the enriched library (11).

After the sequencing process, between 20,000 and 50,000 single nucleotide variations and small insertions and deletions can be detected from WES (17). However, translating the variations into a particular disease phenotype is the main challenge in the clinical area. Bioinformatics analysis from raw sequence data to annotated genomic variants which is the first part of the translation, consists of particular steps.

2.3. WES Data Analysis

WES generates a vast amount of data whose analysis requires robust computational power and expertise. Bioinformatics has been pivotal to the analysis and interpretation of WES data. The study of WES data comprises of a multistep process. A variety of open-source tools and software exist to handle the analysis of WES data, and most of them are developed to perform a particular aspect of the entire study. Because of the diversity of available tools, each laboratory group performing WES has a unique analytical pipeline. Although variations in the process exist, there are standard basic bioinformatics protocols involved in the data analysis.

A typical starting point is to evaluate the quality of raw reads (42). The raw starting material for most common platforms is a FASTQ file format (43). NGS platforms are capable of generating massively parallel sequence reads even in a single run. Yet, the quality of reads may not be perfect due to several kinds of biases introduced in the sequencing experiments (44). These sequence biases can result from some reads with adapters or contaminant, the low-quality bases, especially at the end of the reads, and unknown base calls (45). The correction of these biases is the important step that should be performed before further analysis. Many tools developed based on different algorithms are available to evaluate raw FASTQ data (45–50). These tools generally take FASTQ files as input and generate summary statistics and graphs for a beneficial overview of the raw read quality.

Following the quality check of the raw reads, if required, preprocessing must be performed. The standard preprocessing step consists of trimming of low-quality bases and adapter sequence removal at the end of the reads (42). Adapter sequences are short oligonucleotides that are ligated to the ends of DNA fragments during the library preparation step so that they can be combined with primer sequences for amplification (9). However, when the target DNA fragments are shorter than the sequencing read length run, the adapter sequences are read out with the target DNA fragments. Besides,

the quality of the reads generally becomes lower at the end of the sequencing cycles (51,52). It is crucial to identify the adapter sequence or low-quality bases and trim it to recover the DNA sequences of interest before the alignment step. Several tools with different implementation principles have been developed to perform adapter and quality trimming (53–58).

The raw sequence reads produced by sequencing machines do not include the information about the genomic position, and they must be aligned to reference human genome. The alignment is the most critical step to make an inference whether a sequencing experiment has succeeded (59). Optimal alignment to reference sequences is not an easy computational task and requires a fast and tolerant algorithm to obtain an imperfect alignment due to genomic variations. There are several tools based on different algorithms to achieve the process of alignment using a known reference genome (60). Each base is mapped to the reference sequences under the determined conditions by the mapping software's input parameters. Both GRCh37 (hg19) and GRCh38 (hg38) are widely used as a reference for the human genome.

After the alignment step, it is recommended that processing of aligned reads to improve the quality of downstream variant calling analysis (61). In the variant calling step, the differences between the reference genome and the target genome are calculated. Several tools based on different algorithms have been developed to identify short germline variants, which are inherited variations that exist in the germ cells (62).

After variants are detected, biologically important features such as gene symbols, genomic position, amino acid change, and consequences of variants add to the data in the annotation step. In addition to the ordinary annotation, several tools can be used to integrate the annotations from countless sources. These annotation tools enable to filtration and interpretation of potential disease-causing mutations. The filtration and prioritization of clinically causative mutation among a vast amount of annotated

variations is the most challenging part of the analysis and is not a fully automatized (63).

2.4. Variant Interpretation and Reanalysis of Unsolved WES Cases

An essential consideration when analyzing WES is the strategy of sequencing only the patient who is the first affected individual (proband) versus the proband and biological parents (trio). Studies have shown that the trio approach provides a higher diagnostic yield since it enables more detailed information on inheritance patterns of variants (64). However, WES has about 25-30% diagnostic yield for identifying variants that cause rare monogenic disorders. (14,18,19). Although some of the patients who are in the undiagnosed percentage have a pathogenic mutation in the already sequenced genomic data, it may not be identified in the first analysis (20–22).

There are various reasons proposed that cause to remain a pathogenic variant to be unidentified in the initial exome (23). Firstly, variant calling from short sequence reads and annotating of them to relevant genomic information is not computationally perfect processes (65,66). Many factors at each step affect the quality of variant detection. Another reason is that the knowledge about known gene-disease and variant–disease associations is incomplete; however, the literature databases are growing (67). In addition to these, the phenotype information is the key component in the variant prioritization workflow. Any insufficiency in reporting may lower the probability of prioritization of disease-causing mutation (23). Ultimately, a recent study reported that 10% of the undiagnosed patients could get a precise diagnosis through the reanalysis of the same WES data, using different workflows and with the help of growing knowledge in the literature (23).

3. MATERIALS AND METHODS

In this dissertation, three different cases which had had unsolved WES were studied. WES data re-analysis was performed by employing a variant prioritization workflow described in the dissertation. The overview of the workflow is demonstrated in Figure 3.1.

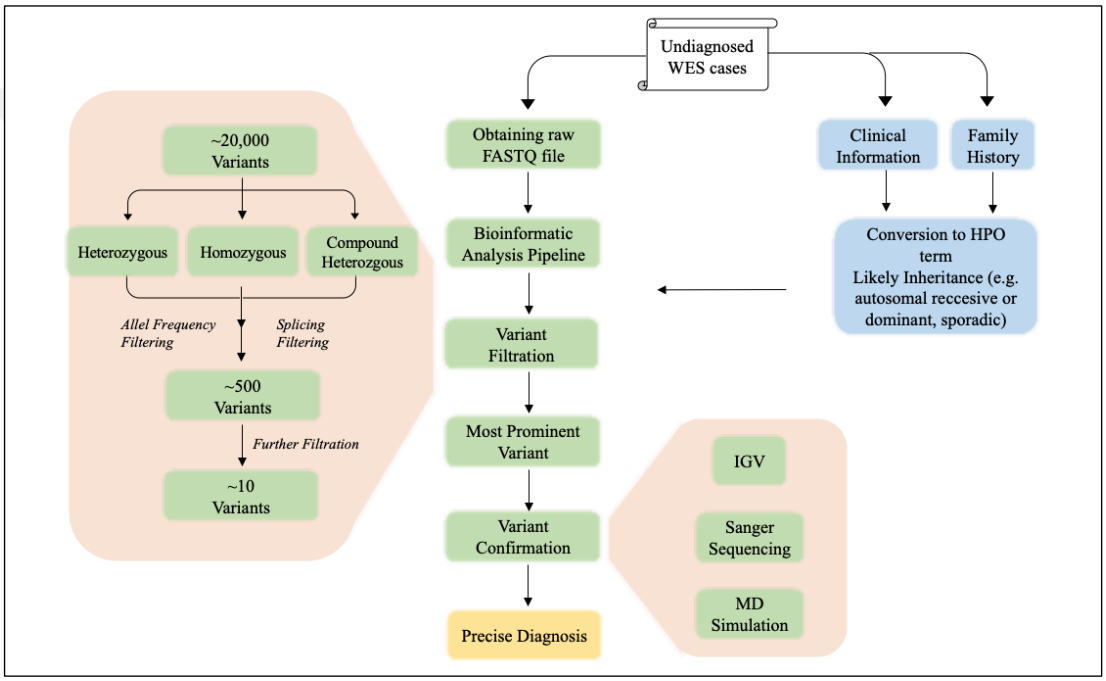


Figure 3.1. Overview of the workflow. Raw FASTQ files are provided from undiagnosed WES cases. WES data is analyzed by implementing the pipeline described in the method section. The vast number of variants obtained from raw data pass to the filtration step described in the thesis. Clinical information and family history are integrated into these analysis process. When the most prominent variant is identified, appropriate confirmation methods are conducted to diagnose a patient precisely.

The materials and methods section can be described in 4 main subtitles:

1. Clinical Presentation of the Cases
2. The Data Analysis Pipeline for WES
3. Variant Filtration
4. Variant Confirmation

Details of each step will be explained in their corresponding subsections.

3.1. Clinical Presentation of the Cases

The reported clinical information is crucial to interpret WES outcomes. If key components of the phenotype are not available, this decreases the probability that a causative variant would be prioritized (23). Therefore, standardized terms corresponding to any phenotypic information about the patient should be integrated into the workflow. Human Phenotype Ontology (HPO) (68) terms are used to describe phenotypic abnormalities. HPO provides a standardized terminology for phenotypic symptoms found in human diseases. Below, clinical information on the three cases will be presented.

3.1.1. Case I

The patient was a 6-year-old female who applied to Istanbul Acibadem Maslak Hospital. The patient was born to non-consanguineous parents at the full-term of gestation by cesarean delivery. Pedigree analysis for case I was shown in Figure 3.2. She had globally delayed development with microcephaly and intellectual disability at the examination. Speech development was also delayed and dysarthric. At the age of 4 months, she was diagnosed with epilepsy. Her electroencephalography (EEG) results indicated generalized, left temporoparietal and right posterior temporal interictal epileptiform discharges. Then, her treatment was started with Levetiracetam. She had involuntary movements including choreoathetosis and dystonia. Her tonus was increased in all four extremities. She could only walk with assistance She also had involuntary eye movements. It was reported that non-specific white matter changes in magnetic resonance imaging (MRI). She also had left lower eyelid hemangioma at birth. She took treatment for reflux. WES had been performed by a genetic diagnosis center, but any pathogenic variant was not reported. She has referred for genetic counseling by pediatric neurology again. Then, her raw data was provided to reanalysis. The research was conducted under a protocol approved by the Acibadem

Mehmet Ali Aydinlar University Institutional Review Board for human subjects research. Written informed consent was obtained for all participants.

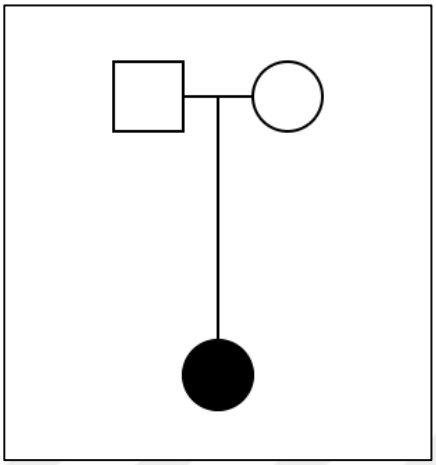


Figure 3.2. Pedigree for the Case I.

3.1.2. Case II

The patient was a 4-year-old female who applied to Istanbul Acibadem Maslak Hospital. There was no known consanguinity between her parents, and pedigree analysis was shown in Figure 3.3. She was born with decreased movement during the last period of 35-week of gestation. She received treatment due to respiratory distress syndrome. Cranial MRI at 1.5-month showed periventricular leukomalacia in white matter. She showed mild-severe global developmental delay with speech disorder and microcephaly at the examination. She was taking medicine for hypothyroidism. She had strabismus and axial hypotonia. GI-II pelviectasis was detected her kidneys. She had some skeletal system abnormalities such as brachydactyly, small jaw, short and wide distal phalanx, and dislocated elbows. Additionally, she was operated two times for hip dislocation. She had distinctive facial features including epichantal folds, bitemporal narrowed forehead, flat and broadened nasal root, flat and broadened nasal tip, wide and long filtrum, thin upper lip, pronounced ears, flattened antihelix. WES

was performed in a genetic diagnosis center, but no reported variation explained her clinical status. Then, her raw data was provided to reanalysis. The study was approved by the Acibadem Mehmet Ali Aydinlar University Institutional Review Board for human subjects research.

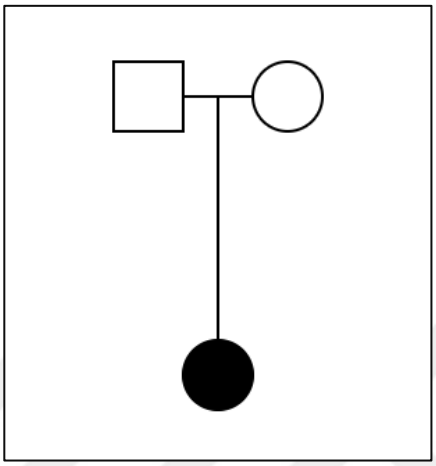


Figure 3.3. Pedigree for the Case II.

3.1.3. Case III

Case III was a 6-year-old male who applied to Erciyes University Faculty of Medicine. The study was approved by the Acibadem Mehmet Ali Aydinlar University Institutional Review Board for human subjects research. The patient was born non-consanguineous parents at the full-term of gestation by normal vaginal delivery. The phenotypic features include generalized hypotonia with global developmental delay, mild facial dysmorphisms such as frontal bossing and low-set ears, speech delay, decreased pain response, hyperactive deep tendon reflexes, strabismus. He had normal evaluations for brain MRI, EEG, chromosomal microarray, and comprehensive biochemical metabolic testing. The pedigree analysis demonstrated two additional healthy brothers and healthy parents (Figure 3.4). Trio WES was performed before,

but any pathogenic variant was not reported. He has referred for genetic counseling again. The raw sequencing data of the family have been obtained for reanalysis.

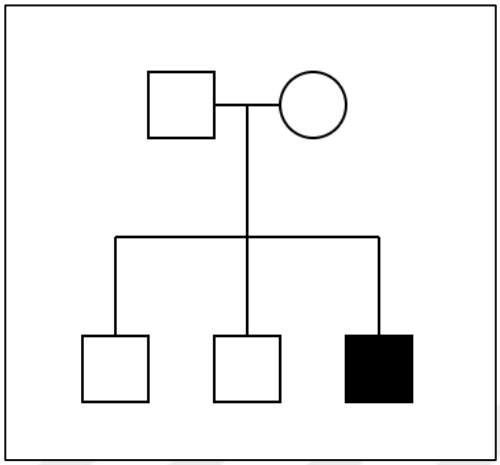


Figure 3.4. Pedigree for the Case III.

3.2. The Data Analysis Pipeline for WES

In this part, the analysis of WES data presented in the dissertation will be explained in detail. An outline of the overall pipeline is shown in Figure 3.5. The pipeline consists of a set of steps: Quality check of raw reads, preprocessing, alignment, post-alignment processing, germline variant calling (haplotyping) & joint genotyping and variant annotation.

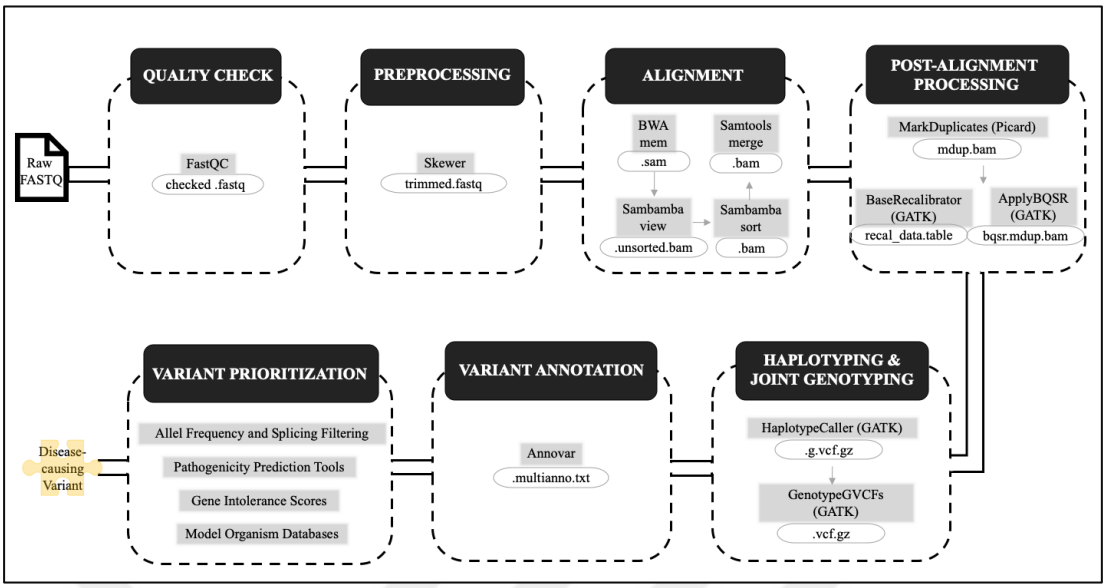


Figure 3.5. An overview WES data analysis pipeline. The pipeline consists of a set of steps: Quality check of raw reads, preprocessing, alignment, post-alignment processing, germline variant calling (haplotyping) & joint genotyping and variant annotation.

Each of these steps, which are processor or memory intensive, is implemented by one or several freely available software tools. Many of the tools permit multithreaded operations that may benefit from the availability of multiple cores. The tools and packages with version numbers used in this dissertation are listed in Table 3.1. The usage of each tool will be described in the corresponding section with a series of commands in a terminal on a standard Unix workstation or server.

Table 3.1. Tools for WES data analysis for detection of germ-line variants

Program/Software version	Tool/Package	Function as used in this dissertation
Conda (v4.8.3)	config	Install and manage packages and environment
FastQC (v0.11.8)	fastqc	Assess the quality of sequence reads
Skewer (v0.2.2)	skewer	Trim adapter and low-quality sequences
BWA (v0.7.17)	mem	Align reads to the reference human genome
Sambamba (v0.6.6)	sambamba view	Extract information from SAM files
Sambamba (v0.6.6)	sambamba sort	Sort BAM files
Samtools (v1.9)	samtools merge	Merges multiple bam files into a single bam output.
Picard (v1.141)	BuildBamIndex	Build bam index (bai)
Picard (v1.141)	MarkDuplicates	Identify duplicate reads.
Picard (v1.141)	FixMateInformation	Verify mate-pair information between reads
GATK (v4.1.4.1)	BaseRecalibrator	Generate recalibration table for Base Quality Score Recalibration (BQSR)
GATK (v4.1.4.1)	ApplyBQSR	Apply base quality score recalibration
GATK (v4.1.4.1)	HaplotypeCaller	Call germline SNPs and small indels via local re-assembly of haplotypes
GATK (v4.1.4.1)	CombineGVCFs	Merge one or more HaplotypeCaller GVCF files into a single GVCF
GATK (v4.1.4.1)	GenotypeGVCFs	Perform joint genotyping on one or more samples pre-called with HaplotypeCaller
Annovar (v20191024)	convert2annovar	Convert the .vcf files to the Annovar input file format
Annovar (v20191024)	table_annovar	Annotate the input file with different options available in Annovar
RStudio (v1.1.453)	VarfromPDB	Mine the genes and variants associated with diseases from literature and multiple public databases.

3.2.1. Quality check of raw reads

In this dissertation, FASTQ files generated by paired-end sequencing of the cases were used to the reanalysis.

FASTQ is a commonly used file format used for storage and sharing raw sequences produced by a sequencing machine (43). FASTQ format stores the nucleotide sequence, the same as in FASTA format, and the Phred quality scores corresponding to those sequences (69,70). FASTQ files with extension ".fq" or ".fastq" can be viewed from the command line on computers had Unix/Linux operating system. Each sequence in the file consists of four lines. The first line begins with a "@" symbol and continues with a sequence identifier and some definitions which can be optionally used. The second line consists of the bases: A, T, G, C, and N (unknown base). The third line starts with a "+" character. The "+" sign specifies the end of the sequence. The fourth line includes the Phred quality scores for the sequence in the second line. Following is an example of a FASTQ file analyzed in the thesis obtained from the Illumina sequence system: (Figure 3.6).

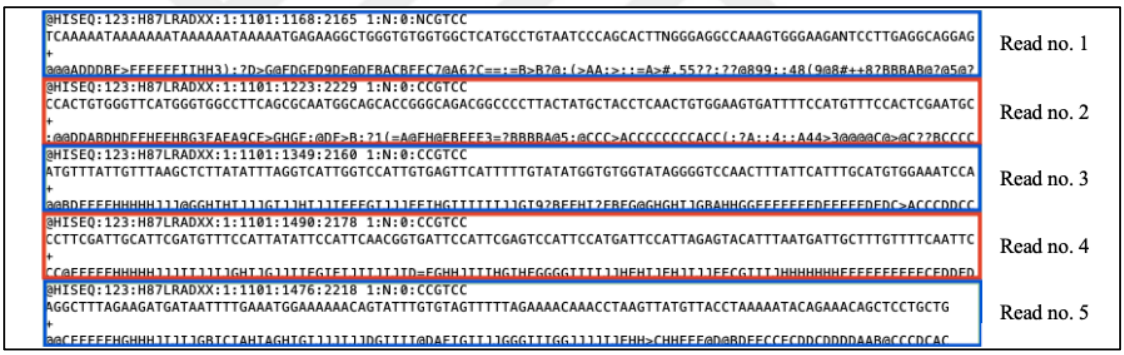


Figure 3.6. An example of a FASTQ file which belongs to a case used in the dissertation.

The first line is the sequencing machine identifier and description items separated by ":". In this example, each item for the read 1 explained as respectively:

HISEQ: sequencer name.

123: run id.

H87LRADXX: flow cell id.

1: flow-cell lane.

1101: the number of the tile.

1168: x-coordinate of the cluster in the tile.

2165: y-coordinate of the cluster in the tile.

1: the member of the pair

N: filter information (Y if the read is filtered, N otherwise).

0: status of control bits (0 for none of the control bits)

NCGTCC: index sequence.

The second line of the example is the sequence contents for the read 1. The third line has only a "+" sign and has no more information added. The fourth line is the Phred quality scores for corresponding bases in the second line. The Phred quality scores range from 0 to 93 stand for sequencing quality for each base in the second line, and the higher the Phred score, the more accurate the base to be determined. Phred scores are encoded with several characters. While the "!" refers to the lowest Phred score, the "~" is for the highest one.

The quality of raw FASTQ data was assessed in the first step of the workflow to ensure no biases. FastQC (v0.11.8), developed by Simon Andrews at Babraham Institute (46), was used to evaluate the quality of FASTQ files. The following command was used:

fastqc -t 16 ${sample lane}_R1.fastq
fastqc -t 16 ${sample lane}_R2.fastq

FastQC takes FASTQ files as input. -t commands the number of threads to use. The tool produces a report as a compressed file and an HTML file for each FASTQ input. The report consists of summary statistics and graphs that enable a useful impression of whether the data has any problems before doing any further analysis. On the below, FastQC output for one of the patients analyzed in this dissertation is seen. The most critical points that must be taken into consideration from the graphs is explained briefly.

The report begins with straightforward information about input FASTQ file including name, type of quality score encoding, the total number of reads, read length and GC content (Figure 3.7)

Measure	Value
Filename	AD_1211_L001_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	12246323
Sequences flagged as poor quality	0
Sequence length	10-110
%GC	42

Figure 3.7. Basic statistics about FASTQ files.

The box and whisker plot in Figure 3.8 shows quality score statistics at each position in the FASTQ file. The yellow box is the inner-quartile range for the 1st to 3rd quartile. The red line within each yellow box represents the median quality score at each base window. The blue line is the mean quality score at each base window. The upper and lower whiskers represent the 10th and 90th percentile scores. The lower median quality scores of the first 5-7 bases are accepted; however, it is expected to rise for the next sequences. It is observed that the average quality score will continuously decrease towards the end of the sequencing.

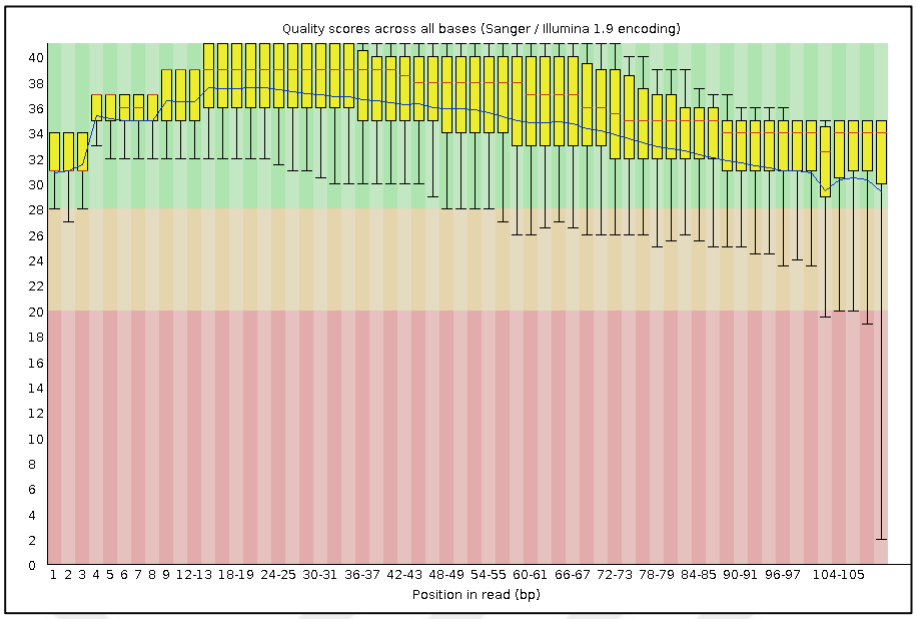


Figure 3.8. Per base sequence quality

The plot in Figure 3.9 shows the deviation from the mean quality for each tile. The colors are coded on a cold to hot scale. While colder colors indicate the quality is at or above the average for that base, colder colors mean a tile had better qualities than other tiles for that base. It is expected that a plot being all blue if there is no problem in the quality related to a part of the flow-cell. In this case, per tile sequence, quality looks a bit low.

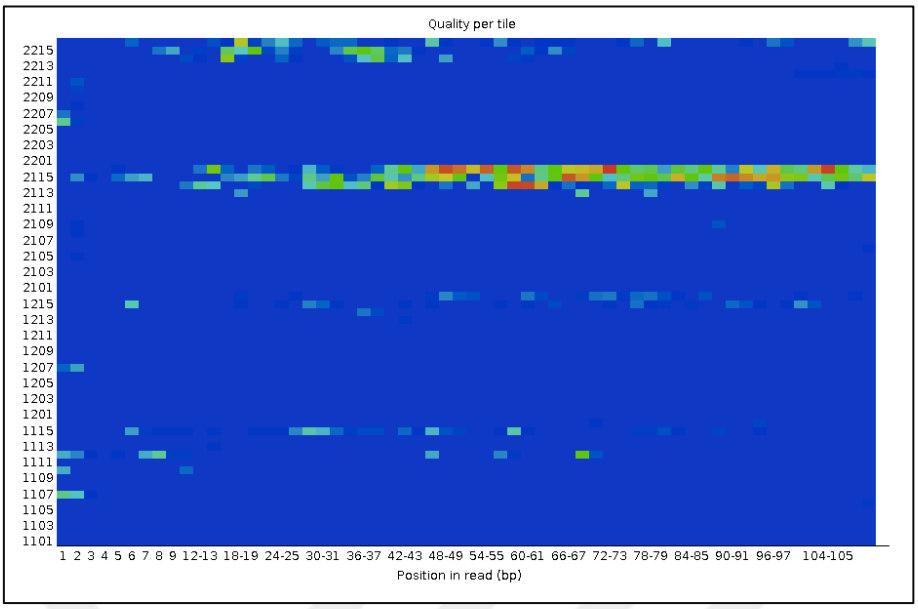


Figure 3.9. Per tile sequence quality

Per sequence quality scores are represented as a plot in Figure 3.10. The plot shows the total number of reads and the average quality score through the reads. The distribution of average quality should be tight in the upper part of the plot. In this case, quality scores look expectedly distributed.

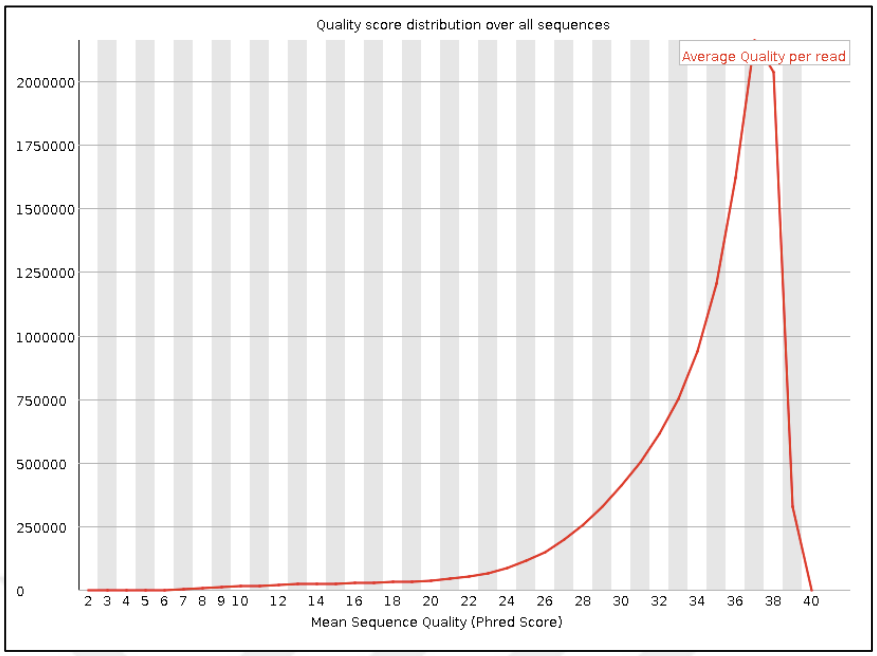


Figure 3.10. Per sequence quality scores

Next plot reports per base sequence content, which is the percent of called bases for each nucleotide at the positions in the FASTQ. It is expected to be seen as straight lines in the plot because the data is assumed that a random sample. So the base content at each position should be identical. Depending on the type of library preparation kit used, the non-uniform distribution of bases for the first 10-15 nucleotides can be expected. In this case, per base sequence content looks quite reasonable (Figure 3.11).

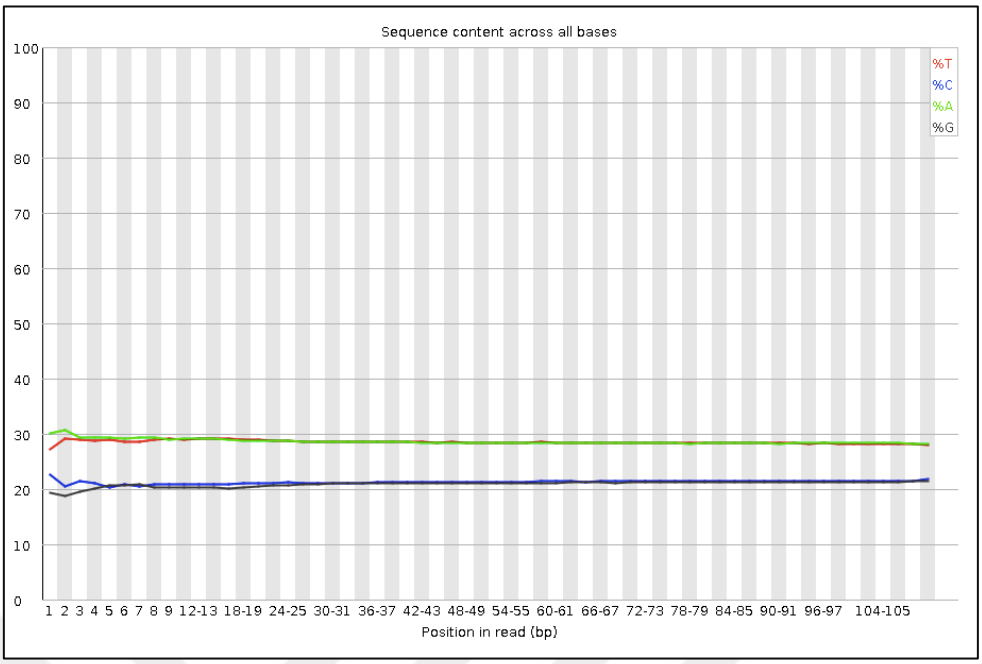


Figure 3.11. Per-base sequence content

Figure 3.12 is the plot of the number of reads and GC% content of the reads. The theoretical distribution assumes a uniform GC content for all reads. It is expected that the observed distribution should not deviate from the theoretical distribution.

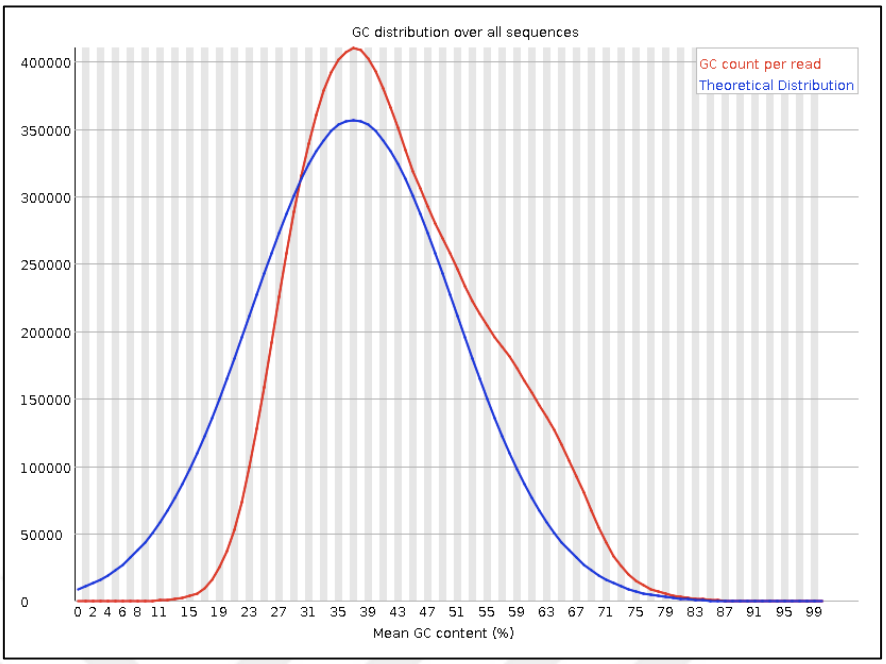


Figure 3.12. Per sequence GC content

Percent of bases at each position without a confident call, ‘N’, is reported in the plot in Figure 3.13. It should not be observed any point of this curve rises above zero. Unknown bases are not seen for the case below.

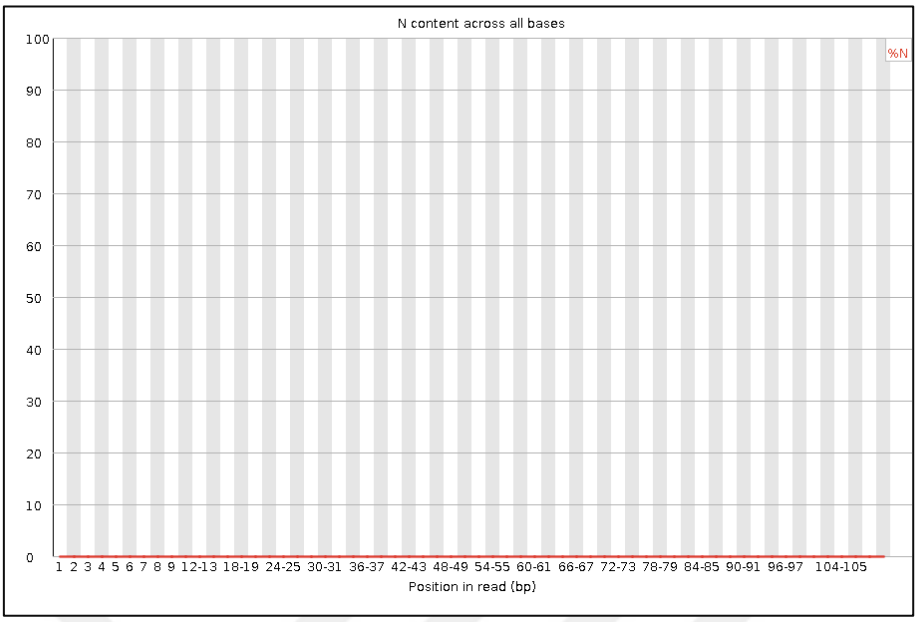


Figure 3.13. Per base N content

The following graph shows the distribution of fragment sizes in the data. As expected, the length of the reads is mostly around 110 bp (Figure 3.14).

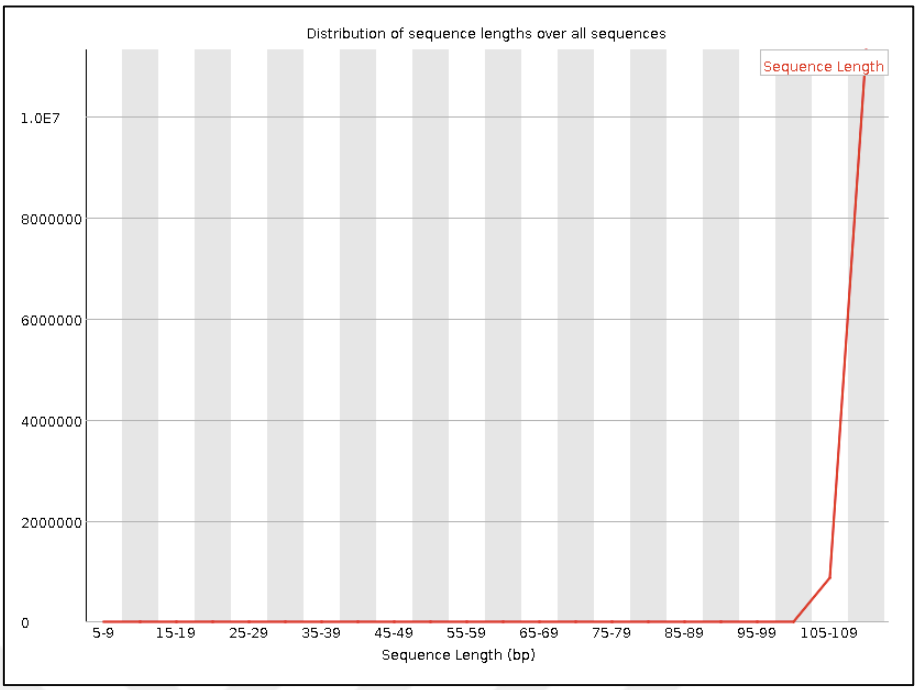


Figure 3.14. Sequence Length Distribution

The next plot shows the percentage of reads of a particular sequence in the FASTQ, which exists a given number of times in the FASTQ (Figure 3.15). Duplicate reads generally resulted from PCR-mediated amplification of some library fragments. These fragments have been over-represented due to bias during the enrichment process. PCR duplicates cause to misrepresent the exact amount of sequences. It is expected that nearly 100% of the reads will be unique. In this case, the duplication level is considered low.

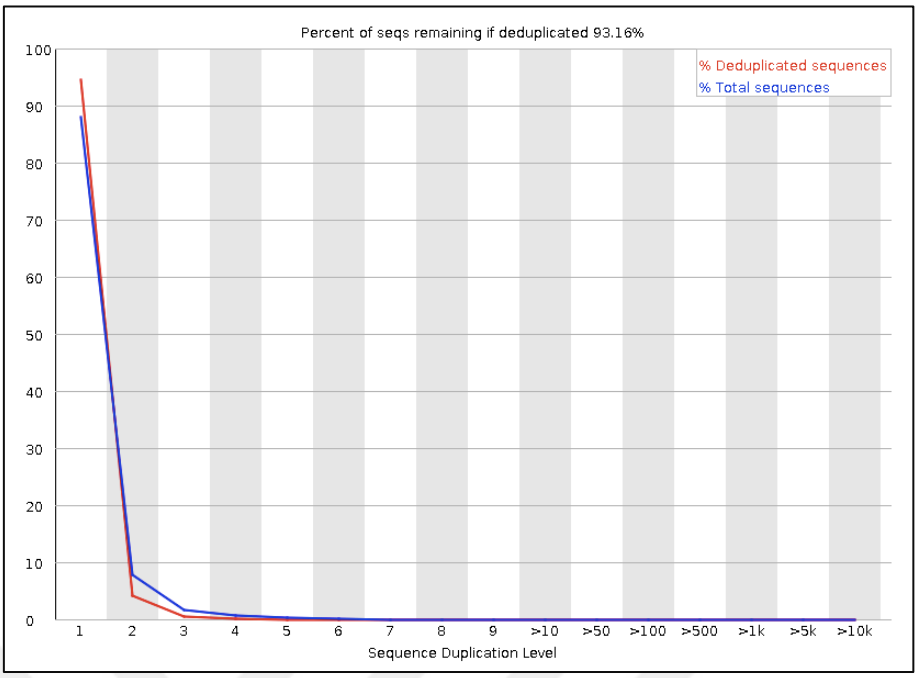


Figure 3.15. Sequence Duplication Level

The last report is the plot of the part of reads where the sequence library adapter sequence is shown at the particular base position. Only adapters specific to the library shown in the legend are searched. In this case, the contamination is not detected for searched adaptors (Figure 3.16).

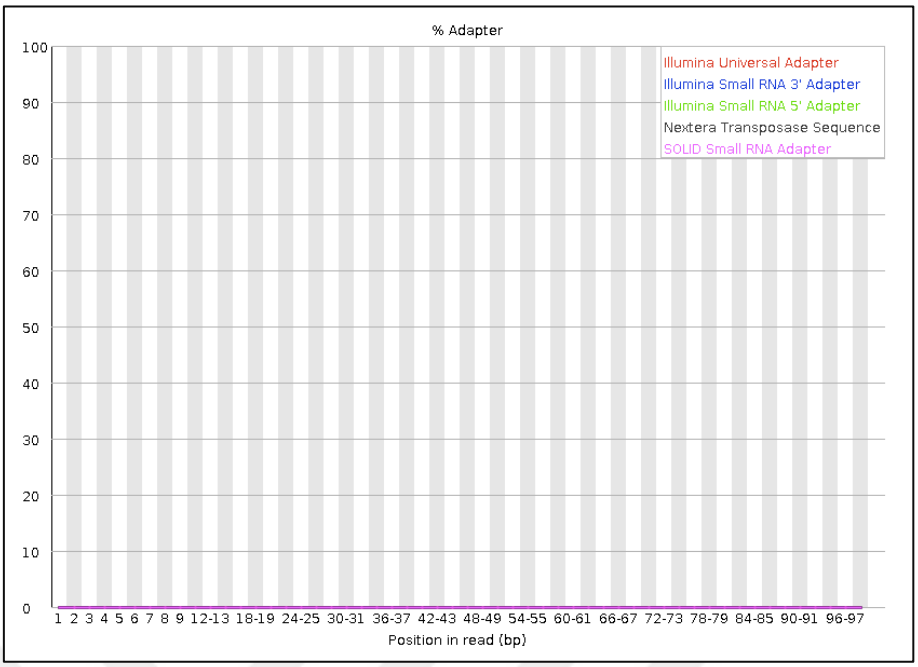


Figure 3.16. Adapter Content

3.2.2. Preprocessing of raw reads

Based on the result of the quality check step, if there is a need, preprocessing is necessary before alignment. Skewer (56) version v0.2.2 was used for trimming adapters and low-quality sequences. Skewer applies a novel dynamic programming algorithm to trim adapters and low-quality sequences based on Phred quality scores. It is specially designed for processing Illumina pair-end reads. The following command uses the Skewer tool, which takes as input paired FASTQ files.

skewer --quiet --threads 16 -z -m pe -q 10 -Q 20 -l 35 -n -x AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT --output trimming/${sample}_L001 ${sample}_L001_R1.fastq ${sample}_L001_R2.fastq

For pair-end reads, the valid mode is ‘pe’ to specify -m option. -q option for 3’ end quality trimming and trim 3’ end until specified or higher quality reached. The value is 10. –Q option specifies the lowest mean quality value allowed before trimming, the value is 20. -l option sets the minimum permitted read length after trimming. The value is 35. -n option for filtering out highly degenerative reads, which are defined as those that above 15% of the nucleotides are ‘N’. The default is no. -x option commands an adapter sequence of file for the reads.

3.2.3. Alignment

After the quality check and preprocessing of raw data, the next step is to align the reads to the reference genome and with high accuracy. The files that include sequencing reads were aligned to the hg19 released from The University of California at Santa Cruz (UCSC) assembly of the human genome using the Burrows-Wheeler Aligner (BWA), version v0.7.17 (71). BWA is a software package for mapping reads against particular reference genomes. BWA-MEM is one of the algorithms in the software designed to align reads ranging from 70bp to 1Mbp. BWA-MEM is a fast and accurate way to map Illumina short reads. The following command takes paired, trimmed read files and FASTA format of the reference genome as input and generates mapped reads file.

bwa mem -R "@RG\tID:${sample_lane}\tPL:ILLUMINA\tLB:${library}\tSM:${sample}" -M -t 16 ucsc.hg19.fasta trimming/"$sample_lane"-trimmed-pair1.fastq.gz trimming/"$sample_lane"-trimmed-pair2.fastq.gz > ${sample_lane}.sam

bwa mem has many different options to adjust the mapping and generates a sequence alignment map (SAM) file with extension “.sam” as an output. SAM file is a tab-delimited text file containing mapped reads to a reference genome and supports

short reads produced by different sequencing platforms (72). Binary alignment map (BAM) is the binary version of a SAM file and has a file extension ".bam". Both BAM files and SAM files have the same information, which including a header, which is optional for the file formats and an alignment section. The alignment section includes the genomic position with relevant descriptive details of each sequence.

The header section generally provides four types of information. Each of them must start with a "@" symbol: header line (@HD), reference sequence dictionary (@SQ), Read group information (@RG), and the mapping program (@PG).

Sambamba tool (73) was used for converting the .sam file to .bam. Sambamba is a fast and robust tool with high performance on processing SAM and BAM files. sambamba view allows accessing SAM header and information about reference sequences through the following command.

sambamba view --nthreads=16 --with-header --sam-input --format=bam --output-filename=${sample_lane}.unsorted.bam ${sample_lane}.sam

sambamba view generates an unsorted BAM file. BAM files can have sort order either 'coordinate', or 'qname'. While 'coordinate' means to sort the file by reference ID corresponding reads by start coordinate, 'qname' orders the reads by lexicographically. Input BAM files can be sorted externally using sambamba sort. The default mode is 'coordinate' because this is the one used for building index later.

sambamba sort --nthreads=16 --out=${sample_lane}.bam ${sample_lane}.unsorted.bam

Multiple sorted alignment files of the same sample should be merged and produced a single sorted output file that contains all the content pf the input file. Samtools (72) was used to achieve this by the following command:

samtools merge -@ 16 -f ${sample}.bam "$sample"_L001.bam "$sample"_L002.bam

Before doing any analysis using a sequence file, the files usually have to be indexed. A BAM index (.bai) file from a BAM file was generated by using BuildBamIndex tool available in Picard (74):

java -jar picard.jar BuildBamIndex INPUT=${sample}.bam VALIDATION_STRINGENCY=LENIENT

3.2.4. Post-alignment processing

Processing of aligned reads is recommended to improve the quality of downstream variant calling analysis (75). This step consists of a set of processes to minimize technical biases.

During the sequencing, a library of DNA fragments from a particular genomic region is prepared using PCR amplification to provide adequate DNA fragments for the sequencing process. Therefore, some amplified fragments could share sequence and the same corresponding alignment position leading to bias in variant detection. These duplicates should be removed to eliminate PCR-introduced bias. MarkDuplicates available in the Picard tools (74) was used to detect read duplicates based on their position on the genome.

java -Xmx4g -jar picard.jar MarkDuplicates I=${sample}.bam O=${sample}.mdup.bam M=${sample}.mdup_metrics.txt CREATE_INDEX=true VALIDATION_STRINGENCY=LENIENT

FixMateInformation available in Picard (74) verifies whether each read and its mate-pair is in sync and fix if needed.

java -Xmx4g -jar picard.jar FixMateInformation I=${sample}.mdup.bam O=${sample}.matefixed.mdup.bam SORT_ORDER=coordinate CREATE_INDEX=true VALIDATION_STRINGENCY=LENIENT

In addition to marking duplicates, base quality is also an essential factor for variant detection. As mentioned before, each sequence read has a Phred quality score generated by the sequencing machine. However, the machine could make systematically biased scores, and these cause to overestimated or underestimated base quality scores. Base Quality Score Recalibration (BQSR) implements an empirically accurate error model to the bases in order to arrange the quality scores. Thus, technical bias is significantly minimized. The critical point in this process is to exclude known variants before BQSR since they are actual genomic variations, and they should not be evaluated as sequencing errors. The recalibration of base qualities was performed following the recommendation of Genome Analysis Toolkit (GATK) (75). The base recalibration process consists of two steps: Generating quality scores and applying the scores to bases. The program creates a covariation model of known variants; then it adjusts the base quality scores in the data based on the model by running the following GATK command:

gatk BaseRecalibrator --reference ucsc.hg19.fasta –input ${sample}.matefixed.mdup.bam --known-sites dbsnp_138.hg19.vcf --known-sites Mills_and_1000G_gold_standard.indels.hg19.sites.vcf --known-sites 1000G_phase1.indels.hg19.sites.vcf –output ${sample}.recal_data.table

Then, the recalibration is applied to the data by running the following GATK command:

gatk ApplyBQSR --reference ucsc.hg19.fasta –input ${sample}.matefixed.mdup.bam --bqsr-recal-file ${sample}.recal_data.table –output ${sample}.bqsr.matefixed.mdup.bam

3.2.5. Variant calling

In the variant calling step, the differences between the reference genome and genome of interest are calculated. After the post-alignment processing step, the identification of germline short variants can be started on an analysis-ready BAM file. The HaplotypeCaller (61) available in GATK was used to call germline single nucleotide variations (SNVs) and small indels simultaneously. These variants are called by the program via the local de-novo assembly of haplotypes in an active region. When the program finds a variation in a particular part, it disposes the present mapping information and reassembles the reads in that region. The HaplotypeCaller achieves this in four steps:

1- Definition of active regions: The program detects areas showing signs of variation and then applies de novo assembly to those regions.

2- Determination of the haplotypes in the active region: The program implements a De Bruijn-like graph for reassembling each active part. Following that, the possible haplotypes in the data are identified. The haplotypes are realigned to the reference haplotype applying the Smith-Waterman algorithm.

3- Determination of probabilities for the haplotypes: Pairwise alignment of each read against reference haplotypes for each active region is performed, implementing the Hidden Markov Model algorithm. A matrix of likelihoods of haplotypes is produced, and they are then marginalized to obtain the probabilities of alleles for each potential variation region.

4- Assignment of sample genotypes: For the potential variation regions, Bayes' rule and calculates the likelihoods are implemented by for each genotype. Then the most likely genotype is assigned to the sample.

This local reassembly approach makes the HaplotypeCaller more accurate when calling regions that contain different types of variants close to each other. It also has a higher performance to call indels than position-based callers.

The tool runs per-sample and takes a .bam files to generate a file, namely GVCF. GVCF is an intermediate file format which can then be used in GenotypeGVCFs (76) for joint genotyping of multiple samples efficiently.

gatk HaplotypeCaller --reference ucsc.hg19.fasta –input ${sample}.bqsr.matefixed.mdup.bam –output ${sample}.g.vcf.gz --emit-ref-confidence GVCF --annotation-group AS_StandardAnnotation

For trio analyses, CombineGVCF (76) available in GATK is used to integrate the data from biological parents to the proband's data. The program combines per-sample gVCF files produced by HaplotypeCaller into a single gVCF file.

gatk CombineGVCFs --reference ucsc.hg19.fasta --variant ${sample}.g.vcf.gz --variant ${sample}.g.vcf.gz --variant ${sample}.g.vcf.gz –output combined.g.vcf.gz --annotation-group AS_StandardAnnotation

Joint genotyping of the samples pre-called with HaplotypeCaller was carried out following by the GenotypeGVCFs available in GATK (76). This tool is developed to perform joint genotyping on a single input containing one or many samples. The GATK4 GenotypeGVCFs tool can take a single-sample GVCF or a single multi-sample GVCF created by CombineGVCFs as input.

gatk GenotypeGVCFs –reference ucsc.hg19.fasta --variant combined.g.vcf.gz --output combined.vcf.gz --annotation-group AS_StandardAnnotation

The tool generates a final Variant Call Format (VCF) in which all samples have been jointly genotyped. A VCF is a text file that stores sequenced variants called by variant callers (77). The file has two main lines, a header line with meta information and a data line. The previous section provides background information about the analysis results and begins with "##" symbol. The first header line always shows the VCF format version followed by lines starting with ##INFO, ##FILTER, and ##FORMAT, which include several descriptions of each component regarding fields of each data line. Data lines in the VCF file contains nine standard columns, which are described below, and continue with one or more sample columns.

CHROM: the name of chromosomes.

POS: the position of the variants.

ID: the identifier of variants.

REF: reference sequence.

ALT: alternate bases.

QUAL: Phred quality score.

FILTER: filter status.

INFO: additional information.

FORMAT: colon separated key and value.

More than one sample columns may be present in a VCF file. The sample field has the values defined in the previous FORMAT field for a sample.

3.2.6. Variant annotation

After variants are detected, biologically important features of variants add to a VCF file in the annotation step. ANNOVAR (78) was used to annotate variants. ANNOVAR takes variant files that includes chr, start, end, ref, alt, and optional fields, as an input. Before the annotation, the .vcf file format must convert to the ANNOVAR

input file format. The convert2annovar.pl command can convert other "genotype calling" format into ANNOVAR format.

perl annovar/convert2annovar.pl -format vcf4 ${sample}.vcf -allsample -withfreq -include > ${sample}_pass.txt | tee log

ANNOVAR annotates functional effects of variants; the location of each variant concerning genes; exonic, intronic, intergenic, splice site, 5' or 3'-UTR, with a predicted impact on human genes from the RefSeq, and with an allele frequency in the gnomAD, 1000 Genomes Project, and Exome Aggregation Consortium control human populations. The table_annovar.pl takes VCF files as an input and generates a tab-delimited output file with columns, each representing one annotation.

perl annovar/table_annovar.pl ${sample}_pass.txt annovar/humandb/ -buildver hg19 -protocol refGene,avsnp150,snp138NonFlagged,clinvar_20190305,gnomad211_exome,popfreq_max_20150413,exac03nontcga,revel,dbnsfp33a,regsnpintron,dbscsnv11 -operation g,f,f,f,f,f,f,f,f,f --argument '--splicing_threshold 40',,,,,,,,, -nastring " --otherinfo | tee run.log

By default, ANNOVAR annotates variant on hg18 coordinate. Since the input file is in hg19 coordinate, -buildver hg19 is added in the code above. The output file includes multiple columns. Each of the columns specified one of the "protocol" that the user determined in the command line. The refGene column is the annotation on how the mutations affect gene structure. It consists of FASTA sequences for all annotated transcripts in RefSeq Gene. The next columns are explained respectively below:

avsnp150: the SNP identifier in the dbSNP version 150.

snp138NonFlagged: dbSNP with ANNOVAR index files, after removing those flagged SNPs (SNPs < 1% MAF or unknown, mapping only once to reference assembly, flagged in dbSnp as "clinically associated".
clinvar_20190305: ClinVar database with separate columns (CLINSIG CLNDBN CLNACC CLNDSDB CLNDSDBID) for each variant. The columns include information about variant clinical significance (unknown, untested, non-pathogenic, probable-non-pathogenic, probable-pathogenic, pathogenic, drug-response, histocompatibility, other) and disease name.
gnomad211_exome: gnomAD exome collection (v2.1.1), with AF, AF_popmax AF_male, AF_female, AF_raw, AF_afr, AF_sas, AF_amr, AF_eas, AF_nfe, AF_fin, AF_asj, AF_oth, non_topmed_AF_popmax, non_neuro_AF_popmax, and non_cancer_AF_popmax controls_AF_popmax headers.
popfreq_max_20150413: A database containing the maximum allele frequency from 1000G, ESP6500, ExAC, and CG46.
exac03nontcga: ExAC on non-TCGA samples. ExAC 65000 exome allele frequency data for ALL, AFR (African), AMR (Admixed American), EAS (East Asian), FIN (Finnish), NFE (Non-finnish European), OTH (other), SAS (South Asian). version 0.3.
Revel: REVEL scores for non-synonymous variants.
dbnsfp33a: The dataset includes SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, MetaSVM, MetaLR, VEST, M-CAP, CADD, GERP++, DANN, fathmm-MKL, Eigen, GenoCanyon, fitCons, PhyloP and SiPhy scores from dbNSFP version 3.3a for WES variants.
regsnpintron: prioritize the disease-causing probability of intronic SNVs.
dbscsnv11: dbscSNV version 1.1 for splice site prediction by AdaBoost and Random Forest, which score how likely that the variant may affect splicing.

The -operation argument specifies which operations to use for each of the protocols: g means gene-based, f means filter-based.

3.3. Variant Filtration

Over 20.000 annotated variants are filtered, taking into consideration many parameters, as shown in Figure 3.17. Variants are evaluated in three categories separately: de novo heterozygous, homozygous and compound heterozygous variants. Common variants with high Minor Allele Frequency (MAF) are filtered at the first step of variant filtration by utilizing several population databases. Following that, variants are restricted to the exonic regions and splice-site junctions. Then, variants are interpreted with extensive database and literature review considering clinical relevance, inheritance pattern, family segregation, disease mechanism, and variant level evidence (for example; evolutionary conservation, computational prediction) and following the guidelines determined by American College of Medical Genetics and Genomics (ACMG) and (79). Details about filtration steps will be explained in their corresponding subsections.

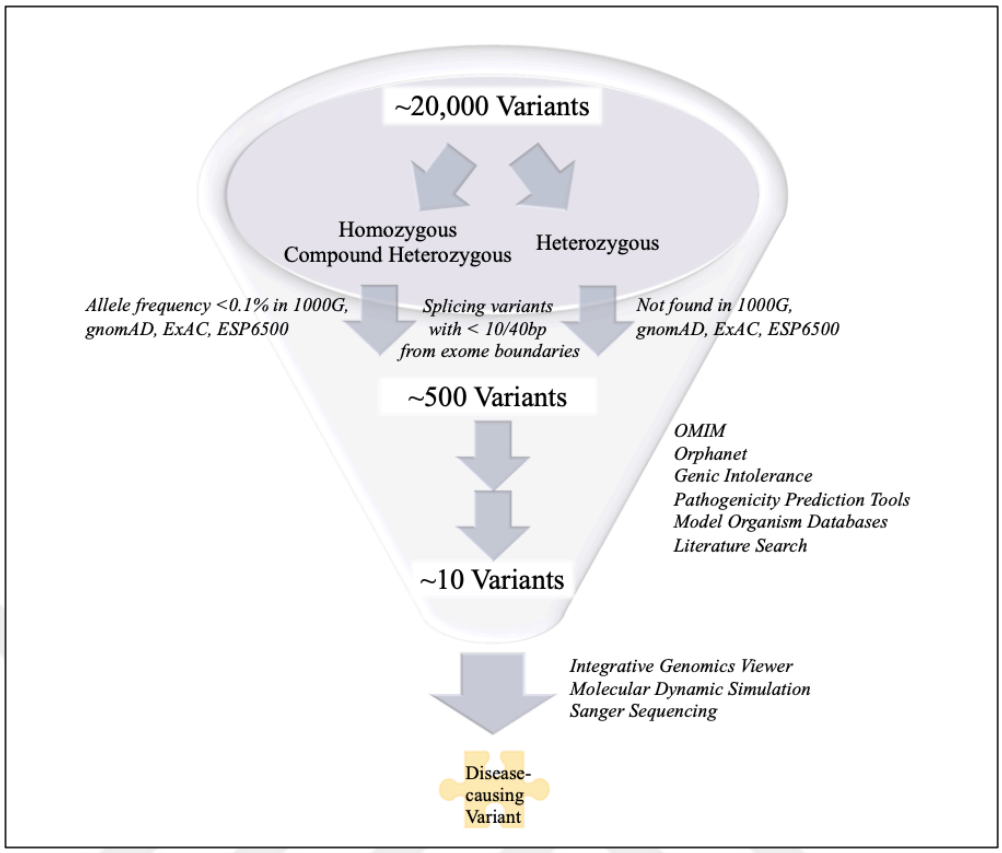


Figure 3.17. An overview of the variant filtration.

3.3.1. Variant frequency filtering

Pathogenic variants have lower allele frequencies than benign variants because of negative selection (80). In the global population database, except for the well-known founder alleles, >5% MAF can be considered as benign (79). A common variant is also not expected to be the cause of RDs. Therefore, allele frequencies are often used as a criterion for predicting pathogenicity. This is usually achieved according to two different approaches. The first approach, called discrete filtering, assumes that a disease-causing variant should not found in these databases (11,81). This approach was used for heterozygous variants. The second approach, called 1%-approach, is based on allele frequency cut-offs that change according to the inheritance model of variants. It is recommended that the threshold of MAF can be set at 1% for the analysis

of autosomal recessive variants in this approach (11). However, Kobayashi et al. showed that 97.3% of pathogenic variants they examined using the Exome Aggregation Consortium (ExAC) dataset have MAF < 0.01% (82). It suggests that a low allele frequency threshold can be adopted to interpret sequence variations, which are in the possible relation with Mendelian diseases. Hence, the MAF threshold was set at 0.1% for homozygous and compound heterozygous variants.

Based on the assumption that variants found in the population with high MAF are not likely to be the cause of RDs, such variants are filtered out by using population databases explained below.

3.3.1.1. 1000 Genome Project Databases 1KGP database

1KGP database provides a comprehensive set of human genetic variations from a diverse set of individuals of several populations. The database includes the reconstructed genomes of 2,504 individuals from 26 populations obtained by combining low-coverage whole-genome sequencing (WGS), WES, and microarray genotyping. The database contains over 88 million variants, consisting of around 84,7 million SNPs and 3,6 million indels (83).

3.3.1.2. The Genome Aggregation Database (gnomAD)

gnomAD is an extensive collection of WES and WGS data from several large-scale sequencing projects. The first release of gnomAD is also known as the ExAC dataset (84). gnomAD short variant v2 release contains 125,748 WES data, and 15,708 WGS data aligned to the GRCh37/hg19 reference sequence. In contrast, the short variant v3 release contains 71,702 WGS data, including most of the WGS data from the v2

version mapped to the GRCh38 reference sequence. Therefore, gnomAD v2 provides higher power for the analysis of the coding regions, while v3 offers a valuable resource for the study of non-coding sequences (85).

The gnomAD dataset was used to calculate metrics describing tolerance to variation for genes in the human genome. The metrics are also available in the constraint table displayed on the gnomAD browser, as shown in Figure 3.18.

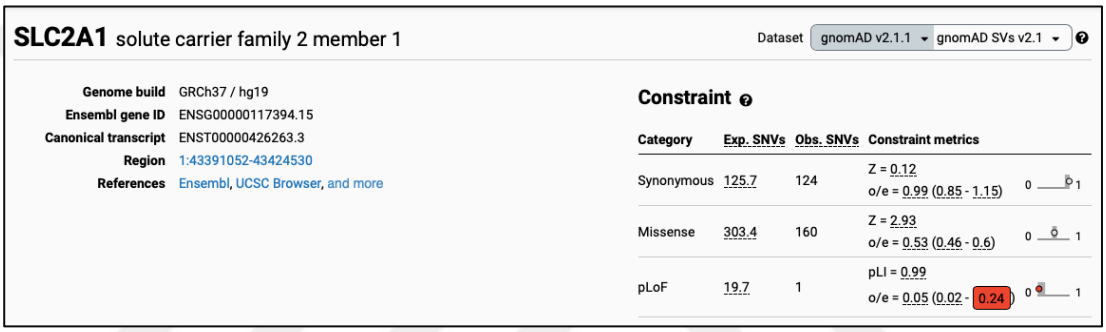


Figure 3.18. The constraint table for SLC2A1 gene.

The metrics are developed to measure the intolerance of a transcript to variation by predicting the number of expected variants that are seen in the gnomAD dataset and comparing those expectations to the number of observed mutations. For synonymous and non-synonymous variations, Z score was created to differentiate observed counts from the expected ones. While positive Z scores mean the transcript is more intolerant of variation, transcripts with more variants than expected have negative Z scores. For protein-truncating changes, it is assumed that there are three categories of genes: null, recessive, and haploinsufficient. The observed and expected variant counts are used to determine the probability of intolerance of loss-of-function variation, which falls into the category of haploinsufficient genes. The likelihood of being loss-of-function intolerant (pLI) scores ≥ 0.9 indicates an extreme intolerance (84). However, it is important to note that the pLI scores developed with the ExAC dataset have been shifted to the observed/expected (o/e) score in gnomAD. Because of the lower

sampling in ExAC, the calculation needs to a transformation of the observed and expected values for the quantity of loss-of-function variants. The revised model with gnomAD includes an increased sample which enables to evaluate more accurately the level of intolerance to loss-of-function variation in each gene. A gene with a low o/e ratio is under a more robust selection for that class of variation than a gene with a higher value (85).

3.3.2. Distance from splicing regions

The majority of pathogenic splice site mutations in the literature are located on ± ten bases far away from the exon/intron boundary. The study performed on 1,059 WES cases by Bergant et al. showed that among all pathogenic splice-site mutations, the majority of them were seen at positions +4 and +5 (63.6%) and ± one and two contributing to 1.2%. Two splice variants were seen at positions −3, −12, and two synonymous variants were predicted to affect splicing (21). Hence, the threshold of splice site distance determined as ± ten base-pair. If any prominent variant is not detected, the length is increased to ± 40 base-pair.

3.3.3. Pathogenicity prediction tools

Even splice distance and population MAF-based filtering, individuals generally have many novel variants that are not reported in databases. According to criteria proposed by some clinical guidelines, most of these variants do not classify definitively as benign or pathogenic. These types of alterations are called as variants of uncertain significance (VUS) (79). Further filtering approaches must use to lower the quantity of VUS as much as possible. For this purpose, numerous pathogenicity prediction tools based on different principles have been developed to evaluate the

variant effect. ACMG (79) and the European Society of Human Genetics (ESHG) (86) guidelines also recommend these in-silico methods to interpret variant pathogenicity.

The methods are based on different principles to predict variant pathogenicity. Evolutionary conservation is among the most beneficial features of such predictions. Some methods, such as SIFT (87) and PROVEAN (88), rely on sequence conservation. For example, as the most widely used algorithm, SIFT compares the alignments of related sequences by performing a PSI-BLAST (Position-Specific Iterated BLAST) (89) search to check if the variant is tolerated in an evolutionary aspect. In addition to sequence conservation, another group of methods that take into account several features such as amino acid physicochemical properties, the context of variation position, and protein structural features through machine learning algorithms are also available. Combined annotation– dependent depletion (CADD) (90), MutationTaster2 (91), PolyPhen-2 (92), DANN (93), and VEST3 (94) are well-known examples of such tools.

The predicted impact of a variation obtained from different tools may not be the same. This problem led to researchers making efforts to develop meta predictors that combine the results from existing tools by using several approaches such as logistic regression, decision trees, random forests, and support vector machines to make their own decisions. MetaSVM and MetaLR (95), The Mendelian clinically applicable pathogenicity (M-CAP) (96), and Rare exome variant ensemble learner (REVEL) (97) are well-known examples of meta-predictors.

Many tools were used for pathogenicity prediction in the dissertation. Below, some background information will be presented about the more inclusive and informative ones.

3.3.3.1. CADD

CADD combines 63 genomic features derived from evolutionary constraint, surrounding sequence context, and functional predictions to evaluate SNVs and short indels. The tool integrates all of these features into a single CADD score. It employs a machine learning approach trained on a binary distinction between simulated variants and variants that have become fixed in human populations since the split between humans and chimpanzees. C scores correlate with pathogenicity of a variant and disease severity (90). Based on the observation on training sets used in CADD, pathogenicity thresholds are set up at different levels. C- score >10 means the 10% most deleterious substitutions, and C-score >20 indicates the 1% most deleterious substitutions.

3.3.3.2. M-CAP

M-CAP uses a supervised learning classifier to interpret genomic variants and especially focus on coding mutations for Mendelian diseases. As a meta-predictor, it uses nine existing tools SIFT (87), PolyPhen-2 (92), CADD (90), MutationTaster (91), MutationAssessor (98), FATHMM (99), LRT (100), MetaLR and MetaSVM (95). It also combines many knowledge came from of genomic and conservation information, from Residual Variation Intolerance Score (RVIS) (101), PhyloP (102), PhastCons (103), SIPHY (104), GERP (105), PAM250 and BLOSUM62 (106). Additionally, M-CAP establishes multiple sequence alignments of 99 primate, mammalian, and vertebrate genomes to the human genome as a new feature (96). The threshold for M-CAP to estimate variant pathogenicity is > 0.025.

3.3.3.3. REVEL

REVEL is an ensemble tool for pathogenicity prediction of the missense variants combining of several tools: MutPred (107), SIFT (87), PROVEAN (88), MutationTaster (91), Poly-Phen (92), VEST (94), MutationAssessor (98), FATHMM (99), LRT (100), phyloP (102), phastCons (103), SiPhy (104), and GERP (105). REVEL was trained with a random forest on the set of variants that have recently reported as pathogenic and rare benign missense variants. The prediction score for variants can range from zero to one in REVEL, reflecting the proportion of trees in the random forest that classified the variant as pathogenic (97). The pathogenicity threshold can be set either as 0,50 or 0,75 to predict pathogenicity. While 75,4% of disease mutations, 10,9% of neutral variants have a score above 0,5; 52,1% of disease mutations, 3,3% of neutral variants, and 4,1% of all missense variants have a score above 0,75.

3.3.4. Genic intolerance

Genic intolerance is a gene-level assessment that has the potential to help in the interpretation of human genetic variants. It has been developed as a scoring system to calculate tolerance of genes to a functional genetic variation using allele frequency information from the 6503 WES data that is available in the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (101). This system predicts the expected common functional variations in the gene and compares them to benign variations found in the gene. The deviation from this prediction is attributed to the intolerance score, namely RVIS. While genes with a positive RVIS score have more common functional variation than expected, genes with negative RVIS scores have less. A negative RVIS score indicates that the gene is intolerant. The scoring system also shows that the genes that cause Mendelian diseases are more strict to variations than genes that cause other diseases.

3.3.5. Model organism databases

The evolutionary conservation of many biological processes among species allows the usage of several different model organisms to study human diseases. Although not all the human genes are conserved in invertebrate models such as worms and fruit flies, vertebrate models such as zebrafish and mouse provide valuable resources to study such genes. So that they serve as a valuable resource during the variant prioritization process. When evaluating the function of a conserved gene in model organisms, it is critical to keep in mind that orthologous genes usually cause different phenotypes in different species, although the gene products have a similar molecular function. The model organism database used in the dissertation will be explained below to provide the related information on the molecular function of query genes.

3.3.5.1. Mouse Genome Informatics (MGI)

MGI is the primary database that integrates genetic, genomic, and biological data for the laboratory mouse. Mouse Genome Database (MGD) and Mouse Gene Expression Database (GXD) are the two most significant contributors to MGI, serving as valuable resources for human disease studies. MGD provides curated phenotypes and functional annotations for mouse genes and alleles, while GXD contains mouse gene expression data with an emphasis on endogenous gene expression during mouse development (108,109). The Human-Mouse Disease Connection tool within MGI is another important feature that facilitates exploring gene-phenotype-disease relationships between human and mouse. By simply searching the human genes on MGI, the algorithm finds matching mouse genes and their homologs and displays the both human and mouse phenotypes associated with the genes of interest. It is shown an example search for the SLC2A1 gene in Figure 3.19. MGI is updated once every week by adding new annotations from the literature.

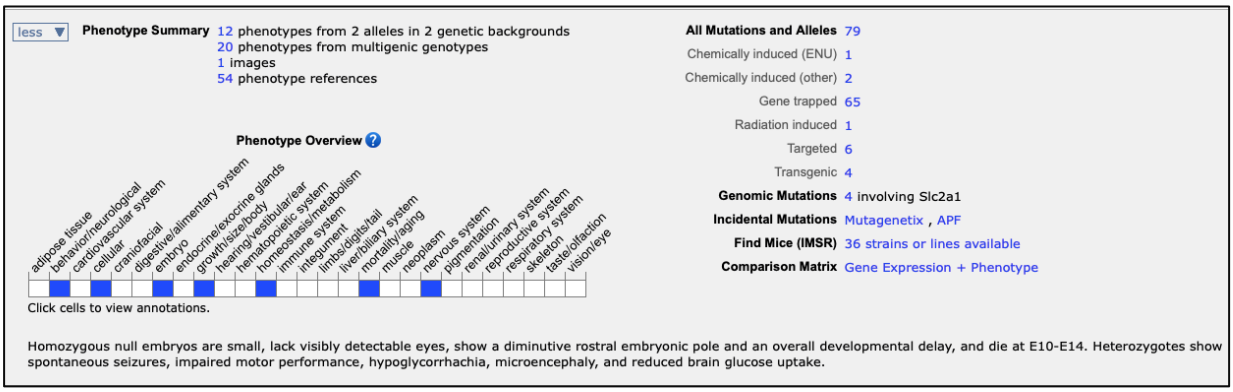


Figure 3.19. Mutant mouse phenotype for SLC2A1 gene.

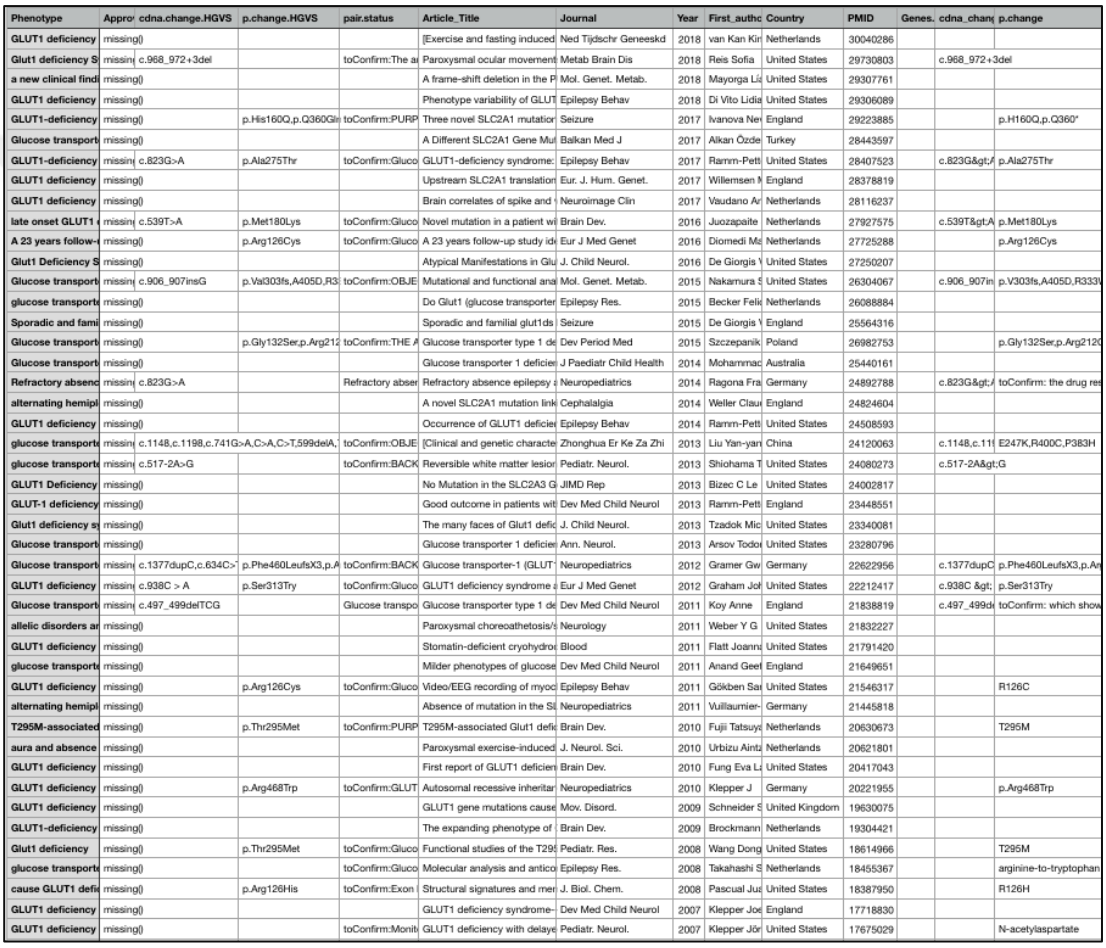
3.3.6. Literature search

The VarfromPDB package (v2.2.10) (110) available in R (111) was used to utilize deep literature search for prioritized variants. VarfromPDB is an automated method to mine the genes and variants related to a Mendelian disorder from several databases which are HPO (68), Human Gene Nomenclature Committee (HGNC) (112), ClinVar (113), Online Mendelian Inheritance in Man (OMIM) (4), Uniprot (114), Orphanet (115), UCSC (116) and literature (PubMed).

The function extract_pubmed captures the information about the keywords from abstracts in PubMed based on text mining. The functions in the R RISmed package (https://www.rdocumentation.org/packages/RISmed) are used to get and search the abstracts (110). The information such as phenotypes, genes, variants, article titles, publication journals, publication years, first authors, and PMIDs can be captured (Figure 3.20)

```
pubmed.phenotype <- extract_pubmed(query = "disease phenotype AND gene AND
mutation",
                                   keyword="disease phenotype")
disease.pubmed <- pubmed.phenotype[[1]]

write.table(disease.pubmed, file = "disease.csv", row.names = F,
                                   sep = ";")
```



Phenotype	Appro	cdna.change.HGVS	p.change.HGVS	pair.status	Article_Title	Journal	Year	First_autho	Country	PMID	Genes.	cdna_chan	p.change
GLUT1 deficiency	missing()				[Exercise and fasting induced	Ned Tijdschr Geneeskd	2018	van Kan Kir	Netherlands	30040286			
Glut1 deficiency S	missin	c.968_972+3del		toConfirm:The a	Paroxysmal ocular movement	Metab Brain Dis	2018	Reis Sofia	United States	29730803		c.968_972+3del	
a new clinical findi	missing()				A frame-shift deletion in the P	Mol. Genet. Metab.	2018	Mayorga Lá	United States	29307761			
GLUT1 deficiency	missing()				Phenotype variability of GLUT	Epilepsy Behav	2018	Di Vito Lidia	United States	29306089			
GLUT1-deficiency	missing()		p.His160Q,p.Q360G	toConfirm:PURP	Three novel SLC2A1 mutation	Seizure	2017	Ivanova Ne	England	29223885			p.H160Q,p.Q360*
Glucose transporti	missing()				A Different SLC2A1 Gene Mu	Balkan Med J	2017	Alkan Özde	Turkey	28443597			
GLUT1-deficiency	missin	c.823G>A	p.Ala275Thr	toConfirm:Gluco	GLUT1-deficiency syndrome:	Epilepsy Behav	2017	Ramm-Petti	United States	28407523		c.823G>A	p.Ala275Thr
GLUT1 deficiency	missing()				Upstream SLC2A1 translation	Eur. J. Hum. Genet.	2017	Willemsen M	England	28378819			
GLUT1 deficiency	missing()				Brain correlates of spike and	Neuroimage Clin	2017	Vaudano Ar	Netherlands	28116237			
late onset GLUT1	missin	c.539T>A	p.Met180Lys	toConfirm:Gluco	Novel mutation in a patient wi	Brain Dev.	2016	Juozapaite	Netherlands	27927575		c.539T>A	p.Met180Lys
A 23 years follow-	missing()		p.Arg126Cys	toConfirm:Gluco	A 23 years follow-up study id	Eur J Med Genet	2016	Diomedi Ma	Netherlands	27725288			p.Arg126Cys
Glut1 Deficiency S	missing()				Atypical Manifestations in Glu	J. Child Neurol.	2016	De Giorgis V	United States	27250207			
Glucose transporti	missin	c.906_907insG	p.Val303fs,A405D,R3	toConfirm:OBJE	Mutational and functional ana	Mol. Genet. Metab.	2015	Nakamura S	United States	26304067		c.906_907in	p.V303fs,A405D,R333V
glucose transporte	missing()				Do Glut1 (glucose transporter	Epilepsy Res.	2015	Becker Felix	Netherlands	26088884			
Sporadic and fami	missing()				Sporadic and familial glut1ds	Seizure	2015	De Giorgis V	England	25564316			
Glucose transporti	missing()		p.Gly132Ser,p.Arg212	toConfirm:THE A	Glucose transporter type 1 de	Dev Period Med	2015	Szczepanik	Poland	26982753			p.Gly132Ser,p.Arg212C
Glucose transporti	missing()				Glucose transporter 1 deficie	J Paediatr Child Health	2014	Mohammad	Australia	25440161			
Refractory absenc	missin	c.823G>A		Refractory abser	Refractory absence epilepsy a	Neuropediatrics	2014	Ragona Fra	Germany	24892788		c.823G>A	toConfirm: the drug res
alternating hemipl	missing()				A novel SLC2A1 mutation link	Cephalalgia	2014	Weller Clau	England	24824604			
GLUT1 deficiency	missing()				Occurrence of GLUT1 deficie	Epilepsy Behav	2014	Ramm-Petti	United States	24508593			
glucose transporte	missin	c.1148,c.1198,c.741G>A,C>A,C>T,599delA,		toConfirm:OBJE	[Clinical and genetic characte	Zhonghua Er Ke Za Zhi	2013	Liu Yan-yan	China	24120063		c.1148,c.11	E247K,R400C,P383H
glucose transporte	missin	c.517-2A>G		toConfirm:BACK	Reversible white matter lesion	Pediatr. Neurol.	2013	Shiohama T	United States	24080273		c.517-2A>G	
GLUT1 Deficiency	missing()				No Mutation in the SLC2A3 G	JIMD Rep	2013	Bizec C Le	United States	24002817			
GLUT-1 deficiency	missing()				Good outcome in patients wit	Dev Med Child Neurol	2013	Ramm-Petti	England	23448551			
Glut1 deficiency s	missing()				The many faces of Glut1 defic	J. Child Neurol.	2013	Tzadok Mic	United States	23340081			
Glucose transporti	missing()				Glucose transporter 1 deficie	Ann. Neurol.	2013	Arsov Todor	United States	23280796			
Glucose transporti	missin	c.1377dupC,c.634C>T	p.Phe460LeufsX3,p.A	toConfirm:BACK	Glucose transporter-1 (GLUT	Neuropediatrics	2012	Gramer Gw	Germany	22622956		c.1377dupC	p.Phe460LeufsX3,p.Ar
GLUT1 deficiency	missin	c.938C > A	p.Ser313Try	toConfirm:Gluco	GLUT1 deficiency syndrome a	Eur J Med Genet	2012	Graham Jol	United States	22212417		c.938C >	p.Ser313Try
Glucose transporti	missin	c.497_499delTCG		Glucose transpo	Glucose transporter type 1 de	Dev Med Child Neurol	2011	Koy Anne	England	21838819		c.497_499d	toConfirm: which show
allelic disorders ar	missing()				Paroxysmal choreoathetosis/s	Neurology	2011	Weber Y G	United States	21832227			
GLUT1 deficiency	missing()				Stomatin-deficient cryohydro	Blood	2011	Flatt Joanna	United States	21791420			
glucose transporte	missing()				Milder phenotypes of glucose	Dev Med Child Neurol	2011	Anand Geet	England	21649651			
GLUT1 deficiency	missing()		p.Arg126Cys	toConfirm:Gluco	Video/EEG recording of myoc	Epilepsy Behav	2011	Gökben Sar	United States	21546317			R126C
alternating hemipl	missing()				Absence of mutation in the SL	Neuropediatrics	2011	Vuillaumier-	Germany	21445818			
T295M-associated	missing()		p.Thr295Met	toConfirm:PURP	T295M-associated Glut1 defic	Brain Dev.	2010	Fujii Tatsuya	Netherlands	20630673			T295M
aura and absence	missing()				Paroxysmal exercise-induced	J. Neurol. Sci.	2010	Urbizu Aintz	Netherlands	20621801			
GLUT1 deficiency	missing()				First report of GLUT1 deficien	Brain Dev.	2010	Fung Eva La	United States	20417043			
GLUT1 deficiency	missing()		p.Arg468Trp	toConfirm:GLUT	Autosomal recessive inheritan	Neuropediatrics	2010	Klepper J	Germany	20221955			p.Arg468Trp
GLUT1 deficiency	missing()				GLUT1 gene mutations cause	Mov. Disord.	2009	Schneider S	United Kingdom	19630075			
GLUT1-deficiency	missing()				The expanding phenotype of	Brain Dev.	2009	Brockmann	Netherlands	19304421			
Glut1 deficiency	missing()		p.Thr295Met	toConfirm:Gluco	Functional studies of the T29	Pediatr. Res.	2008	Wang Dong	United States	18614966			T295M
glucose transporte	missing()			toConfirm:Gluco	Molecular analysis and antico	Epilepsy Res.	2008	Takahashi S	Netherlands	18455367			arginine-to-tryptophan
cause GLUT1 defi	missing()		p.Arg126His	toConfirm:Exon	Structural signatures and mec	J. Biol. Chem.	2008	Pascual Jua	United States	18387950			R126H
GLUT1 deficiency	missing()				GLUT1 deficiency syndrome-	Dev Med Child Neurol	2007	Klepper Joe	England	17718830			
GLUT1 deficiency	missing()			toConfirm:Monit	GLUT1 deficiency with delaye	Pediatr. Neurol.	2007	Klepper Jör	United States	17675029			N-acetylaspartate

Figure 3.20. The output file which is generated by using extract_pubmed function in the VarfromPDB package for GLUT1 deficiency syndrome.

3.4. Variant Confirmation

When WES were integrated into the clinical area for diagnostic purposes, Sanger sequencing has become the confirmation method of NGS results (12–14,117). However, with advances in sequencing technologies and methods for bioinformatics analyses, the accuracy of NGS has parallelly improved. Therefore, laboratories working with NGS data are able to empirically determine quality thresholds for variants without Sanger confirmation (118). In the example study reported by Strom et al., Sanger sequencing was performed on 110 SNVs detected by WES. WES findings were validated for 100% of SNVs with quality scores ≥Q500. The mean coverage for these variants was 116x and ranged from 5x–250x. In the remaining seven variants with quality scores <Q500, six were confirmed by Sanger sequencing (119). Based on this study, they have set a quality threshold of Q500 with approximately 40x coverage for WES. These thresholds were also used in this dissertation. However, Sanger confirmation of low-quality single nucleotide variants and all small indels remains necessary.

As an alternative approach, a manual check of the variants utilizing a visualization tool such as the Integrative Genomics Viewer (IGV) (120,121) is a possible way the corroboration of variants. In addition to IGV visualization, molecular dynamics (MD) simulation provides valuable insight to confirm the variant pathogenicity via simulating the impact of a mutation to protein structure and function.

3.4.1. IGV

IGV is open-source software that enables us to visualize NGS data, including both basic mapped read data and derived results from them, such as read coverage. The tool supports the visualization of many samples and the comparison of these datasets

synchronically. Datasets can be loaded from local or remote sources. IGV takes several file formats as an input. Only .sam or .bam files can be used for sequence alignment data, and bam index files (.bai) are also required. IGV represents NGS data by a simple coverage plot (120,121). This coverage plot was used to evaluate the quality and detect technical issues in the sequencing.

3.4.2. Sanger sequencing

Sanger method described by Frederick Sanger and colleagues is based on the usage of chain-terminating dideoxynucleotide analogs that caused base-specific termination of the DNA synthesis (36). Sanger sequencing was used for the purpose of the confirmation of detected variants and determination of parental segregation for proband-based WES case. In order to perform Sanger sequencing, peripheral blood samples were obtained from the patients and their biological parents following the standard procedures. Genomic DNA was isolated from peripheral blood samples using the Quick-DNA^{TM} Miniprep Plus Kit (Zymo Research, USA) following the manufacturer's protocol. DNA quality was checked by agarose gel electrophoresis and NanoDrop 2000c Spectrophotometer (Thermo Fisher Scientific, USA). PCR primers were designed for each region of interest using Primer3 (122). These regions were amplified employing PCR and run in agarose gel electrophoresis for size and integrity analysis. Sanger sequencing was carried out using standard protocols. Sequence electropherograms were viewed using the 4Peaks (Mekentosj, Amsterdam).

3.4.3. Computational impact prediction of the mutant protein function: MD Simulation

Understanding the impact of a mutation on the three-dimensional (3D) structure and function of the protein is crucial in deducing the pathogenicity. However, studying

with proteins is required tedious and time-consuming experimental methods because of their flexible and dynamic nature, which can play a significant role in their function. Moreover, they can also undergo conformational changes while carrying out their function. Recent advances in structural genomics and computational modeling methods allow simulating the proteins (123). These computational methods enable not only predicting the structure of proteins but also determining the impact of variants on the protein (124–126). MD simulation is a powerful way to measure the impact of the variant on protein 3D structure and function. MD simulation measures the physical movements of several hundreds of atoms at a fixed period of time based on the application of Newton's equation of motion to predict the behavior of proteins (127). Thus, MD simulation makes direct observation of the motion of mutant proteins in their dynamic environment possible at the atomic scale and helps to corroborate the variant pathogenicity.

Computational modeling study for case II was performed by Umut Gerlevik from our research group. For structural modeling, Robetta (128) was used to model the full-length monomer structure of human PRKAR1A protein (UniProt ID: P10644 (114)) by using the crystal structure of mouse orthologous of PRKAR1A (PDB ID: 4DIN , chain B (129)) as a template. Ramachandran plot analysis via PROCHECK (130) was applied to check the quality of the model. Then, to observe the mutation impact on the cAMP binding ability of the regulatory subunit of PKA holoenzyme, two cAMP were docked to the binding regions, as shown in the crystal structure of the bovine orthologous of PRKAR1A protein (PDB ID: 4MX3 (131)). G171E mutation was introduced with the help of Mutator Plugin (v1.3) of Visual Molecular Dynamics (VMD- v1.9.3) (132). The systems were solvated in the TIP3P water box (133) with 0.15 M KCl ions as neutralizing the charge.

In the simulation setup, NAMD 2.13-multicore-CUDA (134) were used to run MD simulations. CHARMM36 all-atom force field (135) was used. After applying 10,000 step minimizations with a conjugate gradient algorithm, the systems were equilibrated

for 2 ns at 298 K under the NVT ensemble. Production simulations were performed along 100 ns at 310 K and 1 atm under the NPT ensemble. For both wild-type and mutant systems, three different productions were performed with the different random seeds to ensure the reproducibility of results. Langevin barostat and thermostat couplings were used for pressure and temperature controls (136). The integration time step was 2 fs. For the computation of long-range Coulomb interactions, the particle-mash Ewald method was used (137).

For trajectory analysis, visualizations and rendering were applied by using VMD (132). In-house TCL scripts were used for aligned backbone root-mean-square deviation (RMSD), aligned Cα atoms root-mean-square fluctuations (RMSF), and radius of gyration of the entire structure. The ggplot2 package (138) in R 4.0.2 (111) was used to plot the results of analyses.

4. RESULTS

In this dissertation, a unique variant prioritization workflow consisting of the WES data analysis pipeline and variant filtering strategy has been employed on previously unsolved WES cases. While two cases were proband-only WES, one of them trio-WES. The workflow created in this dissertation achieved a precise diagnosis by reanalyzing the raw data of all individuals. The diagnosis and details of each case will be explained in their corresponding subsections.

4.1. Case I: GLUT1 Deficiency Syndrome 1

Case I was a 6-year-old female who had symptoms, including global developmental delay, intellectual disability, dysarthria, epilepsy, hemangioma, and involuntary movements (choreoathetosis and dystonia). WES was performed before, but no pathogenic variant has been able to detect. The raw FASTQ file was immediately obtained. At the end of the analysis of initial nondiagnostic WES, 10,800 heterozygous, and 5,500 homozygous passed the depth and quality filter. After the population frequency and splice site filtering, quantity decreased to 402 and 15, respectively. After further filtering described in the method section, three genes were prioritized namely SLC2A1, KMT2C, and CACNA1B (Table 4.1.)

Table 4.1. Summary of prioritized variants for case I.

Gene and Variant	Phenotype	gnomAD Gene constraint	Intolerance Score (RVIS)	CADD REVEL M-CAP	MGI Phenotype
SLC2A1 (OMIM *138140) NM_006516: exon3: c.275+1G>-	GLUT1 deficiency syndrome 1, infantile onset, severe (OMIM #606777)	Missense Z Score = 2.93 Loss of Function pLI Score = 0.99	-1.05	- - -	Heterozygotes embryos show seizures, impaired motor performance, hypoglycorrhachia, microencephaly, and reduced brain glucose uptake. (MGI:95755)
KMT2C (OMIM *606833) NM_170606: exon18: c.C2961G: p.Y987X	Kleefstra syndrome 2 (OMIM #617768)	Missense Z Score = 2.14 Loss of Function pLI Score = 1	-2.52	11.576 0.162 -	Mice homozygous for a knock-out allele display partial embryonic lethality, postnatal growth retardation, and impaired fertility. (MGI:2444959)
CACNA1B (OMIM *601012) NM_000718: exon2: c.390+1->ACGACACGGAGCCCTATTTCATCGGGATCTTTTGCTTCGAGGCAGGGATCAAAA	Neuro-developmental disorder with seizures and nonepileptic hyperkinetic movements (OMIM #618497)	Missense Z Score = 4.52 Loss of Function pLI Score = 1	NA	- - -	Mice deficient in this gene exhibit defects in nociception, memory and learning, hyperactive and aggressive behaviors with defects in the the sleep-wake cycle. (MGI:88296)

Together, these statistical findings, prediction scores, and phenotype information in both human and the model organism provided strong evidence that the novel variant in SLC2A1 causes the observed phenotype. The heterozygous mutation in SLC2A1 (NM_006516) c.275+1G>- was detected with high-depth reading in IGV (Figure 21).

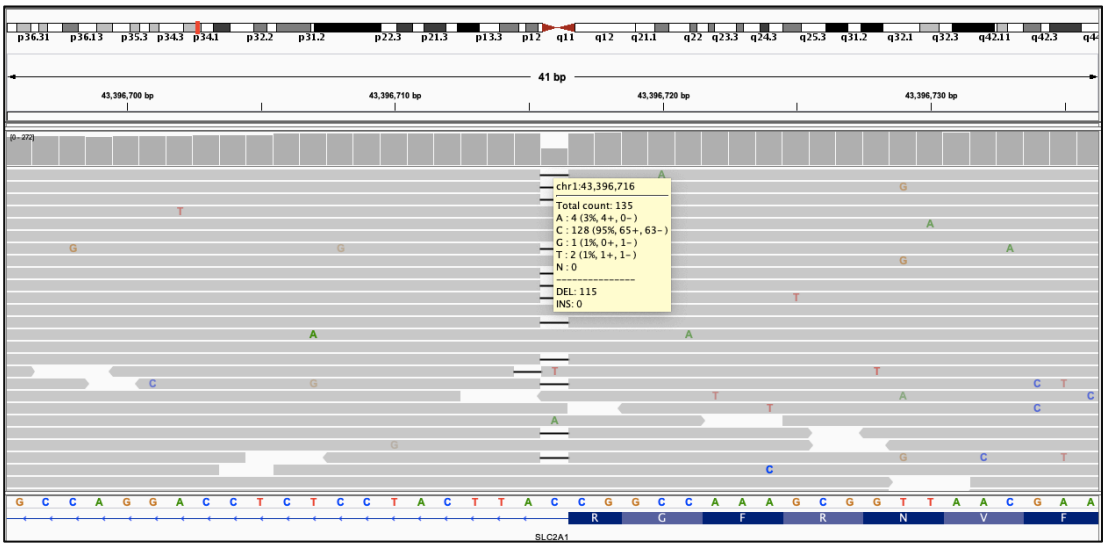


Figure 4.1. IGV visualization of the variant. At the position of 43396716 on Chromosome 1, there are 128 reads for reference (C) and 115 deletions.

Since the patient was a proband-only WES case, it was needed to confirm whether there is parental segregation of the variant or not. Sanger sequencing was carried out to verify parental segregation. In addition to the parental verification, mixed and low-quality traces from the heterozygous deletion point in both forward and reverse sequence direction confirm the IGV visualization (Figure 4.2).

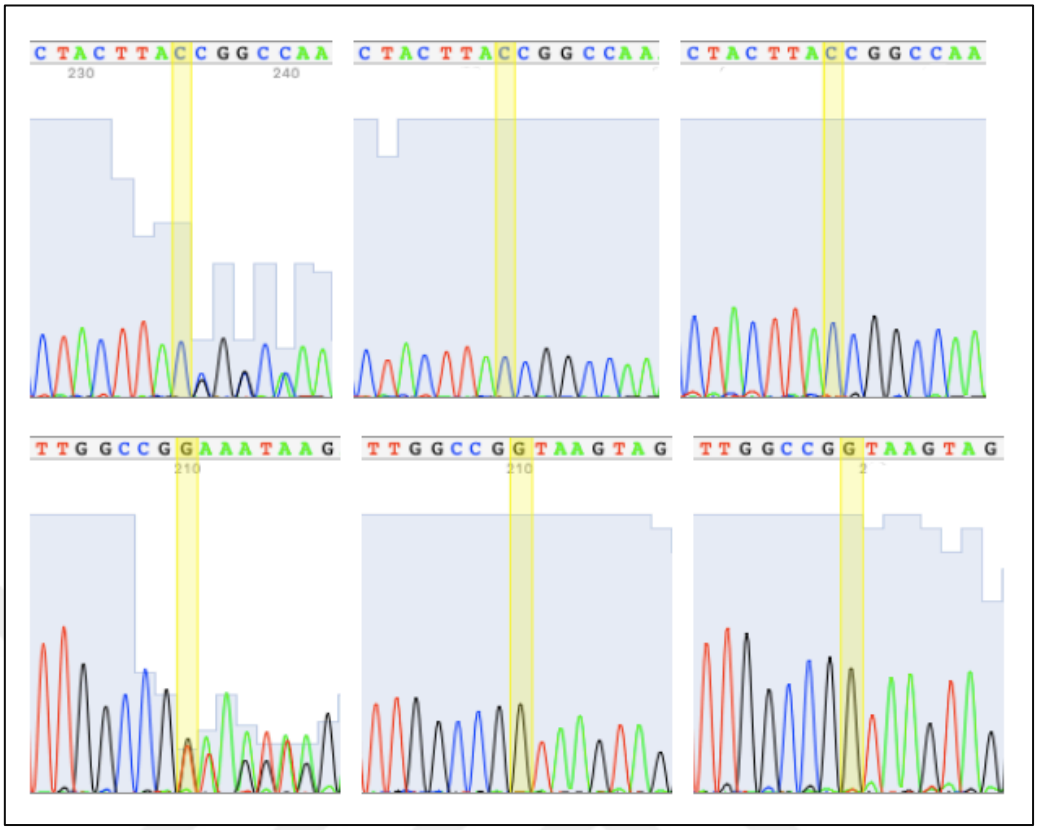


Figure 4.2. Forward and reverse reads from obtained Sanger Sequencing of the patient and her biological parents. While the upper reads come from the forward sequencing, the lower ones come from to reverse sequencing. The reads in the first column belong to the affected child. The second and the third columns belong to the mother and the father, respectively.

The SLC2A1 gene is located on the short arm of chromosome 1, 1p34.2 (139). The gene with ten exons encodes GLUT1 protein composed of 492 amino acids. (140,141). GLUT1 protein is the main glucose transporter responsible for the delivery of glucose in human erythrocytes and blood-tissue barriers (116–118).

Impaired glucose transport at the blood-brain barrier caused by mutations in the SLC2A1 lead to GLUT1DS1 (28). GLUT1DS1 is a rare neurometabolic disorder with the estimated prevalence of 1:80-90,000 (145) and characterized by hypoglycorrhachia, early-onset seizures, delayed development, dysarthria, acquired

microcephaly and movement disorder including spasticity, ataxia, and dystonia (29,146).

Most of the SLC2A1 mutations resulted in GLUT1DS1 occur de-novo with the autosomal dominant condition (147). In familial cases, mutations are usually inherited by autosomal dominant pattern (148,149); however, the presence of autosomal recessive variants have also been shown (150,151). Thus far, it is estimated that more than 140 different pathogenic variants have been shown in the SLC2A1 (152). All mutations, including deletions, missense, nonsense, frameshift, and splice-site variations result in absence or loss of function of one of the SLC2A1 alleles (147). Generally, missense changes can be associated with the whole symptom spectrum of GLUT1DS1, while deletions and null mutations in the SLC2A1 are related to severe forms of the syndrome (153). Yet, genotype-phenotype correlation is complex, and it is not yet clearly defined. On the other hand, it is vital to identify GLUT1DS1 in the early stages since the ketogenic diet therapy is able to provide a reduction in symptoms in GLUT1DS1 patients (154).

4.2. Case II: Acrodysostosis 1, with or without hormone resistance

Case II was a 4-years-old girl with distinctive facial features, skeletal system abnormalities, global developmental delay, and hypothyroidism. WES was performed before, but no pathogenic variant was detected. Raw FASTQ file was obtained for reanalysis. At the end of the WES data analysis, 13741 heterozygous and 9403 homozygous variants were detected. After the population frequency and splice site filtering, the number of variants decreased to 326 and 103 for heterozygous and homozygous, respectively. After further filtering approaches defined in the method section, six genes were highlighted, namely NEB, TRIP12, CHD1, PRKAR1A, RERE, KMT2D (Table 4.2).

Table 4.2. Summary of prioritized variants for case II.

Gene and Variant	Phenotype	gnomAD Gene constraint	Intolerance Score (RVIS)	CADD REVEL M-CAP	MGI Phenotype
NEB (OMIM *161650) NM_004543: exon77: c.T11437G: p.L3813V	Nemaline myopathy 2, autosomal recessive (OMIM # 256030)	Missense Z Score = - 0.04 Loss of Function pLI Score = 0	0.88	-0.520 0.201 0.005	Homozygous inactivation of this gene leads to stunted growth, altered sarcomere structure, reduced contractility in skeletal muscle, muscle weakness. (MGI:97292)
TRIP12 (OMIM *604506) NM_0013483 1: exon22: c.3356+13G>T	Mental retardation, autosomal dominant 49 (OMIM #617752)	Missense Z Score = 4.64 Loss of Function pLI Score = 1	-2.46	- - -	Mice homozygous for a targeted allele exhibit complete embryonic lethality during organogenesis. (MGI:1309481)
CHD1 (OMIM *602118) NM_001270: exon18: c.2718+1G>T	Pilarowski-Bjornsson syndrome (OMIM #617682)	Missense Z Score = 4.21 Loss of Function pLI Score = 1	-0.81	- - -	Mice homozygous for a knock-out allele exhibit complete embryonic lethality. (MGI:88393)
PRKAR1A (OMIM *188830) NM_0012762 9: exon5: c.G512A: p.G171E	Acrodysostosis 1, with or without hormone resistance (OMIM # 101800)	Missense Z Score = 3.12 Loss of Function pLI Score = 1	-0.34	6.879 0.953 0.628	Mice homozygous for a null allele exhibit embryonic lethality during organogenesis. Mice heterozygous for a null allele exhibit background sensitive infertility and increased tumor incidence. (MGI:104878)
RERE (OMIM *605226) NM_0010426 8: exon8: c.C727T: p.H243Y	Neurodevelopmental disorder with or without anomalies of the brain, eye, or heart (OMIM # 616975)	Missense Z Score = 2.03 Loss of Function pLI Score = 1	-2.63	-0.036 0.146 0.034	Mice homozygous for disruptions in this gene display embryonic lethality with abnormalities in neural tube development, somite development, and in the embryonic heart. (MGI:2683486)
KMT2D (OMIM *602113) NM_003482: exon51: c.16053-7C>A	Kabuki syndrome 1 (OMIM #147920)	Missense Z Score = 3.73 Loss of Function pLI Score = 1	-5.29	- - -	Mice homozygous for a gene trap allele exhibit embryonic lethality. (MGI:2682319)

Combining the statistical findings, prediction scores, and phenotype information both in human and the model organism gave strong evidence to conclude that the novel heterozygous variant, c.G512A, p.G171E, in PRKAR1A lead to clinical symptoms of the patient. IGV visualization indicated that the c.G512A mutation was detected with a high-depth read, and 44% heterozygous read ratio in WES analysis (Figure 4.3).

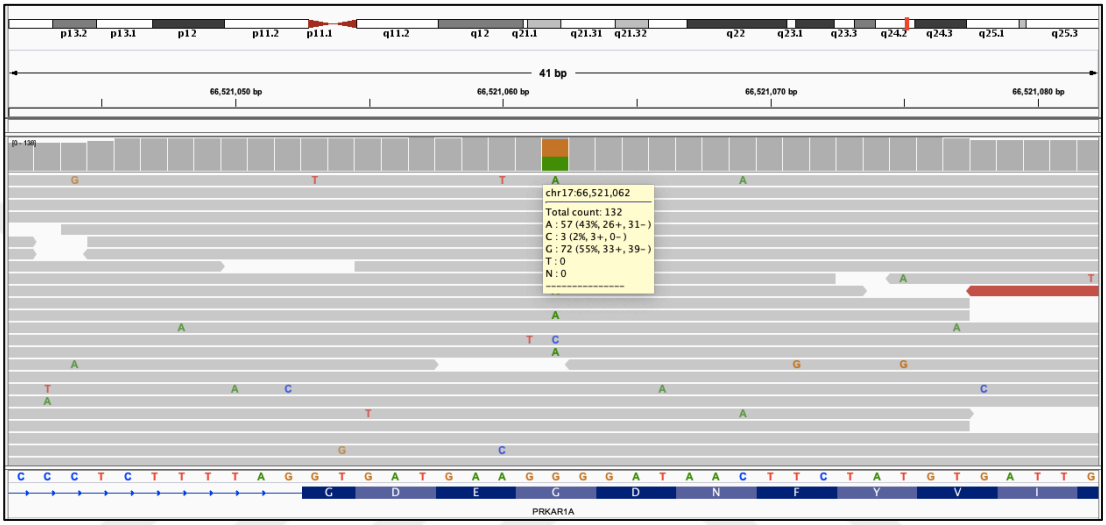


Figure 4.3. IGV visualization of the variant for Case II. At the position of 66521062 on Chromosome 17, there are 72 reads for reference (G) and 57 reads for the mutation (A).

In order to determine the degree of conservation of the affected amino acid, the PRKAR1A protein sequence of 9 species (human, mouse, rat, cattle, cow, pig, rhesus monkey, frog, chimpanzee, and Chinese hamster) was examined. Protein sequences of the species were retrieved via the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/gene/5573/ortholog/?scope=32523). Species and accession numbers are as follows: *Homo sapiens* - NP_002725; *Mus musculus* - NP_001360842; *Rattus norvegicus* - NP_037313; *Bos taurus* - NP_001069826; *Sus scrofa* - NP_999191; *Macaca mulatta* - XP_014975676; *Xenopus tropicalis* - NP_001025530; *Pan troglodytes* - XP_511647; *Cricetulus griseus* - XP_003500360. Multiple sequence alignment was conducted with Clustal Omega (155). The functional

importance of the missense mutation is supported by the fact that the region is quite conserved in several species (Figure 4.4).

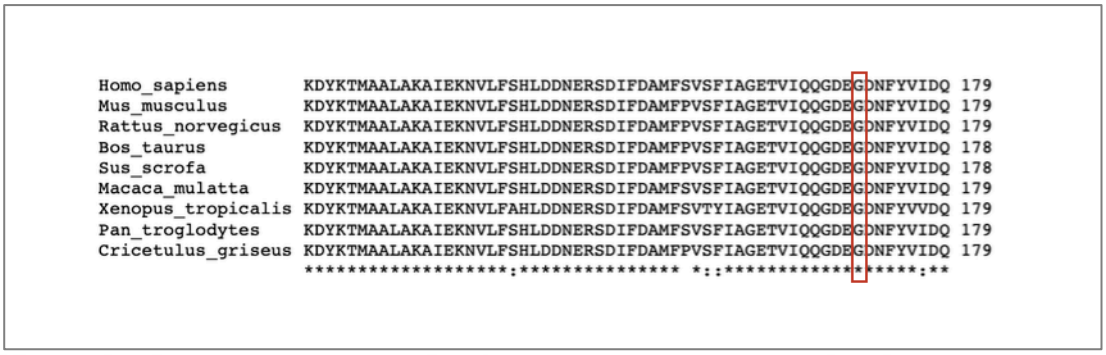


Figure 4.4. Evolutionary conservation of the position of 171 for human PRKAR1A protein.

The PRKAR1A gene is located on chromosome 17q23-q24 and encodes the cAMP-dependent regulatory subunit type I alpha of protein kinase A (PKA) comprising 381 amino acids (156). The cAMP is a seconder messenger in the cells to various hormones, and other signaling chemicals and PKA is the most common effector of cAMP (157). Receptors that activate PKA through cAMP production regulate several vital cellular processes such as metabolism, gene expression, cell proliferation, and differentiation (158). PRKAR1A protein is crucial for the regulation of the PKA system to respond to cAMP (159). In the absence of cAMP, PKA is a tetrameric holoenzyme composed of two regulatory and two catalytic subunits (158). These two subunits dissociate from each other when cAMP binds to the regulatory subunits; thus, PKA can show enzymatic activity phosphorylating downstream targets. PRKAR1A is the most abundantly expressed regulatory subunit (160) and consists of a dimerization domain, an inhibitory site, and two cAMP-binding domains called A and B (161). Heterozygous mutations that disturb the binding interaction between cAMP and PRKAR1A are associated with Acrodysostosis 1 (162). Acrodysostosis 1 refers to a rare type of skeletal dysplasias characterized by facial dysostosis, nasal hypoplasia, brachydactyly, short stature, and mental retardation (163,164). It was shown that some patients with Acrodysosostosis 1 had resistance to multiple hormones, including

thyrotropin, calcitonin, growth hormone-releasing hormone, and gonadotropin (165). According to Orphanet, less than 80 cases with Acrodysostosis have been reported to date (166).

4.3. Case III: Hypotonia, Ataxia, And Delayed Development Syndrome

Case III was a 6-year-old male with global developmental delay, generalized hypotonia, mild facial dysmorphisms including frontal bossing and low-set ears, speech delay, decreased pain response, hyperactive deep tendon reflexes, and strabismus. Trio-WES was performed before, but no pathogenic variant was detected. We immediately obtained raw data of proband and the biological parents. WES data analysis was performed on their FASTQ files. In the WES data analysis, 10011 heterozygous and 4954 homozygous and compound heterozygous variants were detected. After the population frequency and splice site filtering, the number of variants decreased to 346 for heterozygous; 15 for homozygous and compound heterozygous ones. Four genes were prioritized, namely COL6A3, EHMT1, NUTM2B, and EBF3 (Table 4.3).

Table 4.3. Summary of prioritized variants for case III.

Gene and Variant	Phenotype	gnomAD Gene constraint	Intolerance Score (RVIS)	CADD REVEL M-CAP	MGI Phenotype
COL6A3 (OMIM *120250) NM_004369: exon38: c.A8129G: p.Q2710R	Bethlem myopathy 1 (OMIM #158810) Ullrich congenital muscular dystrophy 1 (OMIM #254090)	Missense Z Score = - 0.61 Loss of Function pLI Score = 0	-2.55	0.327 0.118	Mice homozygous for a hypomorphic allele exhibit myopathy, decreased muscle weight, increased collagen deposition in muscles, skeletal muscle interstitial fibrosis and abnormal tendon collagen fibril morphology. (MGI:8846)
EHMT1 (OMIM *607001) NM_024757: exon5: c.824-18->T	Kleefstra syndrome 1 (OMIM #610253)	Missense Z Score = 1.16 Loss of Function pLI Score = 1	-1.58	- - -	Nullizygous embryos die circa E9.5 showing delayed growth and incomplete somite formation and neural groove closure. Heterozygotes show behavioral deficits and synaptic dysfunction. (MGI:1924933)
NUTM2B (OMIM *618639) NM_001278495: exon5: c.G1576A: p.E526K	Oculopharyngeal myopathy with leukoencephalopathy 1 (OMIM #618637)	Missense Z Score = 0.89 Loss of Function pLI Score = 0.02	NA	0.025 0.009 0.001	-
EBF3 (OMIM *607407) NM_001005463: exon6: c.C487T: p.R163W	Hypotonia, ataxia, and delayed development syndrome (OMIM #617330)	Missense Z Score = 3.61 Loss of Function pLI Score = 1	-0.65	33 0.575 0.009	Homozygous mutant mice die perinatally and exhibit impaired olfactory neuron projection. (MGI:894289)

At the end of the analysis of initial nondiagnostic trio-WES, de-novo heterozygous variant c.C487T, p.R163W in EBF3 gene (NM_001005463) was identified as the most

prominent variant IGV visualization verified that the variant found in only the proband, but not in the parents (Figure 4.5).

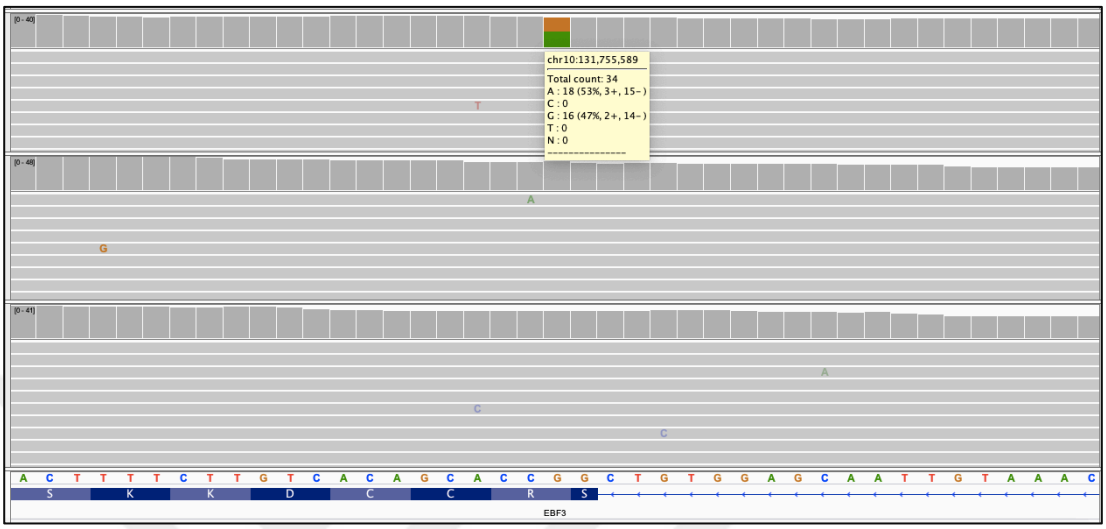


Figure 4.5. IGV visualization of the position of 131755589 on Chromosome 10 for the child, father, and mother, respectively. Affected child has 16 reads for the reference and 18 reads for the variant; the biological father and mother have no variant at this position.

The c.C487T was further confirmed by Sanger Sequencing due to low depth reading. Sanger results indicate that only the affected individual has the c.C487T as well as IGV visualization. The biological parents do not show any variation at this position (Figure 4.6).

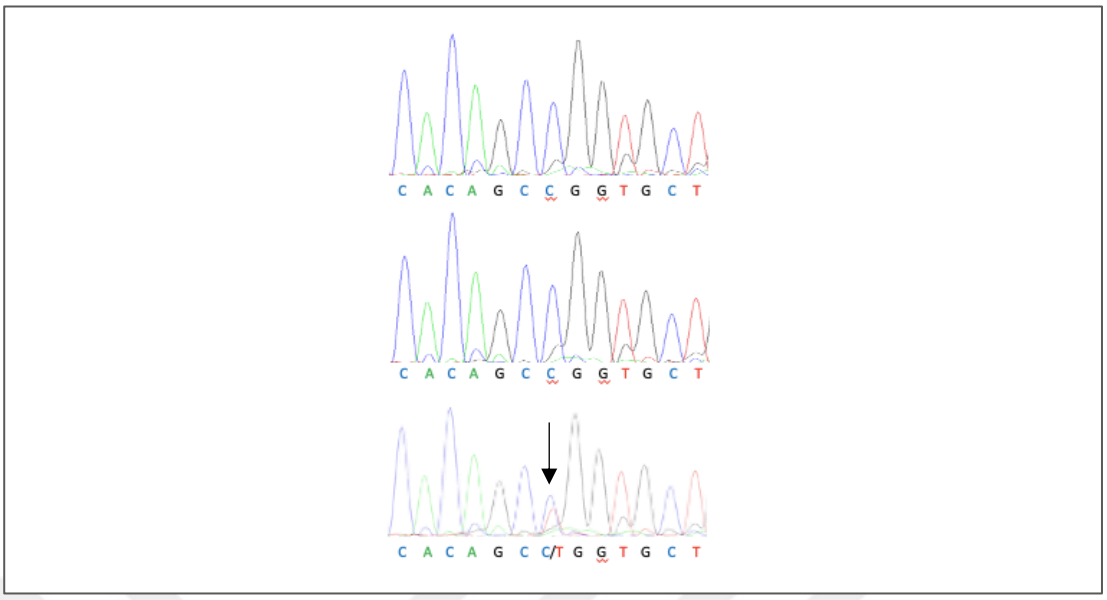


Figure 4.6. Sanger sequencing of the mutation c.C487T of EBF3 in case III. The mutation is only present in the proband, but not in the parents.

EBF3 gene is located on chromosome 10q26.3 encodes a member of the early B-cell factor (EBF) transcription factor family (in other words Olf, COE, or O/E) that has crucial roles in neurogenesis and development (167,168). Heterozygous mutations in EBF3 are the genetic cause behind Hypotonia, Ataxia, And Delayed Development Syndrome; HADDS, which is a neurodevelopmental syndrome characterized by congenital hypotonia, delayed psychomotor development, variable intellectual disability with speech delay, ataxia and variable dysmorphic facial features (169). Although the pathogenic mechanisms of EBF3 mutations remain unclear, missense, nonsense, and intronic variants, and copy number variations are described so far (170). However, no information about the prevalence of the disease has been reported because of insufficient data about incidence and published cases (171).

5. DISCUSSION AND CONCLUSION

The diagnosis of RDs is a complicated journey since they have low prevalence and phenotypic diversity. Associating a particular disease phenotype with genomic variants is a multistep process. The initial steps that include processing raw sequence data are highly automated through the use of many bioinformatics software and tools. However, many of these software and tools focus on a particular aspect of these steps and do not offer a single workflow from start to finish. There are no gold standards in the translation WES to clinical benefit. Besides, the final variant filtration step, which is the most critical step, is not entirely automated. It requires a comprehensive interpretation together with integrative approaches. In this dissertation, it was aimed to create variant prioritization workflow for WES data toward clinical utility. The workflow introduced was tested on three individuals with previously unsolved WES data and was achieved diagnosis for them (Table 5). Detailed information about the bioinformatic analysis, variant interpretation and diagnosis for each case will be explained in the related subsections.

Table 5. Summary of patients with established diagnosis by WES Reanalysis

Case ID	Disease-Causing Gene	Variant	Inheritance	Phenotype
I	SLC2A1	Deletion/ c.275+1G>-	Heterozygous	GLUT1 deficiency syndrome 1, infantile onset, severe
II	PRKAR1A	Non-synonymous SNV/ c.G512A: p.G171E	Heterozygous	Acrodysostosis 1, with or without hormone resistance
III	EBF3	Non-synonymous SNV/ c.C487T: p.R163W	Heterozygous	Hypotonia, ataxia, and delayed development syndrome 1

5.1. Bioinformatic Analysis, Variant Interpretation and Diagnosis of Case I

The clinical symptoms for the case I was global developmental delay, microcephaly, intellectual disability, dysarthria, epilepsy, hemangioma, and involuntary movements (choreoathetosis and dystonia). Re-analysis of the previously unsolved WES data was performed as described in the method section. The workflow revealed the novel heterozygous variant c.275+1G>- in the SLC2A1 gene. This variant has not been reported in the literature available databases. GLUT1DS1 caused by the mutation in the SLC2A1 gene was proposed for the patient as the genetic diagnosis. Since low cerebrospinal fluid (CSF) glucose concentration is a distinctive sign for GLUT1DS1 (29), the genetic diagnosis was confirmed with CSF evaluation. Lumbar puncture (LP) was performed by the pediatric neurologist. While normal CSF-to-blood glucose ratio is about 0.6, in patients with GLUT1DS1, the value ranges from 0.19 to 0.46 (172). LP showed no cells, protein 20.40 mg/dl, glucose 32 mg/dl, simultaneous blood glucose 87 mg/dl. This ratio for case I was 0.36. After extensive neurometabolic and genetic screen, the patient was diagnosed with GLUT1DS1. GLUT1DS1 is a rare neurological disorder characterized by infantile seizures, developmental delay, acquired microcephaly, hypoglycorrhachia, and a complex movement disorder consisting of ataxia and spasticity (29). Impaired glucose transport result from mutations in the SLC2A1 gene lead to these symptoms. The ketogenic diet supplies an alternative fuel for the brain instead of glucose and recovers markedly symptoms (154). The pediatric neurologist informed that Case I responded to the ketogenic diet. She has been seizure-free shortly after the initiation of the diet. She also had decreased involuntary movements; her speech became more understandable after the diet.

In addition to typical symptoms of GLUT1DS1, our patient had a hemangioma as an unusual feature. Based on our knowledge, these two conditions were not previously reported together in any other patients with SLC2A1 variants. Although the hemangioma pathogenesis is not fully clarified, GLUT1 protein has been used as a selective marker to differentiate infantile hemangioma from other vascular

malformations (173,174). It is supposed that the expression of GLUT1 in hemangioma tissue is a potential sign of the pathogenicity because the protein is not expressed in the healthy skin tissue (173). It suggests that this may represent an additional feature associated with the disease. Yet, further research is required to explain whether there is a relationship between Hemangioma and GLUT1DS1.

Interestingly, the missense variant c.275+1G>A, which is in the same position with case I was reported before (175). Although both patients had some phenotypic overlap, including global developmental delay, seizure, signal abnormality in white matter, the case I showed more severe symptoms. She had several complex movement anomalies including dystonia and spasticity. It is thought that the phenotypic severity result from 1 base-pair deletion located on the intron/exon boundary in the SLC2A1 gene for Case I. Since splice sites on the boundary are crucial for gene expression and there are highly conserved dinucleotide sequences at these sites (176,177). It is a possible reason for the difference between our cases and recently reported case by Ismayilova et al. However, there is a need for further experiments at the level of RNA and protein to show the effect of 1 base-pair deletion. These experiments were not able to be conducted since the blood sample of the patient is no longer available.

5.2. Bioinformatic Analysis, Variant Interpretation and Diagnosis of Case II

The clinical symptoms for case II were global developmental delay, skeletal system abnormalities, distinctive facial features, hypothyroidism, and strabismus. Re-analysis of the previously unsolved WES data was performed as described in the method section. The workflow revealed the heterozygous missense variant c.G512A: p.G171E in the PRKAR1A gene. Based on the current knowledge, the variant affecting the cAMP-binding domain A of the PRKAR1A protein was novel (Figure 5.1).

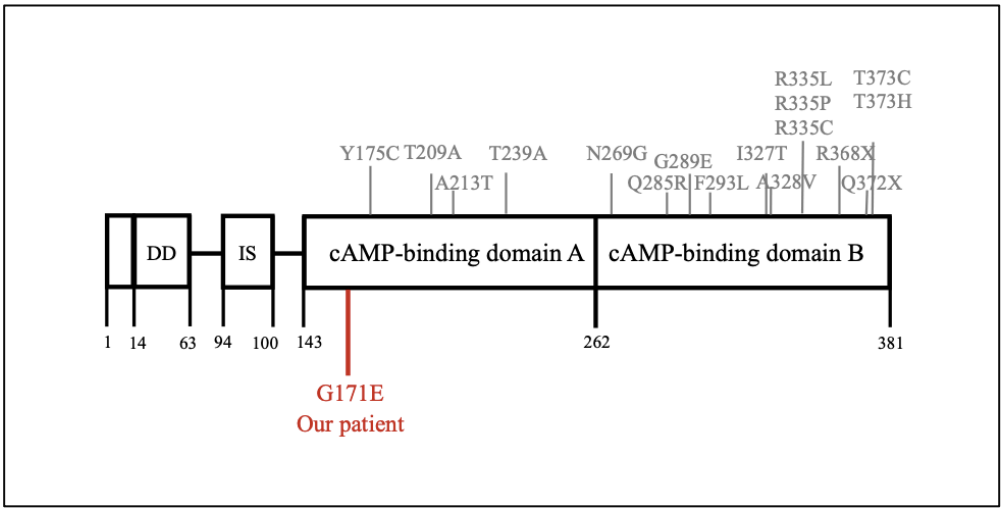


Figure 5.1. The domain structure of PRKAR1A protein and previously reported mutations. The protein is composed of a dimerization domain (DD), an inhibitory site (IS), and two cAMP-binding domains which are called A and B.

MD simulation was performed to enlighten the impact of this novel variant on PRKAR1A protein. MD results demonstrated that G171E mutation causes to increase in the distance between cAMP-binding residues of the protein (Figure 5.2).

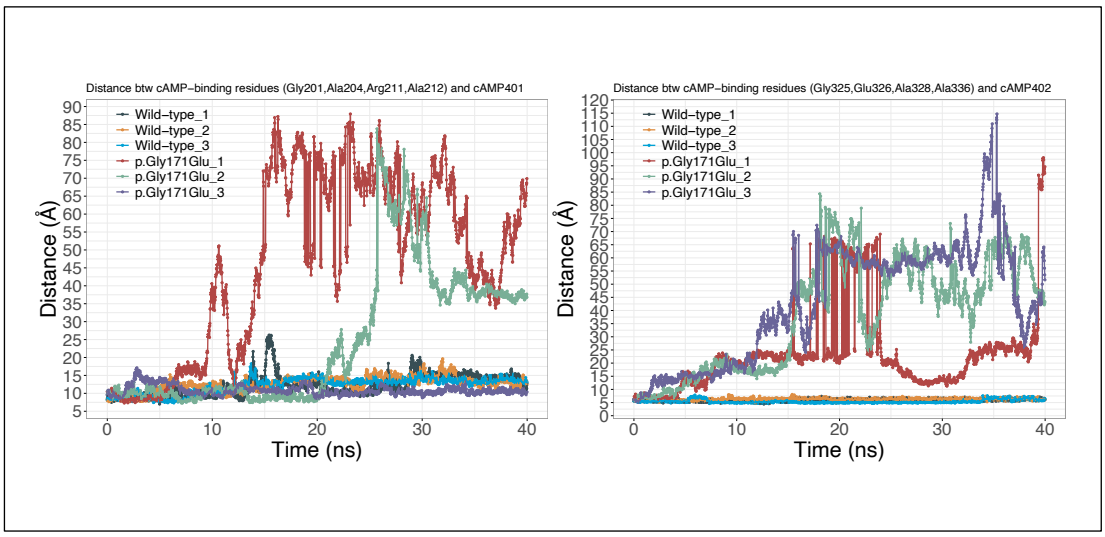


Figure 5.2. The comparison of mutant and wild-type protein in terms of the distance between cAMP binding domains.

In figure Figure 5.3, it is observed that the mutation disturbs the interaction between cAMP molecules, which are shown orange color and PRKAR1A protein. The figure also shows an interaction between PRKAR1A protein and two cAMP molecules at 0 nanoseconds (ns). In the 13th ns, one of the cAMP molecules disassociates from the cAMP binding domain A which is shown pink color. In the 39th ns, it is observed that the decrease in the binding affinity of cAMP to the protein and, ultimately, complete disassociation from cAMP binding domain B, which is shown grey color at the 40th ns.

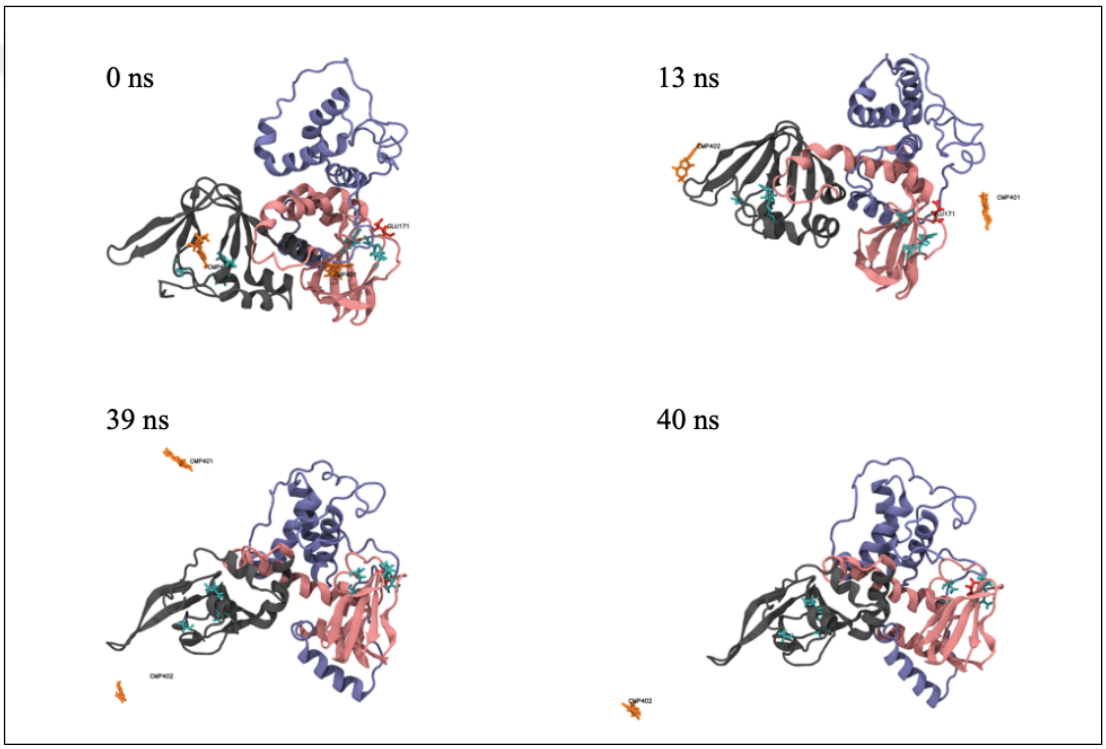


Figure 5.3. The interaction between two cAMP molecules and mutant (p.G171E) PRKAR1A protein at the 0, 13th, 39th, 40th ns.

PRKAR1A is the most commonly expressed regulatory subunit of PKA holoenzyme, which has a vital role in several cellular processes. PRKAR1A composed of two cAMP-binding domains that allow binding cAMP. When cAMP molecules bind these sites, PKA shows enzymatic activity and phosphorylate downstream targets

in cells. In case of the disruption of the binding of cAMP to PRKAR1A, PKA remains as tetrameric holoenzyme and can not perform the function. Ultimately, this results in the impairment of many pathways in the cell and provide a possible pathogenicity mechanism for the variant. In line with these findings, Acrodysostosis 1 caused by the mutation in the PRKAR1A gene was proposed for the patient as the genetic diagnosis. The diagnosis was confirmed by the pediatrician. Acrodysostosis 1 is an extremely rare type of skeletal dysplasias characterized by peripheral dysostosis, including abnormally short and malformed bones of the hands and feet, facial dysostosis, nasal hypoplasia, growth delays, short stature, and mental retardation. It was reported that some patients might have resistance to several hormones (165). This finding in the literature was an explanation for the hypothyroidism phenotype of case II. In addition to the typical symptoms, case II had more severe skeletal system abnormalities, including dislocation on hip and elbows. The functional importance of the variant G171E was a possible reason for this finding; nevertheless, there is a need for further experiments at the level of both in-vitro and in-vivo.

5.3. Bioinformatic Analysis, Variant Interpretation and Diagnosis of Case III

The clinical symptoms for the case III included global developmental delay, generalized hypotonia, mild facial dysmorphisms such as frontal bossing and low-set ears, speech delay, decreased pain response, hyperactive deep tendon reflexes, and strabismus. The workflow described in this dissertation was employed in the previously unsolved trio-WES.

In this case, it is concluded that the trio approach has an obvious advantage for the detection of genetic variants identified by WES. In addition to providing the phase of variants, the trio approach also enables us to filter out rare benign familial variants and quickly identify variants that affect only the child (178). If parents do not have similar symptoms as the child, trio sequencing provides an almost ten-fold decrease of the

candidate variant number compared to the proband sequencing as well as 50% increased diagnostic yield. (64).

After the implementation of the workflow, HADDS caused by the mutation EBF3 gene was proposed as the genetic diagnosis, and it was confirmed by the clinicians. The de-novo missense variant was located at the position c.C487T: p.R163W. Interestingly, all of the reported missense variants are located in the DNA binding domain, which is highly conserved in EBF3 protein (179). As it is shown in Figure 5.4, five of these mutations disturb the same amino acid residue Arg163, which is in the Zn2+ finger Collier/Olf/Ebf (COE) motif (169,180,181). In this respect, a recent study conducted molecular dynamics simulations and demonstrated that p.R163W could cause decreased DNA binding affinity and differential transcriptional activation (181).

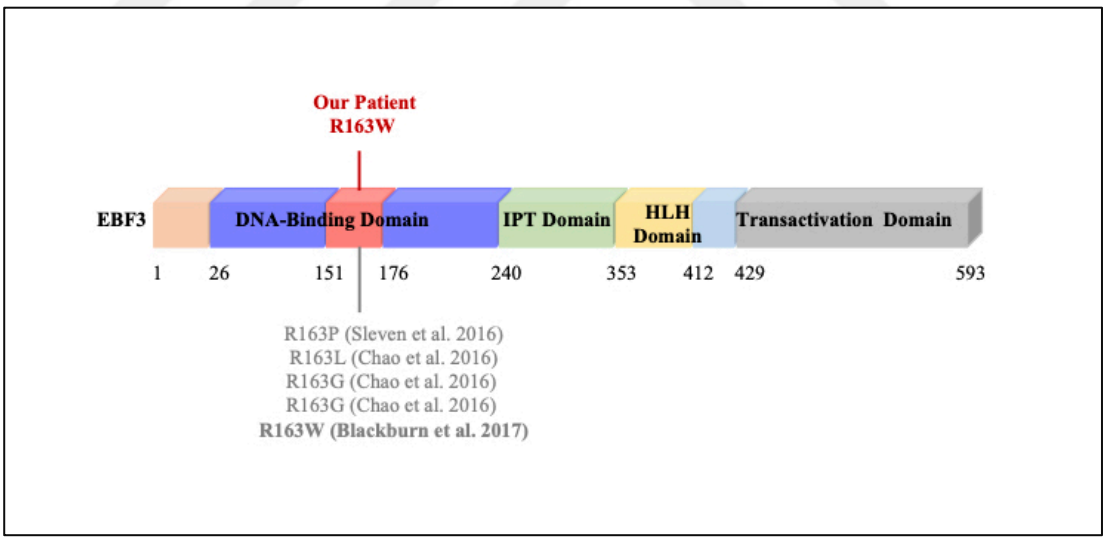


Figure 5.4. The domain structure of EBF3 protein and previously reported mutations.

The variant p.R163W was previously reported in a girl by Blackburn et al (181). The comparison of the case III with the patient in this report revealed substantial

phenotypic overlap. Both patients had global developmental delay, generalized hypotonia, speech delay, mild facial dysmorphisms, strabismus, normal evaluations for comprehensive biochemical metabolic testing, MRI, EEG, chromosomal microarray. Contrary to case III, the reported patient in that article had several urogenital anomalies, including atonic bladder, distal urethral stricture, vesicoureteral reflux, bilateral hydroureter and hydronephrosis, recurrent urinary tract infections and bicornuate uterus (181). Although both patients have the same mutation, case III shows mild symptoms comparing the other patient.

To sum up, by employing the variant prioritization workflow on previously unsolved WES cases, pathogenic de-novo heterozygous variants were identified in SLC2A1, PRKAR1A, and EBF3 genes. Here, it is also highlighted the potential of the reanalysis of WES data via implementing a different workflow for undiagnosed individuals. Finally, it is concluded that the identification of previously undetected disease-causing mutations resulted from improved variant prioritization workflow. The results revealed from the implementation of the workflow also contribute to the RD literature.

7. REFERENCES

1. Pogue RE, Cavalcanti DP, Shanker S, Andrade RV, Aguiar LR, de Carvalho JL, et al. Rare genetic diseases: update on diagnosis, treatment and online resources. Drug Discovery Today. 2018 Jan;23(1):187–95.

2. Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E, et al. Rare Disease Terminology and Definitions—A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. Value in Health. 2015 Sep;18(6):906–14.

3. 6. Priority diseases and reasons for inclusion [Internet]. 2020. Available from: https://www.who.int/medicines/areas/priority_medicines/Ch6_19Rare.pdf

4. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Research. 2019 Jan 8;47(D1):D1038–43.

5. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020 Feb;28(2):165–73.

6. The Global Challenge of Rare Disease Diagnosis [Internet]. 2020. Available from: https://www.shire.com/-/media/shire/shireglobal/shirecom/pdffiles/patient/shire-diagnosis-initiative-pag-leaflet.pdf

7. Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. The American Journal of Human Genetics. 2017 May;100(5):695–705.

8. 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 Oct;14(10):681–91.

9. Voelkerding KV, Dames SA, Durtschi JD. Next-Generation Sequencing: From Basic Research to Diagnostics. Clinical Chemistry. 2009 Apr 1;55(4):641–58.

10. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010 Feb;7(2):111–8.

11. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011 Nov;12(11):745–55.

12. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009 Sep;461(7261):272–6.

13. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010 Jan;42(1):30–5.

14. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N Engl J Med. 2013 Oct 17;369(16):1502–11.

15. Ji J, Shen L, Bootwalla M, Quindipan C, Tatarinova T, Maglinte DT, et al. A semiautomated whole-exome sequencing workflow leads to increased diagnostic yield and identification of novel candidate variants. Cold Spring Harb Mol Case Stud. 2019 Apr;5(2):a003756.

16. Goh G, Choi M. Application of Whole Exome Sequencing to Identify Disease-Causing Variants in Inherited Human Diseases. Genomics Inform. 2012;10(4):214.

17. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet. 2012 May;20(5):490–7.

18. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, et al. Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. JAMA. 2014 Nov 12;312(18):1870.

19. Retterer K, Juusola J, Cho MT, Vitazka P, Millan F, Gibellini F, et al. Clinical application of whole-exome sequencing across clinical indications. Genet Med. 2016 Jul;18(7):696–704.

20. Eldomery MK, Coban-Akdemir Z, Harel T, Rosenfeld JA, Gambin T, Stray-Pedersen A, et al. Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. 2017 Dec;9(1):26.

21. Bergant G, Maver A, Lovrecic L, Čuturilo G, Hodzic A, Peterlin B. Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: a retrospective survey in 1,059 cases. Genet Med. 2018 Mar;20(3):303–12.

22. Salfati EL, Spencer EG, Topol SE, Muse ED, Rueda M, Lucas JR, et al. Re-analysis of whole-exome sequencing data uncovers novel diagnostic variants and improves molecular diagnostic yields for sudden death and idiopathic diseases. Genome Med. 2019 Dec;11(1):83.

23. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med. 2017 Feb;19(2):209–14.

24. Orphanet [Internet]. 2020. Available from: https://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN#AboutRD

25. FDA [Internet]. 2020. Available from: https://www.fda.gov/patients/rare-diseases-fda

26. Quintana-Murci L. Understanding rare and common diseases in the context of human evolution. Genome Biol. 2016 Dec;17(1):225.

27. Tarailo-Graovac M, Shyr C, Ross CJ, Horvath GA, Salvarinova R, Ye XC, et al. Exome Sequencing and the Management of Neurometabolic Disorders. N Engl J Med. 2016 Jun 9;374(23):2246–55.

28. Seidner G, Alvarez MG, Yeh JI, O'Driscoll KR, Klepper J, Stump TS, et al. GLUT-1 deficiency syndrome caused by haploinsufficiency of the blood-brain barrier hexose carrier. Nat Genet. 1998 Feb 18;188-91

29. De Vivo DC, Trifiletti RR, Jacobson RI, Ronen GM, Behmand RA, Harik SI. Defective glucose transport across the blood-brain barrier as a cause of persistent hypoglycorrhachia, seizures, and developmental delay. N Engl J Med. 1991 Sept 5;325(10):703-9

30. Klepper J. Glucose transporter deficiency syndrome (GLUT1DS) and the ketogenic diet. Epilepsia. 2008 Nov;49:46–9.

31. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, et al. Distinguishing protein-coding and noncoding genes in the human genome. Proceedings of the National Academy of Sciences. 2007 Dec 4;104(49):19428–33.

32. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. PNAS. 2009 Nov 10;106(45):19096–101.

33. Hoischen A, van Bon BWM, Gilissen C, Arts P, van Lier B, Steehouwer M, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet. 2010 Jun;42(6):483–5.

34. Wetterstrand KA. DNA sequencing costs: data from the NHGRI genome sequencing program (GSP) [Internet]. Available from: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

35. Maxam AM, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences. 1977 Feb 1;74(2):560–4.

36. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. 1977 Dec 1;74(12):5463–7.

37. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nature. 2006 Sep 15; 437(7057):376–380.

38. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010 Jan;11(1):31–46.

39. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011 Feb;470(7333):198–203.

40. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008 Nov;456(7218):53–9.

41. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome Sequencing: Current and Future Perspectives. G3. 2015 Aug;5(8):1543–50.

42. Zhou X, Rokas A. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. Mol Ecol. 2014 Apr;23(7):1679–700.

43. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2010 Apr;38(6):1767–71.

44. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics. 2012;13(1):341.

45. Zhou Q, Su X, Wang A, Xu J, Ning K. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. Zhang Z, editor. PLoS ONE. 2013 Apr 2;8(4):e60234.

46. The Babraham Bioinformatics group. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

47. Chanumolu SK, Albahrani M, Otu HH. FQStat: a parallel architecture for very high-speed assessment of sequencing quality metrics. BMC Bioinformatics. 2019 Dec;20(1):424.

48. Katta MAVSK, Khan AW, Doddamani D, Thudi M, Varshney RK. NGS-QCbox and Raspberry for Parallel, Automated and Rapid Quality Control Analysis of Large-Scale Next Generation Sequencing (Illumina) Data. Wang J, editor. PLoS ONE. 2015 Oct 13;10(10):e0139868.

49. Thrash A, Arick M, Peterson DG. Quack: A quality assurance tool for high throughput sequence data. Analytical Biochemistry. 2018 May;548:38–43.

50. Lo C-C, Chain PSG. Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinformatics. 2014 Dec;15(1):366.

51. Kircher M, Heyn P, Kelso J. Addressing challenges in the production and analysis of illumina sequencing data. BMC Genomics. 2011 Dec;12(1):382.

52. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. 2011;12(11):R112.

53. Kong Y. Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics. 2011 Aug;98(2):152–3.

54. Ma Y, Xie H, Han X, Irwin DM, Zhang Y-P. QcReads: An Adapter and Quality Trimming Tool for Next-Generation Sequencing Reads. Journal of Genetics and Genomics. 2013 Dec;40(12):639–42.

55. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

56. Jiang H, Lei R, Ding S-W, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014 Dec;15(1):182.

57. Liao X, Li M, Zou Y, Wu F, Pan Y, Wang J. An efficient trimming algorithm based on multi-feature fusion scoring model for NGS data. IEEE/ACM Trans Comput Biol and Bioinf. 2019;1–1.

58. Sun K. Ktrim: an extra-fast and accurate adapter- and quality-trimmer for sequencing data. Birol I, editor. Bioinformatics. 2020 Mar 11;btaa171.

59. Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat Methods. 2009 Nov;6(S11):S6–12.

60. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012 Dec;28(24):3169–77.

61. Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics [Internet]. 2013 Oct [cited 2020 Apr 19];43(1). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1110s43

62. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Sci Rep. 2017 Mar;7(1):43169.

63. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017 Oct;18(10):599–612.

64. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. The Lancet. 2015 Apr;385(9975):1305–14.

65. Cornish A, Guda C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. BioMed Research International. 2015;2015:1–11.

66. McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, asds, et al. Choice of transcripts and software has a large effect on variant annotation. Genome Med. 2014;6(3):26.

67. Baumgartner WA, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics. 2007 Jul 1;23(13):i41–8.

68. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Research. 2019 Jan 8;47(D1):D1018–27.

69. Ewing B, Hillier L, Wendl MC, Green P. Base-Calling of Automated Sequencer Traces Using *Phred.* I. Accuracy Assessment. Genome Res. 1998 Mar 1;8(3):175–85.

70. Ewing B, Green P. Base-Calling of Automated Sequencer Traces Using *Phred.* II. Error Probabilities. Genome Res. 1998 Mar 1;8(3):186–94.

71. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

72. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

73. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015 Jun 15;31(12):2032–4.

74. Broad Institute. Picard [Internet]. 2020. Available from: http://broadinstitute.github.io/picard/

75. DePristo MA, Banks E, Poplin RE, Garimella KV, Maguire JR, Hartl C. A framework for variation discovery and genotyping using next- generation DNA sequencing data. Nat Genet. 2011 May;43(5): 491–498.

76. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples [Internet]. Genomics; 2017 Nov [cited 2020 May 20]. Available from: http://biorxiv.org/lookup/doi/10.1101/201178

77. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

78. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010 Sep 1;38(16):e164–e164.

79. Richards S, Aziz N, Bale S, Bick D, Das 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 May;17(5):405–23.

80. Dudley JT, Kim Y, Liu L, Markov GJ, Gerold K, Chen R, et al. Human genomic disease variants: A neutral evolutionary explanation. Genome Research. 2012 Aug 1;22(8):1383–94.

81. Stitziel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. 2011;12(9):227.

82. Kobayashi Y, Yang S, Nykamp K, Garcia J, Lincoln SE, Topper SE. Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Med. 2017 Dec;9(1):13.

83. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

84. Exome Aggregation Consortium, Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285–91.

85. Genome Aggregation Database Consortium, Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

86. Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet. 2016 Jan;24(1):2–5.

87. Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research. 2003 Jul 1;31(13):3812–4.

88. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. de Brevern AG, editor. PLoS ONE. 2012 Oct 8;7(10):e46688.

89. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997 Sep 1;25(17):3389–402.

90. 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 Mar;46(3):310–5.

91. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014 Apr;11(4):361–2.

92. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

93. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015 Mar 1;31(5):761–3.

94. 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.

95. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human Molecular Genetics. 2015 Apr 15;24(8):2125–37.

96. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016 Dec;48(12):1581–6.

97. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. The American Journal of Human Genetics. 2016 Oct;99(4):877–85.

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

99. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Human Mutation. 2013 Jan;34(1):57–65.

100. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Research. 2009 Sep 1;19(9):1553–61.

101. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. Williams SM, editor. PLoS Genet. 2013 Aug 22;9(8):e1003709.

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

103. Siepel A. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research. 2005 Aug 1;15(8):1034–50.

104. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009 Jun 15;25(12):i54–62.

105. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. Wasserman WW, editor. PLoS Comput Biol. 2010 Dec 2;6(12):e1001025.

106. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences. 1992 Nov 15;89(22):10915–9.

107. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009 Nov 1;25(21):2744–50.

108. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, the Mouse Genome Database Group, et al. Mouse Genome Database (MGD) 2019. Nucleic Acids Research. 2019 Jan 8;47(D1):D801–6.

109. Smith CM, Hayamizu TF, Finger JH, Bello SM, McCright IJ, Xu J, et al. The mouse Gene Expression Database (GXD): 2019 update. Nucleic Acids Research. 2019 Jan 8;47(D1):D774–9.

110. Cao Z, Wang L, Chen Y, Cai R, Lu J, Yu Y, et al. VarfromPDB: An Automated and Integrated Tool to Mine Disease-Gene-Variant Relations from the Public Databases and Literature. J Proteomics Bioinform [Internet]. 2017 [cited 2020 May 8];10(11). Available from: https://www.omicsonline.org/open-access/varfrompdb-an-automated-and-integrated-tool-to-mine-diseasegenevariant-relations-from-the-public-databases-and-literature-jpb-1000455-95973.html

111. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet]. 2020. Available from: https://www.R-project.org/.

112. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, et al. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Research. 2019 Jan 8;47(D1):D786–92.

113. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research. 2018 Jan 4;46(D1):D1062–7.

114. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research. 2019 Jan 8;47(D1):D506–15.

115. Orphanet [Internet]. 2020. Available from: https://www.orpha.net

116. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Research. 2019 Jan 8;47(D1):D853–8.

117. Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, van den Berg MP, et al. Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics. Human Mutation. 2013 Jul;34(7):1035–42.

118. Hegde M, Santani A, Mao R, Ferreira-Gonzalez A, Weck KE, Voelkerding KV. Development and Validation of Clinical Whole-Exome and Whole-Genome Sequencing for Detection of Germline Variants in Inherited Disease. Archives of Pathology & Laboratory Medicine. 2017 Jun;141(6):798–805.

119. Strom SP, Lee H, Das K, Vilain E, Nelson SF, Grody WW, et al. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. Genet Med. 2014 Jul;16(7):510–5.

120. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.

121. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013 Mar 1;14(2):178–92.

122. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Research. 2012 Aug;40(15):e115–e115.

123. van Gunsteren WF, Bakowies D, Baron R, Chandrasekhar I, Christen M, Daura X, et al. Biomolecular Modeling: Goals, Problems, Perspectives. Angew Chem Int Ed. 2006 Jun 19;45(25):4064–92.

124. George Priya Doss C, Rajasekaran R, Sudandiradoss C, Ramanathan K, Purohit R, Sethumadhavan R. A novel computational and structural analysis of nsSNPs in CFTR gene. HUGO J. 2008 Jan;2(1–2):23–32.

125. George Priya Doss C, Rajith B. A New Insight into Structural and Functional Impact of Single-Nucleotide Polymorphisms in PTEN Gene. Cell Biochem Biophys. 2013 Jun;66(2):249–63.

126. Priya Doss CG, Chakraborty C, Chen L, Zhu H. Integrating *In Silico* Prediction Methods, Molecular Docking, and Molecular Dynamics Simulation to Predict the Impact of ALK Missense Mutations in Structural Perspective. BioMed Research International. 2014;2014:1–14.

127. Vasquez M, Nemethyt G. Conformational Energy Calculations on Polypeptides and Proteins. Chem Rev. 1994;94(8):2183-2239.

128. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Research. 2004 Jul 1;32(Web Server):W526–31.

129. Ilouz R, Bubis J, Wu J, Yim YY, Deal MS, Kornev AP, et al. Localization and quaternary structure of the PKA RI holoenzyme. Proceedings of the National Academy of Sciences. 2012 Jul 31;109(31):12443–8.

130. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993 Apr 1;26(2):283–91.

131. Bruystens JGH, Wu J, Fortezzo A, Kornev AP, Blumenthal DK, Taylor SS. PKA RIα Homodimer Structure Reveals an Intermolecular Interface with Implications for Cooperative cAMP Binding and Carney Complex Disease. Structure. 2014 Jan;22(1):59–69.

132. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. Journal of Molecular Graphics. 1996 Feb;14(1):33–8.

133. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics. 1983 Jul 15;79(2):926–35.

134. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem. 2005 Dec;26(16):1781–802.

135. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods. 2017 Jan;14(1):71–3.

136. Grønbech-Jensen N, Farago O. Constant pressure and temperature discrete-time Langevin molecular dynamics. J Chem Phys. 2014 Nov 21;141(19):194108.

137. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. The Journal of Chemical Physics. 1995 Nov 15;103(19):8577–93.

138. Villanueva RAM, Chen ZJ. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). Measurement: Interdisciplinary Research and Perspectives. 2019 Jul 3;17(3):160–7.

139. Wang D, Pascual JM, Yang H, Engelstad K, Jhung S, Sun RP, et al. Glut-1 deficiency syndrome: Clinical, genetic, and therapeutic aspects. Ann Neurol. 2005 Jan;57(1):111–8.

140. Wang D, Kranz-Eble P. Mutational analysis of GLUT1 (SLC2A1) in Glut-1 Deficiency Syndrome. Human Mutation. 2000;16:224-31

141. Baldwin SA. Mammalian passive glucose transporters: members of an ubiquitous family of active and passive transport proteins. Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes. 1993 Jun;1154(1):17–49.

142. Gerhart DZ, Levasseur RJ, Broderius MA, Drewes LR. Glucose transporter localization in brain using light and electron immunocytochemistry. J Neurosci Res. 1989 Apr;22(4):464–72.

143. Harik SI, Kalaria N, Andersson L, Lundahl P. lmmunocytochemical Localization of the Erythroid Glucose Transporter: Abundance in Tissues with Barrier Functions. J Neurosci, 1990 Dec;10(12):3862-72

144. Pardridges M, Boado J, Farrell R. Brain-type Glucose Transporter (GLUT- 1) Is Selectively Localized to the Blood-Brain Barrier. J Biol. 1990 Oct 5;265(29):18035-40

145. Verrotti A, Di Francesco L, Striano P. GLUT1 deficiency and pediatric-onset hereditary spastic paraplegia: A new association. European Journal of Paediatric Neurology. 2019 Mar;23(2):233–4.

146. Larsen J, Johannesen KM, Ek J, Tang S, Marini C, Blichfeldt S, et al. The role of *SLC2A1* mutations in myoclonic astatic epilepsy and absence epilepsy, and the estimated frequency of GLUT1 deficiency syndrome. Epilepsia. 2015 Dec;56(12):e203–8.

147. Verrotti A, D'Egidio C, Agostinelli S, Gobbi G. Glut1 deficiency: When to suspect and how to diagnose? European Journal of Paediatric Neurology. 2012 Jan;16(1):3–9.

148. Brockmann K, Wang D, Korenke CG, Von Moers A, Ho Y-Y, Pascual JM, et al. Autosomal dominant Glut-1 deficiency syndrome and familial epilepsy. Ann Neurol. 2001 Oct;50(4):476–85.

149. Klepper J. Autosomal dominant transmission of GLUT1 deficiency. Human Molecular Genetics. 2001 Jan 1;10(1):63–8.

150. Klepper J, Scheffer H, Elsaid MF, Kamsteeg E-J, Leferink M, Ben-Omran T. Autosomal Recessive Inheritance of GLUT1 Deficiency Syndrome. Neuropediatrics. 2009 Oct;40(05):207–10.

151. Rotstein M, Engelstad K, Yang H, Wang D, Levy B, Chung WK, et al. Glut1 deficiency: Inheritance pattern determined by haploinsufficiency. Ann Neurol. 2010 Dec;68(6):955–8.

152. Di Vito L, Licchetta L, Pippucci T, Baldassari S, Stipa C, Mostacci B, et al. Phenotype variability of GLUT1 deficiency syndrome: Description of a case series with novel SLC2A1 gene mutations. Epilepsy & Behavior. 2018 Feb;79:169–73.

153. Zaman SM, Mullen SA, Petrovski S, Maljevic S, Gazina EV, Phillips AM, et al. Development of a rapid functional assay that predicts GLUT1 disease severity. Neurol Genet. 2018 Dec;4(6):e297.

154. Klepper J, Diefenbach S, Kohlschütter A, Voit T. Effects of the ketogenic diet in the glucose transporter 1 deficiency syndrome. Prostaglandins, Leukotrienes and Essential Fatty Acids. 2004 Mar;70(3):321–7.

155. Madeira F, Park Y mi, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Research. 2019 Jul 2;47(W1):W636–41.

156. Li N, Nie M, Li M, Jiang Y, Xing X, Wang O, et al. The First Mutation Identified in a Chinese Acrodysostosis Patient Confirms a p.G289E Variation of PRKAR1A Causes Acrodysostosis. IJMS. 2014 Jul 29;15(8):13267–74.

157. Butcher RW, Robison GA, Hardman J g., Sutherland EW. The role of cyclic AMP in hormone actions. Advances in Enzyme Regulation. 1968 Jan;6:357–89.

158. Hansson V. Cyclic-AMP-dependent protein kinase (PKA) in testicular cells. Cell specific expression, di□fferential regulation and targeting of subunits of PKA. J Steroid Biochem Mol Biol. 1999;69:367-378

159. Bossis I, Stratakis CA. Minireview: *PRKAR1A* : Normal and Abnormal Functions. Endocrinology. 2004 Dec;145(12):5452–8.

160. Uhler MD, Carmichael DF, Lee DC, Chrivia JC, Krebs EG, McKnight GS. Isolation of cDNA clones coding for the catalytic subunit of mouse cAMP-dependent protein kinase. Proceedings of the National Academy of Sciences. 1986 Mar 1;83(5):1300–4.

161. Solberg R, Sandberg M, Natarajan V, Torjesen PA, Hansson V, Jahnsen T, et al. The Human Gene for the Regulatory Subunit RIα of Cyclic Adenosine 3′,5′-Monophosphate-Dependent Protein Kinase: Two Distinct Promoters Provide Differential Regulation of Alternately Spliced Messenger Ribonucleic Acids [1]. Endocrinology. 1997 Jan;138(1):169–81.

162. Agnès L, Christine M, Alain C, Colette A, Yasemin G, Mathilde C, et al. Recurrent PRKAR1A Mutation in Acrodysostosis with Hormone Resistance. N Engl J Med. 2011 June 9;364(23):2218-26

163. Maroteaux P, Malamut G. [Acrodysostosis]. Presse Med. 1968 Nov 27;76(46):2189–92.

164. Robinow M, Pfeiffer RA, Gorlin RJ, McKusick VA, Renuart AW, Johnson GF, et al. Acrodysostosis. A syndrome of peripheral dysostosis, nasal hypoplasia, and mental retardation. Am J Dis Child. 1971 Mar;121(3):195–203.

165. Linglart A, Menguy C, Couvineau A, Auzan C, Gunes Y, Cancel M, et al. Recurrent PRKAR1A mutation in acrodysostosis with hormone resistance. N Engl J Med. 2011 Jun 9;364(23):2218–26.

166. Acrodysostosis [Internet]. 2020. Available from: https://www.orpha.net/consor/cgi-bin/OC_Exp.php?Lng=GB&Expert=950

167. Zardo G, Tiirikainen MI, Hong C, Misra A, Feuerstein BG, Volik S, et al. Integrated genomic and epigenomic analyses pinpoint biallelic gene inactivation in tumors. Nat Genet. 2002 Nov;32(3):453–8.

168. Dubois L, Vincent A. The COE – Collier/Olf1/EBF – transcription factors: structural conservation and diversity of developmental functions. Mechanisms of Development. 2001 Oct;108(1–2):3–12.

169. Sleven H, Welsh SJ, Yu J, Churchill MEA, Wright CF, Henderson A, et al. De Novo Mutations in EBF3 Cause a Neurodevelopmental Syndrome. The American Journal of Human Genetics. 2017 Jan;100(1):138–50.

170. Lopes F, Soares G, Gonçalves-Rocha M, Pinto-Basto J, Maciel P. Whole Gene Deletion of EBF3 Supporting Haploinsufficiency of This Gene as a Mechanism of Neurodevelopmental Disease. Front Genet. 2017 Oct 9;8:143.

171. HYPOTONIA, ATAXIA, AND DELAYED DEVELOPMENT SYNDROME; HADDS [Internet]. mendelian.co. Available from: https://www.mendelian.co/diseases/hypotonia-ataxia-and-delayed-development-syndrome-hadds

172. Klepper J, Voit T. Facilitated glucose transporter protein type 1 (GLUT1) deficiency syndrome: impaired glucose transport into brain – a review. Eur J Pediatr. 2002 Jun;161(6):295–304.

173. North PE, Waner M, Mizeracki A, Mihm Jr MC. GLUT1: a newly discovered immunohistochemical marker for juvenile hemangiomas. Human Pathology. 2000;31(1):11-22

174. Leon-Villapalos J, Wolfe K, Kangesu L. GLUT-1: an extra diagnostic tool to differentiate between haemangiomas and vascular malformations. British Journal of Plastic Surgery. 2005 Apr;58(3):348–52.

175. Ismayilova N, Hacohen Y, MacKinnon AD, Elmslie F, Clarke A. GLUT-1 deficiency presenting with seizures and reversible leukoencephalopathy on MRI imaging. European Journal of Paediatric Neurology. 2018 Nov;22(6):1161–4.

176. Chang W-C. Alternative splicing of U12-type introns. Front Biosci. 2008;13(13):1681.

177. Parada GE, Munita R, Cerda CA, Gysling K. A comprehensive survey of non-canonical splice sites in the human transcriptome. Nucleic Acids Research. 2014 Aug 14; 42(16): 10564-78.

178. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018 May;19(5):253–68.

179. Tanaka AJ, Cho MT, Willaert R, Retterer K, Zarate YA, Bosanko K, et al. De novo variants in *EBF3* are associated with hypotonia, developmental delay, intellectual disability, and autism. Cold Spring Harb Mol Case Stud. 2017 Nov;3(6):a002097.

180. Chao H-T, Davids M, Burke E, Pappas JG, Rosenfeld JA, McCarty AJ, et al. A Syndromic Neurodevelopmental Disorder Caused by De Novo Variants in EBF3. The American Journal of Human Genetics. 2017 Jan;100(1):128–37.

181. Blackburn PR, Barnett SS, Zimmermann MT, Cousin MA, Kaiwar C, Pinto e Vairo F, et al. Novel de novo variant in *EBF3* is likely to impact DNA binding in a patient with a neurodevelopmental disorder and expanded phenotypes: patient report, in silico functional assessment, and review of published cases. Cold Spring Harb Mol Case Stud. 2017 May;3(3):a001743.

8. APPENDICES

APPENDIX 1. The Ethical Approval Form for Case I

ACIBADEM
ÜNİVERSİTESİ

SAYI: ATADEK-2019/17
KONU: Etik Kurul Kararı

Sayın Tuğçe Bozkurt, Prof.Dr. Uğur Sezerman,

Sorumluluğunu yürüttüğünüz **"GLUT-1 Eksikliği Sendromu Tip 1 ile İlişkili Olası Patojenik bir Varyantın Araştırılması"** başlıklı proje 07.11.2019 tarih 2019/17 Sayılı Atadek Toplantısında görüşülmüş olup 2019-17/36 karar numarası ile tıbbi etik yönden uygun bulunmuştur.

Prof.Dr. Güldal SÜYEN
ATADEK Başkan Yardımcısı

Kerem Aydınlar Kampüsü Kayışdağı Caddesi 32 Içerenköy 34752 Ataşehir İstanbul T 0216 500 44 44 F 0216 576 50 76
www.acibadem.edu.tr

APPENDIX 2. The Ethical Approval Form for Case II

ACIBADEM ÜNİVERSİTESİ

SAYI: ATADEK-2020/13 26.06.2020

KONU: Etik Kurul Kararı

Sayın Prof. Dr. Uğur Sezerman, Tuğçe Bozkurt,

Sorumluluğunu yürüttüğünüz **"Tüm Ekzom Sekanslama Verilerinin Yeniden Analizi ile Gelişimsel Gerilik, Mikrosefali, Çoklu Dislokasyon, Hipotoni, Hipotiroidi ve Çeşitli Yüz Anomalilerine Sahip Vakanın Aydınlatılması"** başlıklı proje 25.06.2020 tarih 2020/13 Sayılı Atadek Toplantısında görüşülmüş olup 2020-13/14 karar numarası ile tıbbi etik yönden uygun bulunmuştur.

Prof. Dr. Güldal Süyen

ATADEK Başkan Yardımcısı

Kerem Aydınlar Kampüsü Kayışdağı Caddesi 32 İçerenköy 34752 Ataşehir İstanbul T 0216 500 4444 F 0216 576 5076

www.acibadem.edu.tr

APPENDIX 3. The Ethical Approval Form for Case III

ACIBADEM
ÜNİVERSİTESİ

SAYI: ATADEK-2020/06
KONU: Etik Kurul Kararı

Sayın Prof.Dr. Uğur Sezerman, Tuğçe Bozkurt,

Sorumluluğunu yürüttüğünüz **"Üçlü Tüm Ekzom Sekanslamanın Yeniden Analizi ile Gelişimsel Gerilik ve Aksiyal Hipotoni Fenotipine Sahip Vakanın Aydınlatılması"** başlıklı proje 30.04.2020 tarih 2020/06 Sayılı Atadek Toplantısında görüşülmüş olup 2020-06/4 karar numarası ile tıbbi etik yönden uygun bulunmuştur.

Prof.Dr. Güldal SÜYEN
ATADEK Başkan Yardımcısı

Kerem Aydınlar Kampüsü Kayışdağı Caddesi 32 İçerenköy 34752 Ataşehir İstanbul T 0216 500 44 44 F 0216 576 50 76
www.acibadem.edu.tr

9. CURRICULUM VITAE

Kişisel Bilgiler

Adı	Tuğçe	**Soyadı**	Bozkurt
Doğum Yeri	Antalya	**Doğum Tarihi**	04.01.1995
Uyruğu	Türk	**Telefon**	545 585 0401
E-mail	bb.tugcee@gmail.com		

Eğitim Düzeyi

	Mezun Olduğu Kurumun Adı	**Mezuniyet Yılı**
Doktora/Uzmanlık		
Yüksek Lisans		
Lisans	Acıbadem Mehmet Ali Aydınlar Üniversitesi	2018
Lise	Antalya Aksu Anadolu Öğretmen Lisesi	2013

İş Deneyimi (Sondan geçmişe doğru sıralayın)

	Görevi	**Kurum**	**Süre (Yıl - Yıl)**
1.	Araştırmacı	Eternans Ltd.	2018-2020
2.			-
3.			-

Yabancı Dilleri	**Okuduğunu Anlama***	**Konuşma***	**Yazma***
İngilizce	İyi	İyi	İyi

***** Çok iyi, iyi, orta, zayıf olarak değerlendirin

	Yabancı Dil Sınav Notu □ 90								
KPDS	ÜDS	IELTS	TOEFL IBT	TOEFL PBT	TOEFL CBT	FCE	CAE	CPE	DİĞER
									YÖKDİL

□ Başarılmış birden fazla sınav varsa, tüm sonuçlar yazılmalıdır

 KPDS: Kamu Personeli Yabancı Dil Sınavı; ÜDS: Üniversitelerarası Kurul Yabancı Dil Sınavı; IELTS: International English Language Testing System; TOEFL IBT: Test of English as a Foreign Language-Internet-Based Test TOEFL PBT: Test of English as a Foreign Language-Paper-Based Test; TOEFL CBT: Test of English as a Foreign Language-Computer-Based Test; FCE: First Certificate in English; CAE: Certificate in Advanced English; CPE: Certificate of Proficiency in English

	Sayısal	**Eşit Ağırlık**	**Sözel**
LES Puanı	75,94028	76,31635	75,30744
(Diğer) Puanı			

Bilgisayar Bilgisi

Program	**Kullanma becerisi**
MS Office	Çok iyi
R	İyi
Bash/Shell	İyi

*****Çok iyi, iyi, orta, zayıf olarak değerlendirin

Uluslararası ve Ulusal Yayınları/Bildirileri/Sertifikaları/Ödülleri/Diğer

Sezerman U, Bozkurt T, Isleyen SF. Integrating Evolutionary Genetics to Medical Genomics: Evolutionary Approaches to Investigate Disease-Causing Variants. Molecular Medicine: IntechOpen; 2020. Available from: https://www.intechopen.com/online-first/integrating-evolutionary-genetics-to-medical-genomics-evolutionary-approaches-to-investigate-disease

Bozkurt T, Bakir-Gungor B, Dogan M, Gumus H, Dundar M, Sezerman OU. Reanalysis of trio whole exome sequencing (WES) data with a novel variant prioritization workflow reveals a de-novo missense variant in EBF3 gene associated with hypotonia and developmental delay. V. Uluslararası Katılımlı Erciyes Tıp Genetik Kongresi, 2020, Kapadokya, Türkiye.

Bozkurt T, Gerlevik U, Sezerman U. Identification of a Novel Missense Variant in the PRKAR1A Gene and Its Pathogenicity Mechanism. The 12th International Symposium on Health Informatics and Bioinformatics, 2019, İzmir, Türkiye.